Technische Universität Berlin

Lecture 2:

Optimal Transport,

Sinkhorn's Algorithm,

SMART

Gabriele Steidl
Applied Mathematics - Imaging Sciences
TU Berlin

Berlin Mathematics Research Center MATH+

Funded under Germany's Excellence Strategy by DFG Deutsche Forschungsgemeinschaft

Outline

1. Basic Notation: Spaces of Measures
2. Monge and Kantorovich Problem of OT
3. Discrete OT and its Dual
4. Regularized Optimal Transport
5. Mirror Descent Algorithm - SMART - Sinkhorn Algorithm

1. Basic Notation: Spaces of Measures

- $X = \mathbb{R}^d$ or $X \subset \mathbb{R}^d$ compact subset of $\mathbb{R}^d$
  (General: Polish space = separable, completely metrizable topol. space)
- $\mathcal{B}(X)$ Borel σ-algebra on X: smallest σ-algebra which contains all open sets
  $= \sigma$-algebra generated by the open sets
  (Remember: σ-algebra Σ: $X \in \Sigma$, $A \in \Sigma \Rightarrow \bar{A} \in \Sigma$, $A_k \in \Sigma \Rightarrow \bigcup_{k\in\mathbb{N}} A_k \in \Sigma$)
- $\mu : \mathcal{B}(X) \to \mathbb{R}$ is a finite, signed measure on $\mathcal{B}(X)$ (Borel measure)
  i.e. for pairwise disjoint sets $A_k \in \mathcal{B}(X)$, $k \in \mathbb{N}$,

$$\mu\Big(\bigcup_{k\in\mathbb{N}} A_k\Big) = \sum_{k\in\mathbb{N}} \mu(A_k) \qquad \sigma - \text{additivity}$$

- $\mathcal{M}(X)$ linear space of Borel measures
  (General: for Radon measures (fulfill inner/outer regularity condition); coincide with Borel measures on Polish spaces, in particular $\mathbb{R}^d$)
- $\mathcal{M}(X)$ is a Banach space with **total variation norm**

$$\|\mu\|_{TV} := |\mu|(X), \quad \text{where}$$

$$|\mu|(A) := \sup_{A=\bigcup A_k, A_i\cap A_j=\emptyset} \sum_{k=1}^{n} |\mu(A_k)|$$

1. Basic Notation: Spaces of Measures

- pre-dual space of $\mathcal{M}(X)$ is $C_0(X)$, i.e. $C_0(X)' = \mathcal{M}(X)$
- weak-* convergence of measures $\mu_n \rightharpoonup \mu$ if

$$\int_X \varphi(x)\, d\mu_n(x) \to \int_X \varphi(x)\, d\mu(x) \quad \text{for all } \varphi \in C_0(X).$$

- $$\|\mu\|_{TV} = \sup_{\|\varphi\|_\infty \leq 1} |\langle \varphi, \mu \rangle|$$

  with dual pairing

$$\langle x, \mu \rangle := \int_X \varphi(x)\, d\mu(x)$$

Further, let

- $\mathcal{M}_+(X)$ subset of non-negative measures on X
- $\mathcal{P}(X)$ subset of probability measures on X,
  i.e. $\mu \in \mathcal{M}_+(X)$ and $\mu(X) = 1$

Remark on Convergence of Measures

Let X locally compact Polish space (separable, complete metric space) and $\mathrm{rba}(X)$ space of finitely additive set functions

- $C_c(X)' \cong \mathcal{M}(X)$ with $T_\mu(\varphi) = \langle \mu, \varphi \rangle := \int_X \varphi \, d\mu$ for all $\varphi \in C_c(X)$.
- $C_0(X)' \cong \mathcal{M}(X)$ with $T_\mu(\varphi) = \langle \mu, \varphi \rangle := \int_X \varphi \, d\mu$ for all $\varphi \in C_0(X)$.
- $C_b(X)' \cong \mathrm{rba}(X)$

Note that $C_0(X) \subset C_b(X)$, but $C_b(X)' \cong \mathrm{rba}(X) \not\subseteq \mathcal{M}(X) \cong C_0(X)'$

- $(\mu_n)_{n\in\mathbb{N}} \subset \mathcal{M}(X)$ bounded Then

$$\mu_n \overset{*}{\rightharpoonup} \mu \text{ in } C_c(X)' \quad \text{if and only if} \quad \mu_n \overset{*}{\rightharpoonup} \mu \text{ in } C_0(X)'.$$

  Counterexample: $\mu_n = n\delta_n$ test against $\varphi(x) = \sin(x)/x \in C_0(\mathbb{R})$
- $(\mu_n)_{n\in\mathbb{N}} \subset \mathcal{M}(X)$ be bounded and tight, then

$$\mu_n \overset{*}{\rightharpoonup} \mu \text{ in } C_c(X)' \quad \text{if and only if} \quad \mu_n \overset{*}{\rightharpoonup} \mu \text{ in } C_b(X)'.$$

  Counterexample: $\mu_n = \delta_n$ test against $\varphi(x) = \sin(x) \in C_b(\mathbb{R})$

$(\mu_n)_n$ is *tight* if for all $\epsilon > 0$, there exists a compact set $K \subset X$ such that $|\mu_n|(X \setminus K) < \epsilon$ for all $n \in \mathbb{N}$.

Theorem (Prokhorov) Let $(\mu_n)_{n\in\mathbb{N}}$ be a tight sequence of probability measures. Then there exists a subsequence $(\mu_{n_k})_{k\in\mathbb{N}}$ which converges weakly to a probability measure μ.

Ref: Plonka, Potts, Steidl, Tasche, Numerical Fourier Analysis, Springer 2023,
Chapter: Fourier Analysis of Measures

Examples

1. Atomic measures (empirical measures if same weights)

$$\mu = \sum_{i=1}^{m} \mu_i \delta_{x_i}, \quad \delta_x(A) := \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{otherwise} \end{cases}$$

If $\mu \in \mathcal{P}(X)$, then $\sum_{i=1}^{m} \mu_i = 1$ and $\mu_i > 0$, $i = 1, \ldots, m$. Then

$$\|\mu\|_{TV} = |\mu_1| + \ldots + |\mu_m|$$

2. Absolutely continuous measures $\mu \in \mathcal{M}(X)$ with density $\varphi \in L^1(X)$

$$\mu(A) = \int_A \varphi(x)\, dx$$

(Remember: $\mu \ll \lambda$ with Lebesgue measure λ, if $\lambda(A) = 0 \Rightarrow \mu(A) = 0 \; \forall A \in \mathcal{B}(X)$)
Then

$$\|\mu\|_{TV} = \int_X |\varphi(x)|\, dx$$

TV-norm is not a good measure for our purposes.

2. Monge Problem (1781)

Given $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$
Find an optimal transport map $\hat{T} : X \to Y$ such that

$$\hat{T} \in \underset{T\text{measurable}}{\operatorname{argmin}} \int_X c\left(x, T(x)\right) \, \mathrm{d}\mu(x) \quad \text{subject to} \quad \nu = T_{\#}\mu$$

with the push forward measure

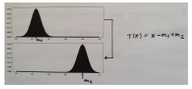
$$T_{\#}\mu := \mu \circ T^{-1}$$

Note that

$$\int_{T^{-1}(A)} h(T(x)) d\mu(x) = \int_A h(y) d \underbrace{(T_{\#}\mu)}_{\nu}(y)$$

and in case of existing densities and a diffeomorphism T:

$$p_{T_{\#}\mu}(y) = p_\mu \left(T^{-1}(y)\right) \left|\det \nabla T^{-1}(y)\right|$$

Example: law of random variables,  , sliced OT, measures on curves

Discrete Monge Problem

Given

$$\mu = \sum_{i=1}^{m} \mu_i \delta_{x_i}, \quad \nu = \sum_{j=1}^{n} \nu_j \delta_{y_j}$$

Find an optimal transport map $\hat{T} : \{x_1, \ldots, x_m\} \to \{y_1, \ldots, y_n\}$ such that

$$\hat{T} \in \operatorname*{argmin}_{T} \sum_{i=1}^{N} c(x_i, T(x_i))\mu_i \quad \text{s.t.} \quad \nu_j = \sum_{T(x_i)=y_j} \mu_i$$

Kantorovich Problem (1942)

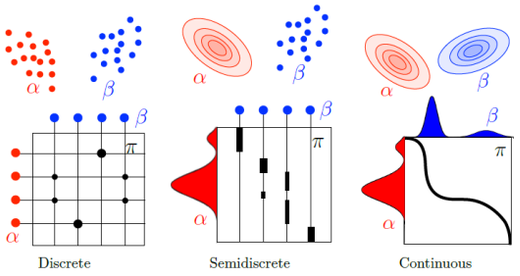
Given $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$
Find an optimal transport plan $\hat{\pi} \in \mathcal{P}(X \times Y)$ with given marginals μ and ν

$$\hat{\pi} \in \operatorname*{argmin}_{\pi \in \Pi(\mu,\nu)} \int_{X \times X} c(x,y) \mathrm{d}\pi(x,y)$$

$$\mathrm{OT}(\mu,\nu) = \min_{\pi \in \Pi(\mu,\nu)} \int_{X \times X} c(x,y) \mathrm{d}\pi(x,y)$$

where

- $\Pi(\mu,\nu) := \{\pi \in \mathcal{P}(X \times X) : (P_1)_{\#}\pi = \mu,\ (P_2)_{\#}\pi = \mu$
- $P_1(x_1,x_2) = x_1,\ P_2(x_1,x_2) = x_2$



: Book Peyré/Cuturi: Computational optimal transport

- Existence of a minimizer is ensured if c is lsc and bounded from below

Discrete Kantorovich Problem

Given

$$\mu = \sum_{i=1}^{m} \mu_i \delta_{x_i}, \quad \nu = \sum_{j=1}^{n} \nu_j \delta_{y_j}$$

