Optimal mass transport as a distance measure between images

Axel Ringh

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

21\textsuperscript{st} of June 2018
INRIA, Sophia-Antipolis, France
Acknowledgements

This is based on joint work with Johan Karlsson\textsuperscript{1}.


I acknowledge financial support from
- Swedish Research Council (VR)
- Swedish Foundation for Strategic Research (SSF)

The code is based on ODL: https://github.com/odlgroup/odl

\textsuperscript{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.
- $\lambda$ 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!
Incorporating prior information in variational schemes

How can one incorporate prior information in such a scheme?

\[
\text{arg min}_{f \in X} G(A(f), g) + \lambda F(f) + \gamma H(\tilde{f}, f)
\]

\(\tilde{f}\) is prior/template, \(H\) defines "closeness" to \(\tilde{f}\).

What is a good choice for \(H\)?

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

\[
\text{arg min}_{f_0, \ldots, f_T \in X_T} \sum_{j=0}^{T} \left[ G(A(f_j), g_j) + \lambda F(f_j) \right] + \sum_{j=1}^{T} \gamma H(f_j - 1, f_j)
\]
How can one incorporate prior information in such a scheme?

One way: consider

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

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

What is a good choice for $H$?
How can one incorporate prior information in such a scheme?

One way: consider

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

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

What is a good choice for $H$?

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

$$\arg \min_{f_0, \ldots, f_T \in X} \sum_{j=0}^{T} \left[ G(A(f_j), g_j) + \lambda F(f_j) \right] + \sum_{j=1}^{T} \gamma H(f_{j-1}, f_j)$$
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 \textit{pointwise}, e.g., using an \(L_p\) metric

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

\textbf{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: formulated optimal mass transport 1781.
Optimal transport of soil for construction of forts and roads.

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

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

\[
\int_X c(x, \phi(x)) f_0(x) \, dx
\]

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

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

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

where $\phi$ is mass preserving map from $f_0$ to $f_1$:

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

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 \( \phi \) is mass preserving map from \( f_0 \) to \( f_1 \):

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

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

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

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:

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

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 $f_0, f_1 : X \rightarrow \mathbb{R}^+$, find a transport plan $M : X \times X \rightarrow \mathbb{R}^+$, 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:

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

s.t. $f_0(x_0) = \int_X M(x_0, x_1) dx_1, \ x_0 \in X$

$f_1(x_1) = \int_X M(x_0, x_1) dx_0, \ x_1 \in X$. 
Define a **distance** between two functions $f_0(x)$ and $f_1(x)$ using **optimal transport**

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

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

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

subject to

$$\sum_{j=1}^{m} m_{ij} = f_0(i), \ i = 1, \ldots, n$$

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

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

subject to

\[
\sum_{j=1}^{m} m_{ij} = f_0(i), \quad i = 1, \ldots, n
\]

\[
\sum_{i=1}^{n} m_{ij} = f_1(j), \quad j = 1, \ldots, m
\]

\[
\min_{M \geq 0} \text{trace}(C^T M)
\]

subject to

\[
M 1_m = f_0
\]

\[
M^T 1_n = f_1
\]
Optimal mass transport - discrete formulation

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

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

subject to

$$
\sum_{j=1}^{m} m_{ij} = f_0(i), \ i = 1, \ldots, n
$$

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

$M \in \mathbb{R}^{256^2 \times 256^2}$ \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],

$T_\epsilon(f_0, f_1) := \min_{M \geq 0} \text{trace}(C^T M) + \epsilon D(M)\]

subject to $f_0 = M1$

$f_1 = M^T 1$.

---

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],

$$T_\epsilon(f_0, f_1) := \min_{M \geq 0} \text{trace}(C^T M) + \epsilon D(M)$$

subject to $f_0 = M1$

$f_1 = M^T 1$.

