A tutorial on optimal transport
Part 1: theory, models, properties

Lénaïc Chizat
INRIA Paris (SIERRA team)

Imaging in Paris - Feb. 8th 2017
What is optimal transport?

**Setting:** Probability measures $P(\mathcal{X})$ on a metric space $(\mathcal{X}, d)$.

**Motive**

Build a metric on $P(\mathcal{X})$ consistent with the geometry of $(\mathcal{X}, d)$. 
What is optimal transport?

**Setting:** Probability measures $P(\mathcal{X})$ on a metric space $(\mathcal{X}, d)$.

**Motive**

Build a metric on $P(\mathcal{X})$ consistent with the geometry of $(\mathcal{X}, d)$.

$$\mu = \delta_{x_1}, \quad \nu = \delta_{y_1}$$

$$W(\mu, \nu) = \ldots$$

$$d(x_1, y_1)$$
What is optimal transport?

**Setting:** Probability measures \( P(\mathcal{X}) \) on a metric space \((\mathcal{X}, d)\).

**Motive**

Build a metric on \( P(\mathcal{X}) \) consistent with the geometry of \((\mathcal{X}, d)\).

\[
\begin{align*}
\mu &= \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}, \quad \nu = \frac{1}{N} \sum_{j=1}^{N} \delta_{y_j} \\
W(\mu, \nu) &= \ldots \\
&= \frac{1}{N^2} \sum_{ij} d(x_i, y_j)
\end{align*}
\]
What is optimal transport?

Setting: Probability measures $P(\mathcal{X})$ on a metric space $(\mathcal{X}, d)$.

Motive
Build a metric on $P(\mathcal{X})$ consistent with the geometry of $(\mathcal{X}, d)$.

$$
\mu = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}, \quad \nu = \frac{1}{N} \sum_{j=1}^{N} \delta_{y_j}
$$

$$
W(\mu, \nu) = \ldots
$$

$$
\min_{\sigma \in \mathfrak{S}_N} \frac{1}{N} \sum_{i} d(x_i, y_{\sigma(i)})
$$
What is optimal transport?

**Setting:** Probability measures $P(\mathcal{X})$ on a metric space $(\mathcal{X}, d)$.

**Motive**
Build a metric on $P(\mathcal{X})$ consistent with the geometry of $(\mathcal{X}, d)$.

$$\mu \in P(\mathcal{X}), \quad \nu \in P(\mathcal{Y})$$

$$W(\mu, \nu) = \ldots$$
Origin and ramifications

Monge Problem (1781)
Move dirt from one configuration to another with least effort

\[ \mu \text{ transport} \rightarrow \nu \]
Origin and ramifications

Monge Problem (1781)

Move dirt from one configuration to another with least effort

\[ \mu \quad \text{transport} \quad \nu \]

Strong modelization power:

Replace “dirt” by:

- probability distribution, empirical distribution
- weighted undistinguishable particles
- density of a gas, a species, a crowd, cells.

Early universe (Brenier et al. ’08)

Color histograms (Delon et al.)

Crowd motion (Roudneff et al., 12’)

Point clouds
Aim of the tutorial

Convey that optimal transport ...

is a rich theory, useful as a theoretical and practical tool;

In part 1: theory

- essentials
- selection of properties and variants;

In part 2: practice

- numerical solvers, entropic regularization
- applications to imaging and machine learning
1. Theoretical facts
   Variational problem
   Special cases
   The metric side

2. A glimpse of applications
   Histogram & shapes processing
   Gradient flows
   Statistical learning

3. Differential properties
   Perturbations
   Wasserstein gradient

4. Unbalanced optimal transport
   Partial OT
   Wasserstein Fisher-Rao
1 Theoretical facts
   Variational problem
   Special cases
   The metric side

