The Stochastic Bregman Primal-Dual Algorithm

Antonio Silveti-Falls, Cesare Molinari, and Jalal Fadili



Main Idea

Solving structured convex optimization problems:

$$\min_{x \in \mathcal{C}} f(x) + g(Tx).$$

Question

How to take advantage of properties of the individual terms?

Main Idea

Solving structured convex optimization problems:

$$\min_{x \in \mathcal{C}} f(x) + g(Tx).$$

Question

How to take advantage of properties of the individual terms?

- Lipschitz-smoothness - $\nabla f(x)$.

Main Idea

Solving structured convex optimization problems:

$$\min_{x \in \mathcal{C}} f(x) + g(Tx).$$

Question

How to take advantage of properties of the individual terms?

- Lipschitz-smoothness - $\nabla f(x)$.
- prox-friendliness - $\mathrm{prox}_g(x) \stackrel{\mathrm{def}}{=} \operatorname*{argmin}_u \left\{ g(u) + \frac{1}{2}\|x - u\|_2^2 \right\}$.

Main Idea

Solving structured convex optimization problems:

$$\min_{x \in \mathcal{C}} f(x) + g(Tx).$$

Question

How to take advantage of properties of the individual terms?

- Lipschitz-smoothness - $\nabla f(x)$.
- prox-friendliness - $\operatorname{prox}_g(x) \stackrel{\text{def}}{=} \operatorname*{argmin}_u \left\{ g(u) + \frac{1}{2}\|x - u\|_2^2 \right\}$.
- Projection onto $\mathcal{C}$ - $P_{\mathcal{C}}(x) \stackrel{\text{def}}{=} \operatorname*{argmin}_{u \in \mathcal{C}} \|x - u\|_2^2$.

Main Idea

Solving structured convex optimization problems:

$$\min_{x \in \mathcal{C}} f(x) + g(Tx).$$

Question

How to take advantage of properties of the individual terms?

- Lipschitz-smoothness - $\nabla f(x)$.
- prox-friendliness - $\operatorname{prox}_g(x) \stackrel{\text{def}}{=} \operatorname*{argmin}_u \left\{ g(u) + \frac{1}{2}\|x-u\|_2^2 \right\}$.
- Projection onto $\mathcal{C}$ - $P_{\mathcal{C}}(x) \stackrel{\text{def}}{=} \operatorname*{argmin}_{u \in \mathcal{C}} \|x-u\|_2^2$.

Changing the geometry?

Main Idea

Solving structured convex optimization problems:

$$\min_{x \in \mathcal{C}} f(x) + g(Tx).$$

Question

How to take advantage of properties of the individual terms?

- ~~Lipschitz~~-smoothness - $\nabla f(x)$.
- prox-friendliness - $\operatorname{prox}_g(x) \stackrel{\text{def}}{=} \operatorname{argmin}_u \left\{ g(u) + \frac{1}{2}\|x-u\|_2^2 \right\}$.
- Projection onto $\mathcal{C}$ - $P_{\mathcal{C}}(x) \stackrel{\text{def}}{=} \operatorname{argmin}_{u \in \mathcal{C}} \|x-u\|_2^2$.

Changing the geometry?

An Example

Consider a matrix $Y \stackrel{\text{def}}{=} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}_{++}^{n \times p}$ with $y_i \in \Delta^p$ and matrices $A_1, \ldots, A_n \in \mathbb{R}_{+}^{p \times m}$ without any zero rows.

An Example

Consider a matrix $Y \stackrel{\text{def}}{=} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^{n\times p}_{++}$ with $y_i \in \Delta^p$ and matrices $A_1, \ldots, A_n \in \mathbb{R}^{p\times m}_{+}$ without any zero rows.

We examine the *trend filtering* problem of recovering a matrix $X \stackrel{\text{def}}{=} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^{n\times m}_{+}$ with $x_i \in \Delta^m$ under the model

$$\mathcal{A}X \approx Y \quad \text{with} \quad \mathcal{A}X \stackrel{\text{def}}{=} \begin{pmatrix} A_1 x_1 \\ \vdots \\ A_n x_n \end{pmatrix} \in \mathbb{R}^{n\times p}_{+}$$

and assuming that the columns of X are piecewise constant.

Trend Filtering - Notation

The Kullback-Leibler divergence

For $u, v \in \mathbb{R}_+$,

$$\mathrm{KL}\left(u, v\right) \stackrel{\text{def}}{=} \begin{cases} u \log\left(\frac{u}{v}\right) - u + v & \text{if } u, v > 0, \\ v & \text{if } u = 0, \\ +\infty & \text{otherwise.} \end{cases}$$

Trend Filtering - Notation

The Kullback-Leibler divergence

For $u, v \in \mathbb{R}_+$,

$$\mathrm{KL}(u, v) \stackrel{\text{def}}{=} \begin{cases} u \log\left(\frac{u}{v}\right) - u + v & \text{if } u, v > 0, \\ v & \text{if } u = 0, \\ +\infty & \text{otherwise.} \end{cases}$$

The row gradient

$\nabla_{\text{row}} : \mathbb{R}^{n \times m} \to \mathbb{R}^{m(n-1)}$. For a matrix $X \in \mathbb{R}^{n \times m}$,

$$\nabla_{\text{row}} X \stackrel{\text{def}}{=} \begin{pmatrix} x_2 - x_1 \\ \vdots \\ x_n - x_{n-1} \end{pmatrix}.$$

Trend Filtering - A Closer Look

Trend filtering

$$\min_{\substack{X \in \mathbb{R}_+^{n \times m} \\ X\mathbb{1}_m = \mathbb{1}_n}} \underbrace{\sum_{i=1}^{n} \mathrm{KL}\left(A_i x_i, y_i\right)}_{f(X)} + \underbrace{\beta \left\|\nabla_{\mathrm{row}} X\right\|_1}_{g \circ \nabla_{\mathrm{row}}(X)}$$

Trend Filtering - A Closer Look

Trend filtering

$$\min_{\substack{X \in \mathbb{R}_+^{n\times m} \\ X\mathbb{1}_m = \mathbb{1}_n}} \underbrace{\sum_{i=1}^{n} \mathrm{KL}\left(A_i x_i, y_i\right)}_{f(X)} + \underbrace{\beta \left\|\nabla_{\mathrm{row}} X\right\|_1}_{g\circ\nabla_{\mathrm{row}}(X)}$$

Obstacles:

- f is not Lipschitz-smooth at the origin.

