Optimal Transport

Abhishek Halder

Department of Aerospace Engineering, Iowa State University
Department of Applied Mathematics, University of California Santa Cruz

Lawrence Livermore National Lab
March 03, 2025

What is Transport

Random variable with given PDF: $X \sim \xi(x)$

New random variable: $Y = f(X)$ for given nonlinear map f

Find new PDF: $Y \sim \eta(y)$

Many names: change of variable, pushforward of probability measure, **transport**

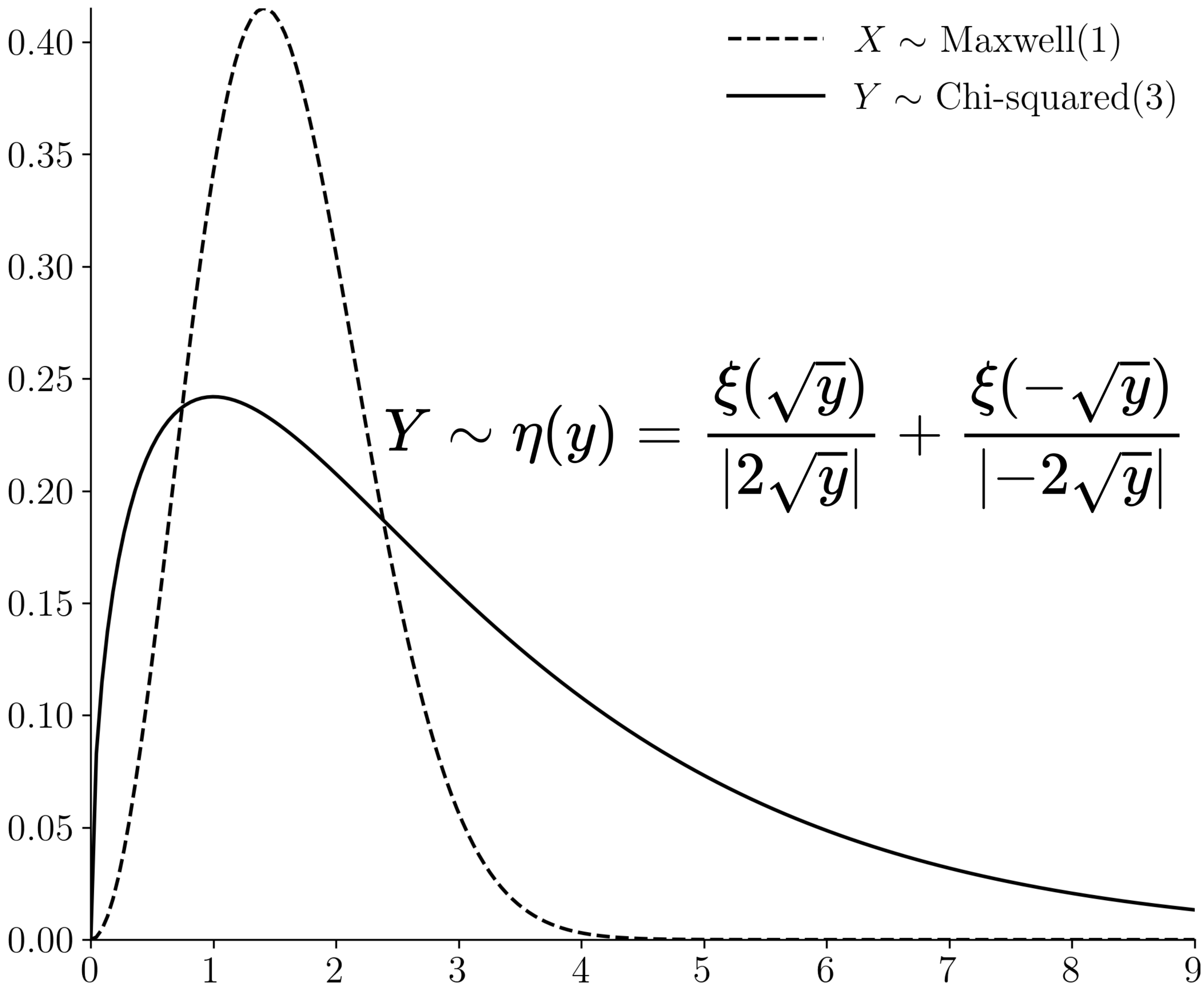
Solution for scalar transport: $$\eta(y) = \sum_{i=1}^{m} \frac{\xi\left(f^{-1}(y)\right)}{\left|f'\left(f^{-1}(y)\right)\right|}$$

m is # of inverses of f

What is Transport: Example

$X \sim \xi(x)$

Pushforward map: $Y = f(X) := X^2$



$$Y \sim \eta(y) = \frac{\xi(\sqrt{y})}{|2\sqrt{y}|} + \frac{\xi(-\sqrt{y})}{|-2\sqrt{y}|} = \frac{\xi(\sqrt{y}) + \xi(-\sqrt{y})}{2\sqrt{y}}$$

Transport vs Optimal Transport

Transport = Forward Problem: Given ξ, f, compute η

Solution for vector transport: $$\eta(\boldsymbol{y}) = \sum_{i=1}^{m} \frac{\xi\left(\boldsymbol{f}^{-1}(\boldsymbol{y})\right)}{\left|\nabla_{\boldsymbol{x}} \boldsymbol{f}\left(\boldsymbol{f}^{-1}(\boldsymbol{y})\right)\right|}$$

Nothing to optimize

Notation: $\eta = \boldsymbol{f}_{\sharp}\xi$

Transport vs Optimal Transport (OT)

Transport = Forward Problem: Given ξ, f, compute η

Solution for vector transport: $\eta(\boldsymbol{y}) = \sum_{i=1}^{m} \frac{\xi\left(\boldsymbol{f}^{-1}(\boldsymbol{y})\right)}{\left|\nabla_{\boldsymbol{x}} \boldsymbol{f}\left(\boldsymbol{f}^{-1}(\boldsymbol{y})\right)\right|}$

Nothing to optimize

Notation: $\eta = \boldsymbol{f}_\sharp \xi$

Optimal transport = Inverse problem: Given ξ, η, compute "best" $\boldsymbol{f}$

$$\underset{\text{Measurable } \boldsymbol{f}:\mathcal{X}\mapsto\mathcal{Y}}{\arg\min} \quad \mathbb{E}_{\boldsymbol{x}}\left[c(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))\right]$$

