Optimal mass transport as a distance measure between images

Axel Ringh[1]

[1]Department of Mathematics, KTH Royal Institute of Technology, Stockholm, Sweden.

21st of June 2018
INRIA, Sophia-Antipolis, France

Acknowledgements

This is based on joint work with Johan Karlsson[1].

[1] J. Karlsson, and A. Ringh. Generalized Sinkhorn iterations for regularizing inverse problems using optimal mass transport. *SIAM Journal on Imaging Sciences*, 10(4), 1935-1962, 2017.

I acknowledge financial support from

- Swedish Research Council (VR)
- Swedish Foundation for Strategic Research (SSF)

Code: `https://github.com/aringh/Generalized-Sinkhorn-and-tomography`
The code is based on ODL: `https://github.com/odlgroup/odl`

[1] Department of Mathematics, KTH Royal Institute of Technology, Stockholm, Sweden

Outline

- Background
  - Inverse problems
  - Optimal mass transport
  - Sinkhorn iterations - solving discretized optimal transport problems
- Sinkhorn iterations as dual coordinate ascent
- Inverse problems with optimal mass transport priors
- Example in computerized tomography

Consider the problem of recovering $f \in X$ from data $g \in Y$, given by

$$g = A(f) + \text{'noise'}$$

Notation:

- X is called the reconstruction space.
- Y is called the data space.
- $A : X \to Y$ is the forward operator.
- $A^* : Y \to X$ denotes the adjoint operator

Consider the problem of recovering $f \in X$ from data $g \in Y$, given by

$$g = A(f) + \text{'noise'}$$

Notation:

- X is called the reconstruction space.
- Y is called the data space.
- $A : X \to Y$ is the forward operator.
- $A^* : Y \to X$ denotes the adjoint operator

Problems of interest are ill-posed inverse problems:

- a solution might not exist,
- the solution might not be unique,
- the solution does not depend continuously on data.

Simply put: A^{-1} does not exist as a continuous bijection!

Comes down to: find approximate inverse $A^{\dagger}$ so that

$$g = A(f) + \text{'noise'} \implies A^{\dagger}(g) \approx f.$$

A common technique to solve ill-posed inverse problems is to use variational regularization:

$$\arg\min_{f \in X} \mathcal{G}(A(f), g) + \lambda \mathcal{F}(f)$$

- $\mathcal{G} : Y \times Y \to \mathbb{R}$, data discrepancy functional.
- $\mathcal{F} : X \to \mathbb{R}$, regularization functional.
- λ is the regularization parameter. Controls trade-off between data matching and regularization.

Common example in imaging is total variation regularization:

- $\mathcal{G}(h, g) = \|h - g\|_2^2$,
- $\mathcal{F}(f) = \|\nabla f\|_1$.

If A is linear this is a convex problem!

How can one incorporate prior information in such a scheme?

How can one incorporate prior information in such a scheme?

One way: consider

$$\arg\min_{f \in X} \mathcal{G}(A(f), g) + \lambda \mathcal{F}(f) + \gamma \mathcal{H}(\tilde{f}, f)$$

- $\tilde{f}$ is prior/template
- $\mathcal{H}$ defines "closeness" to $\tilde{f}$.

What is a good choice for $\mathcal{H}$?

How can one incorporate prior information in such a scheme?

One way: consider

$$\operatorname*{arg\,min}_{f \in X} \mathcal{G}(A(f), g) + \lambda \mathcal{F}(f) + \gamma \mathcal{H}(\tilde{f}, f)$$

- $\tilde{f}$ is prior/template
- $\mathcal{H}$ defines "closeness" to $\tilde{f}$.

What is a good choice for $\mathcal{H}$?

Scenarios where potentially of interest.

- incomplete measurements, e.g. limited angle tomography.
- spatiotemporal imaging:
  - data is a time-series of data sets: $\{g_t\}_{t=0}^{T}$.
    For each set, the underlying image has undergone a deformation.
  - each data set g_t normally "contains less information": $A^{\dagger}(g_t)$ is a poor reconstruction.

Approach: solve coupled inverse problems

$$\operatorname*{arg\,min}_{f_0,\ldots,f_T \in X} \sum_{j=0}^{T} \Big[\mathcal{G}(A(f_j), g_j) + \lambda \mathcal{F}(f_j) \Big] + \sum_{j=1}^{T} \gamma \mathcal{H}(f_{j-1}, f_j)$$

Given two functions $f_0(x)$ and $f_1(x)$, what is a suitable way to measure the distance between the two?

Given two functions $f_0(x)$ and $f_1(x)$, what is a suitable way to measure the distance between the two?

One suggestion: measure it pointwise, e.g., using an L_p metric

$$\|f_0 - f_1\|_p = \left(\int_D |f_0(x) - f_1(x)|^p dx \right)^{1/p}.$$

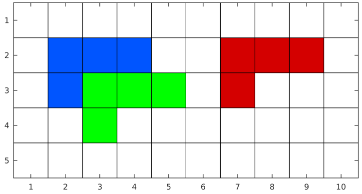
Given two functions $f_0(x)$ and $f_1(x)$, what is a suitable way to measure the distance between the two?

One suggestion: measure it pointwise, e.g., using an L_p metric

$$\|f_0 - f_1\|_p = \left(\int_D |f_0(x) - f_1(x)|^p dx \right)^{1/p}.$$

Draw-backs: for example unsensitive to shifts.

Example:
It gives the same distance from f_0 to f_1 and f_2:
$\|f_0 - f_1\|_1 = \|f_0 - f_2\|_1 = 8$.

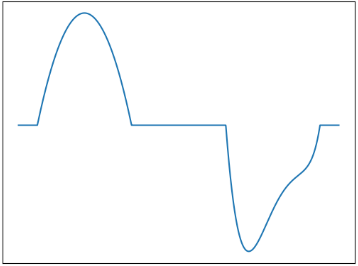


Gaspard Monge: formulated optimal mass transport 1781.
Optimal transport of soil for construction of forts and roads.