Trend Filtering - A Closer Look

Trend filtering

$$\min_{\substack{X \in \mathbb{R}_+^{n \times m} \\ X\mathbb{1}_m = \mathbb{1}_n}} \underbrace{\sum_{i=1}^{n} \mathrm{KL}\left(A_i x_i, y_i\right)}_{f(X)} + \underbrace{\beta \left\|\nabla_{\mathrm{row}} X\right\|_1}_{g \circ \nabla_{\mathrm{row}}(X)}$$

Obstacles:

- f is not Lipschitz-smooth at the origin.
- ∇_{row} makes computing $\mathrm{prox}_{g \circ \nabla_{\mathrm{row}}}$ difficult.

Trend Filtering - A Closer Look

Trend filtering

$$\min_{\substack{X \in \mathbb{R}_+^{n\times m} \\ X\mathbb{1}_m = \mathbb{1}_n}} \underbrace{\sum_{i=1}^{n} \mathrm{KL}\left(A_i x_i, y_i\right)}_{f(X)} + \underbrace{\beta \left\|\nabla_{\mathrm{row}} X\right\|_1}_{g\circ\nabla_{\mathrm{row}}(X)}$$

Obstacles:

- f is not Lipschitz-smooth at the origin.
- ∇_{row} makes computing $\mathrm{prox}_{g\circ\nabla_{\mathrm{row}}}$ difficult.
- prox_f is computable when the A_i are nice but requires special functions (Lambert W-function).

Trend Filtering - A Closer Look

Trend filtering

$$\min_{\substack{X \in \mathbb{R}_+^{n\times m} \\ X\mathbb{1}_m = \mathbb{1}_n}} \underbrace{\sum_{i=1}^{n} \mathrm{KL}\left(A_i x_i, y_i\right)}_{f(X)} + \underbrace{\beta \left\|\nabla_{\mathrm{row}} X\right\|_1}_{g\circ\nabla_{\mathrm{row}}(X)}$$

Obstacles:

- f is not Lipschitz-smooth at the origin.
- ∇_{row} makes computing $\mathrm{prox}_{g\circ\nabla_{\mathrm{row}}}$ difficult.
- prox_f is computable when the A_i are nice but requires special functions (Lambert W-function).
- Projecting (in the euclidean norm) onto the constraint set requires sorting.

Primal-Dual Splitting

Using a primal-dual formulation of the problem, we have

$$\min_{\substack{X \in \mathbb{R}_+^{n\times m} \\ X\mathbb{1}_m = \mathbb{1}_n}} \max_{\mu \in \mathbb{R}^{m(n-1)}} \sum_{i=1}^{n} \mathrm{KL}\left(A_i x_i, y_i\right) + \langle \nabla_{\mathrm{row}} X, \mu \rangle - \iota_{\mathcal{B}_\infty^\beta}(\mu).$$

This formulation solves the issue of computing $\mathrm{prox}_{g \circ \nabla_{\mathrm{row}}}$.

Primal-Dual Splitting

Using a primal-dual formulation of the problem, we have

$$\min_{\substack{X \in \mathbb{R}_+^{n \times m} \\ X\mathbb{1}_m = \mathbb{1}_n}} \max_{\mu \in \mathbb{R}^{m(n-1)}} \sum_{i=1}^{n} \mathrm{KL}\left(A_i x_i, y_i\right) + \langle \nabla_{\mathrm{row}} X, \mu \rangle - \iota_{\mathcal{B}_\infty^\beta}(\mu).$$

This formulation solves the issue of computing $\mathrm{prox}_{g \circ \nabla_{\mathrm{row}}}$.
Remaining obstacles:

- Projection onto the constraint set.

Primal-Dual Splitting

Using a primal-dual formulation of the problem, we have

$$\min_{\substack{X \in \mathbb{R}_+^{n \times m} \\ X\mathbb{1}_m = \mathbb{1}_n}} \max_{\mu \in \mathbb{R}^{m(n-1)}} \quad \sum_{i=1}^{n} \mathrm{KL}\left(A_i x_i, y_i\right) + \langle \nabla_{\mathrm{row}} X, \mu \rangle - \iota_{\mathcal{B}_\infty^\beta}\left(\mu\right).$$

This formulation solves the issue of computing $\mathrm{prox}_{g \circ \nabla_{\mathrm{row}}}$.
Remaining obstacles:

- Projection onto the constraint set.
- Utilize differentiability of $\sum_{i=1}^{n} K\left(A_i x_i, y_i\right)$.

Template Primal-Dual Problem

Let $\mathcal{X}_p$ and $\mathcal{X}_d$ be reflexive Banach spaces.

Primal-dual problem

$$\min_{x\in\mathcal{X}_p} \max_{\mu\in\mathcal{X}_d} \underbrace{f(x) + g(x) + \langle Tx, \mu\rangle - h^*(\mu) - \ell^*(\mu) + \iota_{\mathcal{C}_p}(x) - \iota_{\mathcal{C}_d}(\mu)}_{\mathcal{L}(x,\mu)}$$

Template Primal-Dual Problem

Let $\mathcal{X}_p$ and $\mathcal{X}_d$ be reflexive Banach spaces.

Primal-dual problem

$$\min_{x\in\mathcal{X}_p} \max_{\mu\in\mathcal{X}_d} \quad \underbrace{f(x) + g(x) + \langle Tx, \mu\rangle - h^*(\mu) - \ell^*(\mu) + \iota_{\mathcal{C}_p}(x) - \iota_{\mathcal{C}_d}(\mu)}_{\mathcal{L}(x,\mu)}$$

- $\mathcal{C}_p$ and $\mathcal{C}_d$ are nonempty closed convex subsets.

Template Primal-Dual Problem

Let $\mathcal{X}_p$ and $\mathcal{X}_d$ be reflexive Banach spaces.

