Optimal Transport

Theory and Applications

Chengfeng Wen

Department of Computer Science
Stony Brook University

Research Proficiency Exam, 2016

Outline

Outline

Monge's Problem

- (X, d) Polish space
- $\mu \in \mathscr{P}(X)$, $\nu \in \mathscr{P}(Y)$ with equal total mass i.e.,

$$\int_X \mu = \int_X \nu$$

- Minimize

$$\min_{\psi_{\#}\mu=\nu} \int_X c(x, \psi(x)) d\mu(x) \tag{1}$$

among all transport maps ψ from μ to ν.

- Transport map may not exist.
- Transport map may not be unique

Outline

Kantorovich's Problem

Definition 1 (Admissible)

A probability measure $\gamma \in \mathscr{P}(X \times X)$ is admissible if its marginal measures are μ and ν, i.e.,

$$\pi_{0\#}\gamma = \mu, \ \pi_{1\#}\gamma = \nu$$

Admissible plan always exists, for example, $\gamma = \mu \times \nu$ is admissible.

Kantorovich's Problem

Given a Borel cost function $c : X \times X \to [0, \infty]$, minimize

$$K(\gamma) := \int_{X \times X} c(x, y) d\gamma(x, y)$$

among all admissible γ.
We call admissible γ a transport plan.

transport maps vs transport plans

Proposition 2 (transport maps vs transport plans)

Any Borel transport map $\psi : X \to X$ induces a transport plan defined by

$$\gamma_\psi := (Id \times \psi)_{\#}\mu$$

Conversely, a transport plan is induced by a transport map if γ is concentrated on a γ-measurable graph Γ.

Outline

Probability Space

Definition 3 (Probability Space)

A probability space is a triple $(X, \mathcal{F}, \mathbb{P})$ where

- X is a set
- $\mathcal{F}$ is a σ-algebra of subsets of X
- $\mathbb{P}$ is a function from $\mathcal{F}$ to $[0, 1]$ with $\mathbb{P}(X) = 1$ and if $E_1, E_2, ... \in \mathcal{F}$ are disjoint

$$\mathbb{P}(\cup_{j=1}^{\infty} E_j) = \sum_{j=1}^{\infty} \mathbb{P}(E_j)$$

Coupling

Definition 4 (Coupling)

Let (X, μ) and (Y, ν) be two probability spaces. Coupling of μ and ν is a probability measure $\gamma \in X \times Y$, s.t. γ admits μ and ν as marginals on X and Y respectively.

If γ is a coupling of μ and ν, then for all measurable set $A \subset X, B \subset Y$:

$$\gamma[A \times Y] = \mu[A], \gamma[X \times B] = \nu[B]$$

For any integrable measurable functions ϕ, ψ X, Y:

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

Deterministic Coupling

Definition 5 (Deterministic Coupling)

A coupling π of μ and ν is deterministic if there exists a measurable function $T : X \to Y$ such that $Y = T(X)$.

By changing of variables, we have

$$\int_Y \varphi(y) d\nu(y) = \int_X \varphi(T(x)) d\mu(x)$$

for all ν-integrable functions φ. T is usually called **transport map**, the fact T transports μ to ν can be denoted as $T_{\#}\mu = \nu$.

Lower Semi-Continuous

Definition 6 (Lower Semi-Continuous)

Let (X, τ) be a topological space and $f : X \to \bar{\mathbb{R}}$ be a function on X. f is lower semi-continuous at x if for every open set $(r, +\infty)$ with $f(x) \in (r, +\infty)$, there is an open set $U \subseteq X$, s.t. $x \in U \subseteq f^{-1}(r, +\infty)$.

In metric space (X, d), this definition can be expressed as

$$\text{for every sequence } x_n \to x, f(x) \leq \liminf_{x_n \to x} f(x_n)$$

Weak Convergence

Definition 7 (Weak Convergence)

A sequence $\{x_n\}$ in a Banach space X is said to be weakly converging to x, denoting as $x_n \rightharpoonup x$, if for every $f \in X'$ we have $\langle f, x_n \rangle \to \langle f, x \rangle$.
A sequence $f_n \in X'$ is said to be weakly-* converging to f, denoting as $f_n \overset{*}{\rightharpoonup} f$ if for every $x \in X$ we have $\langle f_n, x \rangle \to \langle f, x \rangle$

Theorem 8

If X is separable and f_n is a bounded sequence in X' then there exists a subsequence f_{n_k} weakly converging to some $f \in X'$.