2 A glimpse of applications
   Histogram & shapes processing
   Gradient flows
   Statistical learning

3 Differential properties
   Perturbations
   Wasserstein gradient

4 Unbalanced optimal transport
   Partial OT
   Wasserstein Fisher-Rao
Optimal transport

**Ingredients**

- Two (complete, separable) metric spaces \(\mathcal{X}\) and \(\mathcal{Y}\)
- Cost function \(c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\infty\}\) (lower bounded, lsc)
- Two probability measures \(\mu \in P(\mathcal{X})\) and \(\nu \in P(\mathcal{Y})\)

**Definition (Optimal transport problem)**

\[
C(\mu, \nu) := \min_{\gamma \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})} \left\{ \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\gamma(x, y) : \pi_x^#\gamma = \mu, \pi_y^#\gamma = \nu \right\}
\]

**Probabilistic**

\[
\min_{(X, Y)} \{ \mathbb{E} [c(X, Y)] : X \sim \mu \text{ and } Y \sim \nu \}
\]
Couplings

**Definition (Set of couplings)**

Positive measures on $\mathcal{X} \times \mathcal{Y}$ with specified marginals:

$$\Pi(\mu, \nu) := \{ \gamma \in M_+(\mathcal{X} \times \mathcal{Y}) : \pi_x^\# \gamma = \mu, \pi_y^\# \gamma = \nu \}$$

**Product coupling**

$$\gamma = \mu \otimes \nu$$

**Deterministic coupling**

$$\gamma = (\text{Id} \times T)_\# \mu$$

*Generalizes:* permutations, discrete matchings

*Properties:* convex, weakly compact
**Definition (Set of couplings)**

Positive measures on $\mathcal{X} \times \mathcal{Y}$ with specified marginals:

$$\Pi(\mu, \nu) := \left\{ \gamma \in M_+(\mathcal{X} \times \mathcal{Y}) : \pi_x^\# \gamma = \mu, \pi_y^\# \gamma = \nu \right\}$$

**Product coupling**

$$\gamma = \mu \otimes \nu$$

**Generalizes:** permutations, discrete matchings

**Properties:** convex, weakly compact

**Cycle-free coupling**
Duality

**Theorem (Kantorovich duality)**

\[
\min_{\gamma \in \mathcal{M}_+(X \times Y)} \left\{ \int_{X \times Y} c(x, y) d\gamma(x, y) : \pi^x_#\gamma = \mu, \pi^y_#\gamma = \nu \right\} \quad \text{(P)}
\]

\[
= \max_{\phi \in L^1(\mu), \psi \in L^1(\nu)} \left\{ \int_X \phi(x) d\mu(x) + \int_Y \psi(y) d\nu(y) : \phi(x) + \psi(y) \leq c(x, y) \right\} \quad \text{(D)}
\]

**Interpretation:** (P) centralized planification, (D) externalized
Duality

**Theorem (Kantorovich duality)**

\[
\begin{align*}
\min_{\gamma \in M_+(X \times Y)} & \left\{ \int_{X \times Y} c(x, y) d\gamma(x, y) : \pi_x^\# \gamma = \mu, \pi_y^\# \gamma = \nu \right\} \\
= & \max_{\phi \in L^1(\mu)} \left\{ \int_X \phi(x) d\mu(x) + \int_Y \psi(y) d\nu(y) : \phi(x) + \psi(y) \leq c(x, y) \right\}
\end{align*}
\]

**Interpretation:** (P) centralized planification, (D) externalized

At optimality

- \( \phi(x) + \psi(y) = c(x, y) \) for \( \gamma \) almost every \((x, y)\)
- \( \gamma \) is concentrated on a \( c \)-cyclically monotone set
Tools from convex analysis

**Definition (Cyclical monotonicity)**

$\Gamma \subset X \times Y$ is $c$-cyclical monotone iff for all $(x_i, y_i)_{i=1}^n \in \Gamma^n$

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

for all permutation $\sigma \in S_n$. 