Primal-dual problem

$$\min_{x \in \mathcal{X}_p} \max_{\mu \in \mathcal{X}_d} \underbrace{f(x) + g(x) + \langle Tx, \mu \rangle - h^*(\mu) - \ell^*(\mu) + \iota_{\mathcal{C}_p}(x) - \iota_{\mathcal{C}_d}(\mu)}_{\mathcal{L}(x,\mu)}$$

- $\mathcal{C}_p$ and $\mathcal{C}_d$ are nonempty closed convex subsets.
- f and h^* are relatively smooth with respect to ϕ_p and ϕ_d, respectively.

Template Primal-Dual Problem

Let $\mathcal{X}_p$ and $\mathcal{X}_d$ be reflexive Banach spaces.

Primal-dual problem

$$\min_{x\in\mathcal{X}_p} \max_{\mu\in\mathcal{X}_d} \quad \underbrace{f(x) + g(x) + \langle Tx, \mu\rangle - h^*(\mu) - \ell^*(\mu) + \iota_{\mathcal{C}_p}(x) - \iota_{\mathcal{C}_d}(\mu)}_{\mathcal{L}(x,\mu)}$$

- $\mathcal{C}_p$ and $\mathcal{C}_d$ are nonempty closed convex subsets.
- f and h^* are relatively smooth with respect to ϕ_p and ϕ_d, respectively.
- T is a bounded linear operator.

A Different Kind of Distance

Bregman divergence

Let $\mathcal{X}$ be a Banach space and define the *Bregman divergence* of a differentiable function $f : \mathcal{C} \subset \mathcal{X} \to \mathbb{R}$, for any $u, v \in \mathcal{C}$,

$$D_f(u, v) \stackrel{\text{def}}{=} f(u) - f(v) - \langle \nabla f(v), u - v \rangle .$$

A Different Kind of Distance

Bregman divergence

Let $\mathcal{X}$ be a Banach space and define the *Bregman divergence* of a differentiable function $f : \mathcal{C} \subset \mathcal{X} \to \mathbb{R}$, for any $u, v \in \mathcal{C}$,

$$D_f(u, v) \stackrel{\text{def}}{=} f(u) - f(v) - \langle \nabla f(v), u - v \rangle .$$

- $D_f(u, v)$ is a sort of distance between u and v. For the euclidean squared norm $f(x) = \frac{1}{2}\|x\|_2^2$, it holds

$$D_f(u, v) = \frac{1}{2}\|u - v\|_2^2 .$$

D-prox Operators

Euclidean prox operator

Given a function $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$, we define the proximal operator

$$\mathrm{prox}_f(u) \stackrel{\text{def}}{=} \underset{v \in \mathcal{H}}{\operatorname{argmin}} \left\{ f(v) + \frac{1}{2} \|v - u\|_2^2 \right\}.$$

D-prox Operators

Euclidean prox operator

Given a function $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$, we define the proximal operator

$$\operatorname{prox}_f(u) \stackrel{\text{def}}{=} \operatorname*{argmin}_{v \in \mathcal{H}} \left\{ f(v) + \frac{1}{2} \|v - u\|_2^2 \right\}.$$

D-prox operator

Bregman divergence D_ϕ for some differentiable $\phi \in \Gamma_0(\mathcal{X})$, define the D-prox operator,

$$\operatorname{prox}_f^{D_\phi}(u) \stackrel{\text{def}}{=} \operatorname*{argmin}_{v \in \mathcal{X}} \left\{ f(v) + D_\phi(v, u) \right\}.$$

Going Beyond Lipschitz-smoothness

Relative smoothness

f is *relatively smooth* [Bauschke et al. 2017], [Lu et al. 2018] with respect to a differentiable function $\phi : \mathcal{C} \subset \mathcal{X} \to \mathbb{R}$ if there exists $L > 0$ such that, for any $u, v \in \mathcal{X}$,

$$D_f(u, v) \leq L D_\phi(u, v)$$

(equivalently, if $L\phi - f$ is a convex function).

Going Beyond Lipschitz-smoothness

Relative smoothness

f is *relatively smooth* [Bauschke et al. 2017], [Lu et al. 2018] with respect to a differentiable function $\phi : \mathcal{C} \subset \mathcal{X} \to \mathbb{R}$ if there exists $L > 0$ such that, for any $u, v \in \mathcal{X}$,

$$D_f(u, v) \leq L D_\phi(u, v)$$

(equivalently, if $L\phi - f$ is a convex function).

- Lipschitz-smooth functions in $\Gamma_0(\mathcal{X})$ are relatively smooth with respect to the euclidean squared norm $\frac{1}{2}\|\cdot\|_2^2$:

$$\begin{aligned} & D_f(u, v) \leq L\|u - v\|_2^2 \\ \Longrightarrow\ & f(u) \leq f(v) + \langle \nabla f(v), u - v\rangle + L\|u - v\|_2^2 \\ \Longrightarrow\ & f \text{ is } L\text{-smooth (Baillon-Haddad Theorem).} \end{aligned}$$

Bregman Primal-Dual Algorithm

Algorithm: Bregman Primal-Dual (BPD)

Input: $x_0 \in \mathcal{C}_p$, $\mu_0 \in \mathcal{C}_d$, $(\lambda_k)_{k\in\mathbb{N}}$, $(\nu_k)_{k\in\mathbb{N}}$,
$\phi_p : \mathcal{X}_p \to \mathbb{R} \cup \{+\infty\}$, $\phi_d : \mathcal{X}_d \to \mathbb{R} \cup \{+\infty\}$.

$k = 0$

repeat

$$x_{k+1} = \underset{x\in\mathcal{C}_p}{\operatorname{argmin}} \Big\{ g(x) + \big\langle \nabla f(x_k) \quad , x \big\rangle + \langle x, T^*\mu_k \rangle + \tfrac{1}{\lambda_k} D_{\phi_p}(x, x_k) \Big\}$$

$$\mu_{k+1} = \underset{\mu\in\mathcal{C}_d}{\operatorname{argmin}} \Big\{ \ell^*(\mu) + \big\langle \nabla h^*(\mu_k) \quad , \mu \big\rangle - \langle T(2x_{k+1} - x_k), \mu \rangle + \tfrac{1}{\nu_k} D_{\phi_d}(\mu, \mu_k) \Big\}$$

$k \leftarrow k + 1$

until *convergence*;

Output: x_k, μ_k.