Find an optimal transport plan $\hat{\pi} = \sum_{i,j=1}^{m,n} \hat{\pi}_{ij} \delta_{x_i, y_j}$, i.e. $(\pi_{ij})_{i,j} \in \mathbb{R}_{\geq 0}^{m,n}$,

$$\hat{\pi} \in \operatorname*{argmin}_{\pi} \sum_{i=1}^{m} \sum_{j=1}^{n} c(x_i, y_j) \pi_{ij}$$

subject to $\pi_{ij} \geq 0$

$$\sum_{i=1}^{m} \pi_{ij} = \nu_j, \quad j = 1, \ldots, n,$$

$$\sum_{j=1}^{n} \pi_{ij} = \mu_i, \quad i = 1, \ldots, m$$

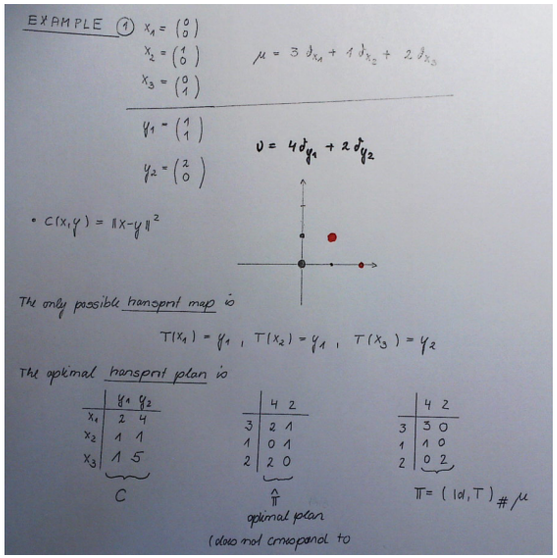
Shorter notation:

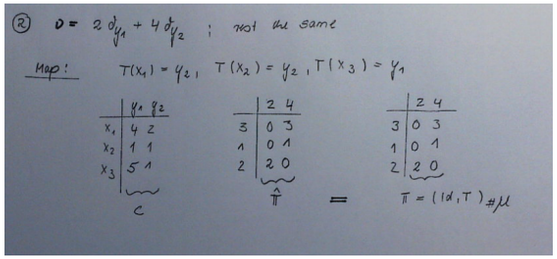
$$\hat{\pi} \in \operatorname*{argmin}_{\pi} \langle c, \pi \rangle \quad \text{subject to} \quad \pi \mathbf{1}_n = \mu,\ \pi^{\mathsf{T}} \mathbf{1}_m = \nu,\ \pi \geq 0$$

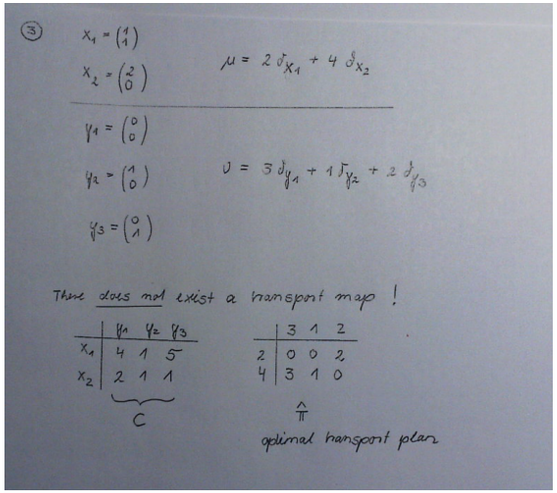
$$P_1 \pi = \mu,\ P_2 \pi = \nu,\ \pi \geq 0$$

$$(1_n \otimes I_m)\pi = \mu,\ (I_n \otimes 1_m)\pi = \nu,\ \pi \geq 0$$

Examples


EXAMPLE ①
$x_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$
$x_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$
$x_3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$
$\mu = 3\delta_{x_1} + 1\delta_{x_2} + 2\delta_{x_3}$
$y_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$
$y_2 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$
$\nu = 4\delta_{y_1} + 2\delta_{y_2}$
• $c(x,y) = \|x-y\|^2$
The only possible transport map is
$T(x_1) = y_1$, $T(x_2) = y_1$, $T(x_3) = y_2$
The optimal transport plan is
optimal plan
(does not correspond to
$\pi = (\mathrm{id}, T) \# \mu$


② $\nu = 2\delta_{y_1} + 4\delta_{y_2}$; not the same
Map: $T(x_1) = y_2$, $T(x_2) = y_2$, $T(x_3) = y_1$
$\pi = (\mathrm{id}, T) \# \mu$


③
$x_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$
$x_2 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$
$\mu = 2\delta_{x_1} + 4\delta_{x_2}$
$y_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$
$y_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$
$y_3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$
$\nu = 3\delta_{y_1} + 1\delta_{y_2} + 2\delta_{y_3}$
There does not exist a transport map !
optimal transport plan

Brenier's Theorem

Theorem Let $\mu, \nu \in \mathcal{P}_2(X)$, where $\mu \ll \lambda$ and $c(x,y) = \|x-y\|^2$. Then

- Kantorovich problem has a unique solution $\hat{\pi}$
- $\hat{\pi} = (I, \hat{T})_{\#}\mu$, where $\hat{T} \in L^2_\mu(X,X)$ is the optimal transport map
- If ν has bounded support, then

$$\hat{T}(x) = x - \nabla\varphi(x) = \nabla\psi(x) \quad for\ \mu - a.e.\ x,$$

  for some lower semi-continuous, convex, differentiable μ-a.e function ψ.
- Conversely, if ψ is lower semi-continuous, convex and differentiable μ-a.e. with $|\nabla\psi| \in L^2_\mu(X,X)$, then

$$T = \nabla\psi$$

  is the optimal transport map from μ to $\nu = T_{\#}\mu \in \mathcal{P}_2(X)$.

Wasserstein Spaces

For $p \in [1, \infty)$, the space of measures with finite p-th moment is defined by

$$\mathcal{P}_p(\mathcal{X}) := \left\{ \mu \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} \|x\|^p \mathrm{d}\mu(x) < \infty \right\}.$$

For $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, the Wasserstein p-distance is given by

$$W_p^p(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \mathrm{d}\pi(x, y),$$

which actually defines a metric.

The metric space $(\mathcal{P}_p(\mathcal{X}), W_p)$ is called the p-th Wasserstein space.

Convergence of $(\mu_n)_n$, $\mu_n \in \mathbb{P}_2(\mathbb{R}^d)$: $W_2(\mu_n, \mu) \to 0$ as $n \to \infty$ if and only if we have weak/narrow convergence

$$\mu_n \overset{*}{\rightharpoonup} \mu \text{ in } C_b(X)'$$

and

$$\int_{\mathbb{R}^d} \|x\|^2 \, \mathrm{d}\mu_n(x) \quad \to \quad \int_{\mathbb{R}^d} \|x\|^2 \, \mathrm{d}\mu(\mathrm{x}) \quad \text{as } \mathrm{n} \to \infty$$

Example: $\mu_1 \sim \mathcal{N}(m_1, \Sigma_1)$, $\mu_2 \sim \mathcal{N}(m_2, \Sigma_2)$ and $\Sigma_{1,2} := \left(\Sigma_2^{\frac{1}{2}} \Sigma_1 \Sigma_2^{\frac{1}{2}} \right)^{\frac{1}{2}}$

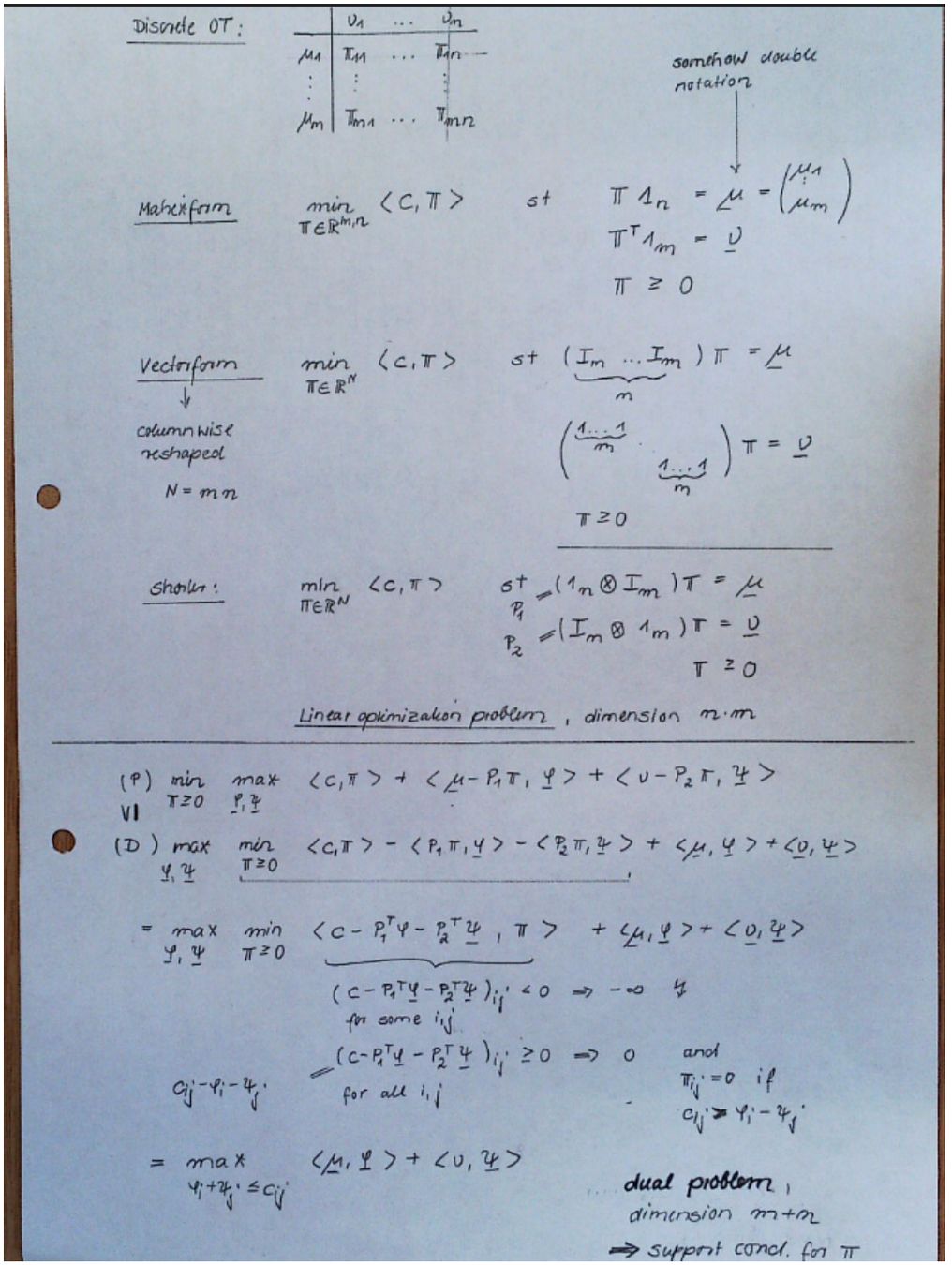
Then

$$W_2^2(\mu_1, \mu_2) = \|m_1 - m_2\|^2 + \mathrm{tr}(\Sigma_1 - 2\Sigma_{1,2} + \Sigma_2)$$