$$\text{subject to} \quad \eta = \boldsymbol{f}_\sharp \xi$$

$c(\,\cdot\,,\,\cdot\,)$ is called ground cost

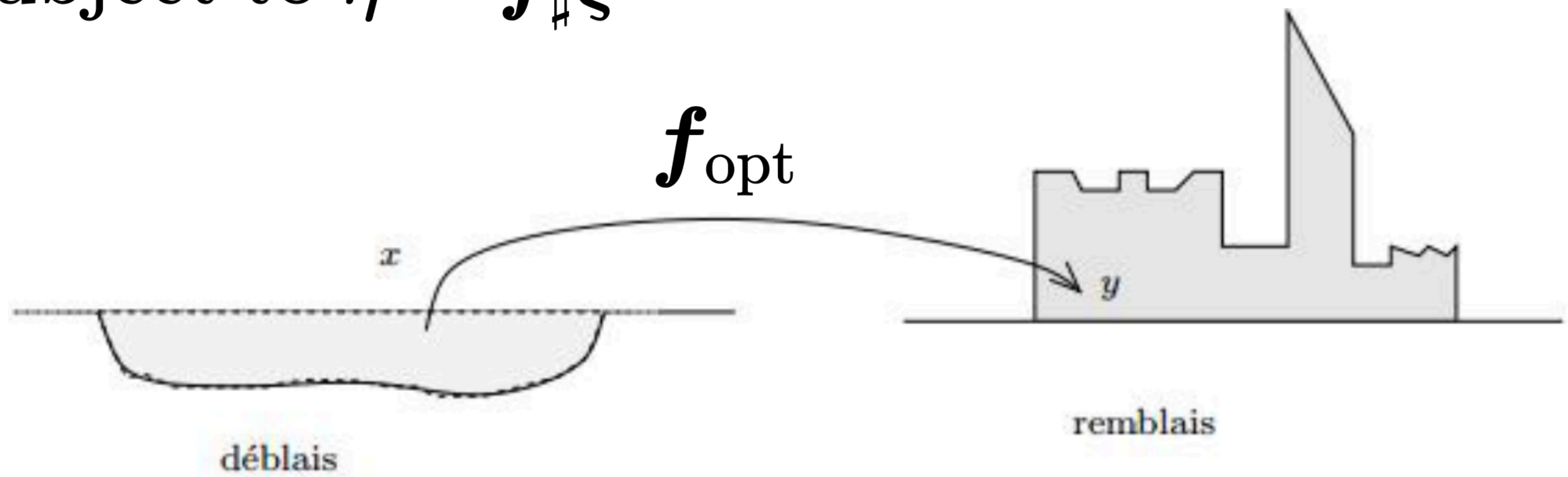
OT Take #1: Monge Formulation

Gaspard Monge
1781

OT map

$$\boldsymbol{f}_{\text{opt}} = \underset{\text{Measurable } \boldsymbol{f}:\mathcal{X}\mapsto\mathcal{Y}}{\arg\min} \int_{\mathcal{X}} c(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))\, \xi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

$$\text{subject to } \eta = \boldsymbol{f}_{\sharp}\xi$$

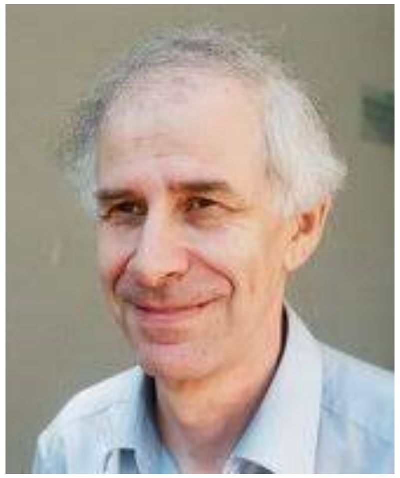


Pushforward constraint is nonlinear and nonconvex in $\boldsymbol{f}$:

$$\left|\det \nabla_{\boldsymbol{x}} \boldsymbol{f}\right| \left(\eta \circ \boldsymbol{f}\right)(\boldsymbol{x}) = \xi(\boldsymbol{x})$$

Monge considered EMD ground cost: $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_1$

OT Take #1: Monge Formulation



Yann Brenier
1991

Brenier's Polar Factorization Thm. (1991)

$$\boldsymbol{f}_{\text{opt}} = (\nabla_{\boldsymbol{x}} \underbrace{\psi}_{\text{convex}}) \circ \underbrace{\boldsymbol{\sigma}}_{\text{measure preserving}}$$

ψ is called **static potential**

For c squared Euclidean, $\boldsymbol{\sigma}$ is identity

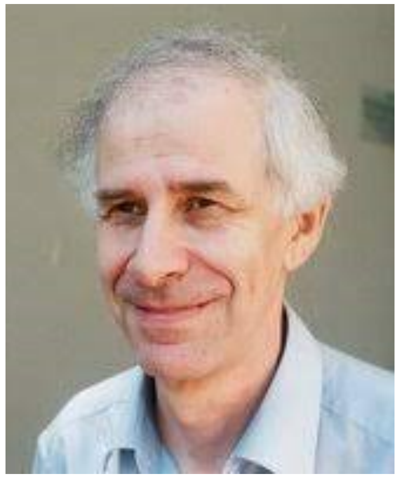
Special cases:

Polar factorization in linear algebra: $\underbrace{\boldsymbol{M}}_{\in \mathrm{GL}(n)} = \underbrace{\boldsymbol{P}}_{\in \mathbb{S}^n_{++}} \underbrace{\boldsymbol{Q}}_{\in \mathrm{O}(n)}$

Helmholtz decomposition of vector field:

$$\underbrace{\boldsymbol{v}}_{\in \mathcal{C}^1(\mathcal{T}\mathbb{R}^n)} = \underbrace{\boldsymbol{s}}_{\text{solenoidal vector field}} + \underbrace{\nabla_{\boldsymbol{x}}\, p}_{\text{gradient vector field}}$$

OT Take #1: Monge Formulation



Yann Brenier
1991

Why not use Polar Factorization Thm. to compute ψ ?

For c squared Euclidean ($\boldsymbol{\sigma}$ is identity)

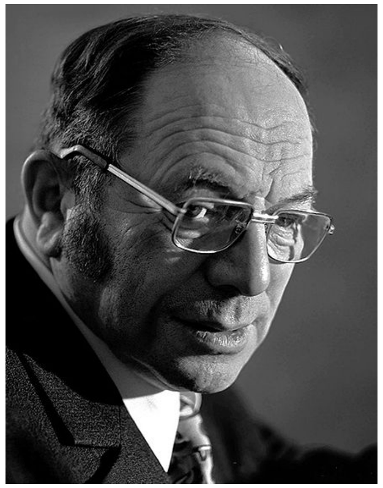
Substituting $\boldsymbol{f}_{\text{opt}} = \nabla_{\boldsymbol{x}}\psi$ in the pushforward constraint gives:

$$\left|\det \text{Hess}_{\boldsymbol{x}}\psi\right| \eta\left(\nabla_{\boldsymbol{x}}\psi\right) = \xi\left(\boldsymbol{x}\right)$$

This is Monge-Ampère PDE to be solved for unknown **convex** ψ

This is 2nd order nonlinear degenerate elliptic PDE …
difficult to solve by finite difference, finite volume etc.