Gaspard Monge

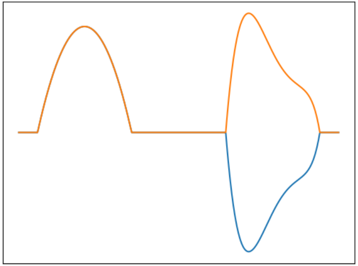
Gaspard Monge: formulated optimal mass transport 1781.
Optimal transport of soil for construction of forts and roads.

Gaspard Monge



Gaspard Monge: formulated optimal mass transport 1781.
Optimal transport of soil for construction of forts and roads.

Gaspard Monge



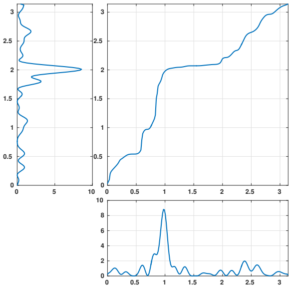
Gaspard Monge: formulated optimal mass transport 1781.
Optimal transport of soil for construction of forts and roads.


Gaspard Monge

Let $c(x_0, x_1) : X \times X \to \mathbb{R}_+$ describes the cost for transporting a unit mass from location x_0 to x_1.

Given two functions f_0, $f_1 : X \to \mathbb{R}_+$, find the function $\phi : X \to X$ minimizing the transport cost

$$\int_X c(x, \phi(x)) f_0(x) dx$$



Background
Optimal mass transport - Monge formulation


Gaspard Monge

Gaspard Monge: formulated optimal mass transport 1781.
Optimal transport of soil for construction of forts and roads.

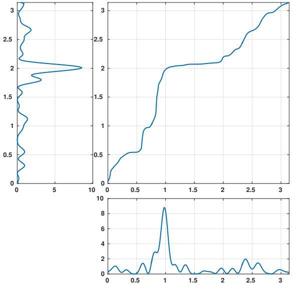
Let $c(x_0, x_1) : X \times X \to \mathbb{R}_+$ describes the cost for transporting a unit mass from location x_0 to x_1.

Given two functions f_0, $f_1 : X \to \mathbb{R}_+$, find the function $\phi : X \to X$ minimizing the transport cost

$$\int_X c(x, \phi(x)) f_0(x) dx$$

where ϕ is mass preserving map from f_0 to f_1:

$$\int_{x \in A} f_1(x) dx = \int_{\phi(x) \in A} f_0(x) dx \quad \text{for all } A \subset X.$$



Gaspard Monge: formulated optimal mass transport 1781.
Optimal transport of soil for construction of forts and roads.

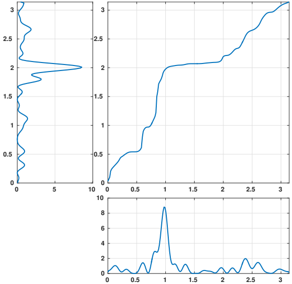
Gaspard Monge

Let $c(x_0, x_1) : X \times X \to \mathbb{R}_+$ describes the cost for transporting a unit mass from location x_0 to x_1.

Given two functions f_0, f_1 : X … $\to X$
minimizing the transport cost …

$$\int_X c(x \ldots$$

Nonconvex problem!



where ϕ is mass preserving map from f_0 to f_1:

$$\int_{x \in A} f_1(x)dx = \int_{\phi(x) \in A} f_0(x)dx \quad \text{for all } A \subset X.$$

Leonid Kantorovich: convex formulation and duality theory 1942.
Nobel Memorial Prize 1975 in Economics for "contributions to the theory of optimum allocation of resources."

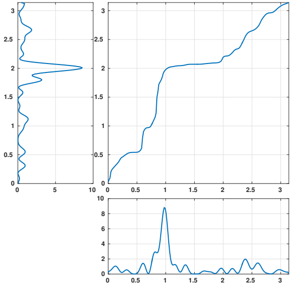

Leonid Kantorovich


Leonid Kantorovich

Leonid Kantorovich: convex formulation and duality theory 1942.
Nobel Memorial Prize 1975 in Economics for "contributions to the theory of optimum allocation of resources."

Again, let $c(x_0, x_1)$ denote the cost of transporting a unit mass from the point x_0 to the point x_1.

Given two functions f_0, $f_1 : X \to \mathbb{R}_+$, find a transport plan $M : X \times X \to \mathbb{R}_+$, where $M(x_0, x_1)$ is the amount of mass moved between x_0 to x_1.

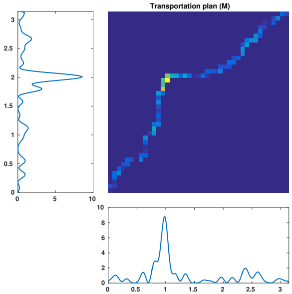


Leonid Kantorovich: convex formulation and duality theory 1942.
Nobel Memorial Prize 1975 in Economics for "contributions to the theory of optimum allocation of resources."

Leonid Kantorovich

Again, let $c(x_0, x_1)$ denote the cost of transporting a unit mass from the point x_0 to the point x_1.

Given two functions f_0, $f_1 : X \to \mathbb{R}_+$, find a transport plan $M : X \times X \to \mathbb{R}_+$, where $M(x_0, x_1)$ is the amount of mass moved between x_0 to x_1.



Background
Optimal mass transport - Kantorovich formulation

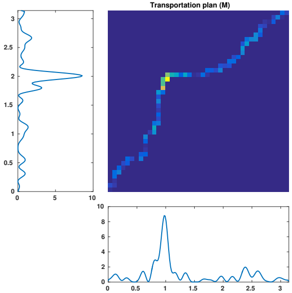

Leonid Kantorovich

Leonid Kantorovich: convex formulation and duality theory 1942.
Nobel Memorial Prize 1975 in Economics for "contributions to the theory of optimum allocation of resources."