Weierstrass theorem for existing of minimizer

Theorem 9 (Weierstrass)

If $f : X \to \mathbb{R} \cup \{+\infty\}$ is l.s.c and X compact, then there exists $\bar{x} \in X$ such that $f(\bar{x}) = \min\{f(x) : x \in X\}$.

Proof.

- Let $I = \inf\{f(x) : x \in X\}$.
- If $I = +\infty$, then $f = +\infty$, any point in X is minimizer.
- Assume $I < +\infty$. One can find a sequence $\{x_n : x_n \in X\}$ such that $f(x_n) \to I$.
  By compactness, we can assume $x_n \to \bar{x} \in X$.
  By l.s.c of f, we have $f(\bar{x}) \leq \liminf_n f(x_n) = I$. On the other hand, $f(\bar{x}) \geq I$ since I is the infimum. Hence $I = f(\bar{x}) \in \mathbb{R}$. Then $\bar{x}$ is the minimizer.

Outline

Existence of Transport Plans

Theorem 10

Let X, Y be compact metric spaces, $\mu \in \mathscr{P}(X), \nu \in \mathscr{P}(Y)$ and $c : X \times Y \to \mathbb{R}$ a continuous function. Then there exists a solution to Kantorovich's problem.

Existence of Transport Plans - Proof

Proof.

- Set of admissible transport plans $\Gamma(\mu, \nu)$ is compact.
  - Take a sequence $\gamma_n \in \Gamma(\mu, \nu)$. γ_n are bounded in the dual space of $C(X \times Y)$
  - By Theorem 8, there exists a subsequence $\gamma_{n_k} \rightharpoonup \gamma$.
  - We show $\gamma \in \Gamma(\mu, \nu)$. Fixing $f \in C(X)$ and using $\int f(x) d\gamma_{n_k} = \int f(x) d\mu$. Taking the limit we have $\int f(x) d\gamma = \int f(x) d\mu$, which shows $(\pi_x)_{\#}\gamma = \mu$. Similarly we have $(\pi_y)_{\#}\gamma = \nu$. Thus $\gamma \in \Gamma(\mu, \nu)$. So $\Gamma(\mu, \nu)$ is compact.
- The map $\gamma \mapsto K(\gamma) = \int c(x, y) d\gamma(x, y)$ is continuous. By Theorem 9, there exists a transport plan γ^* that minimizes Kantorovich's problem.

c-convexity

Definition 11 (c-convexity)

Let X, Y be sets, $c : X \times Y \to (-\infty, +\infty]$. A function $f : X - \mathbb{R} \cup \{+\infty\}$ is that to be c-convex if it is not identical to $+\infty$ and there exists $g : Y \to \mathbb{R} \cup \pm\infty$ such that

$$\forall x \in X \quad f(x) = \sup_{y \in Y}(g(y) - c(x, y))$$

The c-transform of f is the function f^c defined by

$$\forall y \in Y \quad f^c(y) = \inf_{x \in X}(f(x) + c(x, y))$$

THe function f and f^c are said to be c-conjugate.

Definition 12 (c-cyclical monotonicity)

Let $\Gamma \subset X \times X$. Γ is said to be *c-cyclically monotone* if

$$\sum_{i=1}^{n} c(x_i, y_i) \leq \sum_{i=1}^{n} c(x_{\sigma(i)}, y_i)$$

for any permutation $\sigma : \{1, ..., n\} \to \{1, ..., n\}$.

Intuition: suppose we have n source location $x_1, ..., x_n$ and n target location $y_1, ..., y_n$. Consider transport plan $T : x_i \to y_i$. Clearly if T is optimal, then above c-cyclical monotonicity condition holds.

c-subdifferential

Definition 13 (c-subdifferential)

Let $f : X \to \mathbb{R} \cup \{-\infty\}$ be a function. The *c-subdifferential* $\partial_c f$ of f is defined as the pair $(x, y) \in X \times Y$ for which

$$\partial_c f := \{(x, y) | f(x) \leq f(z) + c(z, y) - c(x, y),\ \forall z \in X\}$$

Moreover, we say c-subdifferential of f at point x is

$$\partial_c f(x) = \{y \in Y | (x, y) \in \partial_c f\}$$

Define c-superdifferential of f similarly.