OT Take #2: Kantorovich Formulation



Leonid Kantorovich
1941

OT plan

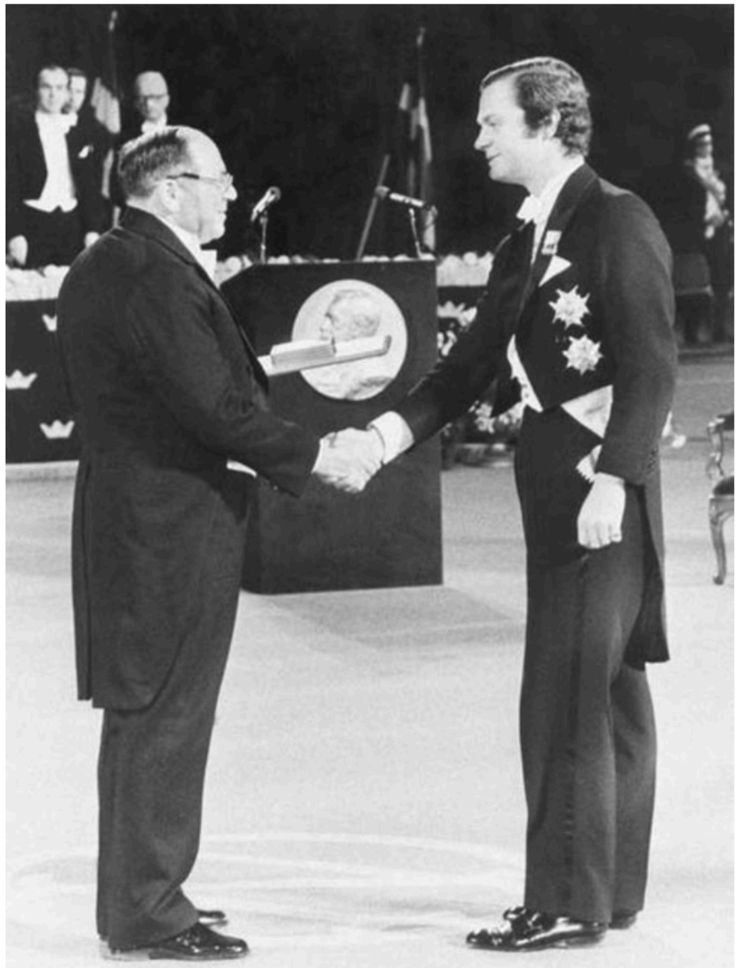
$$\rho_{\text{opt}} = \arg\min_{\rho \geq 0} \int_{\mathcal{X}\times\mathcal{Y}} c(\boldsymbol{x},\boldsymbol{y})\rho(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}$$

$$\text{subject to} \quad \int_{\mathcal{Y}} \rho(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\boldsymbol{y} = \xi(\boldsymbol{x})$$

$$\int_{\mathcal{X}} \rho(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\boldsymbol{x} = \eta(\boldsymbol{y})$$

Linear program!!

1975 Nobel prize in Economics for this work



OT Take #2: Kantorovich Formulation

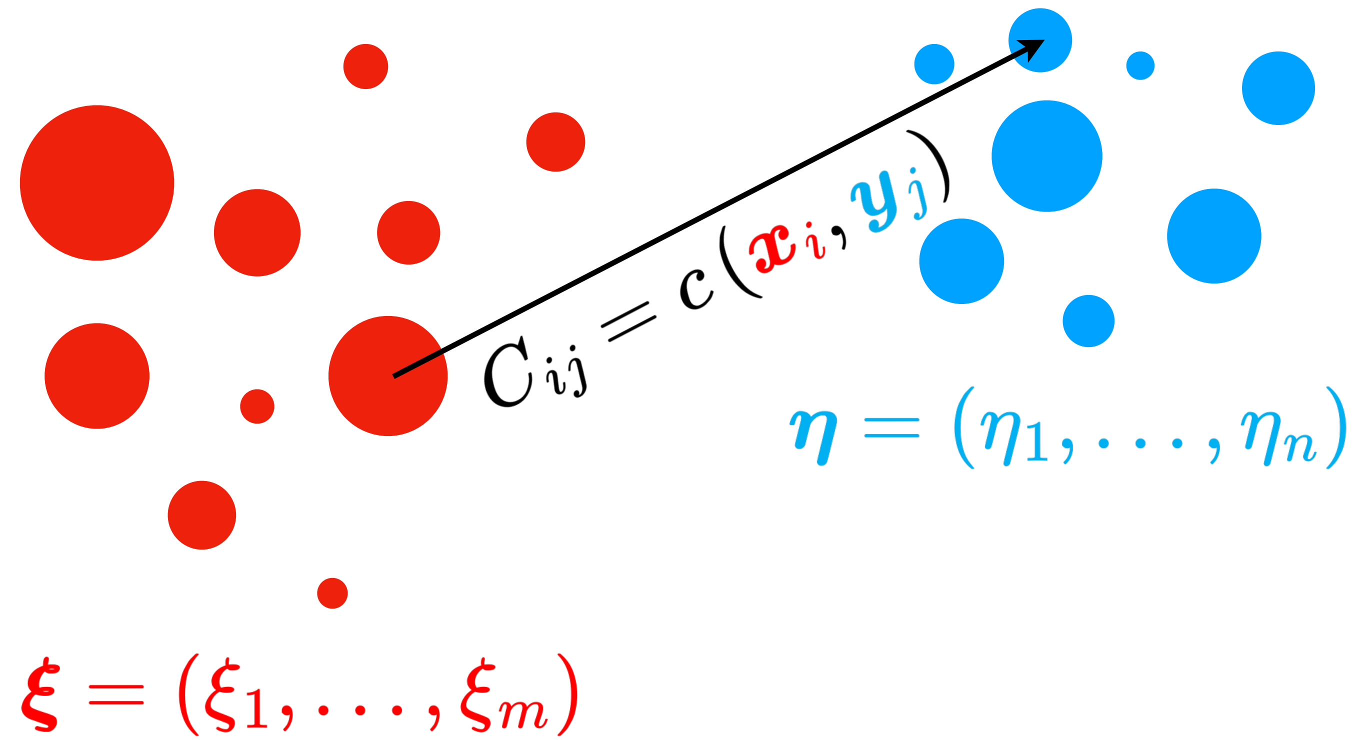
Discrete version

$$\operatorname*{arg\,min}_{[P_{ij}]} \sum_{i=1}^{m} \sum_{j=1}^{n} C_{ij} P_{ij}$$

$$\sum_{j=1}^{n} P_{ij} = \xi_i \quad \forall i = 1, \dots, m$$

$$\sum_{i=1}^{m} P_{ij} = \eta_j \quad \forall j = 1, \dots, n$$

$$P_{ij} \geq 0 \qquad \forall (i,j) \in \{1, \dots, m\} \times \{1, \dots, n\}$$



Difficulty: high computational complexity for large m, n

OT Take #2: Kantorovich Formulation

Regularized discrete version: embrace nonlinearity

Entropy regularization: Strictly convex program (NeurIPS 2013)

$$\begin{aligned} \boldsymbol{P}_{\text{opt}}(\varepsilon) = \underset{\boldsymbol{P}\in\mathbb{R}^{m\times n}}{\arg\min} & \langle \boldsymbol{C} + \varepsilon \log \boldsymbol{P}, \boldsymbol{P} \rangle \\ \text{subject to} \quad & \boldsymbol{P}\mathbf{1} = \boldsymbol{\xi} \\ & \boldsymbol{P}^\top \mathbf{1} = \boldsymbol{\eta} \\ & \boldsymbol{P} \geq 0 \quad \text{elementwise} \end{aligned}$$