Outline

1. Basic Notation: Spaces of Measures
2. Monge and Kantorovich Problem of OT
3. Discrete OT and its Dual
4. Regularized Optimal Transport
5. Mirror Descent Algorithm - SMART - Sinkhorn Algorithm

Discrete OT and its Dual



OT and its Dual

Primal OT:

$$\mathrm{OT}(\mu,\nu) := \min_{\pi\in\Pi(\mu,\nu)} \int_{X^2} c\,\mathrm{d}\pi,$$

Dual OT:

$$\mathrm{OT}(\mu,\nu) = \max_{\substack{(\varphi,\psi)\in C(X)^2 \\ \varphi(x)+\psi(y)\le c(x,y)}} \int_X \varphi\,\mathrm{d}\mu + \int_X \psi\,\mathrm{d}\nu.$$

$(\hat{\varphi},\hat{\psi}) = (\hat{\varphi},\hat{\varphi}^c)$ with *c-transformed function*

$$\varphi^c(y) = \min_{x\in X}\{c(x,y) - \varphi(x)\}.$$

Example: Set $X := [0,1]$, $c(x,y) = |x-y|$, $\mu = \delta_0/2 + \delta_1/2$, $\nu = \delta_{0.1}/2 + \delta_{0.9}/2$.
Then, $\mathrm{OT}(\mu,\nu) = 0.1$ with unique optimal transport plan $\hat{\pi} = \frac{1}{2}\delta_{0,0.1} + \frac{1}{2}\delta_{1,0.9}$.
Optimal dual potentials

$$\hat{\varphi}_1(x) = \begin{cases} 0.1 - x & \text{for } x \in [0,0.1], \\ x - 0.9 & \text{for } x \in [0.9,1], \\ 0 & \text{else,} \end{cases} \quad \text{and} \quad \hat{\varphi}_2(x) = \begin{cases} 0.2 - x & \text{for } x \in [0,0.2], \\ x - 0.9 & \text{for } x \in [0.9,1], \\ 0 & \text{else.} \end{cases}$$

Clearly, these potentials do not differ only by a constant.

Regularized Optimal Transport

Technische Universität Berlin

Primal Problem:

$$\mathrm{OT}_\varepsilon = \min_{\pi \in \Pi(\mu,\nu)} \int_{X \times X} c(x,y) \mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi, \mu \otimes \nu)$$

Discrete setting:

$$\begin{aligned}
\mathrm{OT}_\varepsilon &= \min_{\pi \in \Pi(\mu,\nu)} \langle c, \pi \rangle + \varepsilon \Big(\sum_{i,j} \pi_{i,j} \log \pi_{i,j} - \pi_{i,j} \log(\mu_i \nu_j) - \pi_{i,j} \Big) \\
&= \varepsilon \min_{\pi \in \Pi(\mu,\nu)} \sum_{i,j} \pi_{i,j} \log \pi_{i,j} - \pi_{i,j} \log(\mu_i \nu_j \mathrm{e}^{-c/\varepsilon}) - \pi_{i,j} \\
&= \varepsilon \min_{\pi \in \Pi(\mu,\nu)} \mathrm{KL}(\pi, \underbrace{\mathrm{diag}(\mu)\, \mathrm{e}^{-c/\varepsilon}\, \mathrm{diag}(\nu)}_{K})
\end{aligned}$$

Lagrangian:

$$\min_{\pi} \max_{\varphi,\psi} \mathrm{KL}(\pi, K) + \langle P_1 \pi - \mu, \varphi \rangle + \langle P_2 \pi - \nu, \psi \rangle$$

Dual Problem:

$$\min_{\phi,\psi} \langle K, \mathrm{e}^{-P_1^\mathsf{T} \varphi - P_2^\mathsf{T} \psi} \rangle + \langle \mu, \varphi \rangle + \langle \nu, \psi \rangle$$

Alternating minimization: Sinkhorn algorithm with $u := \mathrm{e}^{-\varphi}$, $v := \mathrm{e}^{-\psi}$

$$\nabla_\varphi = 0 : \operatorname{diag}(\mathrm{e}^{-\varphi})\, K \operatorname{diag}(\mathrm{e}^{-\psi^{(r-1)}})\mathbf{1} = \mu \quad \to \quad u^{(r)} = \frac{\mu}{K v^{(r-1)}}$$

$$\nabla_\psi = 0 : \left(\operatorname{diag}(\mathrm{e}^{-\varphi^{(r)}})\, K \operatorname{diag}(\mathrm{e}^{-\psi})\right)^{\mathsf{T}}\mathbf{1} = \nu \quad \to \quad v^{(r)} = \frac{\nu}{K^{\mathsf{T}} u^{(r)}}$$

Relation to π: (set gradient of Lagrangian wrt π to 0)

$$0 = \log \pi - \log K + P_1^{\mathsf{T}}\varphi + P_2^{\mathsf{T}}\psi \quad \Longleftrightarrow \quad \pi = \operatorname{diag}(u)\, K\, \operatorname{diag}(v)$$

We can also consider the Sinkhorn algorithm wrt π:

$$\pi^{(0)} := K,$$

$$\text{for } r = 0, 1, \ldots$$

$$\pi^{(2r+1)} := \operatorname{diag}\left(\frac{\mu}{\pi^{(2r)}\mathbf{1}}\right)\pi^{(2r)}$$

$$\pi^{(2r+2)} := \pi^{(2r+1)} \operatorname{diag}\left(\frac{\nu}{(\pi^{(2r+1)})^{\mathsf{T}}\mathbf{1}}\right)$$

Convergence of Sinkhorn Algorithm

Linear convergence: in $\mathbb{R}^d_{>0,*}$, $x \sim y$ if $x = \alpha y$ for some $\alpha > 0$

$$d_H(u^{(r+1)}, u^*) \leq \lambda(K)^2 d_H(u^{(r)}, u^*)$$

with

$$\lambda(K) := \sup \left\{ \frac{d_H(Kx, Ky)}{d_H(x, y)} : x \not\sim y \right\}$$

and the Hilbert norm

$$d_H(x, y) := \| \log x - \log y \|_V,$$
$$\|x\|_V := \max_i x_i - \min_i x_i$$

Is there a relation to proximal algorithms?

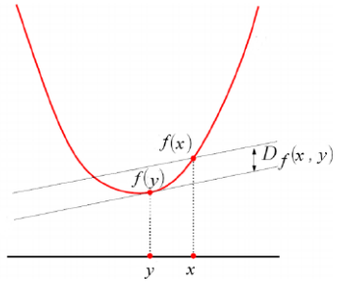
- Mirror Descent Algorithm = linearized proximal algorithm with a general f-divergence instead of the special squared norm
- SMART
- BI-SMART

Bregman Distances

Let $f : \mathbb{R}^d \to R \cup \{+\infty\}$ convex lsc and $\mathrm{dom} f \cap (0, +\infty) \neq \emptyset$ with nonempty subdifferential ∂f

Bregman distance:

$$D_f(x, y) = \{f(x) - f(y) + \langle p, x - y\rangle : p \in \partial f(y)\}$$



Examples:

1. $f(x) = \frac{1}{2}\|x\|^2$

$$D_f(x, y) = \frac{1}{2}\|x\|^2 - \frac{1}{2}\|y\|^2 - \langle y, x - y\rangle = \frac{1}{2}\|x - y\|^2$$

2. $f(x) = x \log x$

$$D_f(x, y) = x \log x - y \log y - \langle \log y + 1, x - y\rangle = x \log x - x \log y - x + y = \mathrm{KL}(x, y)$$

Function Name	$f(x)$	$\mathrm{dom}\, f$	$D_f(x, y)$
Squared Norm	$\frac{1}{2}x^2$	$(-\infty, \infty)$	$\frac{1}{2}(x - y)^2$
Shannon Entropy	$x \log x$	$[0, \infty)$	$x \log \frac{x}{y} - x + y$
Bit Entropy	$x \log x + (1 - x) \log(1 - x)$	$[0, 1]$	$x \log \frac{x}{y} + (1 - x) \log \frac{1-x}{1-y}$
Burg Entropy	$-\log x$	$(0, \infty)$	$\frac{x}{y} - \log \frac{x}{y} - 1$

Properties of Bregman Distances

Properties:

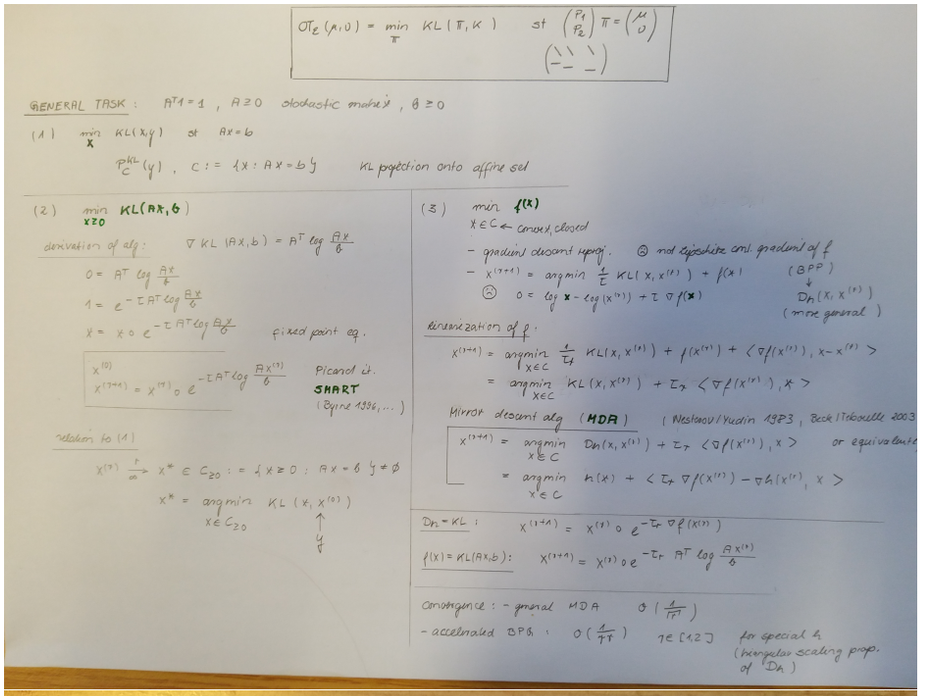
- $D_f(x, y) \geq 0$
- $D_f(x, y) = 0$ iff $x = y$ in case f is strictly convex.
- In general not symmetric and does not fulfill a triangular inequality
- Jointly convex, lsc.
- If f is strictly convex, D_f is strictly convex in the first argument.
- If expressions exist

$$\nabla_x D_f(x, y) = \nabla f(x) - \nabla f(y)$$

Is there a relation to proximal algorithms?