Kantorovich Duality

$\mu = \sum_{i=1}^m \mu_i \delta_{x_i}$, $\nu = \sum_{i=1}^n \nu_i \delta_{y_i}$, $c_{ij} = c(x_i, y_j)$.

Minimize the total cost of moving μ to ν:

$$\sum_{i=1}^m \sum_{j=1}^n c_{ij}\gamma_{ij}$$

subject to $\gamma \in \Gamma(\mu, \nu)$, which becomes a linear programming problem:

$$\sum_i \gamma_{ij} = \nu_j, \quad \sum_j \gamma_{ij} = \mu_i, \quad \gamma_{ij} \geq 0$$

Dual problem: find $(\phi_i)_{i=1,\ldots,m}, (\psi_j)_{j=1,\ldots,n}$ that maximizes

$$\sum_{i=1}^m \phi_i \mu_i + \sum_{j=1}^n \psi_j \nu_j$$

subject to constraints

$$\phi_i + \psi_j \leq c_{ij}$$

Kantorovich Duality

Theorem 14 (Kantorovich Duality)

Let (X, μ) and (Y, ν) be Polish probability spaces. Let cost function $c : X \times Y \to \mathbb{R} \cup +\infty$ be l.s.c, such that

$$\forall (x, y) \in X \times Y, \quad a(x) + b(y) \leq c(x, y)$$

for some real-valued u.s.c functions $a \in L^1(\mu), b \in L^1(\nu)$.

(to be continued...)

Kantorovich Duality - Continued

There is duality

$$
\begin{aligned}
&\min_{\pi\in\Pi(\mu,\nu)} \int_{X\times Y} c(x,y)d\pi(x,y) \\
&\quad = \sup_{(\psi,\phi)\in C_b(X)\times C_b(Y),\phi-\psi\le c} \left(\int_Y \phi(y)d\nu(y) - \int_X \psi(x)d\mu(x)\right) \\
&\quad = \sup_{(\psi,\phi)\in L^1(X)\times L^1(Y),\phi-\psi\le c} \left(\int_Y \phi(y)d\nu(y) - \int_X \psi(x)d\mu(x)\right) \\
&\quad = \sup_{\psi\in L^1(\mu)} \left(\int_Y \psi^c(y)d\nu(y) - \int_X \psi(x)d\mu(x)\right) \\
&\quad = \sup_{\phi\in L^1(\nu)} \left(\int_Y \phi(y)d\nu(y) - \int_X \phi^c(x)d\mu(x)\right)
\end{aligned}
\tag{2}
$$

where ψ c-convex, ϕ c-concave.

Optimal Transport Plan vs Optimal Transport Map

Theorem 15

$\mu, \nu \in \mathscr{P}(X)$, X compact. $c(x, y) = h(x - y)$, h is strictly convex. If μ is absolute continuous w.r.t Lebesgue measure and $\mu(\partial(X)) = 0$, then there exists a unique optimal transport plan γ, of the form $(Id, T)_{\#}\mu$. Moreover, there exists a Kantorovich potential φ, s.t.

$$T(x) = x - (\nabla h)^{-1} \circ \nabla \varphi(x)$$

compactness can be relaxed

Outline

Polar Factorization

- Rearrangement: $\int_X f \circ h = \int_X f \circ \tilde{h}$ for any $f \in L^1(X)$
- measure-preserving map $s \in S(X)$: $s_{\#}\mu = \mu$

Theorem 16 (Brenier's Polar Factorization)

[1] Let (X, μ) be a bounded subset of $\mathbb{R}^n$, $h : X \to \mathbb{R}^n$ be an L^2 vector-valued mapping satisfying the non-degeneracy condition

$$\mu(h^{-1}(E)) = 0, \text{ for each Lebesgue negligible subset } E \subset \mathbb{R}^n$$

Then there exists a unique measure preserving map $s \in S(X)$ and a unique rearrangement $\nabla\varphi$ of h in the class of L^2 gradients of convex functions, such that

$$h = \nabla\varphi \circ s$$

s is the unique L^2 projection of h onto $S(X)$.