Bregman Primal-Dual Algorithm

Algorithm: Bregman Primal-Dual (BPD)

Input: $x_0 \in \mathcal{C}_p$, $\mu_0 \in \mathcal{C}_d$, $(\lambda_k)_{k\in\mathbb{N}}$, $(\nu_k)_{k\in\mathbb{N}}$,
$\phi_p : \mathcal{X}_p \to \mathbb{R} \cup \{+\infty\}$, $\phi_d : \mathcal{X}_d \to \mathbb{R} \cup \{+\infty\}$.

$k = 0$

repeat

$$x_{k+1} = \operatorname*{argmin}_{x\in\mathcal{C}_p} \Big\{ g(x) + \big\langle \nabla f(x_k) \quad , x \big\rangle + \langle x, T^* \mu_k \rangle + \frac{1}{\lambda_k} D_{\phi_p}(x, x_k) \Big\}$$

$$\mu_{k+1} = \operatorname*{argmin}_{\mu\in\mathcal{C}_d} \Big\{ \ell^*(\mu) + \big\langle \nabla h^*(\mu_k) \quad , \mu \big\rangle - \langle T(2x_{k+1} - x_k), \mu \rangle + \frac{1}{\nu_k} D_{\phi_d}(\mu, \mu_k) \Big\}$$

$k \leftarrow k + 1$

until *convergence*;

Output: x_k, μ_k.

Algorithm: Stochastic Bregman Primal-Dual (SBPD)

Input: $x_0 \in \mathcal{C}_p$, $\mu_0 \in \mathcal{C}_d$, $(\lambda_k)_{k\in\mathbb{N}}$, $(\nu_k)_{k\in\mathbb{N}}$,
$\phi_p : \mathcal{X}_p \to \mathbb{R} \cup \{+\infty\}$, $\phi_d : \mathcal{X}_d \to \mathbb{R} \cup \{+\infty\}$.

$k = 0$

repeat

$$x_{k+1} = \operatorname*{argmin}_{x\in\mathcal{C}_p} \Big\{ g(x) + \big\langle \nabla f(x_k) + \delta_k^p, x \big\rangle + \langle x, T^*\mu_k \rangle + \frac{1}{\lambda_k} D_{\phi_p}(x, x_k) \Big\}$$

$$\mu_{k+1} = \operatorname*{argmin}_{\mu\in\mathcal{C}_d} \Big\{ \ell^*(\mu) + \big\langle \nabla h^*(\mu_k) + \delta_k^d, \mu \big\rangle - \langle T(2x_{k+1} - x_k), \mu \rangle + \frac{1}{\nu_k} D_{\phi_d}(\mu, \mu_k) \Big\}$$

$k \leftarrow k + 1$

until *convergence*;

Output: x_k, μ_k.

Interpretation of the Algorithm

Alternatively,

$$x_{k+1} = \underbrace{\left[\nabla \phi_p + \lambda_k \partial g\right]^{-1}}_{\text{Backward step}} \underbrace{\left(\nabla \phi_p (x_k) - \lambda_k \nabla f (x_k) - \lambda_k T^* \mu_k\right)}_{\text{Forward step}};$$

$$\mu_{k+1} = \underbrace{\left[\nabla \phi_d + \nu_k \partial \ell^*\right]^{-1}}_{\text{Backward step}} \underbrace{\left(\nabla \phi_d (\mu_k) - \nu_k \nabla h^* (\mu_k) + \nu_k T (2x_{k+1} - x_k)\right)}_{\text{Forward step}}$$

Interpretation of the Algorithm

Alternatively,

$$x_{k+1} = \underbrace{\left[\nabla\phi_p + \lambda_k \partial g\right]^{-1}}_{\text{Backward step}} \underbrace{\left(\nabla\phi_p\left(x_k\right) - \lambda_k \nabla f\left(x_k\right) - \lambda_k T^* \mu_k\right)}_{\text{Forward step}};$$

$$\mu_{k+1} = \underbrace{\left[\nabla\phi_d + \nu_k \partial \ell^*\right]^{-1}}_{\text{Backward step}} \underbrace{\left(\nabla\phi_d\left(\mu_k\right) - \nu_k \nabla h^*\left(\mu_k\right) + \nu_k T\left(2x_{k+1} - x_k\right)\right)}_{\text{Forward step}}$$

- $\phi_p = \frac{1}{2}\left\|\cdot\right\|_2^2 \implies \nabla\phi_p = \mathrm{Id}$ (likewise for ϕ_d).

Alternatively,

$$x_{k+1} = \underbrace{\left[\nabla\phi_p + \lambda_k \partial g\right]^{-1}}_{\text{Backward step}} \underbrace{\left(\nabla\phi_p\left(x_k\right) - \lambda_k \nabla f\left(x_k\right) - \lambda_k T^* \mu_k\right)}_{\text{Forward step}};$$

$$\mu_{k+1} = \underbrace{\left[\nabla\phi_d + \nu_k \partial \ell^*\right]^{-1}}_{\text{Backward step}} \underbrace{\left(\nabla\phi_d\left(\mu_k\right) - \nu_k \nabla h^*\left(\mu_k\right) + \nu_k T\left(2x_{k+1} - x_k\right)\right)}_{\text{Forward step}}$$

- $\phi_p = \frac{1}{2}\left\|\cdot\right\|_2^2 \implies \nabla\phi_p = \mathrm{Id}$ (likewise for ϕ_d).
- Flavor of mirror descent [Nemirovsky et al. 83], Chambolle-Pock [Chambolle et al. 2011], [Chambolle et al., 2016], NoLips [Bauschke et al. 2017], Bregman Forward-Backward [Nguyen, 2017], etc.

Matching the Geometries

- f is L_p relatively smooth with respect to ϕ_p on $\mathrm{int}\,(\mathcal{C}_p)$.

Matching the Geometries

- f is L_p relatively smooth with respect to ϕ_p on $\operatorname{int}(\mathcal{C}_p)$.
- h^* is L_d relatively smooth with respect to ϕ_d on $\operatorname{int}(\mathcal{C}_d)$.