Fixed regularizer $\varepsilon > 0$

Turns out this is the **static** Schrödinger bridge

OT Take #2: Kantorovich Formulation

Exploit strong duality

Since subtracting a constant ε in the objective cannot change argmin, so consider the Lagrangian

$$L\left(\boldsymbol{P}, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2\right) = \langle \boldsymbol{C} + \varepsilon \log \boldsymbol{P}, \boldsymbol{P} \rangle - \underbrace{\varepsilon}_{=\varepsilon \mathbf{1}^\top \boldsymbol{P} \mathbf{1}} + \langle \boldsymbol{\lambda}_1, \boldsymbol{P}\mathbf{1} - \boldsymbol{\xi} \rangle + \langle \boldsymbol{\lambda}_2, \boldsymbol{P}^\top \mathbf{1} - \boldsymbol{\eta} \rangle$$

Lagrange multipliers

Apply KKT conditions:

$$\left.\frac{\partial L}{\partial P_{ij}}\right|_{\text{opt}} = 0 \Rightarrow \left(P_{\text{opt}}(\varepsilon)\right)_{ij} = \underbrace{\exp\left(-C_{ij}/\varepsilon\right)}_{=:K_{ij}} \underbrace{\exp\left(-\left(\lambda_1\right)_j\right)}_{=:u_j} \underbrace{\exp\left(-\left(\lambda_2\right)_i\right)}_{=:v_i}$$

OT Take #2: Kantorovich Formulation

Therefore, the regularized argmin solves matrix scaling problem

$$\boldsymbol{P}_{\text{opt}}(\varepsilon) = (\text{diag}\, \boldsymbol{v})\boldsymbol{K}\, (\text{diag}\, \boldsymbol{u})$$

Algorithm: **Sinkhorn recursion/IPFP/raking/contingency table**

$$\boldsymbol{u}^{(k+1)} = \boldsymbol{\xi} \oslash \left(\boldsymbol{K}\boldsymbol{v}^{(k)}\right)$$

$$\boldsymbol{v}^{(k+1)} = \boldsymbol{\eta} \oslash \left(\boldsymbol{K}^{\top}\boldsymbol{u}^{(k+1)}\right)$$

A RELATIONSHIP BETWEEN ARBITRARY POSITIVE MATRICES AND DOUBLY STOCHASTIC MATRICES

BY RICHARD SINKHORN

University of Houston

Annals of Mathematical Statistics
1964

Cone preserving nonlinear recursion: nonlinear Perron-Frobenius

Guaranteed linear convergence: contraction w.r.t. Hilbert metric

The $\boldsymbol{u}_{\text{opt}}(\varepsilon), \boldsymbol{v}_{\text{opt}}(\varepsilon)$ are called the Schrödinger potentials

OT Take #2: Kantorovich Formulation

Duality for unregularized OT

Primal LP

$$\rho_{\text{opt}} = \arg\min_{\rho \geq 0} \int_{\mathcal{X}\times\mathcal{Y}} c(\boldsymbol{x},\boldsymbol{y})\rho(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}$$

$$\text{subject to} \quad \int_{\mathcal{Y}} \rho(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\boldsymbol{y} = \xi(\boldsymbol{x})$$

$$\int_{\mathcal{X}} \rho(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\boldsymbol{x} = \eta(\boldsymbol{y})$$

Dual LP

$$(\alpha_{\text{opt}}(\boldsymbol{x}), \beta_{\text{opt}}(\boldsymbol{y})) = \arg\max_{\alpha\in\mathcal{C}_b(\mathcal{X}),\beta\in\mathcal{C}_b(\mathcal{Y})} \int_{\mathcal{X}} \alpha(\boldsymbol{x})\xi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_{\mathcal{Y}} \beta(\boldsymbol{y})\eta(\boldsymbol{y})\mathrm{d}\boldsymbol{y}$$

Kantorovich potentials

$$\text{subject to} \quad \alpha(\boldsymbol{x}) + \beta(\boldsymbol{y}) \leq c(\boldsymbol{x},\boldsymbol{y})$$

OT Take #2: Kantorovich Formulation

Strong duality for unregularized OT

Thm.

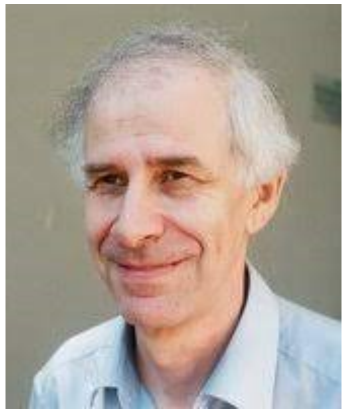
If $\mathcal{X}, \mathcal{Y}$ are polish spaces, and the ground cost $c : \mathcal{X} \times \mathcal{Y} \mapsto \overline{\mathbb{R}}$ is lsc, then strong duality holds.

Furthermore,

- $\alpha_{\text{opt}}(\boldsymbol{x}) + \beta_{\text{opt}}(\boldsymbol{y}) = c(\boldsymbol{x}, \boldsymbol{y})$ for ρ_{opt} a.e. $(\boldsymbol{x}, \boldsymbol{y})$
- $\alpha_{\text{opt}}(\boldsymbol{x}), \beta_{\text{opt}}(\boldsymbol{y})$ are *c*-conjugates of each other

$$\beta_{\text{opt}}(\boldsymbol{y}) = \alpha^c_{\text{opt}}(\boldsymbol{y}) := \inf_{\boldsymbol{x} \in \mathcal{X}} \left\{ c(\boldsymbol{x}, \boldsymbol{y}) - \alpha_{\text{opt}}(\boldsymbol{x}) \right\}$$

OT Take #3: Brenier-Benamou Formulation




Y. Brenier J-D. Benamou

1999

Stochastic control problem

$$\min_{(\rho,\boldsymbol{u})\in\mathcal{P}\times\mathcal{U}} \int_0^1 \int_{\mathcal{X}} \frac{1}{2}\|\boldsymbol{u}\|_2^2\, \rho(t,\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t$$

subject to $\dot{\boldsymbol{x}} = \boldsymbol{u} \Leftrightarrow \dfrac{\partial \rho}{\partial t} + \nabla_{\boldsymbol{x}}\cdot(\rho\boldsymbol{u}) = 0$

$$\rho(t=0,\cdot) = \xi(\cdot), \quad \rho(t=1,\cdot) = \eta(\cdot)$$

Thm.

$$\boldsymbol{u}_{\text{opt}}(t,\boldsymbol{x}) = \nabla_{\boldsymbol{x}}\phi(t,\boldsymbol{x})$$

where $\phi(t,\boldsymbol{x})$ solves the Hamilton-Jacobi-Bellman PDE

$$\frac{\partial \phi}{\partial t} + \frac{1}{2}\|\nabla_{\boldsymbol{x}}\phi\|_2^2 = 0$$

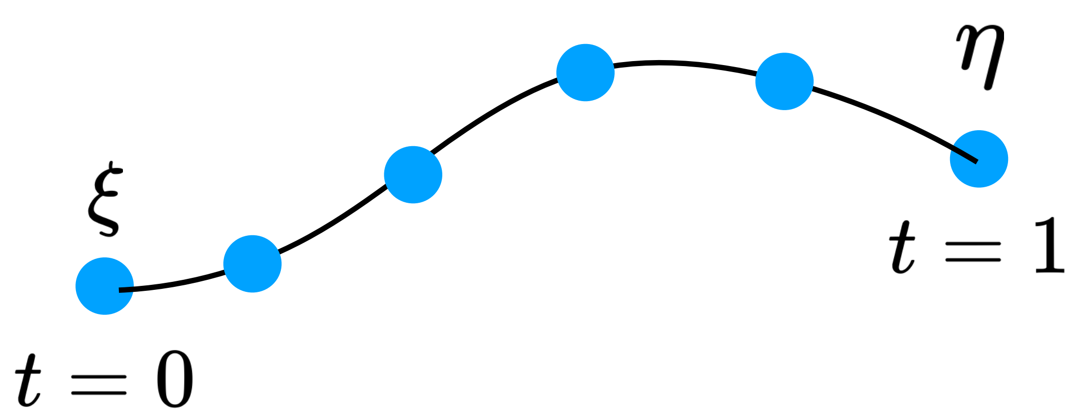
The ϕ is called **dynamic potential**

How are these 3 OT formulations related?