Brenier Theorem

Theorem 17 (Rockefeller)

Any c-cyclically monotone set Γ is contained in the graph of the c-superdifferential of a c-concave function. Conversely, the graph of the c-superdifferential of a c-concave function is c-cyclically monotone.

Theorem 18 (Brenier)

Let $\mu \in \mathscr{P}(\mathbb{R}^n)$ be such that $\int |x|^2 d\mu(x) < +\infty$. Then the following are equivalent:

(i) *for every $\nu \in \mathscr{P}(\mathbb{R}^n)$ with $\int |x|^2 d\nu(x) < +\infty$, there exists only one transport plan from μ to ν and this plan is induced by a map T*

(ii) *μ is regular (measure of graphs of differences of convex functions of $n-1$ variables is 0)*

If (i) or (ii) holds, the optimal map T is the gradient of a convex function.

Polar Factorization - Proof - Existence

Let $\nu = h_{\#}\mu$, $\tilde{s} \in S(X)$, then $\gamma_{\tilde{s}} := (\tilde{s}, h)_{\#}\mu$ is admissible transport plan. Then we have

$$\inf_{\tilde{s}\in S(X)} \int |\tilde{s} - h|^2 d\mu \geq \min_{\gamma \in Adm(\mu,\nu)} \int |x - y|^2 d\gamma(x, y)$$

Next, Let $\bar{\gamma}$ be the unique optimal transport plan, by Theorem 28

$$\bar{\gamma} = (Id, \nabla\varphi)_{\#}\mu = (\nabla\tilde{\varphi}, Id)_{\#}\nu$$

for some convex function $\varphi, \tilde{\varphi}$, s.t. $\nabla\varphi \circ \nabla\tilde{\varphi} = Id$ *a.e*
Let $s = \nabla\tilde{\varphi} \circ h$, then $s_{\#}\mu = \mu$, thus $s \in S(X)$. $h = \nabla\varphi \circ s$ proves the existence of the polar factorization. The identity

$$\begin{aligned} \int |x - y|^2 d\gamma_s(x, y) &= \int |s - h|^2 d\mu = \int |\nabla\tilde{\varphi} \circ h - h|^2 d\mu \\ &= \int |\nabla\tilde{\varphi} - Id|^2 d\nu = \min_{\gamma \in Adm(\mu,\nu)} \int |x - y|^2 d\gamma(x, y) \end{aligned} \tag{3}$$

Polar Factorization - Proof - Uniqueness

Assume $h = \nabla\bar{\varphi} \circ \bar{s}$ is another factorization. $\nabla\bar{\varphi}_{\#}\mu = (\nabla\bar{\varphi} \circ \bar{s})_{\#}\mu = \nu$. Thus $\nabla\bar{\varphi}$ is a transport map from μ to ν and is the gradient of a convex function. Then $\nabla\bar{\varphi}$ is the optimal map, $\nabla\bar{\varphi} = \nabla\varphi$

Polar Factorization on Riemannian Manifold

Theorem 19 (Polar Factorization on Riemannian Manifold)

Let (M, g) be a connected, compact Riemannian manifold, C^3-smooth and without boundary. For $c(x, y) = d^2(x, y)/2$ and two Borel probability measures $\mu \ll$ vol and ν arbitrary on M. Then there exists some potential $\psi : M \to \mathbb{R}$ satisfying $\psi = \psi^{cc}$, and the map $t(x) = exp_x(\nabla(\psi(x)))$ pushes μ forward to ν. t is unique up to μ-negligible set.

Gu's Theory

Gu et. al [2] established variational formulation for discrete version of Brenier's theory, which leads to an efficient algorithm that converges quadratically.

Definition 20

(PL convex function) Let $(p_1, ..., p_k) \in \mathbb{R}^n$ and $h = (h_1, ..., h_k) \in \mathbb{R}^k$, denote $u_h(x)$ the PL convex function defined as

$$u_h(x) = \max_i \{x \cdot p_i + h_i\}$$

Gradient ∇u_h can be obtained piecewisely. One can prove that $W_i(h) = \{x \in \mathbb{R}^n | \nabla u_h(x) = p_i\}$ is a closed convex polytope, maybe empty or unbounded.