Matching the Geometries

- f is L_p relatively smooth with respect to ϕ_p on $\operatorname{int}(\mathcal{C}_p)$.
- h^* is L_d relatively smooth with respect to ϕ_d on $\operatorname{int}(\mathcal{C}_d)$.
- ϕ_p and ϕ_d are Legendre functions with domains $\mathcal{C}_p$ and $\mathcal{C}_d$ and the mappings $[\nabla\phi_p + \lambda_k\partial g]^{-1}$ and $[\nabla\phi_d + \nu_k\partial\ell^*]^{-1}$ are well-defined.

Matching the Geometries

- f is L_p relatively smooth with respect to ϕ_p on $\operatorname{int}(\mathcal{C}_p)$.
- h^* is L_d relatively smooth with respect to ϕ_d on $\operatorname{int}(\mathcal{C}_d)$.
- ϕ_p and ϕ_d are Legendre functions with domains $\mathcal{C}_p$ and $\mathcal{C}_d$ and the mappings $[\nabla\phi_p + \lambda_k \partial g]^{-1}$ and $[\nabla\phi_d + \nu_k \partial \ell^*]^{-1}$ are well-defined.

Note

The geometry of ϕ_p and ϕ_d must match the problem!

Matching the Geometries

- f is L_p relatively smooth with respect to ϕ_p on $\operatorname{int}(\mathcal{C}_p)$.
- h^* is L_d relatively smooth with respect to ϕ_d on $\operatorname{int}(\mathcal{C}_d)$.
- ϕ_p and ϕ_d are Legendre functions with domains $\mathcal{C}_p$ and $\mathcal{C}_d$ and the mappings $[\nabla\phi_p + \lambda_k \partial g]^{-1}$ and $[\nabla\phi_d + \nu_k \partial \ell^*]^{-1}$ are well-defined.

Note

The geometry of ϕ_p and ϕ_d must match the problem!

Note

We do not assume strong convexity of ϕ_p or ϕ_d (cf. [Chambolle et al., 2016], [Van Dung et al., 2021]).

Replacing Young's Inequality

Define

$$M(x_1, x_2, \mu_1, \mu_2) \stackrel{\text{def}}{=} \langle T(x_1 - x_2), \mu_1 - \mu_2 \rangle$$

We require existence of two functions $r_p : \mathcal{X}_p^2 \to \mathbb{R}_+$ and $r_d : \mathcal{X}_d^2 \to \mathbb{R}_+$ and $\epsilon \geq 0$ such that, for all $x_1 \in \mathcal{C}_p \cap \mathrm{dom}\partial g$, $x_2 \in \mathrm{int}(\mathcal{C}_p) \cap \mathrm{dom}\partial g$, $\mu_1 \in \mathcal{C}_d \cap \mathrm{dom}\partial \ell$, and $\mu_2 \in \mathrm{int}(\mathcal{C}_d) \cap \mathrm{dom}\partial \ell$,

$$\frac{\left(\lambda_\infty^{-1} - L_p\right) D_{\phi_p}(x_1, x_2) + \left(\nu_\infty^{-1} - L_d\right) D_{\phi_d}(\mu_1, \mu_2) - M(x_1, x_2, \mu_1, \mu_2)}{r_p(x_1, x_2) + r_d(\mu_1, \mu_2)} \geq \epsilon$$

Replacing Young's Inequality

$$\frac{\left(\lambda_\infty^{-1} - L_p\right) D_{\phi_p}\left(x_1, x_2\right) + \left(\nu_\infty^{-1} - L_d\right) D_{\phi_d}\left(\mu_1, \mu_2\right) - M\left(x_1, x_2, \mu_1, \mu_2\right)}{r_p\left(x_1, x_2\right) + r_d\left(\mu_1, \mu_2\right)} \geq \epsilon$$

Consider $\phi_p = \frac{1}{2}\left\|\cdot\right\|_2^2$ and $\phi_d = \frac{1}{2}\left\|\cdot\right\|_2^2$,

$$\begin{aligned}\frac{\lambda_\infty^{-1} - L_p}{2}\left\|x_1 - x_2\right\|_2^2 + \frac{\nu_\infty^{-1} - L_d}{2}\left\|\mu_1 - \mu_2\right\|_2^2 - \left\langle T\left(x_1 - x_2\right), \mu_1 - \mu_2\right\rangle \\ \geq \underbrace{\frac{\lambda_\infty^{-1} - L_p - \|T\|}{2}\left\|x_1 - x_2\right\|_2^2}_{r_p(x_1, x_2)} + \underbrace{\frac{\nu_\infty^{-1} - L_d - 1}{2}\left\|\mu_1 - \mu_2\right\|_2^2}_{r_d(\mu_1, \mu_2)}\end{aligned}$$

Ergodic Convergence Results - Deterministic Setting

Theorem (Ergodic Convergence Rate)

Define $\bar{x}_k \stackrel{\text{def}}{=} \frac{1}{k}\sum_{i=0}^{k} x_i$, $\bar{\mu}_k \stackrel{\text{def}}{=} \frac{1}{k}\sum_{i=0}^{k} \mu_i$, *and, for* $w \stackrel{\text{def}}{=} (x, \mu)$, $M(w, w') = \langle T(x - x'), \mu - \mu' \rangle$. *Under [assumptions], for each* $k \in \mathbb{N}$, *for every* $w \in \mathcal{C}_p \times \mathcal{C}_d$,

$$\mathcal{L}(\bar{x}_k, \mu) - \mathcal{L}(x, \bar{\mu}_k) \leq \frac{\Lambda_0^{-1} D_{\phi_p,\phi_d}(w, w_0) - M(w, w_0)}{k}.$$

In particular, every weak cluster point of the sequence $(\bar{x}_k, \bar{\mu}_k)_{k\in\mathbb{N}}$ *is a solution to the primal-dual problem.*

Ergodic Convergence Results - Stochastic Setting

Theorem (Ergodic Convergence Rate)

Under [the same assumptions], if the errors δ_k^p and δ_k^d are unbiased conditioned on the the previous iterates, for each $k \in \mathbb{N}$, for every $w \in \mathcal{C}_p \times \mathcal{C}_d$,

$$\mathbb{E}\left[\mathcal{L}\left(\bar{x}_k, \mu\right) - \mathcal{L}\left(x, \bar{\mu}_k\right)\right] \leq \frac{\Lambda_0^{-1} D_{\phi_p,\phi_d}\left(w, w_0\right) - M\left(w, w_0\right)}{k} + \frac{\sum_{i=0}^{k-1} \mathbb{E}\left[\langle \Delta_i, w - w_{i+1} \rangle\right]}{k}.$$

In particular, every almost sure weak cluster point of the sequence $\left(\bar{x}_k, \bar{\mu}_k\right)_{k \in \mathbb{N}}$ is a solution to the primal-dual problem in expectation ($\mathbb{E}\left[(x_\infty, \mu_\infty)\right]$ is a saddle-point).

