First-Order Noneuclidean Splitting Methods for Large-Scale Optimization: Deterministic and Stochastic Algorithms

Antonio Silveti-Falls
Advised by Jalal Fadili and Gabriel Peyré

Theme

Solving structured convex optimization problems:

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

Question

How to take advantage of properties of the individual terms?

GREYC

Theme

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 - $\nabla f(x)$.

Theme

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 - $\operatorname{prox}_g(x) \stackrel{\text{def}}{=} \operatorname{argmin}_u \left\{ g(u) + \frac{1}{2}\|x-u\|_2^2 \right\}$.

Theme

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

GREYC

Theme

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 $\mathcal{C}$ - $P_{\mathcal{C}}(x) \stackrel{\text{def}}{=} \operatorname*{argmin}_{u \in \mathcal{C}} \|x - u\|_2^2$.
- Linear minimization oracle on $\mathcal{C}$ - $\operatorname{lmo}_{\mathcal{C}}(x) \stackrel{\text{def}}{=} \operatorname*{Argmin}_{u \in \mathcal{C}} \langle u, x \rangle$.

GREYC

Theme

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?

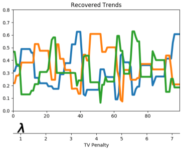
- ~~Lipschitz~~-smoothness - $\nabla f(x)$.
- prox-friendliness - $\mathrm{prox}_g(x) \overset{\text{def}}{=} \operatorname{argmin}_u \left\{ g(u) + \frac{1}{2}\|x-u\|_2^2 \right\}$.
- Projection onto $\mathcal{C}$ - $P_{\mathcal{C}}(x) \overset{\text{def}}{=} \operatorname{argmin}_{u \in \mathcal{C}} \|x-u\|_2^2$.
- Linear minimization oracle on $\mathcal{C}$ - $\mathrm{lmo}_{\mathcal{C}}(x) \overset{\text{def}}{=} \operatorname{Argmin}_{u \in \mathcal{C}} \langle u, x \rangle$.

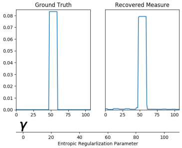
Changing the geometry?

GREYC

Trend Filtering

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



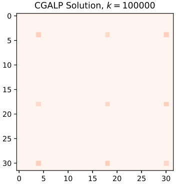


Entropically Regularized Wasserstein Inverse Problems

$$\min_{\substack{\rho \in \mathbb{R}_+^{n} \\ \rho\mathbb{1}_n = 1}} W_\gamma\left(F\rho, \theta\right) + J \circ A\left(\rho\right)$$

Robust Low Rank Sparse Matrix Completion

$$\min_{\substack{X \in \mathbb{R}^{N \times N} \\ \|X\|_* \leq \delta_1 \\ \|X\|_1 \leq \delta_2}} \left\| \Omega X - y \right\|_1$$



Trend Filtering - Notation

The Kullback-Leibler divergence

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

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

GREYC

Trend Filtering - Notation

The Kullback-Leibler divergence

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

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

The row gradient

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

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

GREYC

Trend Filtering - A Closer Look

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

Trend filtering

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

Contributions Part I - Bregman Primal-Dual Splitting

Model problem

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

Contributions

Contributions Part I - Bregman Primal-Dual Splitting

Model problem

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

Contributions

- Reflexive Banach spaces $\mathcal{X}_p$ and $\mathcal{X}_d$.

Contributions Part I - Bregman Primal-Dual Splitting

Model problem

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

Contributions

- Reflexive Banach spaces $\mathcal{X}_p$ and $\mathcal{X}_d$.
- No Lipschitz-smoothness assumption - no problem.

GREYC

Contributions Part I - Bregman Primal-Dual Splitting

Model problem

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

Contributions

- Reflexive Banach spaces $\mathcal{X}_p$ and $\mathcal{X}_d$.
- No Lipschitz-smoothness assumption - no problem.
- Adapted geometry - better constants and complexities.