Gu's Theory

Theorem 21

Let X be a compact convex domain in $\mathbb{R}^n$ and $\{p_1, ..., p_k\}$ a set of distinct points in $\mathbb{R}^n$. $\mu \in \mathscr{P}(X)$ be continuous, $\nu = \sum_{i=1}^k A_i \delta_{p_i}$, satisfy $\int_X \mu = \int_{\mathbb{R}^n} \nu$. The unique solution (up to a constant translation) b is obtained by minimizing a strictly convex functional

$$E(h) = \int_a^h \sum_{i=1}^k \int_{W_i(h)\cap X} \mu(x)dxdh_i - \sum_{i=1}^k h_i A_i$$

on the open convex set
$H = \{h \in \mathbb{R}^k | \sum_i h_i = 0,\ vol(W_i(h) \cap X) > 0,\ \forall i\}$ The optimal transport map is given by ∇u_b, which minimizes the quadratic cost $\int_X |x - T(x)|^2 \mu(x)dx$ among all transport maps $T : (X, \mu dx) \to (\mathbb{R}^n, \nu dx)$

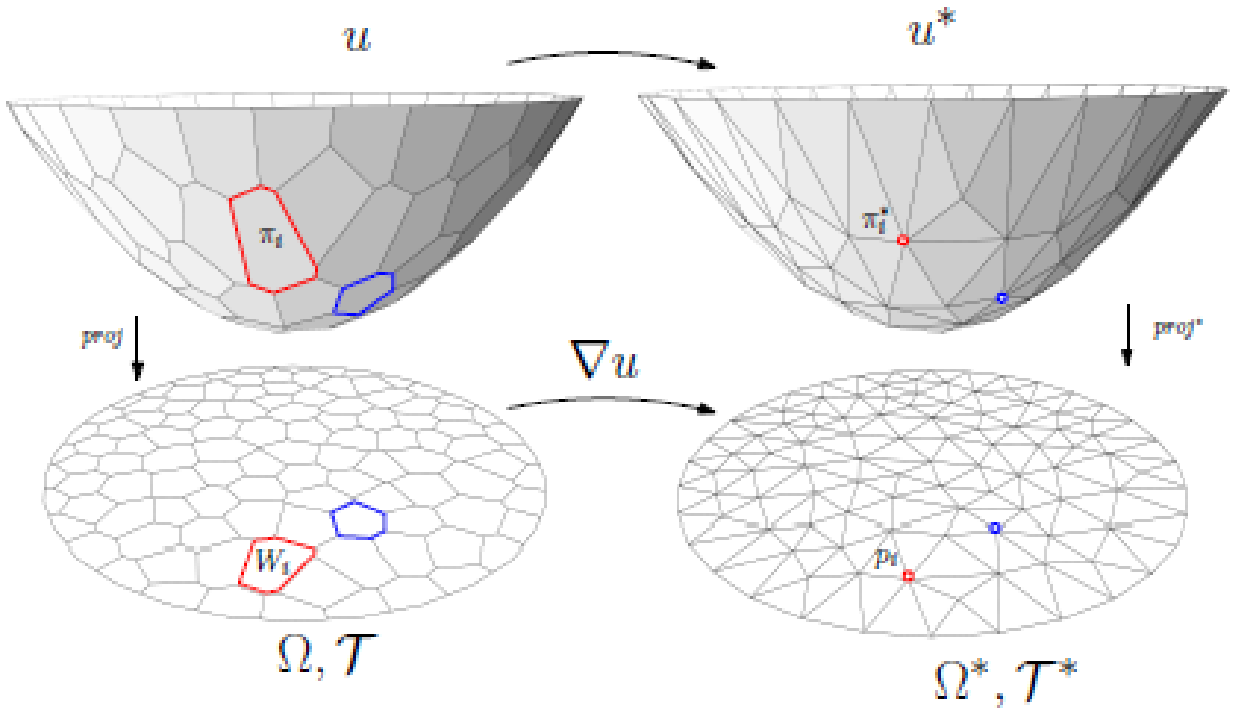


Figure: Discrete optimal transport map

Outline

The minimum cost of Monge or Kantorovich's problem defines a distance between probability measures.