Trend Filtering

Trend filtering problem - primal-dual formulation

$$\min_{\substack{X \in \mathbb{R}_+^{n\times m} \\ X\mathbb{1}_m = \mathbb{1}_n}} \max_{\mu \in \mathbb{R}^{m(n-1)}} \quad \sum_{i=1}^{n} \mathrm{KL}\left(A_i x_i, y_i\right) + \langle \nabla_{\mathrm{row}} X, \mu \rangle - \iota_{\mathcal{B}_\infty^\beta}(\mu).$$

Trend Filtering

Trend filtering problem - primal-dual formulation

$$\min_{\substack{X \in \mathbb{R}_+^{n \times m} \\ X\mathbb{1}_m = \mathbb{1}_n}} \max_{\mu \in \mathbb{R}^{m(n-1)}} \sum_{i=1}^{n} \mathrm{KL}\left(A_i x_i, y_i\right) + \langle \nabla_{\mathrm{row}} X, \mu \rangle - \iota_{\mathcal{B}_\infty^\beta}\left(\mu\right).$$

Apply SBPD with

$$f\left(X\right) = \sum_{i=1}^{n} \mathrm{KL}\left(A_i x_i, y_i\right), \quad g\left(X\right) = \iota_{\mathbb{1}_n}\left(X\mathbb{1}_m\right), \quad \mathcal{C}_p = \mathbb{R}_+^{n \times m},$$

$$T = \nabla_{\mathrm{row}} \quad h^*\left(\mu\right) = 0, \quad \ell^*\left(\mu\right) = \iota_{\mathcal{B}_\infty^\beta}\left(\mu\right) \quad \text{and} \quad \mathcal{C}_d = \mathbb{R}^{m(n-1)}$$

Choosing ϕ_p and ϕ_d

Primal entropy ϕ_p

- $\mathcal{C}_p = \mathbb{R}_+^{n \times m}$

$$\phi_p(X) = \sum_{i=1}^{n} \sum_{j=1}^{m} X_{i,j} \log(X_{i,j}).$$

- Must show $\exists L_p > 0$ such that $L_p \phi_p - f$ is convex.
- Must compute $\operatorname{prox}_{\lambda_k g}^{D_{\phi_p}}(X)$.

Choosing ϕ_p and ϕ_d

Primal entropy ϕ_p

- $\mathcal{C}_p = \mathbb{R}_+^{n\times m}$

$$\phi_p(X) = \sum_{i=1}^{n}\sum_{j=1}^{m} X_{i,j}\log(X_{i,j}).$$

- Must show $\exists L_p > 0$ such that $L_p\phi_p - f$ is convex.
- Must compute $\mathrm{prox}_{\lambda_k g}^{D_{\phi_p}}(X)$.

Dual entropy ϕ_d

- $\mathcal{C}_d = \mathbb{R}^{m(n-1)}$ (trivial constraint)

$$\phi_d(\mu) = \frac{1}{2}\|\mu\|_2^2.$$

- Euclidean prox of $\ell^*(\mu) = \iota_{\mathcal{B}_\infty^\beta}$ is accessible.

New Geometry of ϕ_p

Relative smoothness

For each $i \in \{1, \ldots, n\}$, let $L_i \geq \max\limits_{1 \leq q \leq m} \sum\limits_{j=1}^{p} A_i(j, q)$ and let $L_p = \max\limits_{1 \leq i \leq n} L_i$. Then $L_p \phi_p - f$ is convex on $\operatorname{int}(\mathcal{C}_p)$.

New Geometry of ϕ_p

Relative smoothness

For each $i \in \{1, \ldots, n\}$, let $L_i \geq \max_{1 \leq q \leq m} \sum_{j=1}^{p} A_i(j, q)$ and let $L_p = \max_{1 \leq i \leq n} L_i$. Then $L_p \phi_p - f$ is convex on $\mathrm{int}(\mathcal{C}_p)$.

D-prox under ϕ_p

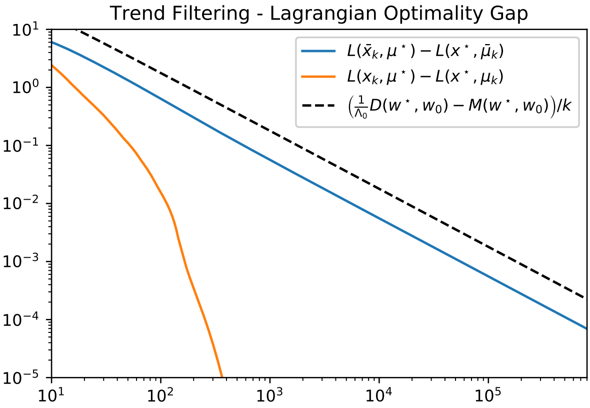
For each $X \in \mathcal{C}_p$,

$$\mathrm{prox}_{\lambda_k g}^{D_{\phi_p}}(X) = \underset{\substack{U \in \mathbb{R}_+^{n \times m} \\ U^T \mathbb{1}_m = \mathbb{1}_n}}{\operatorname{argmin}} \left\{ D_{\phi_p}(U, X) \right\} = \left(\frac{\exp(X_{i,j})}{\sum_{q=1}^{m} \exp(X_{i,q})} \right)_{i,j}$$

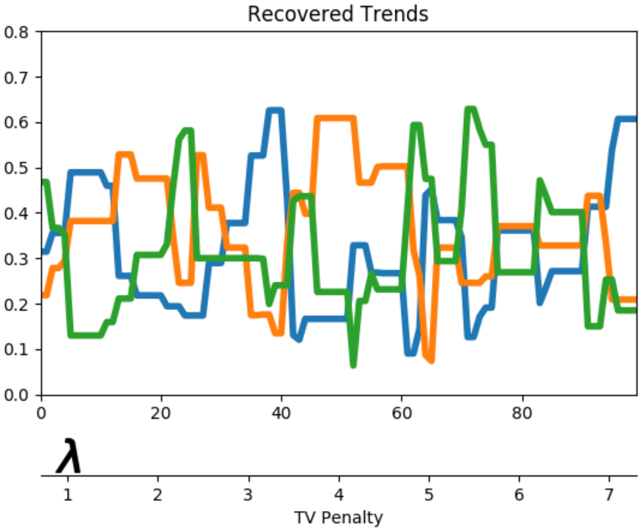
i.e., project each row onto the simplex under D_{ϕ_p}.