---

![Diagram](image-url)
Tools from convex analysis

Definition (Cyclical monotonicity)

\[ \Gamma \subset \mathcal{X} \times \mathcal{Y} \text{ is } c\text{-cyclical monotone iff for all } (x_i, y_i)^n_{i=1} \in \Gamma^n \]

\[ \sum_{i=1}^{n} c(x_i, y_i) \leq \sum_{i=1}^{n} c(x_i, y_{\sigma(i)}) \text{ for all permutation } \sigma \in \mathcal{S}_n. \]
Tools from convex analysis

**Definition (c-conjugacy)**

For $\mathcal{X} = \mathcal{Y}$ and $c : \mathcal{X}^2 \to \mathbb{R}$ symmetric:

$$\phi^c(y) := \inf_{x \in \mathcal{X}} c(x, y) - \phi(x)$$

A function $\phi$ is $c$-concave iff there exists $\psi$ such that $\phi = \psi^c$. 

\[\begin{array}{c}
\mathbb{R} \\
\downarrow \\
\phi(x) \\
\mathcal{X} \\
\rightarrow \\
\mathbb{R} \\
\uparrow \\
c(\cdot, y)
\end{array}\]
Tools from convex analysis

Definition (c-conjugacy)

For $X = Y$ and $c : X^2 \to \mathbb{R}$ symmetric:

$$\phi^c(y) := \inf_{x \in X} c(x, y) - \phi(x)$$

A function $\phi$ is $c$-concave iff there exists $\psi$ such that $\phi = \psi^c$.

- on $\mathbb{R}^n$, for $c(x, y) = x \cdot y$: $\psi$ $c$-concave $\Leftrightarrow$ $\psi$ concave;
- for all $\phi$, $\phi^{ccc} = \phi^c$;
- consequence:

$$C(\mu, \nu) = \max_{\phi \text{ c-concave}} \left\{ \int_X \phi(x) d\mu(x) + \int_Y \phi^c(y) d\nu(y) \right\} \quad \text{(D)}$$
Special cases

- real line
- distance cost
- quadratic cost
The real line

Theorem

If \((\mu, \nu) \in P(\mathbb{R})^2\) and \(c(x, y) = h(y - x)\) with \(h\) strictly convex

- unique optimal coupling \(\gamma^*\): the monotone rearrangement
- denoting \(F^{-1}\) the quantile functions:

\[
C(\mu, \nu) = \int_0^1 h(F^{-1}_{\mu}(s) - F^{-1}_{\nu}(s)) ds
\]

Proof. Here, \(c\)-cyclically monotone \(\iff\) increasing graph. \(\Box\)
If $c(x, y) = d(x, y)$ with $d$ distance

- $\phi$ $c$-concave $\iff$ $\phi$ 1-Lipschitz
- $\phi^c(y) = \inf_x d(x, y) - \phi(x) = -\phi(y)$
- consequence:

$$C(\mu, \nu) = \max_{\phi \text{ 1-Lipschitz}} \left\{ \int_X \phi(x) d(\mu - \nu)(x) \right\} := \|\mu - \nu\|_K \tag{D}$$
Quadratic cost

Context & reformulation

• \((\mu, \nu) \in P(\mathbb{R}^n)^2\) with finite moments of order 2
• cost \(c(x, y) := \frac{1}{2}|y - x|^2\)
• note that \(c(x, y) = (|x|^2 + |y|^2)/2 - x \cdot y\), thus solve:

\[
\max_{\gamma \in M_+(X \times Y)} \left\{ \int_{X \times Y} (x \cdot y) d\gamma(x, y) : \gamma \in \Pi(\mu, \nu) \right\} \tag{P}
\]

Theorem (Brenier)

(i) At optimality, \(\text{supp} \gamma \subset \partial \phi\), where \(\phi : \mathbb{R}^n \to \mathbb{R}\) convex.
(ii) If \(\mu\) has a density, \(T = \nabla \phi\) is the unique optimal map.