GREYC

Contributions Part I - Bregman Primal-Dual Splitting

Model problem

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

Contributions

- Reflexive Banach spaces $\mathcal{X}_p$ and $\mathcal{X}_d$.
- No Lipschitz-smoothness assumption - no problem.
- Adapted geometry - better constants and complexities.
- Pointwise and ergodic convergence results with ergodic rate.

GREYC

Contributions Part I - Bregman Primal-Dual Splitting

Model problem

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

Contributions

- Reflexive Banach spaces $\mathcal{X}_p$ and $\mathcal{X}_d$.
- No Lipschitz-smoothness assumption - no problem.
- Adapted geometry - better constants and complexities.
- Pointwise and ergodic convergence results with ergodic rate.

Related work

[Chambolle et al. 2011], [Chambolle et al, 2016], [Nguyen, 2017]

GREYC

Matrix Completion

Let $\Omega : \mathbb{R}^{N \times N} \rightarrow \mathbb{R}^p$ a masking operator, $y \in \mathbb{R}^p$ a vector of observed entries.

Robust low rank sparse matrix completion

$$\min_{\substack{X \in \mathbb{R}^{N \times N} \\ \|X\|_* \leq \delta_1 \\ \|X\|_1 \leq \delta_2}} \underbrace{\|\Omega X - y\|_1}_{g \circ \Omega(X)}$$

GREYC

Contributions Part II - Generalized Conditional Gradient with Augmented Lagrangian and Proximal step

Model problem

$$\min_{\substack{x \in \mathcal{H} \\ Ax = b}} f(x) + g(Tx) + \iota_{\mathcal{D}}(x)$$

Contributions

Contributions Part II - Generalized Conditional Gradient with Augmented Lagrangian and Proximal step

Model problem

$$\min_{\substack{x \in \mathcal{H} \\ Ax = b}} f(x) + g(Tx) + \iota_{\mathcal{D}}(x)$$

Contributions

- Hilbert space $\mathcal{H}$.

Contributions Part II - Generalized Conditional Gradient with Augmented Lagrangian and Proximal step

Model problem

$$\min_{\substack{x \in \mathcal{H} \\ Ax=b}} f(x) + g(Tx) + \iota_{\mathcal{D}}(x)$$

Contributions

- Hilbert space $\mathcal{H}$.
- Relax assumptions - allow for nonsmooth functions.

GREYC

Contributions Part II - Generalized Conditional Gradient with Augmented Lagrangian and Proximal step

Model problem

$$\min_{\substack{x \in \mathcal{H} \\ Ax=b}} f(x) + g(Tx) + \iota_{\mathcal{D}}(x)$$

Contributions

- Hilbert space $\mathcal{H}$.
- Relax assumptions - allow for nonsmooth functions.
- Affine constraint handles $\bigcap_i \mathcal{D}_i$.

GREYC

Contributions Part II - Generalized Conditional Gradient with Augmented Lagrangian and Proximal step

Model problem

$$\min_{\substack{x \in \mathcal{H} \\ Ax = b}} f(x) + g(Tx) + \iota_{\mathcal{D}}(x)$$

Contributions

- Hilbert space $\mathcal{H}$.
- Relax assumptions - allow for nonsmooth functions.
- Affine constraint handles $\bigcap_i \mathcal{D}_i$.
- Pointwise and ergodic convergence results with rates.

GREYC

Contributions Part II - Generalized Conditional Gradient with Augmented Lagrangian and Proximal step

Model problem

$$\min_{\substack{x\in\mathcal{H}\\ Ax=b}} f(x) + g(Tx) + \iota_{\mathcal{D}}(x)$$

Contributions

- Hilbert space $\mathcal{H}$.
- Relax assumptions - allow for nonsmooth functions.
- Affine constraint handles $\bigcap_i \mathcal{D}_i$.
- Pointwise and ergodic convergence results with rates.

Related work

[Yurtsever et al. 2018], [Gidel et al. 2018], [Argyriou et al. 2014]

Bregman Primal-Dual Splitting (Chapter 5 of thesis)

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

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

- $\mathcal{C}_p$ 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;

GREYC

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;
- T is a bounded linear operator.

