Optimization problems containing optimal transport costs: examples and computational methods

Axel Ringh[1]

[1]Department of Electronic and Computer Engineering,
The Hong Kong University of Science and Technology.

10th of December 2019

Workshop: Computational optimal transport for applications in control and estimation

58th Conference on Decision and Control (CDC), Nice, France

香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

Outline

- Optimal transport recap
- Optimization problems with an optimal transport cost and generalized Sinkhorn iterations
- Variable splitting and the proximal operator of entropy-regularized optimal transport
- Example: variational regularization of inverse problems

Optimal transport recap

Optimal transport distance between two functions $f_0(x)$ and $f_1(x)$ is defined as

$$T(f_0, f_1) := \begin{cases} \min\limits_{M \geq 0} & \int_{X \times X} c(x^{(0)}, x^{(1)}) M(x^{(0)}, x^{(1)}) dx^{(0)} dx^{(1)} \\ \text{s.t.} & f_0(x^{(0)}) = \int_X M(x^{(0)}, x^{(1)}) dx^{(1)}, \ x^{(0)} \in X \\ & f_1(x^{(1)}) = \int_X M(x^{(0)}, x^{(1)}) dx^{(0)}, \ x^{(1)} \in X. \end{cases}$$

for some cost function $c : X \times X \to \mathbb{R}_+$.

Optimal transport recap

Optimal transport distance between two functions $f_0(x)$ and $f_1(x)$ is defined as

$$T(f_0, f_1) := \begin{cases} \min\limits_{M \geq 0} & \int_{X \times X} c(x^{(0)}, x^{(1)}) M(x^{(0)}, x^{(1)}) dx^{(0)} dx^{(1)} \\ \text{s.t.} & f_0(x^{(0)}) = \int_X M(x^{(0)}, x^{(1)}) dx^{(1)}, \; x^{(0)} \in X \\ & f_1(x^{(1)}) = \int_X M(x^{(0)}, x^{(1)}) dx^{(0)}, \; x^{(1)} \in X. \end{cases}$$

for some cost function $c : X \times X \to \mathbb{R}_+$.

Discretized version:

- vectors $f_0 \in \mathbb{R}^n$, $f_1 \in \mathbb{R}^n$
- cost matrix $C = [c_{ij}] \in \mathbb{R}^{n \times n}$, where c_{ij} is the transportation cost $c(x_i, x_j)$
- transportation plan $M = [m_{ij}] \in \mathbb{R}^{n \times n}$.

$$T(f_0, f_1) := \begin{cases} \min\limits_{m_{ij} \geq 0} & \sum\limits_{i=1}^{n} \sum\limits_{j=1}^{n} c_{ij} m_{ij} \\ \text{s.t.} & \sum\limits_{j=1}^{n} m_{ij} = f_0(i), \; i = 1, \ldots, n, \\ & \sum\limits_{i=1}^{n} m_{ij} = f_1(j), \; j = 1, \ldots, n. \end{cases}$$

Optimal transport recap

Optimal transport distance between two functions $f_0(x)$ and $f_1(x)$ is defined as

$$T(f_0, f_1) := \begin{cases} \min\limits_{M \geq 0} & \int_{X \times X} c(x^{(0)}, x^{(1)}) M(x^{(0)}, x^{(1)}) dx^{(0)} dx^{(1)} \\ \text{s.t.} & f_0(x^{(0)}) = \int_X M(x^{(0)}, x^{(1)}) dx^{(1)}, \ x^{(0)} \in X \\ & f_1(x^{(1)}) = \int_X M(x^{(0)}, x^{(1)}) dx^{(0)}, \ x^{(1)} \in X. \end{cases}$$

for some cost function $c : X \times X \to \mathbb{R}_+$.