Proof. (i) \(\phi(x) + \phi^*(y) = x \cdot y\), \(\gamma\)-a.e (ii) \(\nabla \phi\) defined \(\mathcal{L}\)-a.e.
Transport of covariance

Case of a quadratic dual potential $\phi$

**Theorem (Affine transport map)**

Let $c(x, y) = \frac{1}{2}|y - x|^2$ on $\mathbb{R}^n$ and let $A, B \in S^n_+$. It holds

$$\min_{\text{cov}(\mu)=A, \text{cov}(\nu)=B} C(\mu, \nu) = d_b(A, B)^2$$

where $d_b$ is the Bures (geodesic) metric on $S^n_+$. 

- $d_b(A, B)^2 = \text{tr} A + \text{tr} B - 2 \text{tr}(A^{\frac{1}{2}}BA^{\frac{1}{2}})^{\frac{1}{2}}$
- Transport map $T = A^{-1} \# B$ (\cdot\#\cdot geometric mean).
- see, e.g. (Bhatia et al. ’17)
Wasserstein distance

**Theorem**

Let $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a metric. The function

$$W_2(\mu, \nu) := \left\{ \min_{\gamma \in M_+((\mathcal{X}^2), d)} \int_{\mathcal{X}^2} d(x, y)^2 d\gamma(x, y) : \gamma \in \Pi(\mu, \nu) \right\}^{\frac{1}{2}}$$

defines a metric on $P(\mathcal{X})$.

- $W_2$ metrizes weak convergence + 2-nd order moments;
- if $(\mathcal{X}, d)$ is a geodesic space, so is $(P(\mathcal{X}), W_2)$.

Figure: A constant speed geodesic for $W_2$ on $P(\mathbb{R}^2)$
Consider $\mu, \nu$ probability measures on $\mathbb{R}^n$.

Variational characterization of geodesics (Benamou-Brenier)

$$W_2^2(\mu, \nu) = \min_{(\rho_t, \nu_t)_{t \in [0,1]}} \int_0^1 \left( \int_{\mathbb{R}^n} |\nu_t(x)|^2 d\rho_t(x) \right) dt$$

s.t. $\partial_t \rho_t = -\text{div}(\rho_t \nu_t)$

and $(\rho_0, \rho_1) = (\mu, \nu)$

Consequences

- minimizers are geodesics;
- convex in variables $(\rho, \nu \rho)$;
- $W_2$ is similar to a Riemannian metric.
Properties of OT

- rich duality, with concepts from convex analysis
- real line, distance cost, quadratic cost

Properties of the distance $W_2$ on $\mathbb{R}^n$

- optimal plans supported on $\partial \phi$ with $\phi : \mathbb{R}^n \to \mathbb{R}$ convex;
- the space $(\mathcal{P}(\mathbb{R}^n), W_2)$ is a complete geodesic space;
- some explicit cases (real line, linear maps).
Optimal transport
Lénaïc Chizat

Introduction

Theory
Variational problem
Special cases
The metric side

Applications
Histograms
Gradient flows
Statistical learning

Differentiability
Perturbations
Wasserstein gradient

Unbalanced
Partial OT
Wasserstein Fisher-Rao

Conclusion

Outline

1 Theoretical facts
Variational problem
Special cases
The metric side

2 A glimpse of applications
Histogram & shapes processing
Gradient flows
Statistical learning

3 Differential properties
Perturbations
Wasserstein gradient

4 Unbalanced optimal transport
Partial OT
Wasserstein Fisher-Rao
Histogram & shapes processing

Color transfer

\[ \text{color} + \text{target} = \text{OT or unbalanced OT} \]

Barycenters

and much more

- PCA (Seguy, Cuturi’15)
- regression (Bonneel et al’16)

(Benamou et al’15)
Objective: characterize certain evolution EDP as gradient flows of some functional $F : P(\mathbb{R}^n) \to \mathbb{R}$ in the Wasserstein space:

$$\partial_t \mu_t + \text{div}(\mu_t \nu_t) = 0 \quad \text{with} \quad \nu_t = \nabla F'(\mu_t).$$

Interest

- theoretical: existence, uniqueness, convergence...
- numerical: intrinsic mass conservation and positivity