GREYC

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 $u, v \in \mathcal{C}$,

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

GREYC

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

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

GREYC

D-prox Operators

Euclidean prox operator

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

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

D-prox Operators

Euclidean prox operator

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

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

D-prox operator

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

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

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 $u, v \in \mathcal{X}$,

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

(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 $\frac{1}{2}\|\cdot\|_2^2$:

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

GREYC

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

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

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

$k \leftarrow k + 1$

until *convergence*;

Output: x_k, μ_k.

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.

GREYC

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,

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

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

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

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).
- 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.

GREYC

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

Matching the Geometries

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

GREYC

Matching the Geometries

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

Ergodic Convergence Results

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.*

GREYC

Pointwise Convergence Results

Theorem

Under [stricter assumptions], the sequence of iterates $(x_k, \mu_k)_{k \in \mathbb{N}}$ converges weakly to a solution of the primal-dual problem

Trend Filtering

Trend filtering problem - primal-dual formulation

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

Trend Filtering

Trend filtering problem - primal-dual formulation

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

Apply SBPD with

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

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

GREYC

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 $\mathrm{prox}_{\lambda_k g}^{D_{\phi_p}}(X)$.

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.

GREYC

New Geometry of ϕ_p

Relative smoothness

For each $i \in \{1, \dots, n\}$, 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 $L\phi_p - f$ is convex on $\mathrm{int}(\mathcal{C}_p)$.

GREYC

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 $L\phi_p - f$ is convex on $\operatorname{int}(\mathcal{C}_p)$.

D-prox under ϕ_p

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

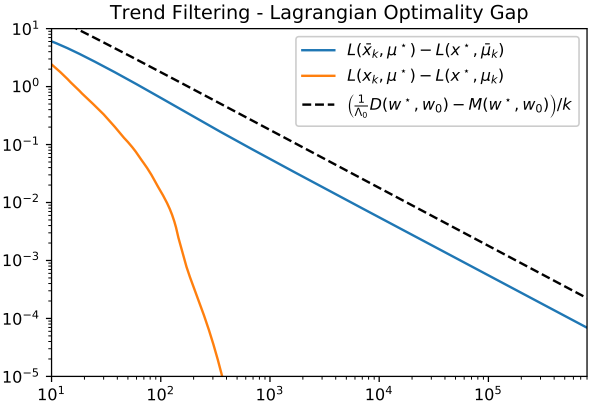
$$\operatorname{prox}_{\lambda_k g}^{D_{\phi_p}}(X) = \underset{\substack{U \in \mathbb{R}_+^{n \times m} \\ U^T \mathbb{1}_m = \mathbb{1}_n}}{\operatorname{argmin}} \left\{ D_{\phi_p}(U, X) \right\} = \left(\frac{\exp(X_{i,j})}{\sum_{q=1}^{m} \exp(X_{i,q})} \right)_{i,j}$$

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

GREYC

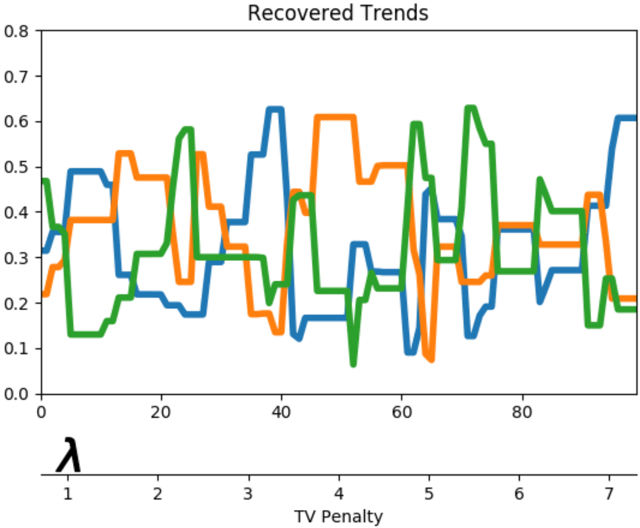
Results - Convergence

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



GREYC

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

Inverse problem

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

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

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

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

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

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

GREYC

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

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

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

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

Dual entropy ϕ_d

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

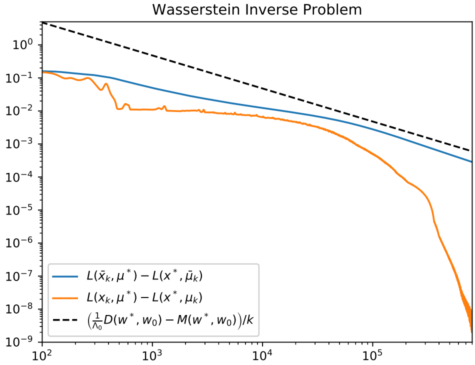
$$\phi_d(\mu) = \frac{1}{2} \|\mu\|_2^2$$