Discretized version:

- vectors $f_0 \in \mathbb{R}^n$, $f_1 \in \mathbb{R}^n$
- cost matrix $C = [c_{ij}] \in \mathbb{R}^{n \times n}$, where c_{ij} is the transportation cost $c(x_i, x_j)$
- transportation plan $M = [m_{ij}] \in \mathbb{R}^{n \times n}$.

$$T(f_0, f_1) := \begin{cases} \min\limits_{m_{ij} \geq 0} & \sum_{i=1}^n \sum_{j=1}^n c_{ij} m_{ij} \\ \text{s.t.} & \sum_{j=1}^n m_{ij} = f_0(i), \ i = 1, \dots, n, \\ & \sum_{i=1}^n m_{ij} = f_1(j), \ j = 1, \dots, n. \end{cases} \iff T(f_0, f_1) := \begin{cases} \min\limits_{M \geq 0} & \text{trace}(C^T M) \\ \text{s.t.} & M\mathbf{1} = f_0 \\ & M^T \mathbf{1} = f_1 \end{cases}$$

Optimal transport recap

Recently proposed to solve via entropy regularization [1]: $D(M) = \sum_{i,j}(m_{ij}\log(m_{ij}) - m_{ij} + 1)$,

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

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

Optimal transport recap

Recently proposed to solve via entropy regularization [1]: $D(M) = \sum_{i,j}(m_{ij}\log(m_{ij}) - m_{ij} + 1)$,

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

- Let $\exp(\cdot)$, $\log(\cdot)$, $./$, $\odot$ denotes the element-wise function.

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

Optimal transport recap

Recently proposed to solve via entropy regularization [1]: $D(M) = \sum_{i,j}(m_{ij}\log(m_{ij}) - m_{ij} + 1)$,

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

- Let $\exp(\cdot)$, $\log(\cdot)$, $./$, $\odot$ denotes the element-wise function.
- For $K = \exp(-C/\epsilon)$, the solution is of the form

$$M = \operatorname{diag}(u_0) K \operatorname{diag}(u_1)$$

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

Optimal transport recap

Recently proposed to solve via entropy regularization [1]: $D(M) = \sum_{i,j}(m_{ij}\log(m_{ij}) - m_{ij} + 1)$,

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

- Let $\exp(\cdot)$, $\log(\cdot)$, $./$, $\odot$ denotes the element-wise function.
- For $K = \exp(-C/\epsilon)$, the solution is of the form

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

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:

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

[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.

Optimization problems with an optimal transport cost

Consider problems of the form:

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

where $\mathcal{G}$ proper, convex and lower semicontinuous.

Optimization problems with an optimal transport cost

Consider problems of the form:

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

where $\mathcal{G}$ proper, convex and lower semicontinuous.

Optimization problems with an optimal transport cost

Consider problems of the form:

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

where $\mathcal{G}$ proper, convex and lower semicontinuous.

Problem well-posed if there exists $f_1 \geq 0$ such that

- $\mathbf{1}^T f_1 = \mathbf{1}^T f_0$,
- $\mathcal{G}(f_1) < \infty$.

In this case, the problem is convex and there exists an optimal solution.

Optimization problems with an optimal transport cost

Consider problems of the form:

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

where $\mathcal{G}$ proper, convex and lower semicontinuous.

Problem well-posed if there exists $f_1 \geq 0$ such that

- $\mathbf{1}^T f_1 = \mathbf{1}^T f_0$,
- $\mathcal{G}(f_1) < \infty$.

In this case, the problem is convex and there exists an optimal solution.

How to solve this problem?

Recap on how to derive the Sinkhorn iterations for

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

Recap on how to derive the Sinkhorn iterations for

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

- Using Lagrangian relaxation gives

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

Recap on how to derive the Sinkhorn iterations for

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

- 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

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

Recap on how to derive the Sinkhorn iterations for

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

- 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 λ_0, λ_1, the minimum m_{ij} is

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

- Solve for m_{ij} to get

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

Recap on how to derive the Sinkhorn iterations for

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

- 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 λ_0, λ_1, the minimum m_{ij} is

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

- Solve for m_{ij} to get

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

- Change of variables: $u_0 = \exp(\lambda_0/\epsilon)$, $u_1 = \exp(\lambda_1/\epsilon)$. The optimal solution is of the form

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

where $K = \exp(-C/\epsilon)$.

One way to interpreted the Sinkhorn iterations: coordinate ascent in the Lagrangian dual.

One way to interpreted the Sinkhorn iterations: coordinate ascent in the Lagrangian dual.

- Lagrangian relaxation gave optimal form of the primal variable

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

One way to interpreted the Sinkhorn iterations: coordinate ascent in the Lagrangian dual.