When ground cost $c = 1/2$ squared Euclidean distance,

optimal value of Take #1 = that of Take #2 = that of Take #3

This optimal value is the 1/2 **squared Wasserstein distance metric**



$$\frac{1}{2} W^2(\xi, \eta)$$

Wasserstein geodesic:

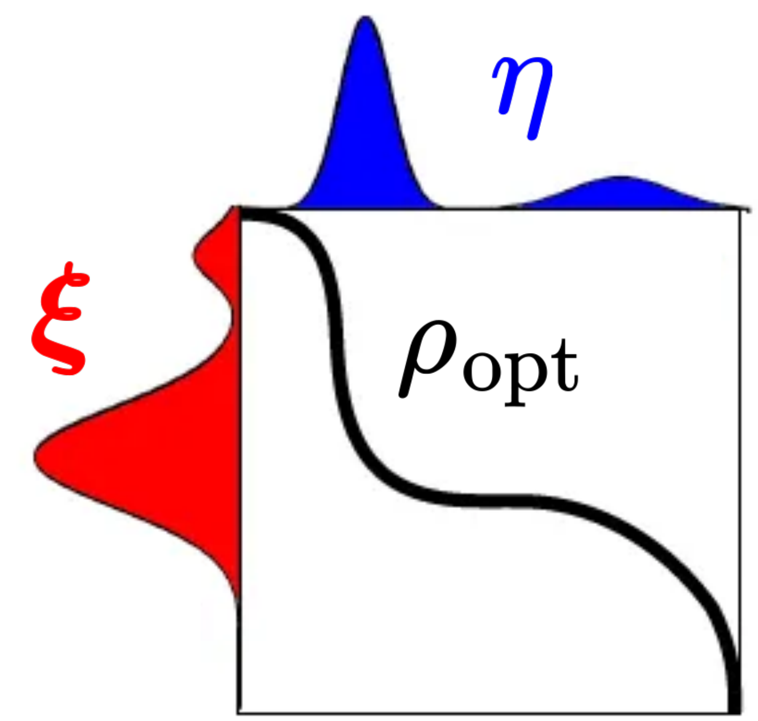
$$\rho_{\text{opt}}(t, \boldsymbol{x}) = \underset{\rho \geq 0, \int \rho = 1}{\arg\min} \left\{ (1-t) W^2(\rho, \xi) + t W^2(\rho, \eta) \right\}, \quad 0 \leq t \leq 1$$

Connections between Take #1 and Take #2

The OT plan ρ_{opt} is supported on

the graph of the OT map $\boldsymbol{f}_{\text{opt}}$

under mild assumptions on problem data



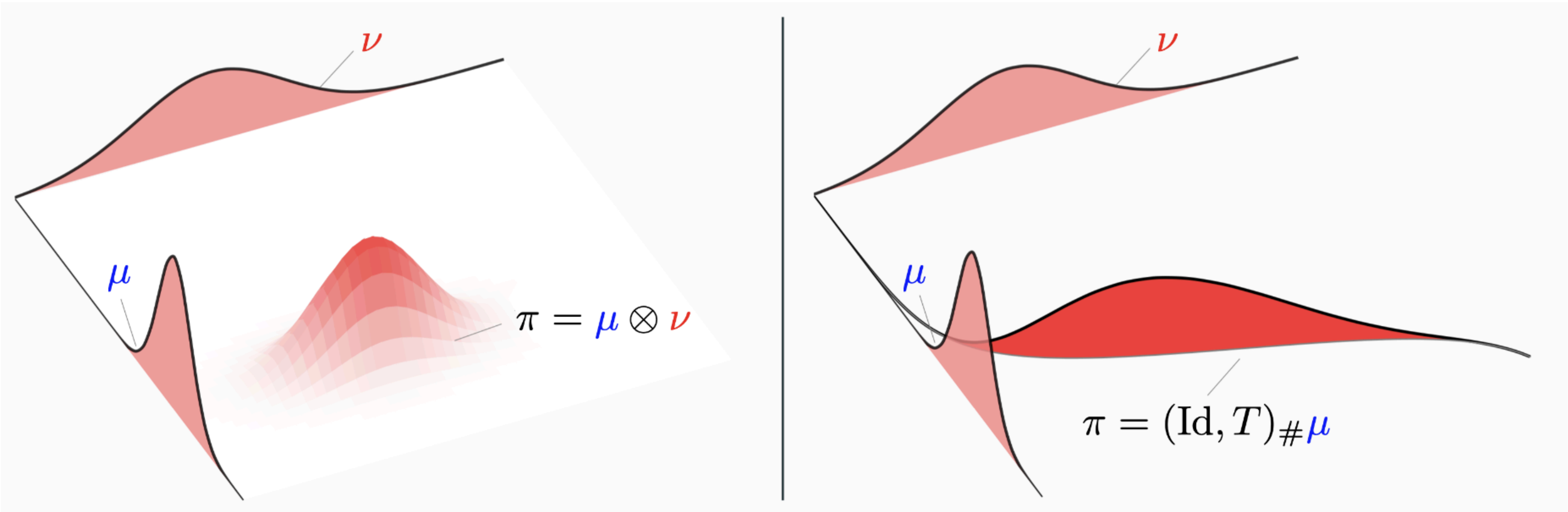


Image credit: L. Chizat

Connections between Take #1 and Take #3

Nonlinear (displacement) interpolation between ξ and η:

$$\rho_{\text{opt}}(t, \boldsymbol{x}) = (\boldsymbol{f}_t)_\sharp\, \xi, \quad 0 \le t \le 1$$

where $\boldsymbol{f}_t = (1-t)\,\text{Id} + t\,\boldsymbol{f}_{\text{opt}}, \quad 0 \le t \le 1$

Connections between Take #1 and Take #3

Nonlinear (displacement) interpolation between ξ and η:

$$\rho_{\text{opt}}(t, \boldsymbol{x}) = (\boldsymbol{f}_t)_\sharp\, \xi, \quad 0 \le t \le 1$$

where $\boldsymbol{f}_t = (1-t)\,\mathrm{Id} + t\,\boldsymbol{f}_{\text{opt}}, \quad 0 \le t \le 1$

Relation between static potential ψ and dynamic potential ϕ:

In Take #1: $\boldsymbol{f}_{\text{opt}} = \nabla_{\boldsymbol{x}} \psi(\boldsymbol{x})$