- Mirror Descent Algorithm = linearized proximal algorithm with a general f-divergence instead of the special squared norm
- SMART (simultaneous multiplicative algebraic reconstruction technique)
- BI-SMART (block iterative SMART)

Blackboard



$$OT_\varepsilon(\mu,\nu) = \min_{\pi} KL(\pi, K) \quad \text{s.t.} \quad \begin{pmatrix} P_1 \\ P_2 \end{pmatrix} \pi = \begin{pmatrix} \mu \\ \nu \end{pmatrix}$$

GENERAL TASK: $A^T 1 = 1$, $A \geq 0$ stochastic matrix, $b > 0$

(1) $\min_x KL(x, y)$ s.t. $Ax = b$

$P_C^{KL}(y)$, $C := \{x : Ax = b\}$ KL projection onto affine set

(2) $\min_{x \geq 0} KL(Ax, b)$

derivation of alg: $\nabla KL(Ax, b) = A^T \log \frac{Ax}{b}$

$0 = A^T \log \frac{Ax}{b}$

$1 = e^{-\tau A^T \log \frac{Ax}{b}}$

$x = x \cdot e^{-\tau A^T \log \frac{Ax}{b}}$ fixed point eq.

$x^{(0)}$

$x^{(r+1)} = x^{(r)} \cdot e^{-\tau A^T \log \frac{Ax^{(r)}}{b}}$ Picard it. SMART (Byrne 1996, ...)

relation to (1)

$x^{(r)} \xrightarrow{r \to \infty} x^* \in C_{20} := \{x \geq 0 : Ax = b\} \neq \emptyset$

$x^* = \operatorname{argmin}_{x \in C_{20}} KL(x, x^{(0)})$ ($x^{(0)}$ = y)

(3) $\min_{x \in C} f(x)$ C convex, closed

- gradient descent typ.: ☹ not Lipschitz cont. gradient of f
- $x^{(r+1)} = \operatorname{argmin}_x \frac{1}{\tau} KL(x, x^{(r)}) + f(x)$ (BPP) ↓ $D_h(x, x^{(r)})$ (more general)
- ☺ $0 = \log x - \log(x^{(r)}) + \tau \nabla f(x)$

linearization of f:

$x^{(r+1)} = \operatorname{argmin}_{x \in C} \frac{1}{\tau_r} KL(x, x^{(r)}) + f(x^{(r)}) + \langle \nabla f(x^{(r)}), x - x^{(r)} \rangle$

$= \operatorname{argmin}_{x \in C} KL(x, x^{(r)}) + \tau_r \langle \nabla f(x^{(r)}), x \rangle$

Mirror descent alg (MDA) (Nesterov/Yudin 1983, Beck/Teboulle 2003

$x^{(r+1)} = \operatorname{argmin}_{x \in C} D_h(x, x^{(r)}) + \tau_r \langle \nabla f(x^{(r)}), x \rangle$ or equivalently

$= \operatorname{argmin}_{x \in C} h(x) + \langle \tau_r \nabla f(x^{(r)}) - \nabla h(x^{(r)}), x \rangle$

$D_h = KL$: $x^{(r+1)} = x^{(r)} \cdot e^{-\tau_r \nabla f(x^{(r)})}$

$f(x) = KL(Ax, b)$: $x^{(r+1)} = x^{(r)} \cdot e^{-\tau_r A^T \log \frac{Ax^{(r)}}{b}}$

Convergence: - general MDA $\mathcal{O}\left(\frac{1}{\sum \tau_r}\right)$

- accelerated BPG: $\mathcal{O}\left(\frac{1}{\sum \tau_r}\right)$ $\gamma \in [1,2]$ for special h (Bregman scaling prop. of D_h)

- SMART for OT_ε: $A = \begin{pmatrix} P_1 \\ P_2 \end{pmatrix}$ rescale by $\frac{1}{2}$ to get $A^T 1 = 1$

$$T^{(r+1)} = \operatorname{diag}\left(\frac{\nu}{(T^{(r)})^T 1}\right)^{1/2} T^{(r)} \operatorname{diag}\left(\frac{\mu}{T^{(r)} 1}\right)^{1/2}$$

- this alg is $\sqrt{4}$ slower than Sinkhorn alg.
- it works also with $\alpha < 1$ (not $\alpha = 1$!) instead of $\alpha = \frac{1}{2}$; then it is $1/2$ slower than Sinkhorn alg. Question 1: Why?

GENERALIZATION

$C = C_1 \cap C_2 \cap \ldots \cap C_N$, $C_k := \{x : A_k x = b_k\}$, $A_k, b_k \geq 0$, $A_k^T 1 = 1$

$\neq \emptyset$

(4) $\operatorname{argmin}_{x \in C} KL(x, y)$

iterative information projection: $x^{(0)} = y$, $x^{(r+1)} = P_{C_{r+1}}^{KL}(x^{(r)})$ → as (1) and (2)

(Csiszar 1989) $C_{r+N} = C_r$ (periodic projection)

- convergence for affine sets; not for general convex sets (Dykstra alg. for convex sets...)

generalized iterative scaling (GIS): $x^{(0)} = y$

BI-SMART: $x^{(r+1)} = x^{(r)} \cdot e^{-\tau A_r^T \log \frac{A_r x^{(r)}}{b_r}}$ one SMART step

- BI-SMART for OT_ε: $\pi^{(r+1)} = \pi^{(r)} \cdot \operatorname{diag} \frac{\mu}{T^{(r)} 1}$
- Sinkhorn alg.: $\pi^{(r+2)} = \operatorname{diag} \frac{\nu}{(T^{(r+1)})^T 1} T^{(r+1)}$

Question 2: BI-SMART for unbalanced OT: $(1-\lambda) KL(x, y) + \lambda KL(Ax, b)$ does not converge in the obvious form

Alternatives?

BI-SMART in case $C = \emptyset$ (old problem, eg. of Byrne)

Comparison

SMART solely performs the multiplicative update specified in (2) with its step-size fixed to $\tau_k = \frac{1}{L}$, where again f is L-smooth relative to φ.

FSMART is based on the iteration suggested in [2], where initially $\theta_0 = 1$ is chosen, which is then subsequently updated via $\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}$, as suggested in [5].

ABPG-e as described in [6, Algorithm 2] was applied to (1) with parameters $\gamma_{\min} = 1$, $\gamma_0 = 5$ and $\delta = 0.05$. The choices for γ_0 and δ deviate slightly from the recommendation in [6], but were chosen to facilitate fastest possible convergence on the selected problem instances. To ensure comparability restarting mechanisms and stopping criteria based on the divergence of iterates were foregone. Updates for θ were conducted via Newton's method.

ABPG-g specified in [6, Algorithm 3] to (1) is used with parameters: $\rho = 1.2$, $\gamma = 2$ and $G_{\min} = 10^{-3}$. Restarting, stopping criteria and updating θ was handled analogously to ABPG-e.

RG is a SMART iteration with Armijo line search for choosing the step size τ_k via the retraction in (29) to iterate according to (37). The line search parameters are $\sigma = 0.5$, $\beta = 0.8$, $\alpha = 5.0$.

PD is the Chambolle-Pock primal dual algorithm [17, Algorithm 1] for solving convex composite structured optimization problems of the form $f(x) = g(x) + h(Ax)$. For $h(y) := \mathrm{KL}(y, b)$ with $y = Ax$ and $g \equiv 0$ we obtain

$$x^{k+1} = x^k e^{-\tau A^\top y^k} \qquad \text{(primal-step)} \qquad (39)$$

$$y^{k+1} = \log\left(\frac{e^{y^k} + \sigma A(2x^{k+1} - x^k)}{\mathbb{1} + \sigma b}\right), \qquad \text{(dual-step)} \qquad (40)$$

Chambolle-Pock 2015:

$$x^{(r+1)} = \operatorname*{argmin}_x \left\{ \langle Ax, y^n \rangle + g(x) + \frac{1}{\tau} D_\varphi\left(x, x^{(r)}\right) \right\}$$

$$y^{(r+1)} = \operatorname*{argmin}_y \left\{ h^*(y) - \langle A(2x^{(r-1)} - x^{(r)}), y \rangle + \frac{1}{\sigma} D_{\varphi^*}\left(y, y^{(r)}\right) \right\}$$

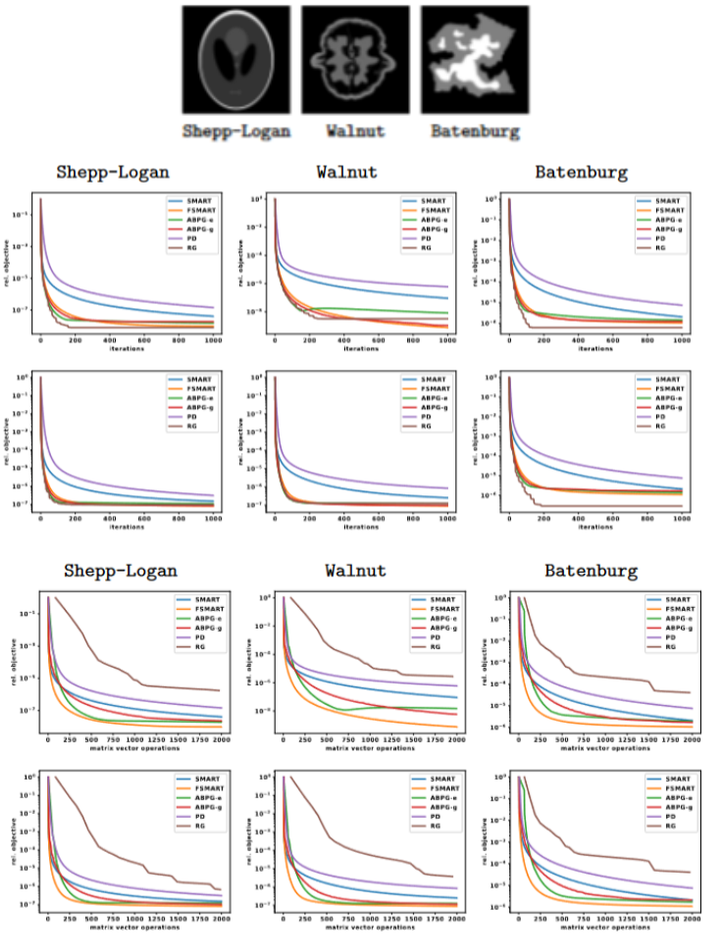
with $\varphi(x) = \langle x, \log x \rangle - \langle 1, x \rangle$ and $\varphi^*(y) = \langle 1, e^y \rangle$

Comparison

CT: parallel beam geometry and equidistant angles in the range $[0, \pi]$.