- Lagrangian relaxation gave optimal form of the primal variable

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

- The Lagrangian dual function:

$$\varphi(u_0, u_1) := \min_{M \geq 0} L(M, u_0, u_1) = L(M^*, u_0, u_1) = \ldots = \epsilon \log(u_0)^T f_0 + \epsilon \log(u_1)^T f_1 - \epsilon u_0^T K u_1 + \epsilon n^2.$$

One way to interpreted the Sinkhorn iterations: coordinate ascent in the Lagrangian dual.

- Lagrangian relaxation gave optimal form of the primal variable

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

- The Lagrangian dual function:

$$\varphi(u_0, u_1) := \min_{M \geq 0} L(M, u_0, u_1) = L(M^*, u_0, u_1) = \ldots = \epsilon \log(u_0)^T f_0 + \epsilon \log(u_1)^T f_1 - \epsilon u_0^T K u_1 + \epsilon n^2.$$

- The dual problem is thus

$$\max_{u_0, u_1 \geq 0} \varphi(u_0, u_1)$$

One way to interpreted the Sinkhorn iterations: coordinate ascent in the Lagrangian dual.

- Lagrangian relaxation gave optimal form of the primal variable

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

- The Lagrangian dual function:

$$\varphi(u_0, u_1) := \min_{M \geq 0} L(M, u_0, u_1) = L(M^*, u_0, u_1) = \ldots = \epsilon \log(u_0)^T f_0 + \epsilon \log(u_1)^T f_1 - \epsilon u_0^T K u_1 + \epsilon n^2.$$

- The dual problem is thus

$$\max_{u_0, u_1 \geq 0} \varphi(u_0, u_1)$$

- Taking the gradient w.r.t u_0 and putting it equal to zero gives

$$\epsilon f_0 ./ u_0 - \epsilon K u_1 = 0$$

One way to interpreted the Sinkhorn iterations: coordinate ascent in the Lagrangian dual.

- Lagrangian relaxation gave optimal form of the primal variable

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

- The Lagrangian dual function:

$$\varphi(u_0, u_1) := \min_{M \geq 0} L(M, u_0, u_1) = L(M^*, u_0, u_1) = \ldots = \epsilon \log(u_0)^T f_0 + \epsilon \log(u_1)^T f_1 - \epsilon u_0^T K u_1 + \epsilon n^2.$$

- The dual problem is thus

$$\max_{u_0, u_1 \geq 0} \varphi(u_0, u_1)$$

- Taking the gradient w.r.t u_0 and putting it equal to zero gives

$$\epsilon f_0 ./ u_0 - \epsilon K u_1 = 0 \qquad \rightsquigarrow \qquad u_0 = f_0 ./ (K u_1),$$

One way to interpreted the Sinkhorn iterations: coordinate ascent in the Lagrangian dual.

- Lagrangian relaxation gave optimal form of the primal variable

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

- The Lagrangian dual function:

$$\varphi(u_0, u_1) := \min_{M \geq 0} L(M, u_0, u_1) = L(M^*, u_0, u_1) = \ldots = \epsilon \log(u_0)^T f_0 + \epsilon \log(u_1)^T f_1 - \epsilon u_0^T K u_1 + \epsilon n^2.$$

- The dual problem is thus

$$\max_{u_0, u_1 \geq 0} \varphi(u_0, u_1)$$

- Taking the gradient w.r.t u_0 and putting it equal to zero gives

$$\epsilon f_0 ./ u_0 - \epsilon K u_1 = 0 \qquad \rightsquigarrow \qquad u_0 = f_0 ./ (K u_1),$$

and w.r.t u_1 gives

$$\epsilon f_1 ./ u_1 - \epsilon \left(u_0^T K \right)^T = 0 \qquad \rightsquigarrow \qquad u_1 = f_1 ./ (K^T u_0).$$

These are the Sinkhorn iterations! (cf. [1])

[1] P. Tseng. Dual ascent methods for problems with strictly convex costs and linear constraints: A unified approach. *SIAM Journal on Control and Optimization*, 28(1), 214–242, 1990.