Again, let $c(x_0, x_1)$ denote the cost of transporting a unit mass from the point x_0 to the point x_1.

Given two functions $f_0, f_1 : X \to \mathbb{R}_+$, find a transport plan $M : X \times X \to \mathbb{R}_+$, where $M(x_0, x_1)$ is the amount of mass moved between x_0 to x_1. Look for M that minimizes transportation cost:

$$\min_{M \geq 0} \int_{X \times X} c(x_0, x_1) M(x_0, x_1) dx_0 dx_1$$



Background
Optimal mass transport - Kantorovich formulation

Leonid Kantorovich

Leonid Kantorovich: convex formulation and duality theory 1942.
Nobel Memorial Prize 1975 in Economics for "contributions to the theory of optimum allocation of resources."

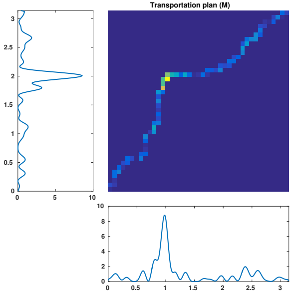
Again, let $c(x_0, x_1)$ denote the cost of transporting a unit mass from the point x_0 to the point x_1.

Given two functions $f_0, f_1 : X \to \mathbb{R}_+$, find a transport plan $M : X \times X \to \mathbb{R}_+$, where $M(x_0, x_1)$ is the amount of mass moved between x_0 to x_1. Look for M that minimizes transportation cost:

$$\min_{M \geq 0} \int_{X \times X} c(x_0, x_1) M(x_0, x_1) dx_0 dx_1$$

$$\text{s.t.} \quad f_0(x_0) = \int_X M(x_0, x_1) dx_1, \; x_0 \in X$$

$$f_1(x_1) = \int_X M(x_0, x_1) dx_0, \; x_1 \in X.$$



Define a distance between two functions $f_0(x)$ and $f_1(x)$ using optimal transport

$$T(f_0, f_1) := \begin{cases} \min_{M \geq 0} & \int_{X \times X} c(x_0, x_1) M(x_0, x_1) dx_0 dx_1 \\ \text{s.t.} & f_0(x_0) = \int_X M(x_0, x_1) dx_1, \; x_0 \in X \\ & f_1(x_1) = \int_X M(x_0, x_1) dx_0, \; x_1 \in X. \end{cases}$$

If $d(x, y)$ metric on X and $c(x, y) = d(x, y)^p$, then $T(f_0, f_1)^{1/p}$ is a metric on the space of measures.

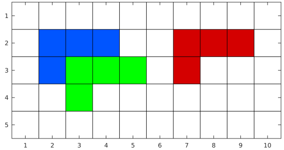
Define a distance between two functions $f_0(x)$ and $f_1(x)$ using optimal transport

$$T(f_0, f_1) := \begin{cases} \min_{M \geq 0} & \int_{X \times X} c(x_0, x_1) M(x_0, x_1) dx_0 dx_1 \\ \text{s.t.} & f_0(x_0) = \int_X M(x_0, x_1) dx_1, \ x_0 \in X \\ & f_1(x_1) = \int_X M(x_0, x_1) dx_0, \ x_1 \in X. \end{cases}$$

If $d(x,y)$ metric on X and $c(x,y) = d(x,y)^p$, then $T(f_0, f_1)^{1/p}$ is a metric on the space of measures.

Example revisited: let $c(x,y) = \|x - y\|_2^2$

$T(f_0, f_1) = 4 \cdot (\sqrt{2})^2 = 8,$
$T(f_0, f_2) = 4 \cdot 5^2 = 100$



This indicates that optimal transport is a more natural distance between two images than L_p, at least if one is a deformation of the other.

How to solve the optimal transport problem? Here: "discretize then optimize"

How to solve the optimal transport problem? Here: "discretize then optimize"

A linear programming problem: for vectors $f_0 \in \mathbb{R}^n$ and $f_1 \in \mathbb{R}^m$ and a cost matrix $C = [c_{ij}] \in \mathbb{R}^{n \times m}$, where c_{ij} defines the transportation cost between pixels x_i and x_j, find the transportation plan $M = [m_{ij}] \in \mathbb{R}^{n \times m}$, m_{ij} is the mass transported between pixels x_i and x_j, such that

$$\min_{m_{ij} \geq 0} \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} m_{ij}$$

$$\text{subject to} \quad \sum_{j=1}^{m} m_{ij} = f_0(i), \; i = 1, \ldots, n$$

$$\sum_{i=1}^{n} m_{ij} = f_1(j), \; j = 1, \ldots, m$$

How to solve the optimal transport problem? Here: "discretize then optimize"

A linear programming problem: for vectors $f_0 \in \mathbb{R}^n$ and $f_1 \in \mathbb{R}^m$ and a cost matrix $C = [c_{ij}] \in \mathbb{R}^{n\times m}$, where c_{ij} defines the transportation cost between pixels x_i and x_j, find the transportation plan $M = [m_{ij}] \in \mathbb{R}^{n\times m}$, m_{ij} is the mass transported between pixels x_i and x_j, such that

$$\begin{aligned} \min_{m_{ij}\geq 0} & \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij} m_{ij} \\ \text{subject to} \quad & \sum_{j=1}^{m} m_{ij} = f_0(i),\ i = 1,\dots,n \\ & \sum_{i=1}^{n} m_{ij} = f_1(j),\ j = 1,\dots,m \end{aligned} \iff \begin{aligned} \min_{M\geq 0} \quad & \text{trace}(C^T M) \\ \text{subject to} \quad & M\mathbf{1}_m = f_0 \\ & M^T\mathbf{1}_n = f_1 \end{aligned}$$