In Take #3: $\boldsymbol{u}_{\text{opt}}(t, \boldsymbol{x}) = \nabla_{\boldsymbol{x}} \phi(t, \boldsymbol{x})$

Hopf-Lax representation formula:

Infimal convolution

$$\phi(t, \boldsymbol{x}) = \min_{\boldsymbol{y}} \left\{ \phi_0(\boldsymbol{x}) + \frac{1}{2t} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \right\}, \; 0 \le t \le 1$$

where $\phi_0(\boldsymbol{x}) := \psi(\boldsymbol{x}) - \dfrac{1}{2}\|\boldsymbol{x}\|_2^2$

Analytically Solvable OT Problems

Problem	OT value W^2	OT map $\boldsymbol{f}_{\text{opt}}$
1D OT with CDFs: $F(x), G(y)$	$\int_0^1 \left(F^{-1}(u) - G^{-1}(u)\right)^2 \mathrm{d}u$	$G \circ F^{-1}(\boldsymbol{x})$
Multivariate normals: $\xi = \mathcal{N}\left(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x\right)$ $\eta = \mathcal{N}\left(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y\right)$	$\Vert\boldsymbol{\mu}_x - \boldsymbol{\mu}_y\Vert_2^2 + \operatorname{tr}\left(\boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y - 2\left(\boldsymbol{\Sigma}_y^{\frac{1}{2}}\boldsymbol{\Sigma}_x\boldsymbol{\Sigma}_y^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)$	$\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ where $\boldsymbol{A} = \boldsymbol{\Sigma}_y^{\frac{1}{2}}\left(\boldsymbol{\Sigma}_y^{\frac{1}{2}}\boldsymbol{\Sigma}_x\boldsymbol{\Sigma}_y^{\frac{1}{2}}\right)^{-\frac{1}{2}}\boldsymbol{\Sigma}_y^{\frac{1}{2}}$ $\boldsymbol{b} = \boldsymbol{\mu}_y - \boldsymbol{\mu}_x$

Wasserstein Gradient Flows

$$\frac{\partial \mu}{\partial t} = -\nabla^{W_2} F(\mu) := \nabla \cdot \left(\mu \nabla \frac{\delta F}{\delta \mu} \right) \qquad (\star)$$

Wasserstein gradient

Minimizer of $\underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\arg\inf} F(\mu)$ ⇜ Stationary solution of $(\star)$

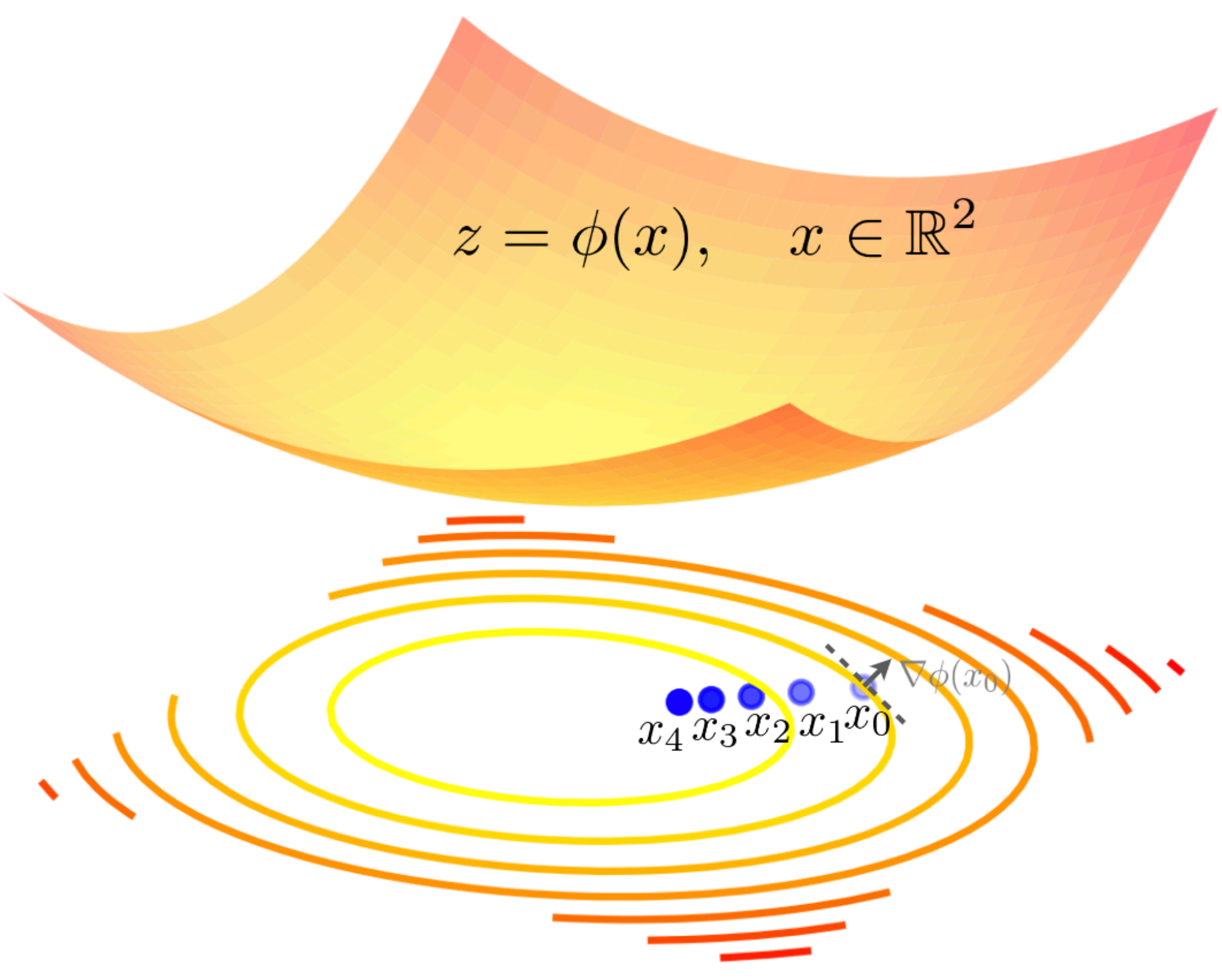
Transient solution of $(\star)$ ⇝ Discrete time-stepping realizing grad. descent of $\underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\arg\inf} F(\mu)$