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

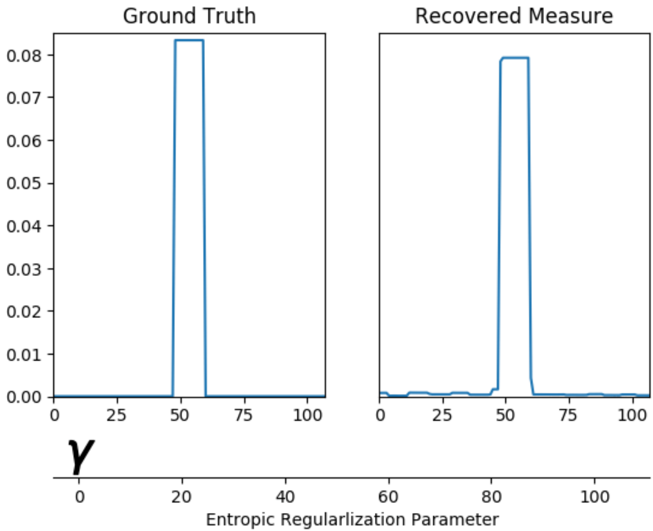
GREYC

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 - Different Values of γ - Entropic Regularization Parameter

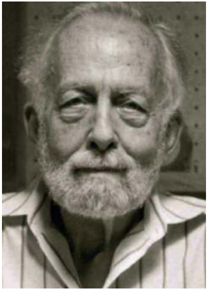


Part II

Generalized Conditional Gradient with Augmented Lagrangian and Proximal step (Chapter 3 of thesis, [Silveti et al., 2020])

GREYC

- 1956 Marguerite Frank and Philip Wolfe: *An algorithm for quadratic programming.*

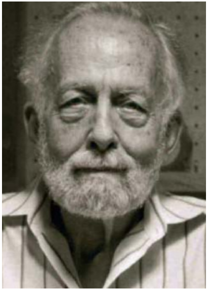


History and Motivation

- 1956 Marguerite Frank and Philip Wolfe: *An algorithm for quadratic programming.*
- Considered the following problem:

$$\min_{x \in \mathcal{D} \subset \mathbb{R}^n} f(x)$$

- $\mathcal{D}$ is a convex, compact set and f is Lipschitz-smooth.



The Frank-Wolfe Algorithm

Algorithm: Frank-Wolfe (Conditional Gradient)

Input: $x_0 \in \mathcal{D}$.

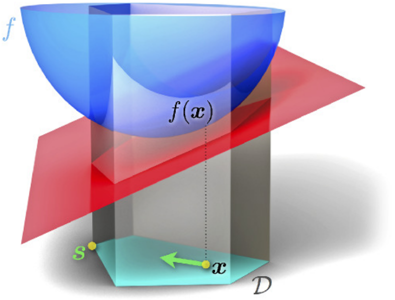
$k = 0$

repeat

$$\gamma_k = \frac{1}{k+2}$$
$$s_k \in \operatorname*{Argmin}_{s \in \mathcal{D}} \langle \nabla f(x_k), s \rangle$$
$$x_{k+1} = x_k - \gamma_k (x_k - s_k)$$
$$k \leftarrow k + 1$$