Crowd motions (Roudneff-Chupin et al.'14)
Statistical learning

- $W_p$ loss for regression (Frogner et al.’15):
  Learn predictor $f_\theta : X \to Y := P(\{1, \ldots, k\})$

$$\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{(X,Y) \sim \mu} \left[ W_2^2(f_\theta(X), Y) \right].$$

- $W_p$ loss for generative models:
  Given $\mu \in P(X)$, $\nu \in P(Y)$, learn map $f_\theta : X \to Y$

$$\min_{\theta \in \mathbb{R}^d} W_2^2((f_\theta)^\# \mu, \nu)$$

- Barycenters for multiscale learning (Srivastava et al.’17), transfer learning (Courty et al.’17), convergence of Langevin MC (Dalalyan’17)...
And much more...

- **applied analysis**: incompressible flows (Euler), sticky particules
- **metric geometry**: Ricci curvature, perimetric inequalities
- **mathematical physics**: density functional theory, Schröedinger bridge
- **mathematical economy**: matching problems, principal agent, MFG, finance (martingale transport)...
And much more...

- **applied analysis**: incompressible flows (Euler), sticky particles
- **metric geometry**: Ricci curvature, perimetric inequalities
- **mathematical physics**: density functional theory, Schrödinger bridge
- **mathematical economy**: matching problems, principal agent, MFG, finance (martingale transport)

Recurring needs:

- differential properties
- unbalanced OT
Optimal transport

Lénaïc Chizat

Introduction

Theory
  Variational problem
  Special cases
  The metric side

Applications
  Histograms
  Gradient flows
  Statistical learning

Differentiability
  Perturbations
  Wasserstein gradient

Unbalanced
  Partial OT
  Wasserstein Fisher-Rao

Conclusion

Outline

1. Theoretical facts
   Variational problem
   Special cases
   The metric side

2. A glimpse of applications
   Histogram & shapes processing
   Gradient flows
   Statistical learning

3. Differential properties
   Perturbations
   Wasserstein gradient

4. Unbalanced optimal transport
   Partial OT
   Wasserstein Fisher-Rao
Reminder

Optimal transport between \( \mu, \nu \in P(\mathbb{R}^n) \) with cost \( c \):

\[
C(\mu, \nu) = \sup_{(\varphi, \psi) \text{ admissible}} \int_{\mathbb{R}^n} \varphi \, d\mu + \int_{\mathbb{R}^n} \psi \, d\nu
\]
Vertical perturbations

Reminder

Optimal transport between $\mu, \nu \in P(\mathbb{R}^n)$ with cost $c$:

$$C(\mu, \nu) = \sup_{(\varphi, \psi) \text{ admissible}} \int_{\mathbb{R}^n} \varphi \, d\mu + \int_{\mathbb{R}^n} \psi \, d\nu$$

Perturbed marginal: $\mu + \epsilon \delta$

Vertical perturbations

Let $\delta$ a signed measure with $\int \delta = 0$. If optimal $\varphi$ unique, then:

$$\frac{d}{d \epsilon} C(\mu + \epsilon \delta, \nu) |_{\epsilon = 0} = \int_{\mathbb{R}^n} \varphi \, d\delta$$

If $\varphi$ nonunique (up to a constant) $\Rightarrow$ subdifferential.
Reminder

Optimal transport between $\mu, \nu \in P(\mathbb{R}^n)$ with cost $c$:

$$C(\mu, \nu) = \sup_{(\varphi, \psi) \text{ admissible}} \int_{\mathbb{R}^n} \varphi \, d\mu + \int_{\mathbb{R}^n} \psi \, d\nu$$

Vertical perturbation

Let $\delta$ a signed measure with $\int \delta = 0$. If optimal $\varphi$ unique,

$$\frac{d}{d\epsilon} C(\mu + \epsilon \delta, \nu)\bigg|_{\epsilon=0} = \int_{\mathbb{R}^n} \varphi \, d\delta$$