Using Lagrangian relaxation gives

$$L(M, \lambda_0, \lambda_1) = \text{trace}(C^T M) + \epsilon D(M) + \lambda_0^T (f_0 - M1) + \lambda_1^T (f_1 - M^T 1).$$

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],

$$T_\epsilon(f_0, f_1) := \min_{M \geq 0} \text{trace}(C^T M) + \epsilon D(M)$$

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

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

- Using **Lagrangian relaxation** gives

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

- Given dual variables $\lambda_0, \lambda_1$, the minimum $m_{ij}$ is

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

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],
\[
T_\epsilon(f_0, f_1) := \min_{M \geq 0} \quad \text{trace}(C^T M) + \epsilon D(M)
\]
subject to
\[
f_0 = M 1 \quad f_1 = M^T 1.
\]

- Using Lagrangian relaxation gives
\[
L(M, \lambda_0, \lambda_1) = \text{trace}(C^T M) + \epsilon D(M) + \lambda_0^T (f_0 - M 1) + \lambda_1^T (f_1 - M^T 1).
\]

- Given dual variables \( \lambda_0, \lambda_1 \), 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 \( m_{ij} = e^{\lambda_0(i) / \epsilon} e^{-c_{ij} / \epsilon} e^{\lambda_1(j) / \epsilon} \)
\[
M = \text{diag}(e^{\lambda_0^T / \epsilon}) K \text{diag}(e^{\lambda_1 / \epsilon})
\]
where \( K = \exp(-C / \epsilon) \). Here and in what follows \( \exp(\cdot), \log(\cdot), ./, \odot \) denotes the element-wise function.

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],

$$T_{\epsilon}(f_0, f_1) := \min_{M \geq 0} \text{trace}(C^T M) + \epsilon D(M)$$
subject to $f_0 = M1$
$f_1 = M^T 1$.

Using Lagrangian relaxation gives

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

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

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

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

Background
Optimal mass transport - Sinkhorn iterations

How to find the values of $\lambda_0$ and $\lambda_1$?

Theorem (Sinkhorn iterations [2])
For any matrix $K$ with positive elements there are diagonal matrices $\text{diag}(u_0)$, $\text{diag}(u_1)$ such that $M = \text{diag}(u_0) K \text{diag}(u_1)$ has prescribed row- and column-sums $f_0$ and $f_1$. The vectors $u_0$ and $u_1$ can be obtained by alternating marginalization:

$$u_0 = f_0 / (K u_1)$$

$$u_1 = f_1 / (K^T u_0)$$

Each iteration only requires the multiplications $K u_1$ and $K^T u_0$. This is the bottleneck.

Linear convergence rate. Thus highly computationally efficient, allowing for solving large problems.


How to find the values of $\lambda_0$ and $\lambda_1$?

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

$$\text{diag}(u_0)K\text{diag}(u_1)1 = f_0 \text{ and } \text{diag}(u_1)K^T\text{diag}(u_0)1 = f_1.$$
How to find the values of $\lambda_0$ and $\lambda_1$?

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

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

**Theorem (Sinkhorn iterations [2])**

For any matrix $K$ with positive elements there are diagonal matrices $\text{diag}(u_0), \text{diag}(u_1)$ such that $M = \text{diag}(u_0)K\text{diag}(u_1)$ 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.


Several ways to motivate Sinkhorn iterations [1]

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

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

Lagrangian relaxation gave optimal form of the primal variable

\[ M^* = \text{diag}(u_0)K\text{diag}(u_1) \]
Sinkhorn iterations as dual coordinate ascent

- Lagrangian relaxation gave optimal form of the primal variable
  \[ M^* = \text{diag}(u_0)K\text{diag}(u_1) \]

- The Lagrangian dual function:
  \[
  \varphi(u_0, u_1) := \min_{M \geq 0} L(M, u_0, u_1) = L(M^*, u_0, u_1) = \ldots
  = \epsilon \log(u_0)^T f_0 + \epsilon \log(u_1)^T f_1 - \epsilon u_0^T Ku_1 + \epsilon nm.
  \]
Sinkhorn iterations as dual coordinate ascent