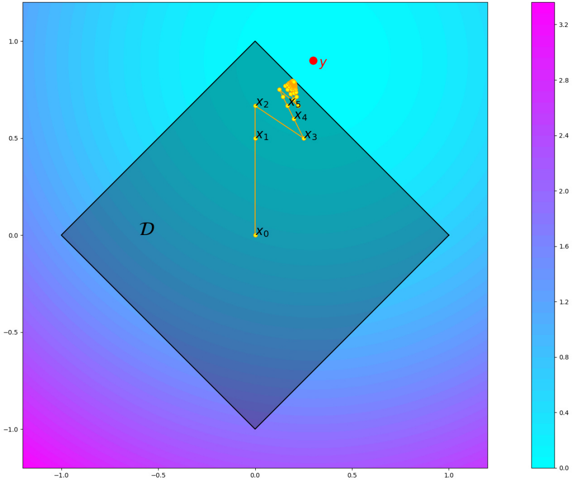
until *convergence*;

Output: x_{k+1}.



(Credit: Stephanie Stutz/Wikipedia)

Frank-Wolfe for Sparse Optimizaiton



$$\min_{\|x\|_1 \leq 1} \|x - y\|_p \,, \quad p > 1$$

GREYC

Advantages of Frank-Wolfe

Question

Why not just do projected gradient descent?

Advantages of Frank-Wolfe

Question

Why not just do projected gradient descent?

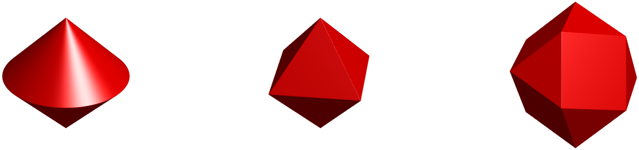
- The set $\mathcal{D}$ might not admit easy projections.
  - Nuclear norm $\|\cdot\|_*$ of a matrix (ℓ^1 norm on singular values).

Advantages of Frank-Wolfe

Question

Why not just do projected gradient descent?

- The set $\mathcal{D}$ might not admit easy projections.
  - Nuclear norm $\|\cdot\|_*$ of a matrix (ℓ^1 norm on singular values).
- The updates of Frank-Wolfe maintain structure.
  - Useful when $\mathcal{D}$ is *atomically generated*, i.e. $\mathcal{D} = \overline{\mathrm{conv}}\,(a_1, \ldots a_j)$.
  - Sparsity, low-rank, etc.



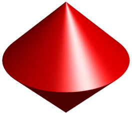
GREYC

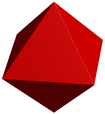
Advantages of Frank-Wolfe

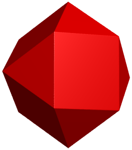
Question

Why not just do projected gradient descent?

- The set $\mathcal{D}$ might not admit easy projections.
  - Nuclear norm $\|\cdot\|_*$ of a matrix (ℓ^1 norm on singular values).
- The updates of Frank-Wolfe maintain structure.
  - Useful when $\mathcal{D}$ is *atomically generated*, i.e. $\mathcal{D} = \overline{\mathrm{conv}}\,(a_1, \ldots a_j)$.
  - Sparsity, low-rank, etc.
- The iterates are always feasible, i.e. $x_k \in \mathcal{D}$ for all $k \in \mathbb{N}$.







Limitations of Frank-Wolfe

- Lipschitz-smoothness can be a strong assumption.

Limitations of Frank-Wolfe

- Lipschitz-smoothness can be a strong assumption.
- Not able to handle nonsmooth problems.

Limitations of Frank-Wolfe

- Lipschitz-smoothness can be a strong assumption.
- Not able to handle nonsmooth problems.
- Affine constraints are not handled in a straightforward way if the intersection of the affine constraint and the set $\mathcal{D}$ is not simple.

Limitations of Frank-Wolfe

- Lipschitz-smoothness can be a strong assumption.
- Not able to handle nonsmooth problems.
- Affine constraints are not handled in a straightforward way if the intersection of the affine constraint and the set $\mathcal{D}$ is not simple.
- Unable to handle intersection $\bigcap_i \mathcal{D}_i$ in a separable way.

Modern Problem

Classical problem ($\mathbb{R}^n$):

$$\min_{x \in \mathcal{D}} f(x)$$

- f is Lipschitz-smooth.
- $\mathcal{D} \subset \mathbb{R}^n$ is convex, compact.