Wasserstein proximal recursion à la Jordan-Kinderlehrer-Otto (JKO) scheme

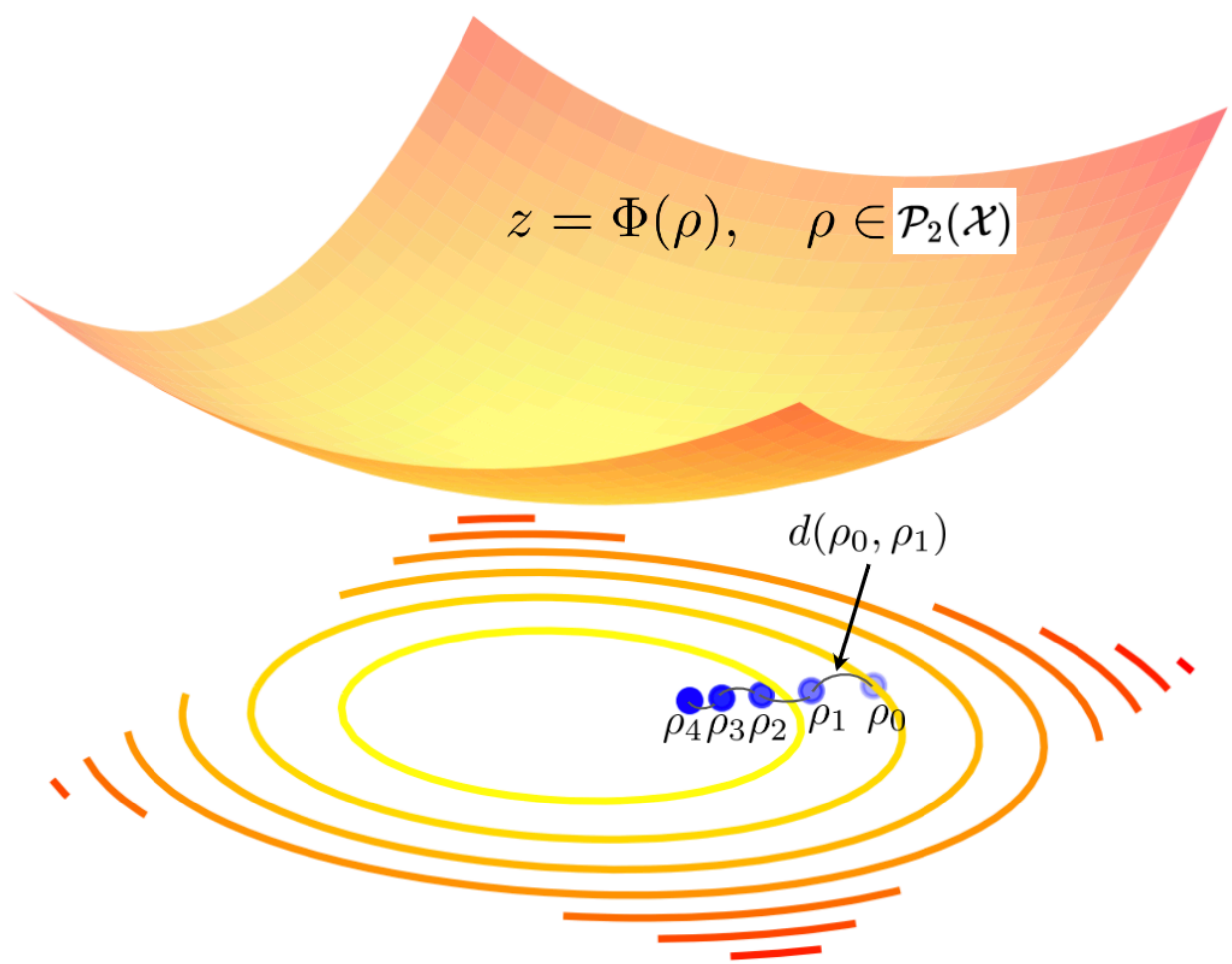
Wasserstein Gradient Flows

PDE solution as gradient descent on the metric space $(\mathcal{P}_2(\mathcal{X}), W)$

Gradient Flow in $\mathcal{X}$



Gradient Flow in $\mathcal{P}_2(\mathcal{X})$



Wasserstein Gradient Flows

Gradient Flow in $\mathcal{X}$	Gradient Flow in $\mathcal{P}_2(\mathcal{X})$
$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = -\nabla\varphi(\boldsymbol{x}), \quad \boldsymbol{x}(0) = \boldsymbol{x}_0$	$\frac{\partial\rho}{\partial t} = -\nabla^W\Phi(\rho), \quad \rho(\boldsymbol{x}, 0) = \rho_0$
Recursion: $\boldsymbol{x}_k = \boldsymbol{x}_{k-1} - h\nabla\varphi(\boldsymbol{x}_k) = \underset{\boldsymbol{x}\in\mathcal{X}}{\arg\min}\left\{\frac{1}{2}\lVert\boldsymbol{x}-\boldsymbol{x}_{k-1}\rVert_2^2 + h\varphi(\boldsymbol{x})\right\} =: \mathrm{prox}_{h\varphi}^{\lVert\cdot\rVert_2}(\boldsymbol{x}_{k-1})$	**Recursion:** $\rho_k = \rho(\cdot, t = kh) = \underset{\rho\in\mathcal{P}_2(\mathcal{X})}{\arg\min}\left\{\frac{1}{2}W^2(\rho, \rho_{k-1}) + h\Phi(\rho)\right\} =: \mathrm{prox}_{h\Phi}^{W^2}(\rho_{k-1})$
Convergence: $\boldsymbol{x}_k \to \boldsymbol{x}(t = kh) \quad \text{as} \quad h \downarrow 0$	**Convergence:** $\rho_k \to \rho(\cdot, t = kh) \quad \text{as} \quad h \downarrow 0$
φ as Lyapunov function: $\frac{\mathrm{d}}{\mathrm{d}t}\varphi = -\lVert\nabla\varphi\rVert_2^2 \;\leq\; 0$	**Φ as Lyapunov functional:** $\frac{\mathrm{d}}{\mathrm{d}t}\Phi = -\mathbb{E}_\rho\left[\left\lVert\nabla\frac{\delta\Phi}{\delta\rho}\right\rVert_2^2\right] \;\leq\; 0$

Wasserstein Gradient Flows

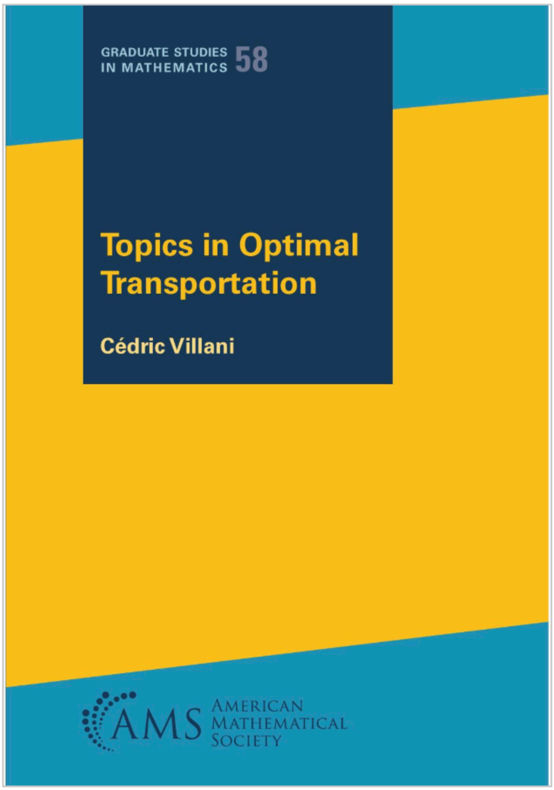
PDE	Free energy Φ	Specific instances
McKean-Vlasov-Fokker-Planck-Kolmogorov PDEs with gradient/mixed conservative-dissipative drift	$\mathbb{E}_\rho\left[V + \beta^{-1}\log\rho + U * \rho\right]$ Potential energy Internal energy Nonlocal interaction energy	Fokker-Planck-Kolmogorov PDE Mean field dynamics: crowd, overparameterized neural networks
Nonlinear diffusion PDEs	$\mathbb{E}_\rho\left[\dfrac{\beta^{-1}}{m-1}\rho^{m-1}\right]$	Power law diffusion with $\Delta\rho^m,\ m > 1$
Vlasov-Poisson-Fokker-Planck PDEs	$\mathbb{E}_\rho\left[\dfrac{\lVert v\rVert_2^2}{2} + U_0(x) + \beta^{-1}\log\rho\right] + \dfrac{1}{2\lambda}\int \lVert E(t,x)\rVert_2^2\,\mathrm{d}x$	Plasma dynamics Astrophysics Bacterial chemotaxis