- Lagrangian relaxation gave optimal form of the primal variable
  \[ M^* = \text{diag}(u_0)K\text{diag}(u_1) \]

- The Lagrangian dual function:
  \[ \varphi(u_0, u_1) := \min_{M \geq 0} L(M, u_0, u_1) = L(M^*, u_0, u_1) = \ldots \]
  \[ = \epsilon \log(u_0)^T f_0 + \epsilon \log(u_1)^T f_1 - \epsilon u_0^T Ku_1 + \epsilon nm. \]

- 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
  \[ M^* = \text{diag}(u_0)K\text{diag}(u_1) \]

- The Lagrangian dual function:
  \[
  \varphi(u_0, u_1) := \min_{M \geq 0} L(M, u_0, u_1) = L(M^*, u_0, u_1) = \ldots \\
  = \epsilon \log(u_0)^T f_0 + \epsilon \log(u_1)^T f_1 - \epsilon u_0^T Ku_1 + \epsilon nm.
  \]

- 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
  \[ 0 = f_0 ./ u_0 - Ku_1, \]
  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

$$\min_{f_1} T_\epsilon(f_0, f_1) + g(f_1) = \min_{M \geq 0, f_1} \text{trace}(C^T M) + \epsilon D(M) + g(f_1)$$

subject to

$$f_0 = M \mathbf{1}$$
$$f_1 = M^T \mathbf{1}.$$
Sinkhorn iterations as dual coordinate ascent

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

$$\min_{f_1} T_\epsilon(f_0, f_1) + g(f_1) = \min_{M \geq 0, f_1} \text{trace}(C^T M) + \epsilon D(M) + g(f_1)$$

subject to

$$f_0 = M\mathbf{1}$$

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

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
- $g(\cdot) = \|\cdot\|_2^2$
Consider the inverse problems

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

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

Consider the inverse problems

\[ \min_{f_1 \geq 0} \| \nabla f_1 \|_1 + "f_1 \text{ close to } f_0" \]

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 \( f_0 \)
Consider the inverse problems

$$\min_{f_1 \geq 0} \|\nabla f_1\|_1 + \gamma T_\epsilon(f_0, f_1)$$

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

\[
\min_{f_1 \geq 0} \| \nabla f_1 \|_1 + \gamma \mathcal{E}(f_0, f_1)
\]

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 \( f_0 \)


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

\[
\text{Prox}^\sigma_{\mathcal{F}}(h) = \arg \min_f \mathcal{F}(f) + \frac{1}{2\sigma} \| f - h \|^2_2.
\]

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

\[
\text{Prox}_{T_\epsilon(f_0, \cdot)}(h) = \arg\min_{f_1} T_\epsilon(f_0, f_1) + \frac{1}{2\sigma} \|f_1 - h\|_2^2.
\]

Thus we want to solve

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

subject to

\[
\begin{align*}
f_0 &= M \mathbf{1} \\
f_1 &= M^T \mathbf{1}.
\end{align*}
\]
We want to compute the proximal of $T_\epsilon(f_0, \cdot)$, given by

$$\text{Prox}_{T_\epsilon(f_0, \cdot)}^\sigma(h) = \arg\min_{f_1} T_\epsilon(f_0, f_1) + \frac{1}{2\sigma} \|f_1 - h\|_2^2.$$ 

Thus we want to solve

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

subject to $f_0 = M 1$

$$f_1 = M^T 1.$$ 

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 $\omega$ 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. $u_1 = f_1 ./ (K^T u_0)$
Inverse problems with optimal mass transport priors
Generalized Sinkhorn iterations

We want to compute the proximal of $T_\epsilon(f_0, \cdot)$, given by
\[
\text{Prox}_T^{\sigma}(h) = \arg\min_{f_1} T_\epsilon(f_0, f_1) + \frac{1}{2\sigma} \|f_1 - h\|_2^2.
\]
Thus we want to solve
\[
\min_{M \geq 0, f_1} \text{trace}(C^T M) + \epsilon D(M) + \frac{1}{2\sigma} \|f_1 - h\|_2^2
\]
subject to $f_0 = M 1$
$f_1 = M^T 1$.