Results - Convergence

We take $n = 100$, $m = 3$ and $\beta = 1$ with synthetic (randomly generated) data Y and $A_i = \mathrm{Id}$.



Results - Different Values of β



Entropically Regularized Wasserstein Inverse Problems

Simplest case: discrete measures ρ and θ with ground cost matrix $C \in \mathbb{R}_+^{n\times m}$.

Entropically regularized Wasserstein distance

$$W_\gamma(\rho,\theta) = \inf_{\pi\in\Pi(\rho,\theta)} \left\{\gamma \mathrm{KL}\left(\pi, \exp\left(-\gamma^{-1}C\right)\right)\right\}.$$

where $\Pi(\rho,\theta) \stackrel{\text{def}}{=} \left\{\pi \in \mathbb{R}_+^{n\times m} : \pi \mathbb{1}_m = \rho, \pi^T \mathbb{1}_n = \theta\right\}$

Entropically Regularized Wasserstein Inverse Problems

Simplest case: discrete measures ρ and θ with ground cost matrix $C \in \mathbb{R}_+^{n\times m}$.

Entropically regularized Wasserstein distance

$$W_\gamma(\rho,\theta) = \inf_{\pi\in\Pi(\rho,\theta)} \left\{\gamma \mathrm{KL}\left(\pi, \exp\left(-\gamma^{-1}C\right)\right)\right\}.$$

where $\Pi(\rho,\theta) \stackrel{\text{def}}{=} \left\{\pi \in \mathbb{R}_+^{n\times m} : \pi \mathbb{1}_m = \rho, \pi^T \mathbb{1}_n = \theta\right\}$

Inverse problem

$$\min_{\rho\in\Delta^n} \underbrace{\inf_{\pi\in\Pi(F\rho,\theta)} \left\{\gamma \mathrm{KL}\left(\pi, \exp\left(-\gamma^{-1}C\right)\right)\right\}}_{W_\gamma(F\rho,\theta)} + J\circ A(\rho),$$

where $J \in \Gamma_0(\mathbb{R}^p)$, $F : \Delta^n \to \Delta^m$ is linear, and $A \in \mathbb{R}^{n\times p}$.

Primal-Dual Splitting for OT

Since $W_\gamma(F\rho, \theta)$ is itself a minimization problem, we introduce a dual variable τ and use Lagrangian duality to have

$$\min_{\rho \in \Delta^n} \max_{\tau \in \mathbb{R}^m} \langle \tau, F\rho \rangle - \gamma \sum_{j=1}^{m} \theta_j \log \left(\sum_{i=1}^{m} \exp \left(\frac{\tau_i - C_{i,j}}{\gamma} \right) \right) + J \circ A(\rho)$$

Primal-Dual Splitting for OT

Since $W_\gamma(F\rho, \theta)$ is itself a minimization problem, we introduce a dual variable τ and use Lagrangian duality to have

$$\min_{\rho \in \Delta^n} \max_{\tau \in \mathbb{R}^m} \langle \tau, F\rho \rangle - \gamma \sum_{j=1}^{m} \theta_j \log \left(\sum_{i=1}^{m} \exp\left(\frac{\tau_i - C_{i,j}}{\gamma} \right) \right) + J \circ A(\rho)$$

We can further dualize to split $J \circ A$,

$$\min_{\rho \in \Delta^n} \max_{\substack{\tau \in \mathbb{R}^m \\ \zeta \in \mathbb{R}^p}} \left\langle \begin{pmatrix} \tau \\ \zeta \end{pmatrix}, \begin{pmatrix} F\rho \\ A\rho \end{pmatrix} \right\rangle - \gamma \sum_{j=1}^{m} \theta_j \log \left(\sum_{i=1}^{m} \exp\left(\frac{\tau_i - C_{i,j}}{\gamma} \right) \right) - J^*(\zeta)$$

and now we can apply SBPD.

Splitting the Inverse Problem

Inverse problem - primal-dual formulation

$$\min_{\rho\in\Delta^n} \max_{\substack{\tau\in\mathbb{R}^m \\ \zeta\in\mathbb{R}^p}} \left\langle \begin{pmatrix} \tau \\ \zeta \end{pmatrix}, \begin{pmatrix} F\rho \\ A\rho \end{pmatrix} \right\rangle - \gamma \sum_{j=1}^{m} \theta_j \log\left(\sum_{i=1}^{m} \exp\left(\frac{\tau_i - C_{i,j}}{\gamma} \right) \right) - J^*(\zeta)$$

Splitting the Inverse Problem

Inverse problem - primal-dual formulation

$$\min_{\rho\in\Delta^n} \max_{\substack{\tau\in\mathbb{R}^m\\ \zeta\in\mathbb{R}^p}} \left\langle \begin{pmatrix}\tau\\ \zeta\end{pmatrix}, \begin{pmatrix}F\rho\\ A\rho\end{pmatrix} \right\rangle - \gamma \sum_{j=1}^m \theta_j \log\left(\sum_{i=1}^m \exp\left(\frac{\tau_i - C_{i,j}}{\gamma}\right)\right) - J^*(\zeta)$$

Apply SBPD with

$$f(\rho) = 0, \quad g(\rho) = \iota_{\{1\}}\left(\rho^T \mathbb{1}_n\right), \quad \mathcal{C}_p = \mathbb{R}^n_+,$$

$$T(\rho) = \begin{pmatrix}F\rho\\ A\rho\end{pmatrix}, \quad h^*(\mu) = h^*(\tau) = \gamma \sum_{j=1}^m \theta_j \log\left(\sum_{i=1}^m \exp\frac{\tau_i - C_{i,j}}{\gamma}\right),$$

$$\ell^*(\mu) = \ell^*(\zeta) = J^*(\zeta), \quad \text{and} \quad \mathcal{C}_d = \mathbb{R}^{m+p}.$$

Choosing ϕ_p and ϕ_d

Primal entropy ϕ_p

- $\mathcal{C}_p = \mathbb{R}^n_+$

$$\phi_p(\rho) = \sum_{i=1}^{n} \rho_i \log(\rho_i).$$

- $\mathrm{prox}_{\lambda_k g}^{D_{\phi_p}}$ is same as in trend filtering (consider 1 row).