Definition 22 (Wasserstein distance)

Assume μ and ν are two probability measures on $X \subset \mathbb{R}^n$ and cost $c(x, y) = |x - y|^p, 0 < p < +\infty$. Wasserstein distance is defined as

$$W_p(\mu, \nu) := \left(\inf_{\gamma} \int_{X \times X} |x - y|^p d\gamma(x, y), \gamma \in \Gamma(\mu, \nu) \right)^{1/p}$$

When μ is absolutely continuous w.r.t dx, we have

$$W_p(\mu, \nu) := \left(\inf_{T} \int_{X} |x - T(x)|^p d\mu(x), T_{\#}\mu = \nu \right)^{1/p}$$

Wasserstein Distance

$W_p(\mu, \nu)$ is a distance:

- First, $W_p(\mu, \nu) = W_p(\nu, \mu)$.
- Next assume $W_p(\mu, \nu) = 0$. From the definition, there exists a transport plan which is concentrated on the diagonal (x=y) in $X \times X$. so $\nu = Id_{\#}\mu = \mu$.
- triangle inequality holds (see paper)

Wasserstein Space

Definition 23 (Wasserstein Space)

$$\mathbb{W}_p(X) := \{\mu \in \mathscr{P}(X) | \int_X d(x_0, x)\mu(x)dx < +\infty\}$$

- $\mu = \sum_i \lambda_i \mu_i$ is not well defined: no associativity.
- $y = \sum_i \lambda_i x_i \Longleftrightarrow y = \arg_y \min \sum_i \lambda_i |y - x_i|^2$
- Wasserstein Barycenter

$$\min_{\mu \in \mathscr{P}(X)} \sum_i \lambda_i W_2^2(\mu, \mu_i)$$

unique if μ_i absolutely continuous

Wasserstein Barycentric Coordinates

Bonneel et. al in [3] use Wasserstein barycenter to perform histogram regression. Since optimal transport takes underline geometry into account, it's meaningful to use Wasserstein distance in applications such as image or shape processing, color or material modification.

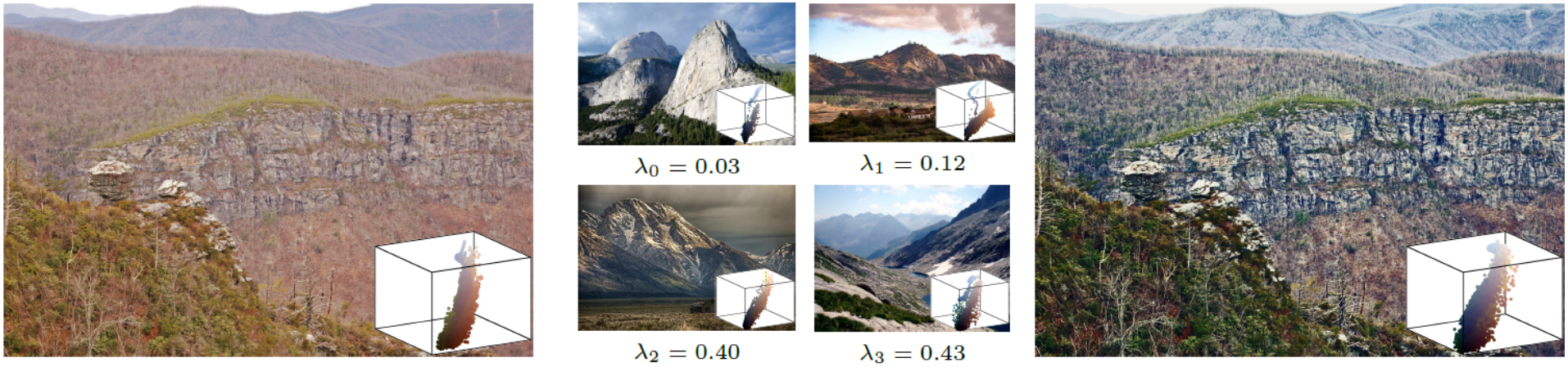


Figure: Wasserstein barycentric coordinates

Area-Preserving Map

Su et. al. [4] applied this approach to compute area-preserving map, or more generally, area-controllable map; used Wasserstein distance to perform shape classification [5].