Using dual coordinate ascent, with $g(\cdot) = \frac{1}{2\sigma} \|\cdot - h\|_2^2$, we get the algorithm:
1. $u_0 = f_0/(K u_1)$
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 $\omega$ denotes the (elementwise) Wright omega function, i.e., $x = \log(\omega(x)) + \omega(x)$.
   - Solved elementwise. Bottleneck is still computation of $K u_1, K^T u_0$.

**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 \( \Rightarrow \) incoming and outgoing intensities gives information about the object.
- Simplest model

\[
\int_{L_{r,\theta}} f(x) \, dx = \log \left( \frac{l_0}{l} \right),
\]

- \( f(x) \) 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,
- \( l_0 \) and \( l \) are the the incoming and outgoing intensities.

Illustration from Wikipedia
Example in computerized tomography

Parallel beam 2D CT example:
- Reconstruction space: 256 \times 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 $\ell_2^2$ prior:

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

subject to $\|Af_1 - w\|_2 \leq \kappa$.

TV-regularization and optimal transport prior:

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

subject to $\|Af_1 - w\|_2 \leq \kappa$.

(f) Shepp-Logan phantom  
(g) Prior
Example in computerized tomography

TV-regularization and $\ell_2^2$ prior:
\[
\min_{f_1} \gamma \|f_0 - f_1\|_2^2 + \|\nabla f_1\|_1
\]
subject to $\|Af_1 - w\|_2 \leq \kappa$.

TV-regularization and optimal transport prior:
\[
\min_{f_1} \gamma T_c(f_0, f_1) + \|\nabla f_1\|_1
\]
subject to $\|Af_1 - w\|_2 \leq \kappa$.

(k) Shepp-Logan phantom

(l) Prior

(m) TV-regularization

(n) TV-regularization and $\ell_2^2$-prior ($\gamma = 10$)

(o) TV-regularization and optimal transport prior ($\gamma = 4$)
Example in computerized tomography

Comparing different regularization parameters for the problem with $\ell_2^2$ prior.

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

subject to $\| Af_1 - w \|_2 \leq \kappa$.

Figure: Reconstructions using $\ell_2$ prior with different regularization parameters $\gamma$. 

(p) $\gamma = 1$    (q) $\gamma = 10$    (r) $\gamma = 100$    (s) $\gamma = 1000$    (t) $\gamma = 10000$. 

Example in computerized tomography

Parallel beam 2D CT example:
- Reconstruction space: 256 $\times$ 256 pixels
- Angles: 30 in $[0, \pi]$ [0x0]
- Detector partition: uniform 350 bins
- Noise level 3% [234x120]

(a) Phantom  
(b) Prior
Example in computerized tomography

Parallel beam 2D CT example:
- Reconstruction space: $256 \times 256$ 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 $\ell_2^2$-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

Potential future directions:

\[
\arg\min_{f_0, \ldots, f_T} \sum_{j=0}^{T} [G(A(f_j), g_j) + \lambda F(f_j)] + \sum_{j=1}^{T} \gamma H(f_j - 1, f_j)
\]

- More efficient ways of solving the dual problem?
- Learning for inverse problems using optimal transport as a loss function
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:
  \[
  \arg \min_{f_0, \ldots, f_T \in X} \sum_{j=0}^{T} \left[ G(A(f_j), g_j) + \lambda F(f_j) \right] + \sum_{j=1}^{T} \gamma \mathcal{H}(f_{j-1}, f_j)
  \]

- 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?