How to solve the optimal transport problem? Here: "discretize then optimize"

A linear programming problem: for vectors $f_0 \in \mathbb{R}^n$ and $f_1 \in \mathbb{R}^m$ and a cost matrix $C = [c_{ij}] \in \mathbb{R}^{n \times m}$, where c_{ij} defines the transportation cost between pixels x_i and x_j, find the transportation plan $M = [m_{ij}] \in \mathbb{R}^{n \times m}$, m_{ij} is the mass transported between pixels x_i and x_j, such that

$$\begin{aligned} \min_{m_{ij} \geq 0} & \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} m_{ij} \\ \text{subject to} & \sum_{j=1}^{m} m_{ij} = f_0(i), \; i = 1, \ldots, n \\ & \sum_{i=1}^{n} m_{ij} = f_1(j), \; j = 1, \ldots, m \end{aligned} \iff \begin{aligned} \min_{M \geq 0} \quad & \operatorname{trace}(C^T M) \\ \text{subject to} \quad & M \mathbf{1}_m = f_0 \\ & M^T \mathbf{1}_n = f_1 \end{aligned}$$

Number of variables is $n \cdot m$. To compare the distance between to images of size $n = m = 256 \times 256$:

$$M \in \mathbb{R}^{256^2 \times 256^2} \implies \text{approximately } 4 \cdot 10^9 \text{variables. Prohibitively large!}$$

Original problem too computationally demanding.

Solution: introduce an entropy barrier/regularization term $D(M) = \sum_{i,j}(m_{ij}\log(m_{ij}) - m_{ij} + 1)$ [1],

$$\begin{aligned} T_\epsilon(f_0, f_1) \ := \quad & \min_{M \geq 0} && \mathrm{trace}(C^T M) + \epsilon D(M) \\ & \text{subject to} && f_0 = M\mathbf{1} \\ & && f_1 = M^T\mathbf{1}. \end{aligned}$$

[1] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

Original problem too computationally demanding.

Solution: introduce an entropy barrier/regularization term $D(M) = \sum_{i,j}(m_{ij}\log(m_{ij}) - m_{ij} + 1)$ [1],

$$\begin{aligned} T_\epsilon(f_0, f_1) \ := \ \min_{M \geq 0} \quad & \mathrm{trace}(C^T M) + \epsilon D(M) \\ \text{subject to} \quad & f_0 = M\mathbf{1} \\ & f_1 = M^T\mathbf{1}. \end{aligned}$$

- Using Lagrangian relaxation gives

$$L(M, \lambda_0, \lambda_1) = \mathrm{trace}(C^T M) + \epsilon D(M) + \lambda_0^T(f_0 - M\mathbf{1}) + \lambda_1^T(f_1 - M^T\mathbf{1}).$$

[1] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

Original problem too computationally demanding.

Solution: introduce an entropy barrier/regularization term $D(M) = \sum_{i,j}(m_{ij}\log(m_{ij}) - m_{ij} + 1)$ [1],

$$\begin{aligned} T_\epsilon(f_0, f_1) \; := \quad & \min_{M\geq 0} \quad && \operatorname{trace}(C^T M) + \epsilon D(M) \\ & \text{subject to} && f_0 = M\mathbf{1} \\ & && f_1 = M^T\mathbf{1}. \end{aligned}$$

- Using Lagrangian relaxation gives

$$L(M, \lambda_0, \lambda_1) = \operatorname{trace}(C^T M) + \epsilon D(M) + \lambda_0^T(f_0 - M\mathbf{1}) + \lambda_1^T(f_1 - M^T\mathbf{1}).$$

- Given dual variables λ_0, λ_1, the minimum m_{ij} is

$$0 = \frac{\partial L(M, \lambda_0, \lambda_1)}{\partial m_{ij}} = c_{ij} + \epsilon\log(m_{ij}) - \lambda_0(i) - \lambda_1(j)$$

[1] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

Original problem too computationally demanding.

Solution: introduce an entropy barrier/regularization term $D(M) = \sum_{i,j}(m_{ij}\log(m_{ij}) - m_{ij} + 1)$ [1],

$$\begin{aligned} T_\epsilon(f_0, f_1) \; := \quad & \min_{M\geq 0} \quad \mathrm{trace}(C^T M) + \epsilon D(M) \\ & \text{subject to} \quad f_0 = M\mathbf{1} \\ & \qquad\qquad\quad f_1 = M^T\mathbf{1}. \end{aligned}$$

- Using Lagrangian relaxation gives

$$L(M, \lambda_0, \lambda_1) = \mathrm{trace}(C^T M) + \epsilon D(M) + \lambda_0^T(f_0 - M\mathbf{1}) + \lambda_1^T(f_1 - M^T\mathbf{1}).$$

- Given dual variables λ_0, λ_1, the minimum m_{ij} is

$$0 = \frac{\partial L(M, \lambda_0, \lambda_1)}{\partial m_{ij}} = c_{ij} + \epsilon \log(m_{ij}) - \lambda_0(i) - \lambda_1(j)$$

- Expressed as $m_{ij} = e^{\lambda_0(i)/\epsilon} e^{-c_{ij}/\epsilon} e^{\lambda_1(j)/\epsilon}$

$$M = \mathrm{diag}(e^{\lambda_0^T/\epsilon}) K \mathrm{diag}(e^{\lambda_1/\epsilon})$$

where $K = \exp(-C/\epsilon)$. Here and in what follows $\exp(\cdot)$, $\log(\cdot)$, ./, $\odot$ denotes the elementwise function.

[1] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

Background
Optimal mass transport - Sinkhorn iterations

Original problem too computationally demanding.