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

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

Lagrangian dual problem

$$\max_{u_0, u_1} \varphi(u_0, u_1) = \max_{u_0, u_1} \epsilon \log(u_0)^T f_0 - \mathcal{G}^*(-\epsilon \log(u_1)) - \epsilon u_0^T K u_1 + \epsilon n^2,$$

where $\mathcal{G}^*(u) := \sup_f u^T f - \mathcal{G}(f)$ is the Fenchel dual.

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

Lagrangian dual problem

$$\max_{u_0, u_1} \varphi(u_0, u_1) = \max_{u_0, u_1} \epsilon \log(u_0)^T f_0 - \mathcal{G}^*(-\epsilon \log(u_1)) - \epsilon u_0^T K u_1 + \epsilon n^2,$$

where $\mathcal{G}^*(u) := \sup_f u^T f - \mathcal{G}(f)$ is the Fenchel dual.

Can be solve by dual coordinate ascent

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

if the second inclusion can be solved efficiently.

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

Lagrangian dual problem

$$\max_{u_0, u_1} \varphi(u_0, u_1) = \max_{u_0, u_1} \epsilon \log(u_0)^T f_0 - \mathcal{G}^*(-\epsilon \log(u_1)) - \epsilon u_0^T K u_1 + \epsilon n^2,$$

where $\mathcal{G}^*(u) := \sup_f u^T f - \mathcal{G}(f)$ is the Fenchel dual.

Can be solve by dual coordinate ascent

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

if the second inclusion can be solved efficiently.

The second inclusion can be efficiently solved when $\partial\mathcal{G}^*(\cdot)$ is component-wise.
Example of such cases:

- $\mathcal{G}(\cdot) = \mathcal{I}_{\tilde{f}}(\cdot)$ indicator function on $\{\tilde{f}\}$ $\rightsquigarrow$ $\mathcal{G}^*(\cdot) = \cdot^T \tilde{f}$ $\rightsquigarrow$ $\partial\mathcal{G}^*(\cdot) = \tilde{f}$
- $\mathcal{G}(\cdot) = \|\cdot\|_2^2$ $\rightsquigarrow$ $\mathcal{G}^*(\cdot) = \frac{1}{4}\|\cdot\|_2^2$ $\rightsquigarrow$ $\partial\mathcal{G}^*(\cdot) = \frac{1}{2}\cdot$

How to solve

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

for more general functions $\mathcal{G}$?

Intermission: ADMM and variable splitting

Consider

$$\begin{aligned} \min_{y,z} \quad & \mathcal{H}(y) + \mathcal{G}(z) \\ \text{subject to} \quad & Ax + Bz = c \end{aligned}$$

Intermission: ADMM and variable splitting

Consider

$$\begin{aligned} \min_{y,z} \quad & \mathcal{H}(y) + \mathcal{G}(z) \\ \text{subject to} \quad & Ax + Bz = c \end{aligned}$$

Can be solved by ADMM [1]: for $\rho > 0$

$$\begin{aligned} y^{k+1} &= \arg\min_{y} \mathcal{H}(y) + \frac{2}{\rho}\|Ay + Bz^k - c + u^k\|_2^2 \\ z^{k+1} &= \arg\min_{z} \mathcal{G}(z) + \frac{2}{\rho}\|Ay^{k+1} + Bz - c + u^k\|_2^2 \\ u^{k+1} &= u^k + Ay^{k+1} + Bz^{k+1} - c \end{aligned}$$

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1), 1-122, 2011.

Intermission: ADMM and variable splitting

Consider

$$\begin{aligned} \min_{y,z} \quad & \mathcal{H}(y) + \mathcal{G}(z) \\ \text{subject to} \quad & Ax + Bz = c \end{aligned}$$

Can be solved by ADMM [1]: for $\rho > 0$

$$\begin{aligned} y^{k+1} &= \arg\min_{y} \mathcal{H}(y) + \frac{2}{\rho}\|Ay + Bz^k - c + u^k\|_2^2 \\ z^{k+1} &= \arg\min_{z} \mathcal{G}(z) + \frac{2}{\rho}\|Ay^{k+1} + Bz - c + u^k\|_2^2 \\ u^{k+1} &= u^k + Ay^{k+1} + Bz^{k+1} - c \end{aligned}$$

Special case:

$$\min_{y} \mathcal{H}(y) + \mathcal{G}(y)$$

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1), 1-122, 2011.