If $\varphi$ nonunique (up to a constant) $\Rightarrow$ subdifferential.
Horizontal perturbations

Reminder

Optimal transport between $\mu, \nu \in P(\mathbb{R}^n)$ with cost $c$: 

$$C(\mu, \nu) = \inf_{\gamma \text{ admissible}} \int_{(\mathbb{R}^n)^2} c(x, y) \, d\gamma(x, y)$$
Horizontal perturbations

Reminder

Optimal transport between \( \mu, \nu \in P(\mathbb{R}^n) \) with cost \( c \):

\[
C(\mu, \nu) = \inf_{\gamma \text{ admissible}} \int_{(\mathbb{R}^n)^2} c(x, y) \, d\gamma(x, y)
\]

Perturbed cost: \( c(x + \epsilon v(x), y) \approx c(x, y) + \epsilon \nabla_x c(x, y) \cdot v(x) \)
Horizontal perturbations

Reminder

Optimal transport between $\mu, \nu \in P(\mathbb{R}^n)$ with cost $c$:

$$C(\mu, \nu) = \inf_{\gamma \text{ admissible}} \int_{(\mathbb{R}^n)^2} c(x, y) \, d\gamma(x, y)$$

Perturbed cost: $c(x + \epsilon v(x), y) \approx c(x, y) + \epsilon \nabla_x c(x, y) \cdot v(x)$

Horizontal perturbation

Let $v : \mathbb{R}^n \to \mathbb{R}^n$ a velocity field. If optimal $\gamma$ unique,

$$\frac{d}{d\epsilon} C((\text{id} + \epsilon v) \# \mu, \nu)|_{\epsilon=0} = \int_{(\mathbb{R}^n)^2} \nabla_x c(x, y) \cdot v(x) \, d\gamma(x).$$

Corresponds to the vertical perturbation $\partial_\epsilon \mu = -\text{div}(v \mu)$
Special case of $\mathcal{W}_2$

**Setting:** quadratic cost on $\mathbb{R}^n$, $\nu: \mathbb{R}^n \to \mathbb{R}^n$ velocity field.

**Differentiability of $\mathcal{W}_2$**

If unique optimal transport plan $\gamma$, then

$$
\frac{d}{d\epsilon} \mathcal{W}_2^2((\text{id} + \epsilon \nu)\# \mu, \nu)|_{\epsilon=0} = \int_{(\mathbb{R}^n)^2} 2(y - x) \cdot \nu(x) d\gamma(x, y)
$$

Next talk: regularized $\mathcal{W}_2$, always differentiable.
Special case of $W_2$

Setting: quadratic cost on $\mathbb{R}^n$, $\nu : \mathbb{R}^n \to \mathbb{R}^n$ velocity field.

Differentiability of $W_2$

If unique optimal transport plan $\gamma$, then

$$\frac{d}{d\epsilon} W_2^2((\text{id} + \epsilon \nu)\#\mu, \nu)|_{\epsilon = 0} = \int_{(\mathbb{R}^n)^2} 2(y - x) \cdot \nu(x)d\gamma(x, y)$$

Next talk: regularized $W_2$, always differentiable.
Euclidean Gradient

**Goal:** defining the gradient though metric quantities only.
Euclidean Gradient

**Goal:** defining the gradient though metric quantities only.

### Proximal operator

Let $F : \mathbb{R}^n \to \mathbb{R}$ a (semiconvex) function. The proximal operator assigns to each $x \in \mathbb{R}^n$

$$x^\tau := \arg \min_{y \in \mathbb{R}^n} \left( \frac{|x - y|^2}{2\tau} + F(y) \right)$$

### Definition (Euclidean gradient)

$$\text{grad} F(x) := \lim_{\tau \to 0} \frac{(x - x^\tau)}{\tau} \in \mathbb{R}^n$$
Wasserstein Gradient