Undersampling rate was chosen to be 20

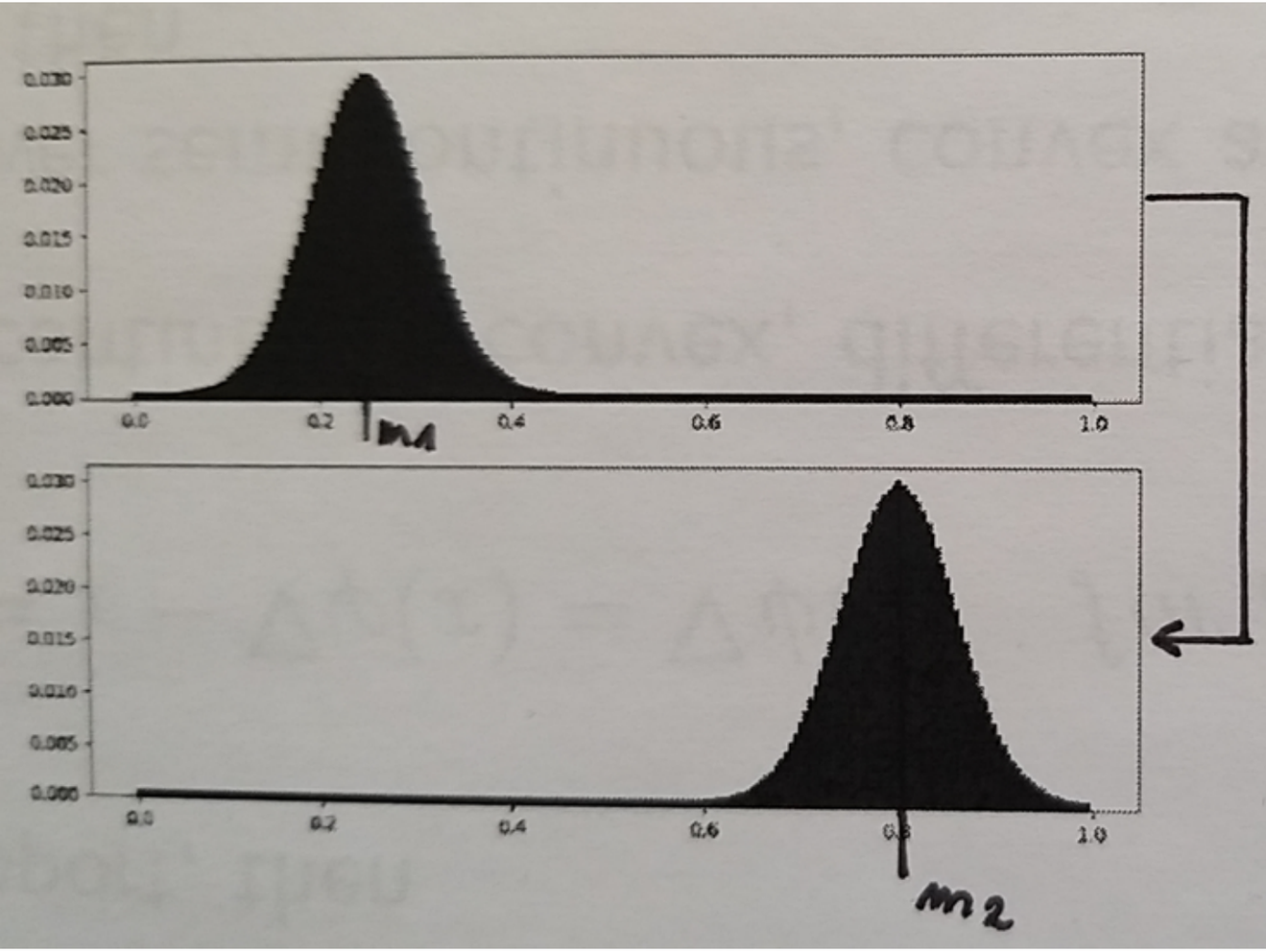
Setting without noise and with Poisson noise of SNR $= 20$ db.



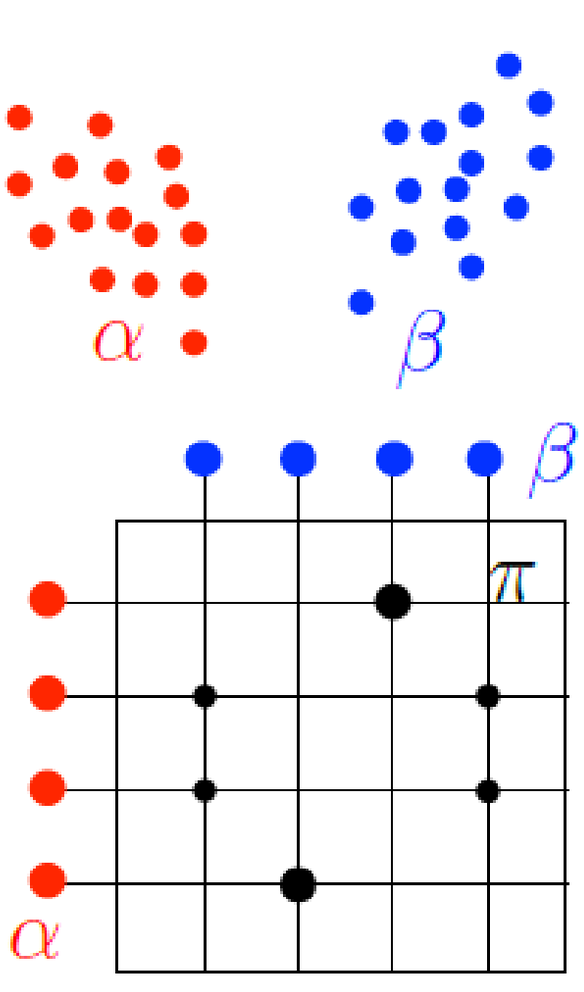
M. Kahl, S. Petra , Ch. Scnörr, G. Steidl , M. Zisler: On the remarkable efficiency of SMART, 2023

Berlin Mathematics Research Center

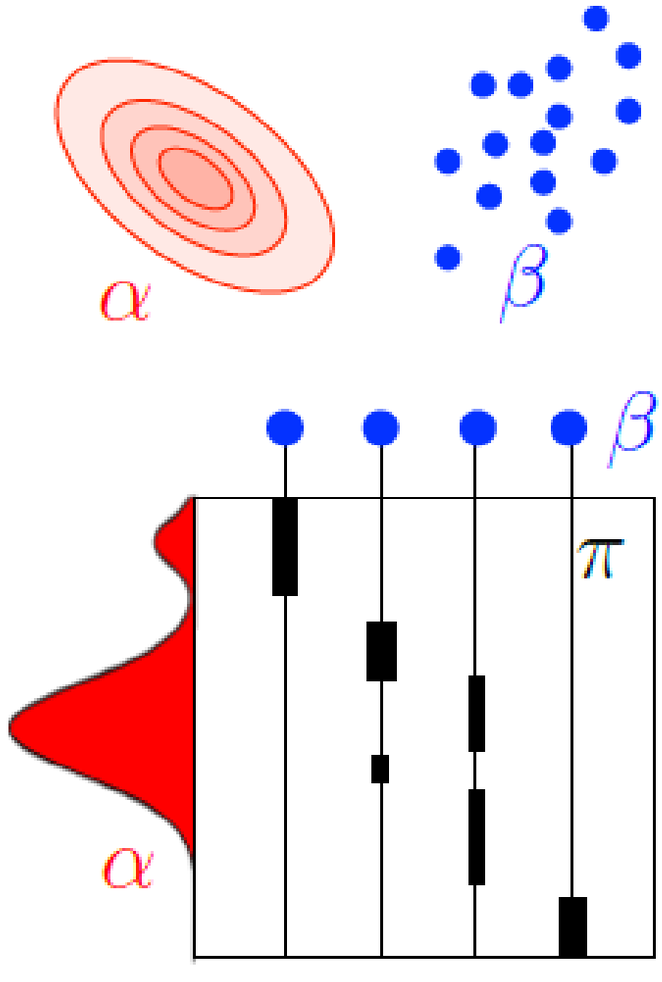
Funded under Germany's Excellence Strategy by
DFG Deutsche Forschungsgemeinschaft



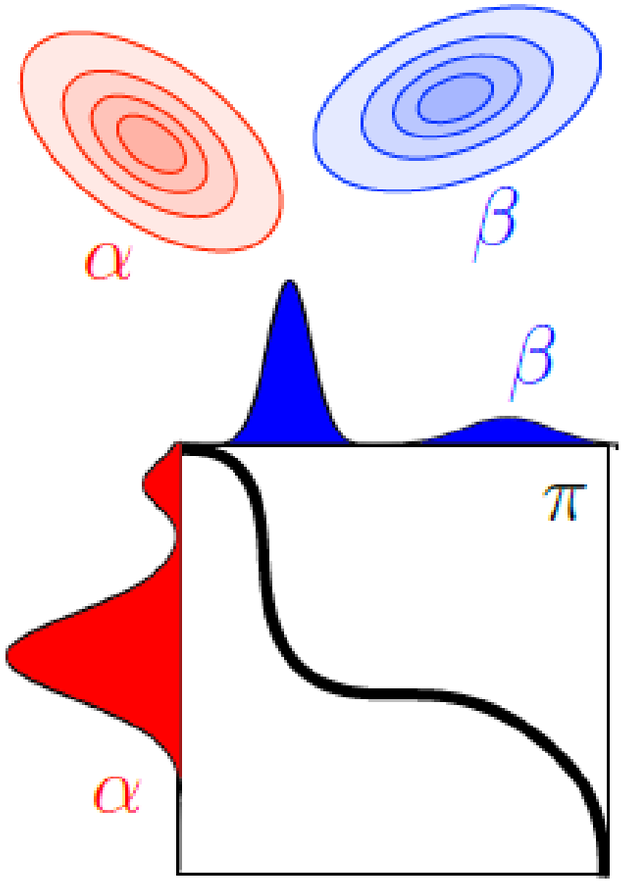
$$T(x) = x - m_1 + m_2$$



Discrete



Semidiscrete



Continuous

EXAMPLE (1) $x_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

$x_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

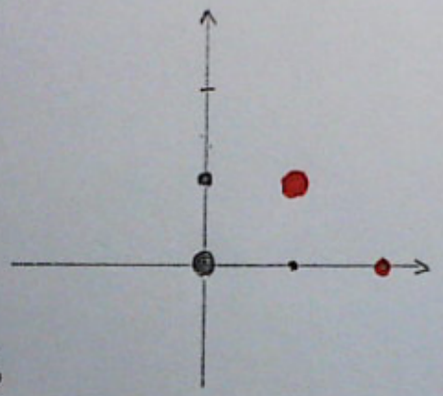
$x_3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