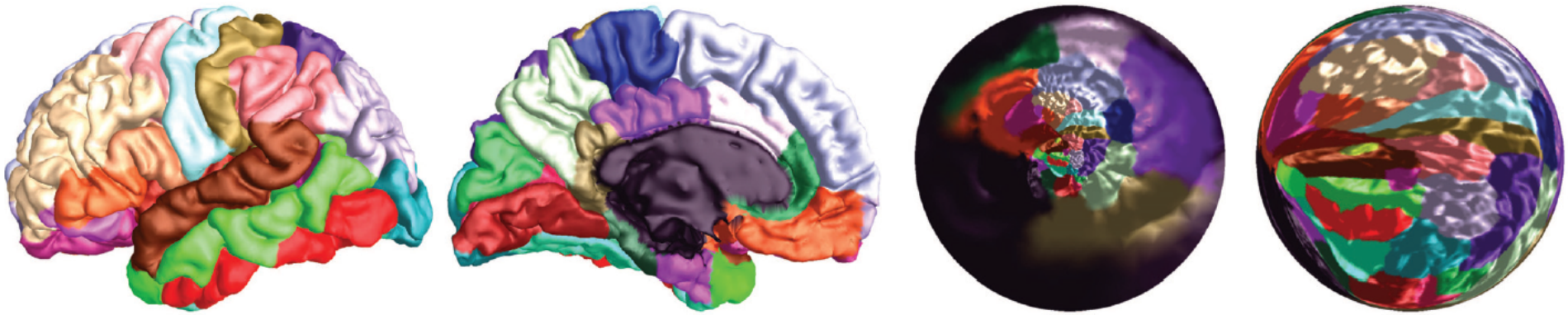
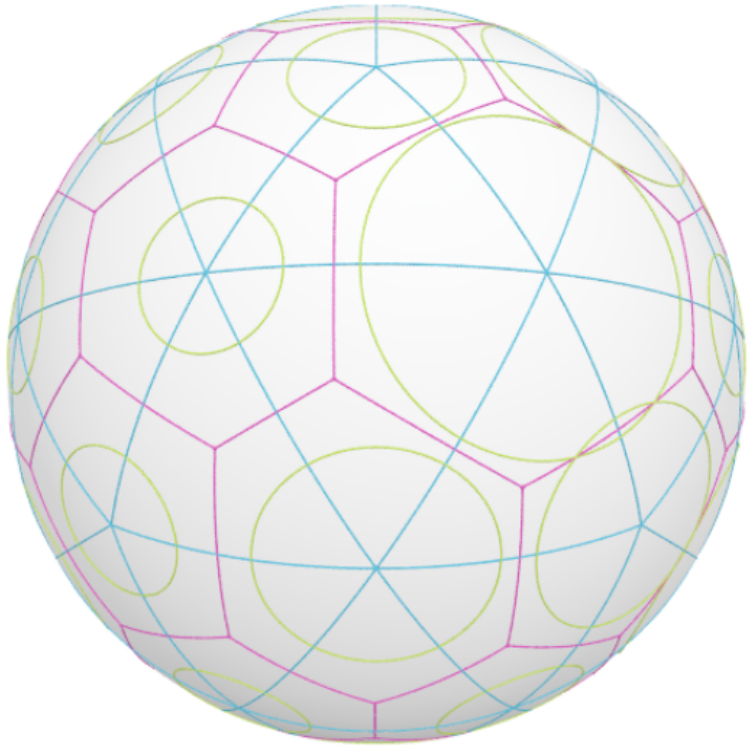


Figure: Area preserving map

Spherical Area-Preserving Map

Gu's discrete theory and algorithm works for arbitrary dimension of Euclidean space. The spherical variation of Gu's algorithm works as well as Euclidean case (Theorem 21), the variational formulation, quadratic convergence.