**Proximal map**: let $F : P(\mathbb{R}^n) \rightarrow \mathbb{R}$ a functional, $\mu \in P^{ac}(\mathbb{R}^n)$.

$$\mu^\tau = \arg \min_{\nu \in P(\mathbb{R}^n)} \frac{W_2^2(\mu, \nu)}{2\tau} + F(\nu)$$
Wasserstein Gradient

**Proximal map:** let $F : P(\mathbb{R}^n) \to \mathbb{R}$ a functional, $\mu \in P^{ac}(\mathbb{R}^n)$.

\[
\mu^\tau = \arg \min_{\nu \in P(\mathbb{R}^n)} \frac{W_2^2(\mu, \nu)}{2\tau} + F(\nu)
\]

**OT with quadratic cost:** with $\varphi$ dual variable w.r.t. $\mu^\tau$ it holds

\[
\mu = T \# \mu^\tau \quad \text{where} \quad T(x) = x - \nabla \varphi(x).
\]
Wasserstein Gradient

Proximal map: let $F : P(\mathbb{R}^n) \rightarrow \mathbb{R}$ a functional, $\mu \in P^{ac}(\mathbb{R}^n)$.

$$\mu^\tau = \arg \min_{\nu \in P(\mathbb{R}^n)} \frac{W_2^2(\mu, \nu)}{2\tau} + F(\nu)$$

OT with quadratic cost: with $\varphi$ dual variable w.r.t. $\mu^\tau$ it holds

$$\mu = T\#\mu^\tau$$

where $T(x) = x - \nabla \varphi(x)$.

First order optimality condition (vertical perturbation):

$$\frac{\varphi}{\tau} + F'(\mu^\tau) = cst \Rightarrow \frac{id - T}{\tau} + \nabla F'(\mu^\tau) = 0$$
**Wasserstein Gradient**

**Proximal map:** let $F : P(\mathbb{R}^n) \to \mathbb{R}$ a functional, $\mu \in P^{ac}(\mathbb{R}^n)$.

$$\mu^\tau = \arg \min_{\nu \in P(\mathbb{R}^n)} \frac{W_2^2(\mu, \nu)}{2\tau} + F(\nu)$$

OT with quadratic cost: with $\varphi$ dual variable w.r.t. $\mu^\tau$ it holds

$$\mu = T \# \mu^\tau \quad \text{where} \quad T(x) = x - \nabla \varphi(x).$$

First order optimality condition (vertical perturbation):

$$\frac{\varphi}{\tau} + F'(\mu^\tau) = \text{cst} \Rightarrow \frac{\text{id} - T}{\tau} + \nabla F'(\mu^\tau) = 0$$

**Wasserstein gradient (limit $\tau \to 0$)**

$$\text{grad} \ F(\mu) = \text{div}(\nabla F'(\mu) \mu)$$
Wasserstein Gradient

Proximal map: let $F : P(\mathbb{R}^n) \to \mathbb{R}$ a functional, $\mu \in P^{ac}(\mathbb{R}^n)$.

$$\mu^\tau = \arg \min_{\nu \in P(\mathbb{R}^n)} \frac{W_2^2(\mu, \nu)}{2\tau} + F(\nu)$$

OT with quadratic cost: with $\phi$ dual variable w.r.t. $\mu^\tau$ it holds

$$\mu = T\#\mu^\tau \quad \text{where} \quad T(x) = x - \nabla \phi(x).$$

First order optimality condition (vertical perturbation):

$$\frac{\phi}{\tau} + F'(\mu^\tau) = \text{cst} \Rightarrow \frac{\text{id} - T}{\tau} + \nabla F'(\mu^\tau) = 0$$

Wasserstein gradient (limit $\tau \to 0$)

$$\text{grad} \ F(\mu) = \text{div}(\nabla F'(\mu)\mu)$$

Fundamental exemple: with $F(\mu) = \int \mu \log(d\mu/d\mathcal{L})$, one has

$$\text{grad} \ F(\mu) = \Delta \mu.$$
1 Theoretical facts
   Variational problem
   Special cases
   The metric side

2 A glimpse of applications
   Histogram & shapes processing
   Gradient flows
   Statistical learning

3 Differential properties
   Perturbations
   Wasserstein gradient

4 Unbalanced optimal transport
   Partial OT
   Wasserstein Fisher-Rao
Unbalanced OT

OT comes with an intrinsic constraint:

$$\mu(\mathcal{X}) = \nu(\mathcal{Y})$$

What if $\mu(\mathcal{X}) \neq \nu(\mathcal{Y})$?
Unbalanced OT

OT comes with an intrinsic constraint:

$$\mu(X) = \nu(Y)$$

What if $$\mu(X) \neq \nu(Y)$$?