$\mu = 3\,\delta_{x_1} + 1\,\delta_{x_2} + 2\,\delta_{x_3}$

$y_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

$y_2 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$

$\nu = 4\,\delta_{y_1} + 2\,\delta_{y_2}$



- $c(x,y) = \|x-y\|^2$

The only possible transport map is

$$T(x_1) = y_1\,,\quad T(x_2) = y_1\,,\quad T(x_3) = y_2$$

The optimal transport plan is

	y_1	y_2
x_1	2	4
x_2	1	1
x_3	1	5

c

	4	2
3	2	1
1	0	1
2	2	0

$\hat{\pi}$

optimal plan

(does not correspond to

	4	2
3	3	0
1	1	0
2	0	2

$\pi = (\mathrm{id}, T)_{\#}\mu$

② $\nu = 2\delta_{y_1} + 4\delta_{y_2}$: not the same

Map: $T(x_1) = y_2,\ T(x_2) = y_2,\ T(x_3) = y_1$

	y_1	y_2
x_1	4	2
x_2	1	1
x_3	5	1

C

	2	4
3	0	3
1	0	1
2	2	0

$\hat{\pi}$

$=$

	2	4
3	0	3
1	0	1
2	2	0

$\pi = (Id, T)_{\#}\mu$

(3)

$x_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

$x_2 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$

$\mu = 2\,\delta_{x_1} + 4\,\delta_{x_2}$

$y_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

$y_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

$y_3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

$\nu = 3\,\delta_{y_1} + 1\,\delta_{y_2} + 2\,\delta_{y_3}$

There <u>does not</u> exist a transport map !

	y_1	y_2	y_3
x_1	4	1	5
x_2	2	1	1

C

	3	1	2
2	0	0	2
4	3	1	0

$\hat{\pi}$

optimal transport plan

Discrete OT:

	ν_1	...	ν_n
μ_1	π_{11}	...	π_{1n}
⋮	⋮		⋮
μ_m	π_{m1}	...	π_{mn}

somehow double notation

Matrixform

$$\min_{\pi \in \mathbb{R}^{m,n}} \langle C, \pi \rangle \quad \text{s.t.} \quad \pi 1_n = \underline{\mu} = \begin{pmatrix} \mu_1 \\ \mu_m \end{pmatrix}$$
$$\pi^T 1_m = \underline{\nu}$$
$$\pi \geq 0$$

Vectorform ↓ column wise reshaped, $N = mn$

$$\min_{\pi \in \mathbb{R}^N} \langle c, \pi \rangle \quad \text{s.t.} \quad (\underbrace{I_m \dots I_m}_{n}) \pi = \underline{\mu}$$
$$\begin{pmatrix} \underbrace{1 \dots 1}_{m} & & \\ & \ddots & \\ & & \underbrace{1 \dots 1}_{m} \end{pmatrix} \pi = \underline{\nu}$$
$$\pi \geq 0$$

Shorter:

$$\min_{\pi \in \mathbb{R}^N} \langle c, \pi \rangle \quad \text{s.t.} \quad \underset{P_1}{=} (1_n \otimes I_m) \pi = \underline{\mu}$$
$$\underset{P_2}{=} (I_m \otimes 1_m) \pi = \underline{\nu}$$
$$\pi \geq 0$$

Linear optimization problem, dimension $n \cdot m$

$$(P) \quad \min_{\pi \geq 0} \max_{\varphi, \psi} \langle c, \pi \rangle + \langle \underline{\mu} - P_1 \pi, \underline{\varphi} \rangle + \langle \underline{\nu} - P_2 \pi, \underline{\psi} \rangle$$

VI

$$(D) \quad \max_{\underline{\varphi}, \underline{\psi}} \min_{\pi \geq 0} \langle c, \pi \rangle - \langle P_1 \pi, \underline{\varphi} \rangle - \langle P_2 \pi, \underline{\psi} \rangle + \langle \underline{\mu}, \underline{\varphi} \rangle + \langle \underline{\nu}, \underline{\psi} \rangle$$

$$= \max_{\underline{\varphi}, \underline{\psi}} \min_{\pi \geq 0} \underbrace{\langle c - P_1^T \varphi - P_2^T \psi, \pi \rangle} + \langle \underline{\mu}, \underline{\varphi} \rangle + \langle \underline{\nu}, \underline{\psi} \rangle$$

$(c - P_1^T \underline{\varphi} - P_2^T \underline{\psi})_{i,j} < 0 \Rightarrow -\infty$ for some i, j

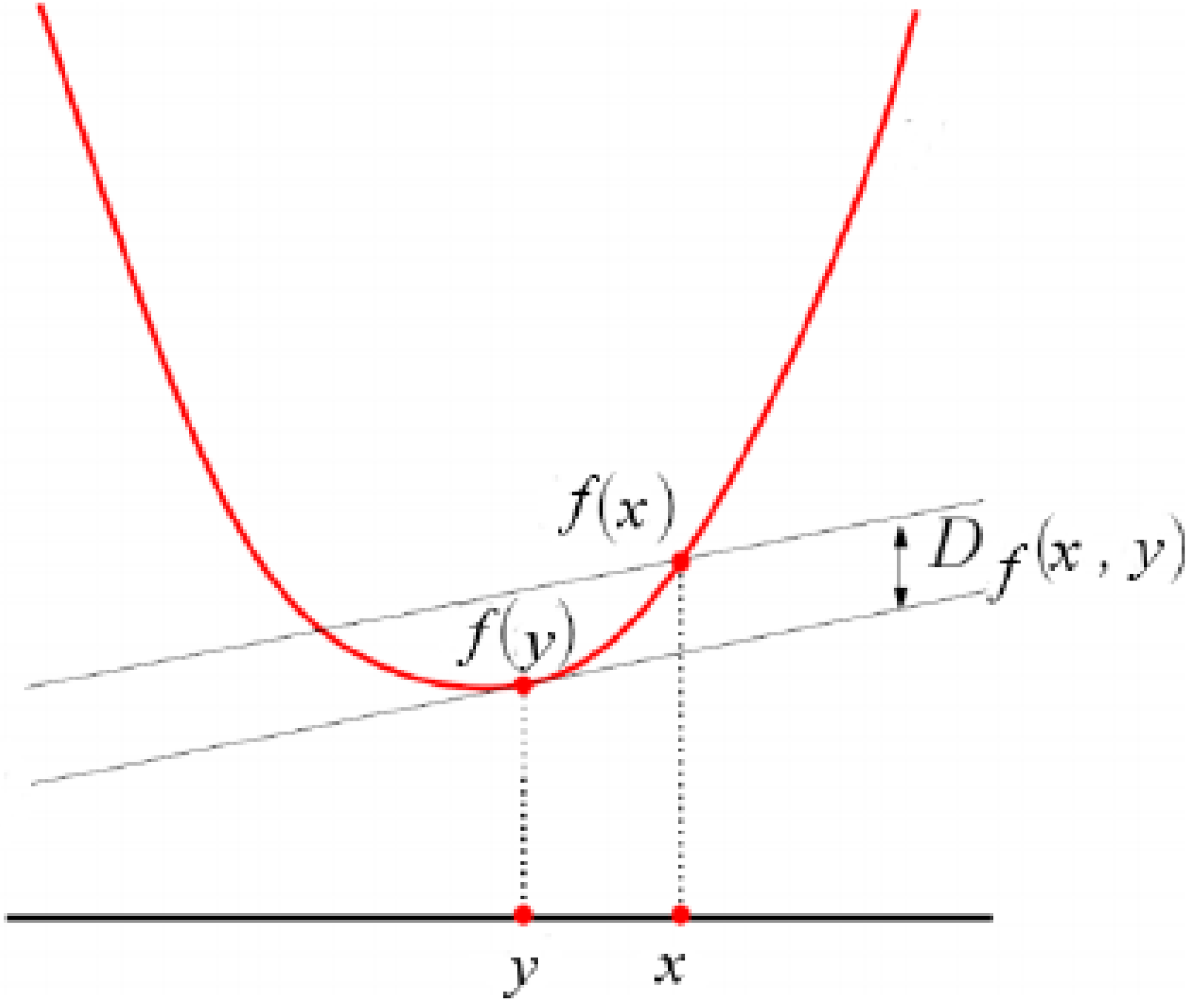
$(c - P_1^T \underline{\varphi} - P_2^T \underline{\psi})_{i,j} \geq 0 \Rightarrow 0$ for all i, j and $\pi_{i,j} = 0$ if $c_{i,j} > \varphi_i - \psi_j$

$c_{ij} - \varphi_i - \psi_j$

$$= \max_{\varphi_i + \psi_j \leq c_{ij}} \langle \underline{\mu}, \underline{\varphi} \rangle + \langle \underline{\nu}, \underline{\psi} \rangle$$

dual problem, dimension $m+n$

$\Rightarrow$ support concl. for π


f(x)
D
f(x , y)
f(y)
y
x

Function Name	$f(x)$	$\operatorname{dom} f$	$D_f(x, y)$
Squared Norm	$\frac{1}{2}x^2$	$(-\infty, \infty)$	$\frac{1}{2}(x-y)^2$
Shannon Entropy	$x \log x$	$[0, \infty)$	$x \log \frac{x}{y} - x + y$
Bit Entropy	$x \log x + (1-x) \log(1-x)$	$[0, 1]$	$x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y}$
Burg Entropy	$-\log x$	$(0, \infty)$	$\frac{x}{y} - \log \frac{x}{y} - 1$

$$OT_\varepsilon(\mu,\nu) = \min_{\pi} KL(\pi, K) \quad \text{st} \quad \begin{pmatrix} P_1 \\ P_2 \end{pmatrix} \pi = \begin{pmatrix} \mu \\ \nu \end{pmatrix}$$

GENERAL TASK : $A^T 1 = 1$, $A \geq 0$ stochastic matrix , $b > 0$

(1) $\min_x KL(x, y)$ st $Ax = b$

$P_C^{KL}(y)$, $C := \{x : Ax = b\}$ KL projection onto affine set

---

(2) $\min_{x \geq 0} KL(Ax, b)$

derivation of alg : $\nabla KL(Ax, b) = A^T \log \frac{Ax}{b}$

$0 = A^T \log \frac{Ax}{b}$

$1 = e^{-\tau A^T \log \frac{Ax}{b}}$

$x = x \circ e^{-\tau A^T \log \frac{Ax}{b}}$ fixed point eq.