Intermission: ADMM and variable splitting

Consider

$$\begin{aligned} \min_{y,z} \quad & \mathcal{H}(y) + \mathcal{G}(z) \\ \text{subject to} \quad & Ax + Bz = c \end{aligned}$$

Can be solved by ADMM [1]: for $\rho > 0$

$$\begin{aligned} y^{k+1} &= \arg\min_{y} \mathcal{H}(y) + \frac{2}{\rho}\|Ay + Bz^k - c + u^k\|_2^2 \\ z^{k+1} &= \arg\min_{z} \mathcal{G}(z) + \frac{2}{\rho}\|Ay^{k+1} + Bz - c + u^k\|_2^2 \\ u^{k+1} &= u^k + Ay^{k+1} + Bz^{k+1} - c \end{aligned}$$

Special case:

$$\min_{y} \mathcal{H}(y)+\mathcal{G}(y) \quad \rightsquigarrow \quad \begin{aligned} \min_{y,z} \quad & \mathcal{H}(y) + \mathcal{G}(z) \\ \text{subject to} \quad & y - z = 0 \end{aligned}$$

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1), 1-122, 2011.

Intermission: ADMM and variable splitting

Consider

$$\begin{aligned} \min_{y,z} \quad & \mathcal{H}(y) + \mathcal{G}(z) \\ \text{subject to} \quad & Ax + Bz = c \end{aligned}$$

Can be solved by ADMM [1]: for $\rho > 0$

$$\begin{aligned} y^{k+1} &= \arg\min_{y} \mathcal{H}(y) + \frac{2}{\rho}\|Ay + Bz^k - c + u^k\|_2^2 \\ z^{k+1} &= \arg\min_{z} \mathcal{G}(z) + \frac{2}{\rho}\|Ay^{k+1} + Bz - c + u^k\|_2^2 \\ u^{k+1} &= u^k + Ay^{k+1} + Bz^{k+1} - c \end{aligned}$$

Special case:

$$\min_{y} \mathcal{H}(y)+\mathcal{G}(y) \quad \rightsquigarrow \quad \begin{aligned} \min_{y,z} \quad & \mathcal{H}(y) + \mathcal{G}(z) \\ \text{subject to} \quad & y - z = 0 \end{aligned} \quad \rightsquigarrow \quad \begin{aligned} y^{k+1} &= \arg\min_{y} \mathcal{H}(y) + \frac{2}{\rho}\|y - z^k + u^k\|_2^2 \\ z^{k+1} &= \arg\min_{z} \mathcal{G}(z) + \frac{2}{\rho}\|y^{k+1} - z + u^k\|_2^2 \\ u^{k+1} &= u^k + y^{k+1} - z^{k+1} \end{aligned}$$

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1), 1-122, 2011.

Variable splitting

ADMM is a special case of so-called variable splitting.

Common for large convex optimization problem with several terms.

Examples of methods

- ADMM [1]
- primal-dual hybrid gradient algorithm (Chambolle-Pock) [2]
- primal-dual Douglas-Rachford [3]

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1), 1-122, 2011.

[2] A. Chambolle, and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1), 120-145, 2011.

[3] R.I. Boţ, and C. Hendrich. A Douglas–Rachford type primal-dual method for solving inclusions with mixtures of composite and parallel-sum type monotone operators. *SIAM Journal on Optimization*, 23(4), 2541-2565, 2013.

Variable splitting

ADMM is a special case of so-called variable splitting.

Common for large convex optimization problem with several terms.

Examples of methods

- ADMM [1]
- primal-dual hybrid gradient algorithm (Chambolle-Pock) [2]
- primal-dual Douglas-Rachford [3]

Common tool in these algorithms: the proximal operator of the involved functions $\mathcal{H}$ and $\mathcal{G}$ [4, 5]

$$\mathrm{Prox}^{\sigma}_{\mathcal{H}}(h) \;=\; \arg\min_{f} \; \mathcal{H}(f) + \frac{1}{2\sigma}\|f-h\|_2^2.$$

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1), 1-122, 2011.

[2] A. Chambolle, and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1), 120-145, 2011.

[3] R.I. Boţ, and C. Hendrich. A Douglas–Rachford type primal-dual method for solving inclusions with mixtures of composite and parallel-sum type monotone operators. *SIAM Journal on Optimization*, 23(4), 2541-2565, 2013.