Solution: introduce an entropy barrier/regularization term $D(M) = \sum_{i,j}(m_{ij}\log(m_{ij}) - m_{ij} + 1)$ [1],

$$
\begin{aligned}
T_\epsilon(f_0, f_1) \ := \quad & \min_{M \geq 0} \quad && \operatorname{trace}(C^T M) + \epsilon D(M) \\
& \text{subject to} && f_0 = M\mathbf{1} \\
& && f_1 = M^T\mathbf{1}.
\end{aligned}
$$

- Using Lagrangian relaxation gives

> The solution $M = \operatorname{diag}(e^{\lambda_0^T/\epsilon}) K \operatorname{diag}(e^{\lambda_1/\epsilon})$ is specified by $n + m$ unknowns!

- Gi

$$0 = \frac{\partial L(M, \lambda_0, \lambda_1)}{\partial m_{ij}} = c_{ij} + \epsilon \log(m_{ij}) - \lambda_0(i) - \lambda_1(j)$$

- Expressed as $m_{ij} = e^{\lambda_0(i)/\epsilon} e^{-c_{ij}/\epsilon} e^{\lambda_1(j)/\epsilon}$

$$M = \operatorname{diag}(e^{\lambda_0^T/\epsilon}) K \operatorname{diag}(e^{\lambda_1/\epsilon})$$

where $K = \exp(-C/\epsilon)$. Here and in what follows $\exp(\cdot)$, $\log(\cdot)$, $./$, $\odot$ denotes the elementwise function.

[1] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

How to find the values of λ_0 and λ_1?

[1] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

How to find the values of λ_0 and λ_1?

Let $u_0 = \exp(\lambda_0/\epsilon)$, $u_1 = \exp(\lambda_1/\epsilon)$. The optimal solution $M = \mathrm{diag}(u_0)K\mathrm{diag}(u_1)$ needs to satisfy

$$\mathrm{diag}(u_0)K\mathrm{diag}(u_1)\mathbf{1} = f_0 \text{ and } \mathrm{diag}(u_1)K^T\mathrm{diag}(u_0)\mathbf{1} = f_1.$$

[1] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

How to find the values of λ_0 and λ_1?

Let $u_0 = \exp(\lambda_0/\epsilon)$, $u_1 = \exp(\lambda_1/\epsilon)$. The optimal solution $M = \mathrm{diag}(u_0)K\mathrm{diag}(u_1)$ needs to satisfy

$$\mathrm{diag}(u_0)K\mathrm{diag}(u_1)\mathbf{1} = f_0 \text{ and } \mathrm{diag}(u_1)K^T\mathrm{diag}(u_0)\mathbf{1} = f_1.$$

Theorem (Sinkhorn iterations [2])

For any matrix K with positive elements there are diagonal matrices $diag(u_0), diag(u_1)$ such that $M = diag(u_0)K diag(u_1)$ has prescribed row- and column-sums f_0 and f_1. The vectors u_0 and u_1 can be obtained by alternating marginalization:

$$u_0 = f_0./(Ku_1)$$
$$u_1 = f_1./(K^T u_0).$$

- Each iteration only requires the multiplications Ku_1 and $K^T u_0$. This is the bottle neck.
- Linear convergence rate.

Thus highly computationally efficient, allowing for solving large problems.

[1] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

[2] R. Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4), 402–405, 1967.

Sinkhorn iterations as dual coordinate ascent

Several ways to motivate Sinkhorn iterations [1]

- Diagonal matrix scaling
- Bregman projections
- Dykstra's algorithm

[1] J.D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2), A1111-A1138, 2015.

Sinkhorn iterations as dual coordinate ascent

Several ways to motivate Sinkhorn iterations [1]

- Diagonal matrix scaling
- Bregman projections
- Dykstra's algorithm

Here we will introduce yet another interpretation:

- Use a dual formulation
- The Sinkhorn iteration corresponds to dual coordinate ascent
- This allows us to generalize Sinkhorn iterations
- Approach for addressing inverse problem with optimal transport term

[1] J.D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2), A1111-A1138, 2015.

Sinkhorn iterations as dual coordinate ascent

- Lagrangian relaxation gave optimal form of the primal variable

$$M^* = \mathrm{diag}(u_0) K \mathrm{diag}(u_1)$$

Sinkhorn iterations as dual coordinate ascent

- Lagrangian relaxation gave optimal form of the primal variable

$$M^* = \mathrm{diag}(u_0) K \mathrm{diag}(u_1)$$

- The Lagrangian dual function:

$$\begin{aligned}\varphi(u_0, u_1) &:= \min_{M \geq 0} L(M, u_0, u_1) = L(M^*, u_0, u_1) = \dots \\ &= \epsilon \log(u_0)^T f_0 + \epsilon \log(u_1)^T f_1 - \epsilon u_0^T K u_1 + \epsilon nm.\end{aligned}$$

Sinkhorn iterations as dual coordinate ascent

- Lagrangian relaxation gave optimal form of the primal variable

$$M^* = \mathrm{diag}(u_0) K \mathrm{diag}(u_1)$$

- The Lagrangian dual function:

$$\begin{aligned} \varphi(u_0, u_1) &:= \min_{M \geq 0} L(M, u_0, u_1) = L(M^*, u_0, u_1) = \dots \\ &= \epsilon \log(u_0)^T f_0 + \epsilon \log(u_1)^T f_1 - \epsilon u_0^T K u_1 + \epsilon nm. \end{aligned}$$

- The dual problem is thus

$$\max_{u_0, u_1} \varphi(u_0, u_1)$$

Sinkhorn iterations as dual coordinate ascent

- Lagrangian relaxation gave optimal form of the primal variable
$$M^* = \mathrm{diag}(u_0) K \mathrm{diag}(u_1)$$
- The Lagrangian dual function:
$$\varphi(u_0, u_1) := \min_{M \geq 0} L(M, u_0, u_1) = L(M^*, u_0, u_1) = \dots$$
$$= \epsilon \log(u_0)^T f_0 + \epsilon \log(u_1)^T f_1 - \epsilon u_0^T K u_1 + \epsilon nm.$$
- The dual problem is thus
$$\max_{u_0, u_1} \varphi(u_0, u_1)$$
- Taking the gradient w.r.t u_0 and putting it equal to zero gives
$$0 = f_0./u_0 - K u_1,$$
and w.r.t u_1 gives
$$0 = f_1./u_1 - \left(u_0^T K\right)^T.$$
These are the Sinkhorn iterations!