Modern Problem

Classical problem ($\mathbb{R}^n$):

$$\min_{x \in \mathcal{D}} f(x)$$

- f is Lipschitz-smooth.
- $\mathcal{D} \subset \mathbb{R}^n$ is convex, compact.

Modern problem (Hilbert space):

$$\min_{Ax=b} f(x) + (g \circ T)(x) + \iota_{\mathcal{D}}(x)$$

Modern Problem

Classical problem ($\mathbb{R}^n$):

$$\min_{x \in \mathcal{D}} f(x)$$

- f is Lipschitz-smooth.
- $\mathcal{D} \subset \mathbb{R}^n$ is convex, compact.

Modern problem (Hilbert space):

$$\min_{Ax=b} f(x) + (g \circ T)(x) + \iota_{\mathcal{D}}(x)$$

- f satisfies a *relative* smoothness condition.

Modern Problem

Classical problem ($\mathbb{R}^n$):

$$\min_{x \in \mathcal{D}} f(x)$$

- f is Lipschitz-smooth.
- $\mathcal{D} \subset \mathbb{R}^n$ is convex, compact.

Modern problem (Hilbert space):

$$\min_{Ax=b} f(x) + (g \circ T)(x) + \iota_{\mathcal{D}}(x)$$

- f satisfies a *relative* smoothness condition.
- prox_g is accessible.

Modern Problem

Classical problem ($\mathbb{R}^n$):

$$\min_{x \in \mathcal{D}} f(x)$$

- f is Lipschitz-smooth.
- $\mathcal{D} \subset \mathbb{R}^n$ is convex, compact.

Modern problem (Hilbert space):

$$\min_{Ax=b} f(x) + (g \circ T)(x) + \iota_{\mathcal{D}}(x)$$

- f satisfies a *relative* smoothness condition.
- prox_g is accessible.
- T and A are bounded linear operators.

GREYC

The CGALP Algorithm

Algorithm: Conditional Gradient with Augmented Lagrangian and Proximal-step (CGALP)

Input: $x_0 \in \mathcal{D} = \mathrm{dom}\,(h)$; $\mu_0 \in \mathrm{ran}(A)$; $(\gamma_k)_{k\in\mathbb{N}}$, $(\beta_k)_{k\in\mathbb{N}}$, $(\theta_k)_{k\in\mathbb{N}}, (\rho_k)_{k\in\mathbb{N}} \in \ell_+$.

$k = 0$.

repeat

until *convergence*;

Output: x_{k+1}.

The CGALP Algorithm

Algorithm: Conditional Gradient with Augmented Lagrangian and Proximal-step (CGALP)

Input: $x_0 \in \mathcal{D} = \mathrm{dom}\,(h)$; $\mu_0 \in \mathrm{ran}(A)$; $(\gamma_k)_{k\in\mathbb{N}}$, $(\beta_k)_{k\in\mathbb{N}}$, $(\theta_k)_{k\in\mathbb{N}}\,, (\rho_k)_{k\in\mathbb{N}} \in \ell_+$.

$k = 0$.

repeat

$$y_k = \mathrm{prox}_{\beta_k g}\left(Tx_k\right)$$

$$z_k = \nabla f(x_k) + T^*\left(Tx_k - y_k\right)/\beta_k + A^*\mu_k + \rho_k A^*\left(Ax_k - b\right)$$

until *convergence*;

Output: x_{k+1}.

The CGALP Algorithm

Algorithm: Conditional Gradient with Augmented Lagrangian and Proximal-step (CGALP)

Input: $x_0 \in \mathcal{D} = \mathrm{dom}\,(h)$; $\mu_0 \in \mathrm{ran}(A)$; $(\gamma_k)_{k\in\mathbb{N}}$, $(\beta_k)_{k\in\mathbb{N}}$, $(\theta_k)_{k\in\mathbb{N}}, (\rho_k)_{k\in\mathbb{N}} \in \ell_+$.