$x^{(0)}$

$x^{(\gamma+1)} = x^{(\gamma)} \circ e^{-\tau A^T \log \frac{Ax^{(\gamma)}}{b}}$ Picard it.

SMART

(Byrne 1996, ...)

relation to (1)

$x^{(\gamma)} \xrightarrow{\gamma \to \infty} x^* \in C_{\geq 0} := \{x \geq 0 : Ax = b\} \neq \emptyset$

$x^* = \underset{x \in C_{\geq 0}}{\operatorname{argmin}} KL(x, x^{(0)})$ ↑ y

---

(3) $\min f(x)$

$x \in C \leftarrow$ convex, closed

- gradient descent reproj. ☹ not Lipschitz cont. gradient of f
- $x^{(\gamma+1)} = \underset{x}{\operatorname{argmin}} \frac{1}{\tau} KL(x, x^{(\gamma)}) + f(x)$ (BPP) ↓ $D_h(x, x^{(\gamma)})$ (more general)

☹ $0 = \log x - \log(x^{(\gamma)}) + \tau \nabla f(x)$

linearization of f :

$$x^{(\gamma+1)} = \underset{x \in C}{\operatorname{argmin}} \frac{1}{\tau_\gamma} KL(x, x^{(\gamma)}) + f(x^{(\gamma)}) + \langle \nabla f(x^{(\gamma)}), x - x^{(\gamma)} \rangle$$

$$= \underset{x \in C}{\operatorname{argmin}} KL(x, x^{(\gamma)}) + \tau_\gamma \langle \nabla f(x^{(\gamma)}), x \rangle$$

Mirror descent alg (**MDA**) (Nesterov/Yudin 1983, Beck/Teboulle 2003

$$x^{(\gamma+1)} = \underset{x \in C}{\operatorname{argmin}} D_h(x, x^{(\gamma)}) + \tau_\gamma \langle \nabla f(x^{(\gamma)}), x \rangle \quad \text{or equivalent}$$

$$= \underset{x \in C}{\operatorname{argmin}} h(x) + \langle \tau_\gamma \nabla f(x^{(\gamma)}) - \nabla h(x^{(\gamma)}), x \rangle$$

$D_h = KL$: $x^{(\gamma+1)} = x^{(\gamma)} \circ e^{-\tau_\gamma \nabla f(x^{(\gamma)})}$

$f(x) = KL(Ax, b)$: $x^{(\gamma+1)} = x^{(\gamma)} \circ e^{-\tau_\gamma A^T \log \frac{Ax^{(\gamma)}}{b}}$

---

convergence : - general MDA $O(\frac{1}{\gamma^\tau})$

- accelerated BPG : $O(\frac{1}{\gamma^\tau})$ $\tau \in [1, 2]$ for special h (triangular scaling prop. of D_h)

- SMART for OT_ε : $A = \begin{pmatrix} P_1 \\ P_2 \end{pmatrix}$ rescale by $\frac{1}{2}$ to get $A^T 1 = 1$

$$\pi^{(\tau+1)} = \operatorname{diag}\left(\frac{v}{(\pi^{(\tau)})^T 1}\right)^{1/2} \pi^{(\tau)} \operatorname{diag}\left(\frac{\mu}{\pi^{(\tau)} 1}\right)^{1/2}$$

- this alg is 1/4 slower than Sinkhorn alg.
- it works also with $\alpha < 1$ (not $\alpha = 1$!) instead of $\alpha = \frac{1}{2}$; then it is 1/2 slower than Sinkhorn alg.

Question 1 : Why ?

GENERALIZATION

$C = C_1 \cap C_2 \cap \dots \cap C_N$, $C_k := \{x : A_k x = b_k\}$, $A_k, b_k \geq 0$, $A_k^T 1 = 1$
$\neq \emptyset$

(4) $\operatorname{argmin}_{x \in C} KL(x, y)$

Iterative information projection : $x^{(0)} = y$ → see (1) and (2)
(Csiszar 1989) $x^{(\tau+1)} = P^{KL}_{C_\tau}(x^{(\tau)})$ $C_{\tau+eN} = C_\tau$ (periodic projection)

- convergence for affine sets ; not for general convex sets (Dijstra alg. for convex sets...)

Generalized iterative scaling (GIS) : $x^{(0)} = y$
BI-SMART : $x^{(\tau+1)} = x^{(\tau)} \circ e^{-\tau A_\tau^T \log \frac{A_\tau x^{(\tau)}}{b_\tau}}$ one SMART step

- BI-SMART for OT_ε : $\pi^{(\tau+1)} = \pi^{(\tau)} \circ \operatorname{diag} \frac{\mu}{\pi^{(\tau)} 1}$
- Sinkhorn alg. $\pi^{(\tau+2)} = \operatorname{diag} \frac{v}{(\pi^{(\tau+1)})^T 1} \pi^{(\tau+1)}$

Question 2: BI-SMART for unbalanced OT : $(1-\lambda) KL(x, y) + \lambda KL(Ax, b)$
does not converge in the obvious form
Alternatives ?
BI-SMART in case $C = \emptyset$ (old problem, eg. of Byrne)

SMART solely performs the multiplicative update specified in (2) with its step-size fixed to $\tau_k = \frac{1}{L}$, where again f is L-smooth relative to φ.

FSMART is based on the iteration suggested in [2], where initially $\theta_0 = 1$ is chosen, which is then subsequently updated via $\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}$, as suggested in [5].

ABPG-e as described in [6, Algorithm 2] was applied to (1) with parameters $\gamma_{\min} = 1$, $\gamma_0 = 5$ and $\delta = 0.05$. The choices for γ_0 and δ deviate slightly from the recommendation in [6], but were chosen to facilitate fastest possible convergence on the selected problem instances. To ensure comparability restarting mechanisms and stopping criteria based on the divergence of iterates were foregone. Updates for θ were conducted via Newton's method.

ABPG-g specified in [6, Algorithm 3] to (1) is used with parameters: $\rho = 1.2$, $\gamma = 2$ and $G_{\min} = 10^{-3}$. Restarting, stopping criteria and updating θ was handled analogously to ABPG-e.

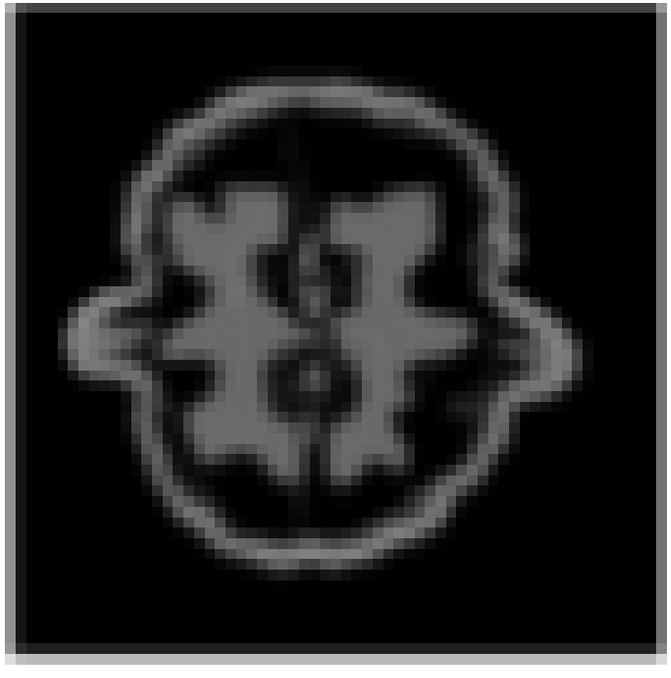
RG is a SMART iteration with Armijo line search for choosing the step size τ_k via the retraction in (29) to iterate according to (37). The line search parameters are $\sigma = 0.5$, $\beta = 0.8$, $\alpha = 5.0$.

PD is the Chambolle-Pock primal dual algorithm [17, Algorithm 1] for solving convex composite structured optimization problems of the form $f(x) = g(x) + h(Ax)$. For $h(y) := \mathrm{KL}(y, b)$ with $y = Ax$ and $g \equiv 0$ we obtain

$$x^{k+1} = x^k e^{-\tau A^\top y^k} \qquad \text{(primal-step)} \qquad (39)$$

$$y^{k+1} = \log\left(\frac{e^{y^k} + \sigma A(2x^{k+1} - x^k)}{\mathbb{1} + \sigma b}\right), \qquad \text{(dual-step)} \qquad (40)$$