Sinkhorn iterations as dual coordinate ascent

For g proper, convex and lower semicontinuous, we can now consider problems of the form

$$\min_{f_1} T_\epsilon(f_0, f_1) + g(f_1)$$

Sinkhorn iterations as dual coordinate ascent

For g proper, convex and lower semicontinuous, we can now consider problems of the form

$$\min_{f_1} T_\epsilon(f_0, f_1) + g(f_1) \quad = \quad \begin{array}{ll} \min\limits_{M \geq 0,\ f_1} & \mathrm{trace}(C^T M) + \epsilon D(M) + g(f_1) \\ \text{subject to} & f_0 = M\mathbf{1} \\ & f_1 = M^T\mathbf{1}. \end{array}$$

Sinkhorn iterations as dual coordinate ascent

For g proper, convex and lower semicontinuous, we can now consider problems of the form

$$\min_{f_1} T_\epsilon(f_0, f_1) + g(f_1) \quad = \quad \begin{array}{ll} \displaystyle\min_{M\geq 0,\ f_1} & \mathrm{trace}(C^T M) + \epsilon D(M) + g(f_1) \\ \text{subject to} & f_0 = M\mathbf{1} \\ & f_1 = M^T\mathbf{1}. \end{array}$$

Can be solve by dual coordinate ascent

$$u_0 = f_0./(Ku_1)$$
$$0 \in \partial g^*(-\epsilon \log(u_1))\frac{1}{u_1} - K^T u_0,$$

if the second inclusion can be solved efficiently.

Can be solvable when $\partial g^*(\cdot)$ is component-wise. Example of such cases:

- $g(\cdot) = \mathcal{I}_{\tilde{f}}(\cdot)$ indicator function on $\{\tilde{f}\}$ $\rightsquigarrow$ original optimal transport problem
- $g(\cdot) = \|\cdot\|_2^2$

Inverse problems with optimal mass transport priors

Consider the inverse problems

$$\min_{f_1 \geq 0} \quad \|\nabla f_1\|_1$$
$$\text{subject to} \quad \|Af_1 - g\|_2 \leq \kappa.$$

- TV-regularization term: $\|\nabla f_1\|_1$
- Forward model A, data g, and data mismatch term: $\|Af_1 - g\|_2$

Inverse problems with optimal mass transport priors

Consider the inverse problems

$$\min_{f_1 \geq 0} \quad \|\nabla f_1\|_1 + \text{“}f_1 \text{ close to } f_0\text{”}$$
$$\text{subject to } \ \|Af_1 - g\|_2 \leq \kappa.$$

- TV-regularization term: $\|\nabla f_1\|_1$
- Forward model A, data g, and data mismatch term: $\|Af_1 - g\|_2$
- Prior f_0

Inverse problems with optimal mass transport priors

Consider the inverse problems

$$\min_{f_1 \geq 0} \quad \|\nabla f_1\|_1 + \gamma T_\epsilon(f_0, f_1)$$
$$\text{subject to} \quad \|Af_1 - g\|_2 \leq \kappa.$$

- TV-regularization term: $\|\nabla f_1\|_1$
- Forward model A, data g, and data mismatch term: $\|Af_1 - g\|_2$
- Prior f_0

Inverse problems with optimal mass transport priors

Consider the inverse problems

$$\min_{f_1 \geq 0} \quad \|\nabla f_1\|_1 + \gamma T_\epsilon(f_0, f_1)$$
$$\text{subject to} \quad \|Af_1 - g\|_2 \leq \kappa.$$

- TV-regularization term: $\|\nabla f_1\|_1$
- Forward model A, data g, and data mismatch term: $\|Af_1 - g\|_2$
- Prior f_0

Large convex optimization problem with several terms, variable splitting common: ADMM, primal-dual hybrid gradient algorithm, primal-dual Douglas-Rachford algorithm.

Common tool in these algorithms: the proximal operator of the involved functions $\mathcal{F}$.

$$\mathrm{Prox}^{\sigma}_{\mathcal{F}}(h) = \operatorname*{argmin}_f \mathcal{F}(f) + \frac{1}{2\sigma}\|f - h\|_2^2.$$

[1] R.T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5), 877-898, 1976.

We want to compute the proximal of $T_\epsilon(f_0, \cdot)$, given by

$$\mathrm{Prox}^{\sigma}_{T_\epsilon(f_0,\cdot)}(h) = \operatorname*{argmin}_{f_1} T_\epsilon(f_0, f_1) + \frac{1}{2\sigma}\|f_1 - h\|_2^2.$$

Thus we want to solve

$$\min_{M\geq 0, f_1} \mathrm{trace}(C^T M) + \epsilon D(M) + \frac{1}{2\sigma}\|f_1 - h\|_2^2$$

$$\text{subject to } \quad f_0 = M\mathbf{1}$$

$$f_1 = M^T\mathbf{1}.$$

We want to compute the proximal of $T_\epsilon(f_0, \cdot)$, given by

$$\mathrm{Prox}^{\sigma}_{T_\epsilon(f_0,\cdot)}(h) = \operatorname*{argmin}_{f_1} T_\epsilon(f_0, f_1) + \frac{1}{2\sigma}\|f_1 - h\|_2^2.$$

Thus we want to solve

$$\min_{M \geq 0, f_1} \ \mathrm{trace}(C^T M) + \epsilon D(M) + \frac{1}{2\sigma}\|f_1 - h\|_2^2$$

$$\text{subject to} \quad f_0 = M\mathbf{1}$$

$$f_1 = M^T\mathbf{1}.$$

Using dual coordinate ascent, with $g(\cdot) = \frac{1}{2\sigma}\| \cdot - h\|_2^2$, we get the algorithm:

1. $u_0 = f_0./(Ku_1)$
2. $u_1 = \exp\left(\frac{h}{\sigma\epsilon} - \omega\left(\frac{h}{\sigma\epsilon} + \log\left(K^T u_0\right)\right) + \log(\sigma\epsilon)\right)$

- Here ω denotes the (elementwise) Wright omega function, i.e., $x = \log(\omega(x)) + \omega(x)$.
- Solved elementwise. Bottleneck is still computation of Ku_1, $K^T u_0$.