$k = 0$.

repeat

$$y_k = \mathrm{prox}_{\beta_k g}\,(Tx_k)$$

$$z_k = \nabla f(x_k) + T^*\,(Tx_k - y_k)\,/\beta_k + A^*\mu_k + \rho_k A^*\,(Ax_k - b)$$

$$s_k \in \mathrm{Argmin}_{s\in\mathcal{D}}\,\langle z_k, s\rangle$$

until *convergence*;

Output: x_{k+1}.

The CGALP Algorithm

Algorithm: Conditional Gradient with Augmented Lagrangian and Proximal-step (CGALP)

Input: $x_0 \in \mathcal{D} = \mathrm{dom}\,(h)$; $\mu_0 \in \mathrm{ran}(A)$; $(\gamma_k)_{k\in\mathbb{N}}$, $(\beta_k)_{k\in\mathbb{N}}$, $(\theta_k)_{k\in\mathbb{N}}, (\rho_k)_{k\in\mathbb{N}} \in \ell_+$.

$k = 0$.

repeat

$$y_k = \mathrm{prox}_{\beta_k g}(Tx_k)$$

$$z_k = \nabla f(x_k) + T^*(Tx_k - y_k)/\beta_k + A^*\mu_k + \rho_k A^*(Ax_k - b)$$

$$s_k \in \mathrm{Argmin}_{s\in\mathcal{D}} \langle z_k, s\rangle$$

$$x_{k+1} = x_k - \gamma_k(x_k - s_k)$$

until *convergence*;

Output: x_{k+1}.

Algorithm: Conditional Gradient with Augmented Lagrangian and Proximal-step (CGALP)

Input: $x_0 \in \mathcal{D} = \mathrm{dom}\,(h)$; $\mu_0 \in \mathrm{ran}(A)$; $(\gamma_k)_{k\in\mathbb{N}}$, $(\beta_k)_{k\in\mathbb{N}}$, $(\theta_k)_{k\in\mathbb{N}}, (\rho_k)_{k\in\mathbb{N}} \in \ell_+$.

$k = 0$.

repeat

$$y_k = \mathrm{prox}_{\beta_k g}\,(Tx_k)$$

$$z_k = \nabla f(x_k) + T^*\,(Tx_k - y_k)\,/\beta_k + A^*\mu_k + \rho_k A^*\,(Ax_k - b)$$

$$s_k \in \mathrm{Argmin}_{s\in\mathcal{D}}\,\langle z_k, s\rangle$$

$$x_{k+1} = x_k - \gamma_k\,(x_k - s_k)$$

$$\mu_{k+1} = \mu_k + \theta_k\,(Ax_{k+1} - b)$$

$$k \leftarrow k+1$$

until *convergence*;

Output: x_{k+1}.

The ICGALP Algorithm

Algorithm: Inexact Conditional Gradient with Augmented Lagrangian and Proximal-step (ICGALP)

Input: $x_0 \in \mathcal{D} = \mathrm{dom}\,(h)$; $\mu_0 \in \mathrm{ran}(A)$; $(\gamma_k)_{k\in\mathbb{N}}$, $(\beta_k)_{k\in\mathbb{N}}$, $(\theta_k)_{k\in\mathbb{N}}, (\rho_k)_{k\in\mathbb{N}} \in \ell_+$.

$k = 0$.

repeat

$$y_k = \mathrm{prox}_{\beta_k g}(Tx_k)$$

$$z_k = \nabla f(x_k) + T^*(Tx_k - y_k)/\beta_k + A^*\mu_k + \rho_k A^*(Ax_k - b) + \lambda_k^z$$

$$s_k \in \mathrm{Argmin}^{\lambda_k^s}_{s\in\mathcal{D}} \langle z_k, s\rangle$$

$$x_{k+1} = x_k - \gamma_k(x_k - s_k)$$

$$\mu_{k+1} = \mu_k + \theta_k(Ax_{k+1} - b)$$

$$k \leftarrow k+1$$

until *convergence*;

Output: x_{k+1}.

Asymptotic Feasibility

Theorem

Let $(x_k)_{k \in \mathbb{N}}$ be the sequence of primal iterates generated by CGALP. Then,

- *Ax_k converges strongly to b, i.e.,*

$$\lim_{k \to \infty} \|Ax_k - b\| = 0$$

Pointwise Rates

Let $\Gamma_k = \sum_{i=0}^{k} \gamma_i$; usually $\Gamma_k \approx O\left((k+2)^{1/3}\right)$.