Unbalanced OT:

- often comes up in applications
- normalization is generally a poor choice
- are there approaches that stand out?
Unbalanced OT

OT comes with an intrinsic constraint:

\[ \mu(\mathcal{X}) = \nu(\mathcal{Y}) \]

What if \( \mu(\mathcal{X}) \neq \nu(\mathcal{Y}) \)?

**Unbalanced OT:**

- often comes up in applications
- normalization is generally a poor choice
- are there approaches that stand out?

**Strategy**

- preserve key properties of optimal transport
- combine two geometries:
  - *horizontal* (transport) and *vertical* (linear)
Optimal partial transport

Setting: \( \mu \in M_+(\mathcal{X}) \) and \( \nu \in M_+(\mathcal{Y}) \) nonnegative measures.

Variational problem

Choose \( 0 < m \leq \min\{\mu(\mathbb{R}^n), \nu(\mathbb{R}^n)\} \) and solve

\[
\min_{\gamma} \int c(x, y) d\gamma(x, y)
\]

subject to

\[
\pi^x_\# \gamma \leq \mu
\]
\[
\pi^y_\# \gamma \leq \nu
\]
\[
\gamma(\mathbb{R}^n \times \mathbb{R}^n) = m
\]

- simple modification of the OT problem
- “equivalent” formulations: dynamic, entropy-transport
- alternatively, add a sink/source reachable at a certain cost
Wasserstein Fisher-Rao

Setting: \( \mu \in M_+(\mathcal{X}) \) and \( \nu \in M_+(\mathcal{Y}) \) nonnegative measures.

**Definition**

The natural generalization of \( W_2 \) to this setting is

\[
\hat{W}_2^2(\mu, \nu) = \min_{\gamma \in M_+(\mathcal{X} \times \mathcal{Y})} KL(\pi^x_\# \gamma | \mu) + KL(\pi^y_\# \gamma | \nu) + \int c_\ell(x, y) d\gamma(x, y)
\]

where \( c_\ell(x, y) = -\log \cos^2(\min\{|y - x|, \pi/2\}) \).
Setting: $\mu \in M_+(\mathcal{X})$ and $\nu \in M_+(\mathcal{Y})$ nonnegative measures.

Definition

The natural generalization of $W_2$ to this setting is

$$\widehat{W}_2^2(\mu, \nu) = \min_{\gamma \in M_+(\mathcal{X} \times \mathcal{Y})} KL(\pi^x_\# \gamma | \mu) + KL(\pi^y_\# \gamma | \nu) + \int c_\ell(x, y) d\gamma(x, y)$$

where $c_\ell(x, y) = -\log \cos^2(\min\{|y-x|, \pi/2\})$.

Main properties

- geodesic space, Riemannian-like structure
- growth and displacement intertwined
- various explicit formulations: lifted problem, dynamic problem with velocity and rate of growth...

References: (Liero et al’15), (Monsaingeon et al’15), (Chizat et al’15), my PhD thesis.
End of part 1

In part 1: theory
- essentials
- selection of properties and variants;

In part 2: practice
- numerical solvers, entropic regularization
- applications to imaging and machine learning

Reference textbooks
- Santambrogio, *OT for applied mathematicians*
- Villani, *OT, Old and New*
- Ambrosio, Gigli, Savaré, *Gradient flows in metric spaces and in the space of probability measures*
- Peyré and Cuturi, *Computational OT* (upcoming)