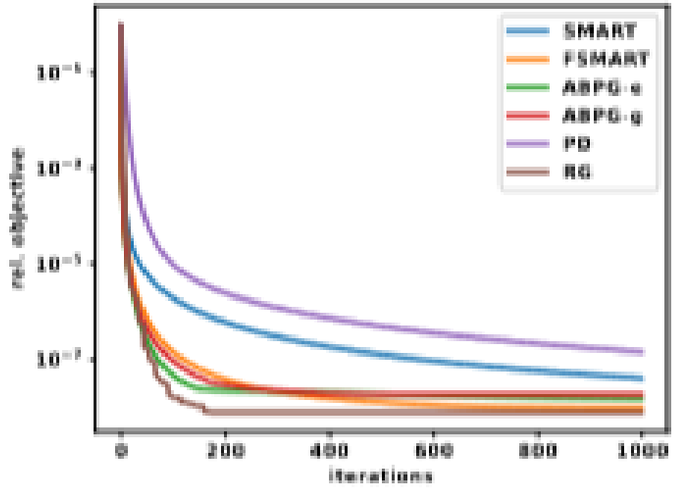
Shepp-Logan

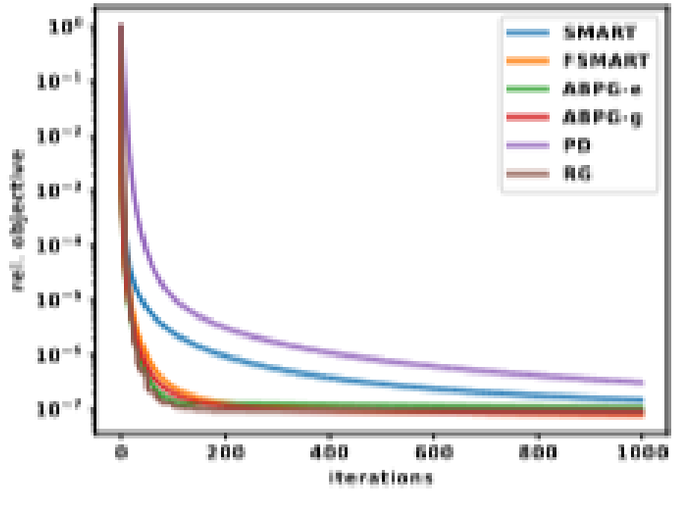


Walnut

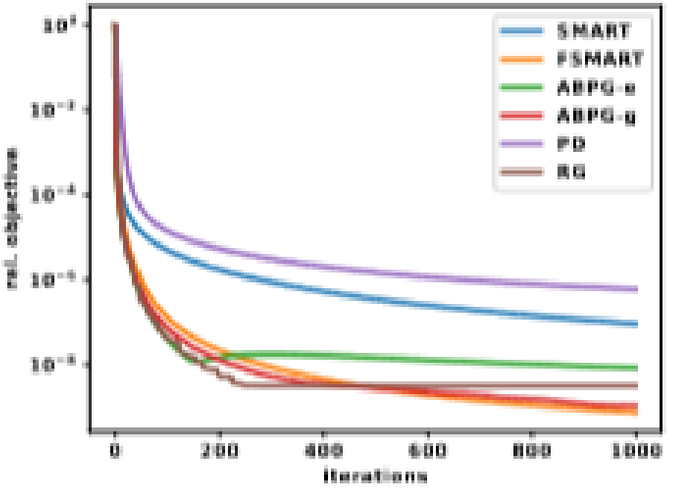
Batenburg

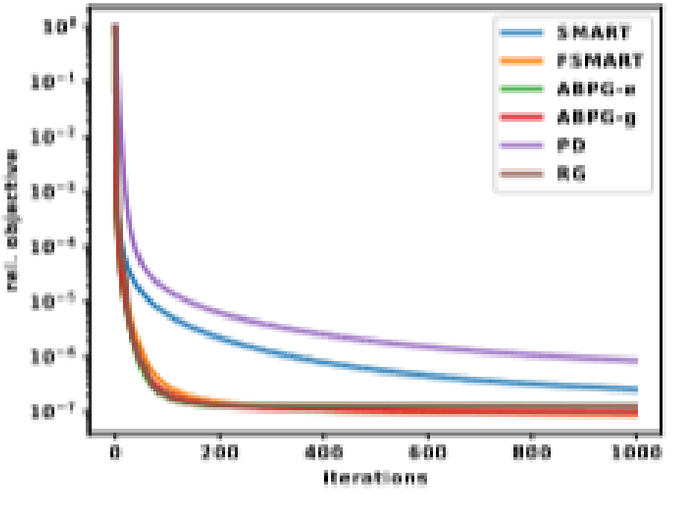
Shepp-Logan



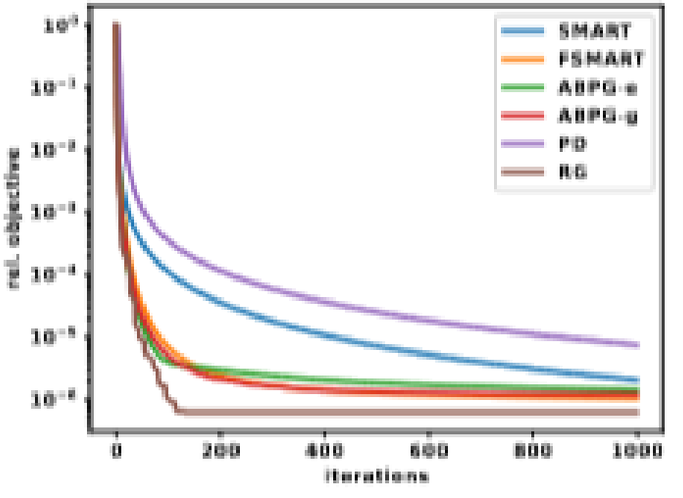


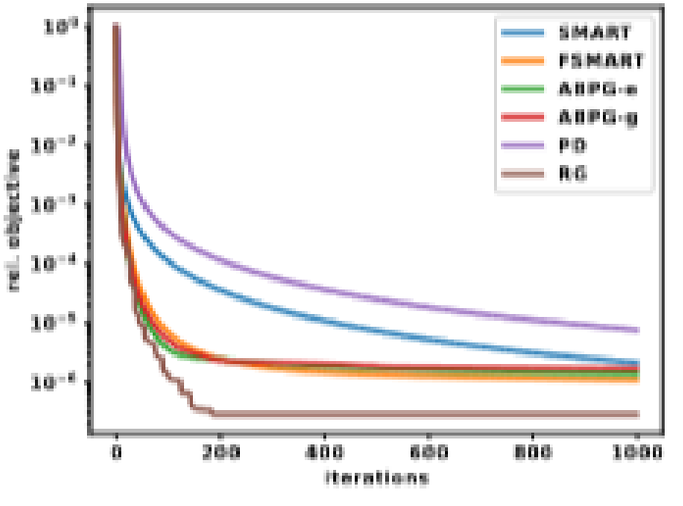
Walnut



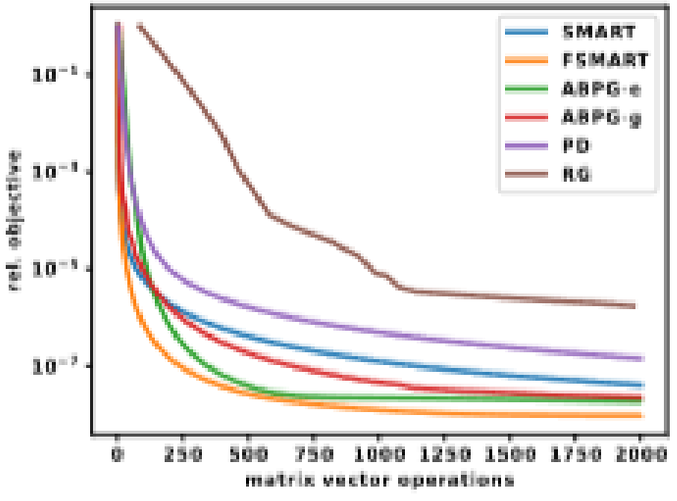


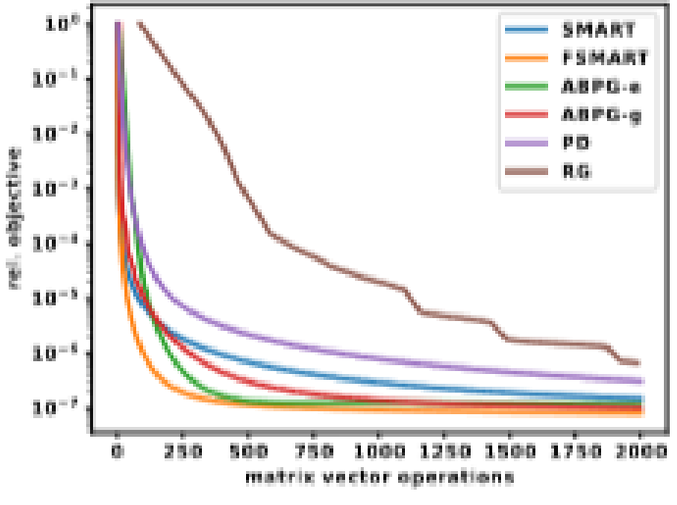
Batenburg



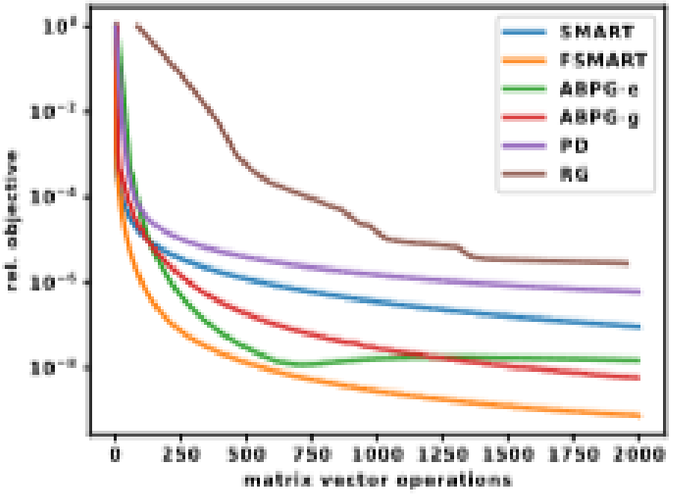


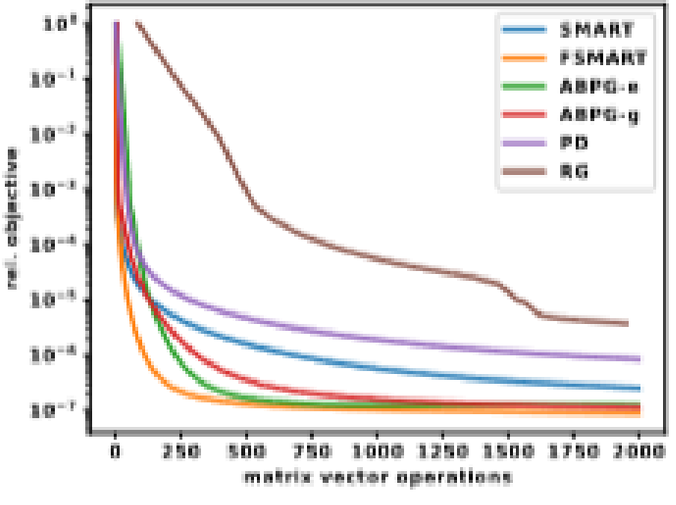
Shepp-Logan





Walnut





Batenburg