optimality not proved



Spherical Area-Preserving Map - Examples

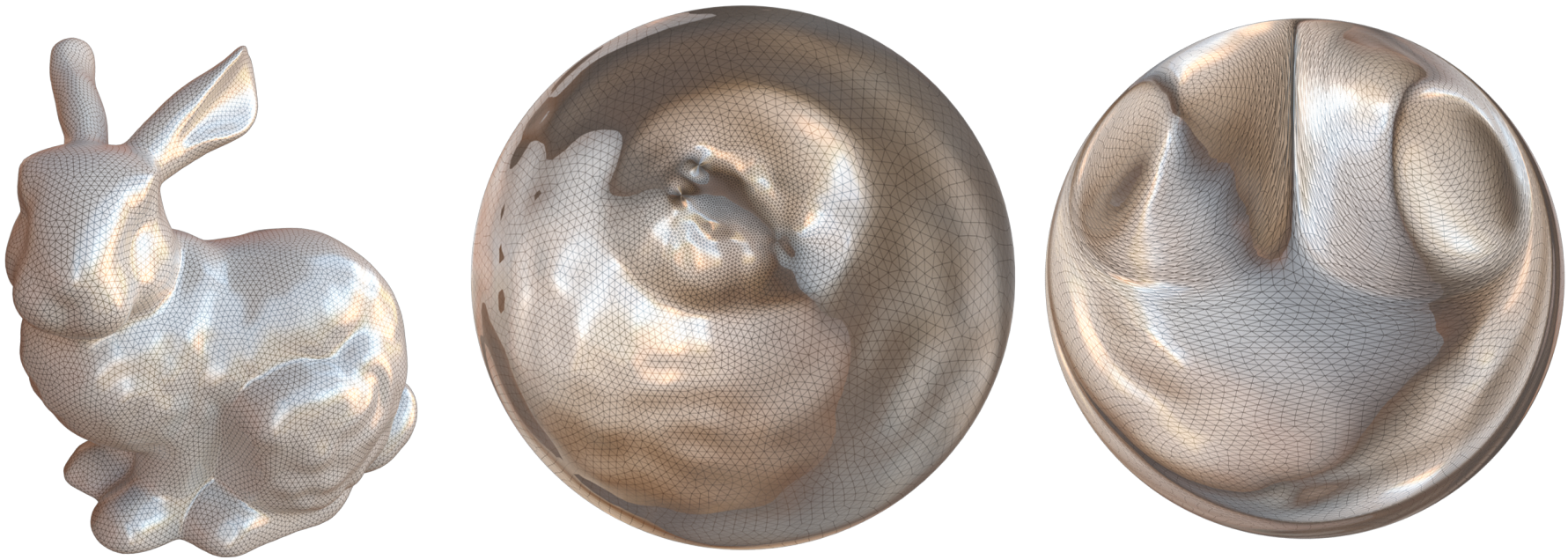


Figure: Spherical area-preserving parameterization. left: bunny model (70k faces); middle: initial conformal parameterization; right: our area-preserving parameterization.

Spherical Area-Preserving Map - Efficiency

Compare with [6] on spherical area-preserving map.

model	# vertices	# iter	time (s)	time [6] (s)	ratio
squirrel	2.5k	15	3	971	323.7
gargoyle	10k	29	37	1451	39.2
maxplanck	12.5k	20	19	2460	129.5
skull	20k	18	21	2132	101.5
bunny	35k	30	181	3093	17.1

Table: Performance statistics. 4th column is running time of our algorithm, 5th column is running time of Dominitz and Tannanbaum's algorithm.

Summary

- Monge-Kantorovich problem
- general theory of optimal transport
- Applications of optimal transport map and Wasserstein distance

Main References I

Brenier, Y.:
Polar factorization and monotone rearrangement of vector-valued functions.
Communications on Pure and Applied Mathematics **44**(4) (1991) 375–417

Gu, X., Luo, F., Sun, J., Yau, S.T.:
Variational principles for minkowski type problems, discrete optimal transport, and discrete monge-ampere equations.
Asian Journal of Mathematics(AJM) **2**(20) (Apr. 2016) 383–398

Bonneel, N., Peyré, G., Cuturi, M.:
Wasserstein barycentric coordinates: Histogram regression using optimal transport.
ACM Transactions on Graphics (Proceedings of SIGGRAPH 2016) **35**(4) (2016)

Main References II


Su, Z., Zeng, W., Shi, R., Wang, Y., Sun, J., Gu, X.:
Area preserving brain mapping.
In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2013) 2235–2242


Su, Z., Zeng, W., Wang, Y., Lu, Z.L., Gu, X.:
Shape classification using wasserstein distance for brain morphometry analysis.
In: Information Processing in Medical Imaging, Springer (2015) 411–423


Dominitz, A., Tannenbaum, A.:
Texture mapping via optimal mass transport.
Visualization and Computer Graphics, IEEE Transactions on **16**(3) (2010) 419–433