Choosing ϕ_p and ϕ_d

Primal entropy ϕ_p

- $\mathcal{C}_p = \mathbb{R}^n_+$

$$\phi_p(\rho) = \sum_{i=1}^{n} \rho_i \log(\rho_i).$$

- $\operatorname{prox}^{D_{\phi_p}}_{\lambda_k g}$ is same as in trend filtering (consider 1 row).

Dual entropy ϕ_d

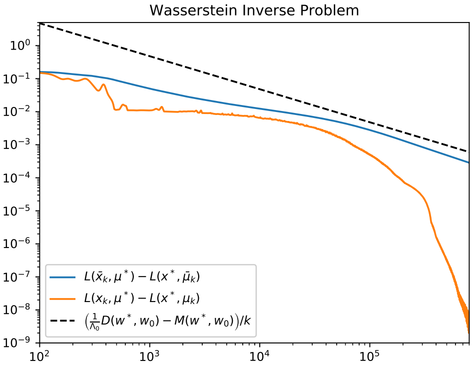
- $\mathcal{C}_d = \mathbb{R}^{m+p}$ (trivial constraint)

$$\phi_d(\mu) = \frac{1}{2}\|\mu\|_2^2$$

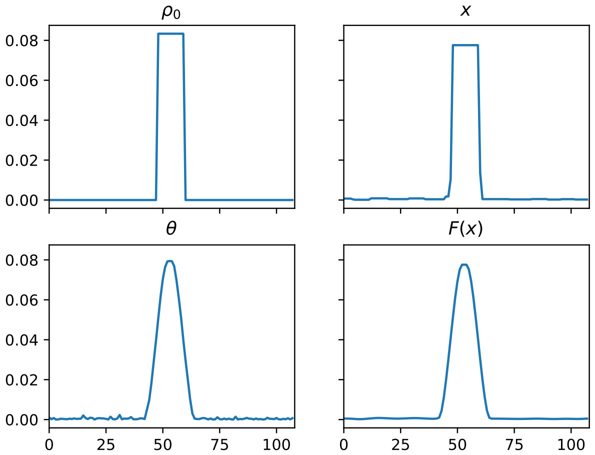
- Must show that h^* is Lipschitz-smooth (straightforward).

An Example Problem

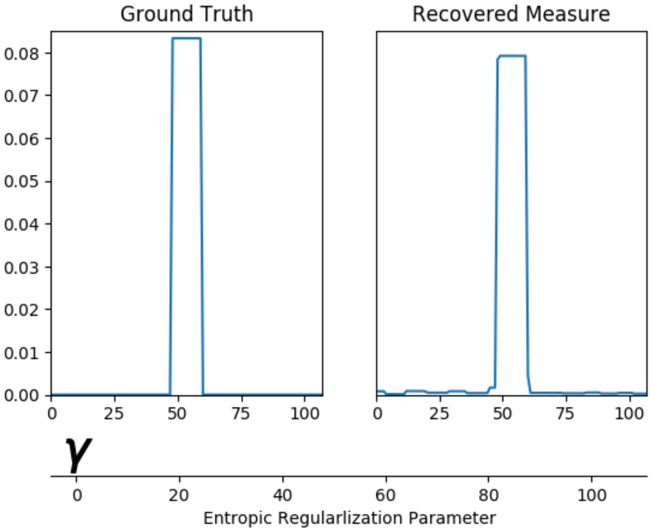
- $n = 108$,
- $C_{i,j} = \frac{1}{2} \|i - j\|_2^2$,
- F - convolution operator (bump function),
- $J \circ A = \|\cdot\|_1 \circ \nabla$.



Results - Recovered Probability Measure



Results - Different Values of γ - Entropic Regularization Parameter



Let's Recap

Contributions

- No Lipschitz-smoothness assumptions: $\nabla \mathrm{KL}$ vs $\mathrm{prox}_{\mathrm{KL}}$.

Let's Recap

Contributions

- No Lipschitz-smoothness assumptions: $\nabla\mathrm{KL}$ vs $\mathrm{prox}_{\mathrm{KL}}$.
- Improved constants: $\|x - x_0\|_2^2$ vs $D_{\phi_p}(x, x_0)$.

Let's Recap

Contributions

- No Lipschitz-smoothness assumptions: $\nabla \mathrm{KL}$ vs $\mathrm{prox}_{\mathrm{KL}}$.
- Improved constants: $\|x - x_0\|_2^2$ vs $D_{\phi_p}(x, x_0)$.
- Improved complexities: sorting vs softmax.

Let's Recap

Contributions

- No Lipschitz-smoothness assumptions: $\nabla \mathrm{KL}$ vs $\mathrm{prox}_{\mathrm{KL}}$.
- Improved constants: $\|x - x_0\|_2^2$ vs $D_{\phi_p}(x, x_0)$.
- Improved complexities: sorting vs softmax.
- Stochasticity permitted.

Let's Recap

Contributions

- No Lipschitz-smoothness assumptions: $\nabla \mathrm{KL}$ vs $\mathrm{prox}_{\mathrm{KL}}$.
- Improved constants: $\|x - x_0\|_2^2$ vs $D_{\phi_p}(x, x_0)$.
- Improved complexities: sorting vs softmax.
- Stochasticity permitted.
- Infinite-dimensional problems (Reflexive Banach spaces) permitted.

Note

Code (NumPy) is available on github soon.

Thanks for Listening

Thanks for listening.

Full paper available on arxiv: https://arxiv.org/abs/ 2112.11928

"A Stochastic Bregman Primal-Dual Splitting Algorithm for Composite Optimization" - Antonio Silveti-Falls, Cesare Molinari, Jalal Fadili.