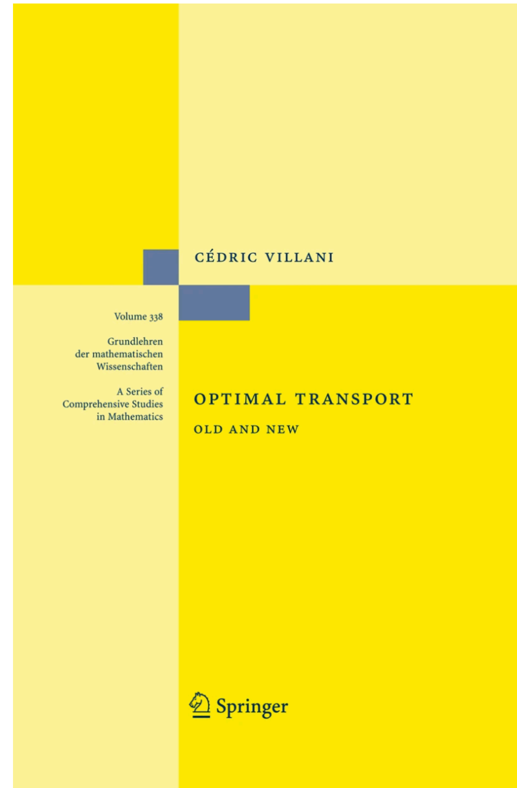
Caveat Emptor

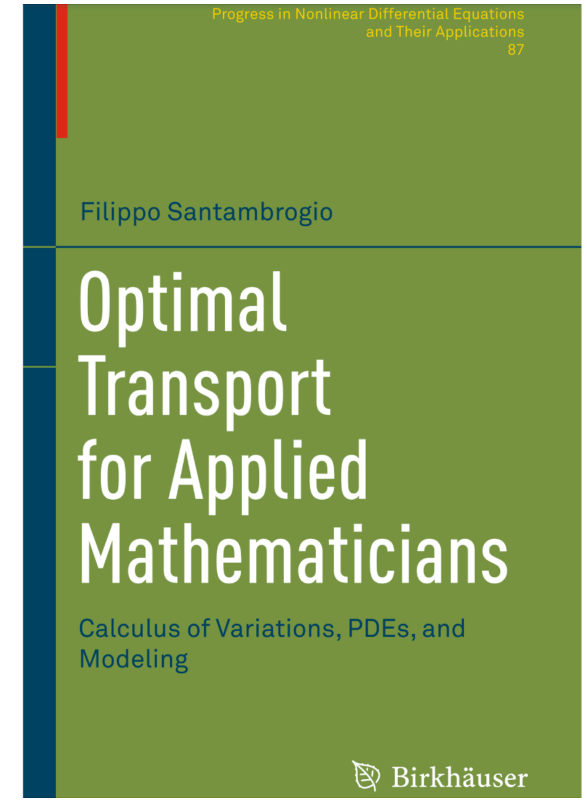
Potentials galore:

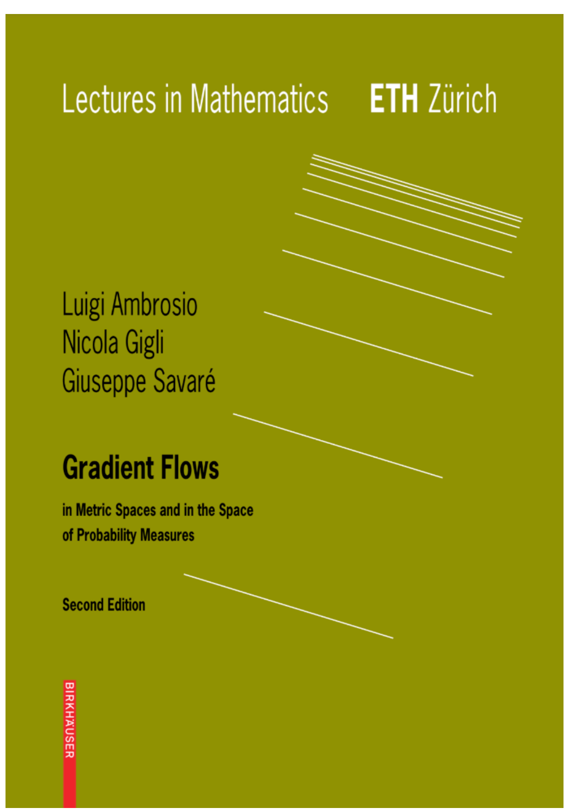
- static (Monge) OT potential $\psi(\boldsymbol{x})$
- dynamic (Brenier-Benamou) OT potential $\phi(t, \boldsymbol{x})$
- static Kantorovich (dual) potentials $\alpha_{\text{opt}}(\boldsymbol{x}), \beta_{\text{opt}}(\boldsymbol{y})$
- static Schrödinger (regularized dual) potentials $\boldsymbol{u}_{\text{opt}}(\varepsilon), \boldsymbol{v}_{\text{opt}}(\varepsilon)$

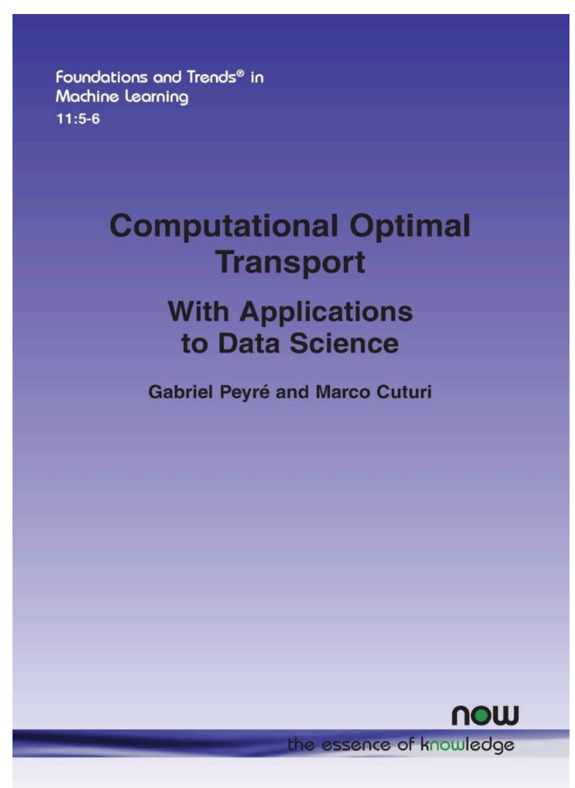
OT References











Thank You