Compare to

1. $u_0 = f_0./(Ku_1)$
2. $u_1 = f_1./(K^T u_0)$

We want to compute the proximal of $T_\epsilon(f_0, \cdot)$, given by

$$\mathrm{Prox}^{\sigma}_{T_\epsilon(f_0,\cdot)}(h) = \operatorname*{argmin}_{f_1} T_\epsilon(f_0, f_1) + \frac{1}{2\sigma}\|f_1 - h\|_2^2.$$

Thus we want to solve

$$\min_{M\geq 0, f_1} \ \mathrm{trace}(C^T M) + \epsilon D(M) + \frac{1}{2\sigma}\|f_1 - h\|_2^2$$

$$\text{subject to } \ f_0 = M\mathbf{1}$$

$$f_1 = M^T\mathbf{1}.$$

Using dual coordinate ascent, with $g(\cdot) = \frac{1}{2\sigma}\| \cdot - h\|_2^2$, we get the algorithm:

1. $u_0 = f_0./(Ku_1)$
2. $u_1 = \exp\left(\frac{h}{\sigma\epsilon} - \omega\left(\frac{h}{\sigma\epsilon} + \log\left(K^T u_0\right)\right) + \log(\sigma\epsilon)\right)$

- Here ω denotes the (elementwise) Wright omega function, i.e., $x = \log(\omega(x)) + \omega(x)$.
- Solved elementwise. Bottleneck is still computation of Ku_1, $K^T u_0$.

Compare to

1. $u_0 = f_0./(Ku_1)$
2. $u_1 = f_1./(K^T u_0)$

Theorem

The algorithm is globally convergent, and with linear convergence rate.

Computerized Tomography (CT): imaging modality used in many areas, e.g., medicine.

- The object is probed with X-rays.
- Different materials attenuates X-rays differently $\Longrightarrow$ incoming and outgoing intensities gives information about the object.
- Simplest model

$$\int_{L_{r,\theta}} f(x)dx = \log\left(\frac{I_0}{I}\right),$$

- $f(x)$ is the attenuation in the point x, which is what we want to reconstruct,
- $L_{r,\theta}$ is the line along which the X-ray beam travels,
- I_0 and I are the the incoming and outgoing intensities.

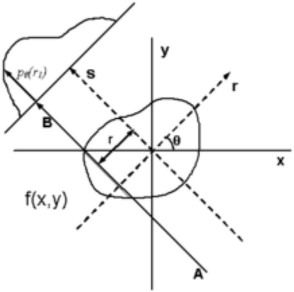
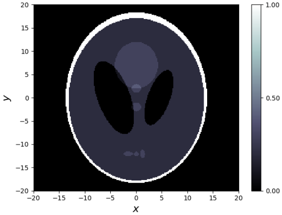


Illustration from Wikipedia

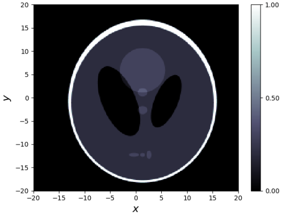
Example in computerized tomography

Parallel beam 2D CT example:

- Reconstruction space: 256×256 pixels
- Angles: 30 in $[\pi/4, 3\pi/4]$ (limited angle)
- Detector partition: uniform 350 bins
- Noise level 5%



(a) Shepp-Logan phantom



(b) Prior

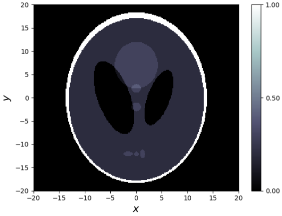
Example in computerized tomography

TV-regularization and ℓ_2^2 prior:

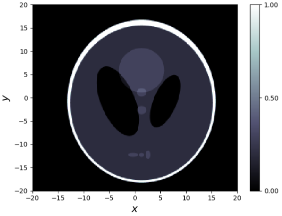
$$\min_{f_1} \quad \gamma \|f_0 - f_1\|_2^2 + \|\nabla f_1\|_1$$
$$\text{subject to} \quad \|A f_1 - w\|_2 \leq \kappa.$$

TV-regularization and optimal transport prior:

$$\min_{f_1} \quad \gamma T_\epsilon(f_0, f_1) + \|\nabla f_1\|_1$$
$$\text{subject to} \quad \|A f_1 - w\|_2 \leq \kappa.$$



(f) Shepp-Logan phantom



(g) Prior

Example in computerized tomography

TV-regularization and ℓ_2^2 prior:

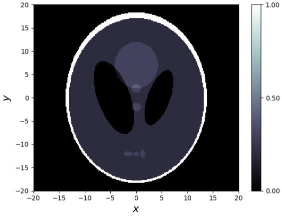
$$\min_{f_1} \quad \gamma \| f_0 - f_1 \|_2^2 + \| \nabla f_1 \|_1$$

$$\text{subject to} \quad \| A f_1 - w \|_2 \leq \kappa.$$

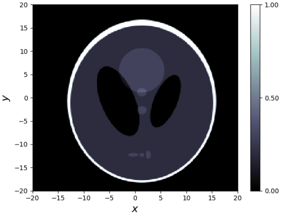
TV-regularization and optimal transport prior:

$$\min_{f_1} \quad \gamma T_\epsilon(f_0, f_1) + \| \nabla f_1 \|_1$$

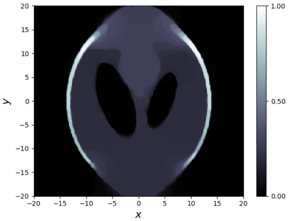
$$\text{subject to} \quad \| A f_1 - w \|_2 \leq \kappa.$$



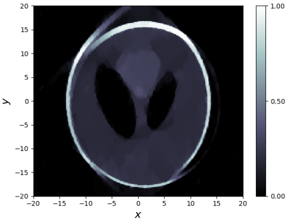
(k) Shepp-Logan phantom



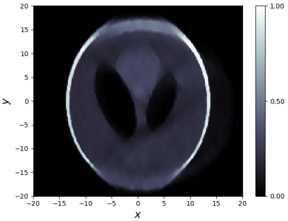
(l) Prior



(m) TV-regularization



(n) TV-regularization and ℓ_2^2-prior ($\gamma = 10$)



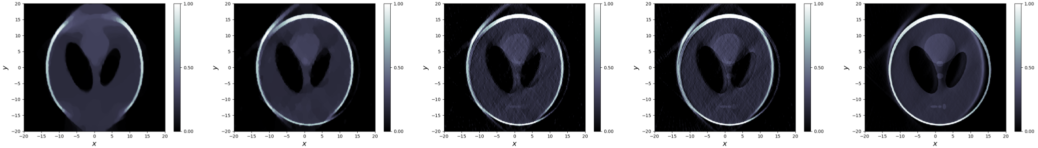
(o) TV-regularization and optimal transport prior ($\gamma = 4$)

Example in computerized tomography

Comparing different regularization parameters for the problem with ℓ_2^2 prior.

$$\min_{f_1} \quad \gamma \| f_0 - f_1 \|_2^2 + \| \nabla f_1 \|_1$$

$$\text{subject to } \| A f_1 - w \|_2 \leq \kappa.$$



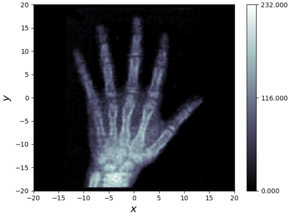
(p) $\gamma = 1$ (q) $\gamma = 10$ (r) $\gamma = 100$ (s) $\gamma = 1000$ (t) $\gamma = 10\,000$.

Figure: Reconstructions using ℓ_2 prior with different regularization parameters γ.

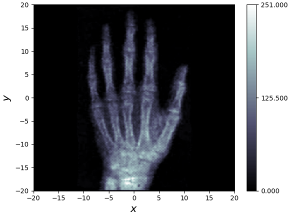
Example in computerized tomography

Parallel beam 2D CT example:

- Reconstruction space: 256×256 pixels
- Angles: 30 in $[0, \pi]$
- Detector partition: uniform 350 bins
- Noise level 3%



(a) Phantom

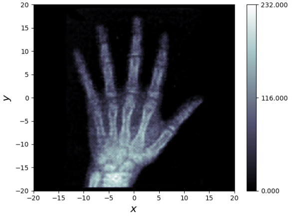


(b) Prior

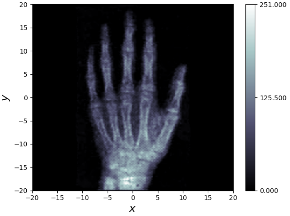
Example in computerized tomography

Parallel beam 2D CT example:

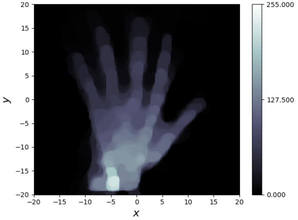
- Reconstruction space: 256×256 pixels
- Angles: 30 in $[0, \pi]$
- Detector partition: uniform 350 bins
- Noise level 3%



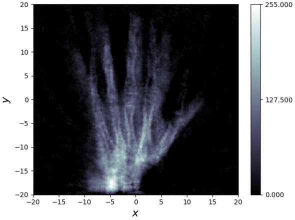
(f) Phantom



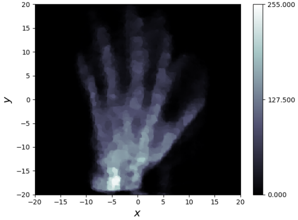
(g) Prior



(h) TV-regularization



(i) TV-regularization and ℓ_2^2-prior ($\gamma = 10$)



(j) TV-regularization and optimal transport prior ($\gamma = 4$)

Conclusions and further work

Conclusions

- Optimal mass transport - a viable framework for imaging applications
- Generalized Sinkhorn iteration for computing the proximal operator of optimal transport cost
- Use variable splitting for solving the inverse problem
- Application to CT reconstruction using optimal transport priors

Conclusions and further work

Conclusions

- Optimal mass transport - a viable framework for imaging applications
- Generalized Sinkhorn iteration for computing the proximal operator of optimal transport cost
- Use variable splitting for solving the inverse problem
- Application to CT reconstruction using optimal transport priors

Potential future directions:

- Application to spatiotemporal image reconstruction:

$$\operatorname*{arg\,min}_{f_0,\dots,f_T \in X} \sum_{j=0}^{T} \Big[\mathcal{G}(A(f_j), g_j) + \lambda \mathcal{F}(f_j) \Big] + \sum_{j=1}^{T} \gamma \mathcal{H}(f_{j-1}, f_j)$$

- More efficient ways of solving the dual problem?
- Learning for inverse problems using optimal transport as a loss function

Thank you for your attention!

Questions?