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

[5] H.H. Bauschke and P.L. Combettes. Convex analysis and monotone operator theory in Hilbert spaces. Springer, New York, 2011.

Evaluating proximal operator of $T_\epsilon(f_0, \cdot)$

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

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

Evaluating proximal operator of $T_\epsilon(f_0, \cdot)$

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

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

Thus we want to solve

$$\begin{aligned} \min_{M \geq 0, f_1} \ & \mathrm{trace}(C^T M) + \epsilon D(M) + \frac{1}{2\sigma}\|f_1 - h\|_2^2 \\ \text{subject to} \ & f_0 = M\mathbf{1} \\ & f_1 = M^T\mathbf{1}. \end{aligned}$$

Use dual coordinate ascent with $\mathcal{G}(\cdot) = \frac{1}{2\sigma}\|\cdot - h\|_2^2$

Evaluating proximal operator of $T_\epsilon(f_0, \cdot)$

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

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

Thus we want to solve

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

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

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

Use dual coordinate ascent with $\mathcal{G}(\cdot) = \frac{1}{2\sigma}\|\cdot - h\|_2^2$

This gives 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. $u_1 = f_1./(K^T u_0)$

Evaluating proximal operator of $T_\epsilon(f_0, \cdot)$

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

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

Thus we want to solve

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

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

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

Use dual coordinate ascent with $\mathcal{G}(\cdot) = \frac{1}{2\sigma}\|\cdot - h\|_2^2$

This gives 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.

UPDATE OR REMOVE THIS SLIDE!

Potentially add slides on how to deal with structured cost matrix C

- uniform discretization and $c(\cdot, \cdot)$ translation invariant $\rightsquigarrow$ Toeplitz-block-Toeplitz.
- $c(\cdot, \cdot)$ that decomposes in each dimension.

Example: variational regularization of inverse problems

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.

Example: variational regularization of inverse problems

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.

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.

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

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

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

Example: variational regularization of inverse problems

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

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

- $\tilde{f}$ 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

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

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: $\|Af_1 - g\|_2$

Consider the inverse problems

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

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

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

Computed 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.

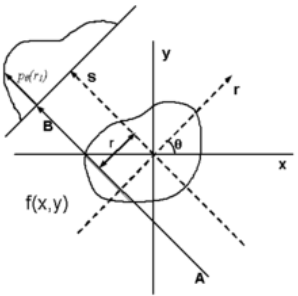
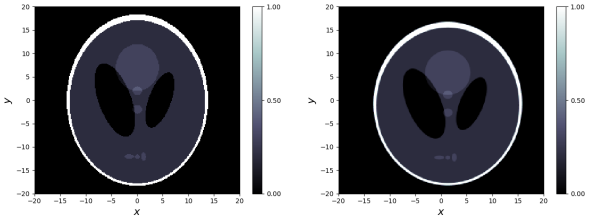


Illustration from Wikipedia

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%



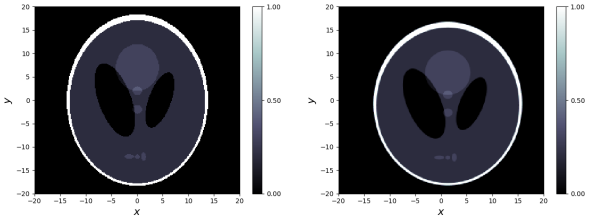
(a) Shepp-Logan phantom

(b) Prior

TV-regularization and optimal transport prior:

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

$$\text{subject to} \quad \|Af_1 - g\|_2 \leq \kappa.$$



(f) Shepp-Logan phantom

(g) Prior

TV-regularization and optimal transport prior:

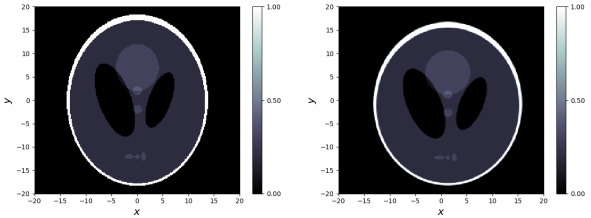
$$\min_{f_1} \quad \|\nabla f_1\|_1 + \gamma T_\epsilon(f_0, f_1)$$

$$\text{subject to} \quad \|Af_1 - g\|_2 \leq \kappa.$$

TV-regularization and ℓ_2^2 prior:

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

$$\text{subject to} \quad \|Af_1 - g\|_2 \leq \kappa.$$



(k) Shepp-Logan phantom

(l) Prior

TV-regularization and optimal transport prior:

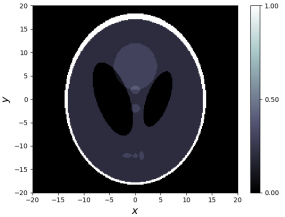
$$\min_{f_1} \quad \|\nabla f_1\|_1 + \gamma T_\epsilon(f_0, f_1)$$

$$\text{subject to} \quad \|Af_1 - g\|_2 \leq \kappa.$$

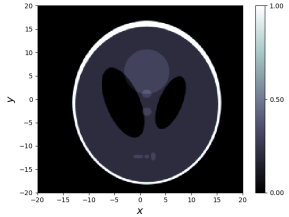
TV-regularization and ℓ_2^2 prior:

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

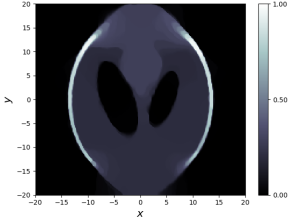
$$\text{subject to} \quad \|Af_1 - g\|_2 \leq \kappa.$$



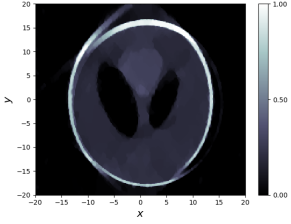
(p) Shepp-Logan phantom



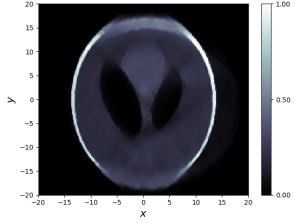
(q) Prior



(r) TV-regularization



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

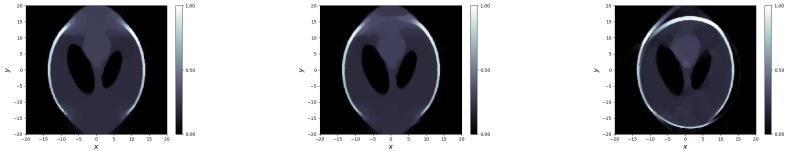


(t) TV-regularization and optimal transport prior ($\gamma = 4$)

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

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

$$\text{subject to} \quad \|Af_1 - g\|_2 \leq \kappa.$$



(u) $\gamma = 0$ (v) $\gamma = 1$ (w) $\gamma = 10$

(x) $\gamma = 100$ (y) $\gamma = 1000$ (z) $\gamma = 10\,000$.

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

Summary and conclusions

Conclusions:

- Sinkhorn iterations can be interpreted as coordinate ascent in the dual.

Summary and conclusions

Conclusions:

- Sinkhorn iterations can be interpreted as coordinate ascent in the dual.
- Generalizes to methods for solving $\min_{f_1} T_\epsilon(f_0, f_1) + \mathcal{G}(f_1)$, where $\mathcal{G}$ is "simple".

Summary and conclusions

Conclusions:

- Sinkhorn iterations can be interpreted as coordinate ascent in the dual.
- Generalizes to methods for solving $\min_{f_1} T_\epsilon(f_0, f_1) + \mathcal{G}(f_1)$, where $\mathcal{G}$ is "simple".
- Iterative method to compute the proximal operator of $T_\epsilon(f_0, \cdot)$
  $\rightsquigarrow$ can solve more advanced problems using variable splitting.

Summary and conclusions

Conclusions:

- Sinkhorn iterations can be interpreted as coordinate ascent in the dual.
- Generalizes to methods for solving $\min_{f_1} T_\epsilon(f_0, f_1) + \mathcal{G}(f_1)$, where $\mathcal{G}$ is "simple".
- Iterative method to compute the proximal operator of $T_\epsilon(f_0, \cdot)$
  $\rightsquigarrow$ can solve more advanced problems using variable splitting.

Optimal transport - a viable framework for many applications!

Thank you for your attention!

Questions?