Asymptotic Feasibility

Pointwise rate:

$$\inf_{0 \leq i \leq k} \|Ax_i - b\|^2 = O\left(\frac{1}{\Gamma_k}\right) \approx O\left(\frac{1}{(k+2)^{1/3}}\right)$$

Furthermore, $\exists$ *a subsequence* $\left(x_{k_j}\right)_{j \in \mathbb{N}}$ *such that*

$$\|Ax_{k_j} - b\|^2 \leq \frac{1}{\Gamma_{k_j}}.$$

GREYC

Ergodic Rates

Asymptotic Feasibility

Ergodic rate: let $\bar{x}_k = \sum_{i=0}^{k} \gamma_i x_i / \Gamma_k$. *Then*

$$\|A\bar{x}_k - b\|^2 = O\left(\frac{1}{\Gamma_k}\right) \approx O\left(\frac{1}{(k+2)^{1/3}}\right)$$

GREYC

Convergence to Optimality

Theorem

Let $(x^\star, \mu^\star)$ be a saddle-point. Under [assumptions], it holds

GREYC

Convergence to Optimality

Theorem

Let $(x^\star, \mu^\star)$ be a saddle-point. Under [assumptions], it holds

- *Convergence of the Lagrangian:*

$$\lim_{k\to\infty} \mathcal{L}(x_k, \mu^\star) = \mathcal{L}(x^\star, \mu^\star)$$

Convergence to Optimality

Theorem

Let $(x^\star, \mu^\star)$ be a saddle-point. Under [assumptions], it holds

- *Convergence of the Lagrangian:*

$$\lim_{k \to \infty} \mathcal{L}(x_k, \mu^\star) = \mathcal{L}(x^\star, \mu^\star)$$

- *Every weak cluster point $\tilde{x}$ of $(x_k)_{k \in \mathbb{N}}$ is a solution of the primal problem, and $(\mu_k)_{k \in \mathbb{N}}$ converges strongly to $\tilde{\mu}$ a solution of the dual problem, i.e., $(\tilde{x}, \tilde{\mu})$ is a saddle point of $\mathcal{L}$.*

GREYC

Lagrangian Convergence Rate

Optimality

Pointwise rate:

$$\inf_{0\leq i\leq k} \mathcal{L}\left(x_i, \mu^\star\right) - \mathcal{L}\left(x^\star, \mu^\star\right) = O\left(\frac{1}{\Gamma_k}\right) \approx O\left(\frac{1}{(k+2)^{1/3}}\right)$$

Furthermore, $\exists$ *a subsequence* $\left(x_{k_j}\right)_{j\in\mathbb{N}}$ *such that*

$$\mathcal{L}\left(x_{k_j+1}, \mu^\star\right) - \mathcal{L}\left(x^\star, \mu^\star\right) \leq \frac{1}{\Gamma_{k_j}}$$

GREYC

Lagrangian Convergence Rate

Optimality

Ergodic rate: let $\bar{x}_k = \sum_{i=0}^{k} \gamma_i x_{i+1} / \Gamma_k$. *Then*

$$\mathcal{L}(\bar{x}_k, \mu^\star) - \mathcal{L}(x^\star, \mu^\star) = O\left(\frac{1}{\Gamma_k}\right) \approx O\left(\frac{1}{(k+2)^{1/3}}\right)$$

Rates on the Objective

Recall the primal problem

$$\min_{Ax=b} f(x) + g(Tx) + \iota_{\mathcal{D}}(x)$$

Denote the primal objective

$$\Phi(x) \stackrel{\text{def}}{=} f(x) + g(Tx) + \iota_{\mathcal{D}}(x)$$

GREYC

Rates on the Objective

Recall the primal problem

$$\min_{Ax=b} f(x) + g(Tx) + \iota_{\mathcal{D}}(x)$$

Denote the primal objective

$$\Phi(x) \stackrel{\text{def}}{=} f(x) + g(Tx) + \iota_{\mathcal{D}}(x)$$

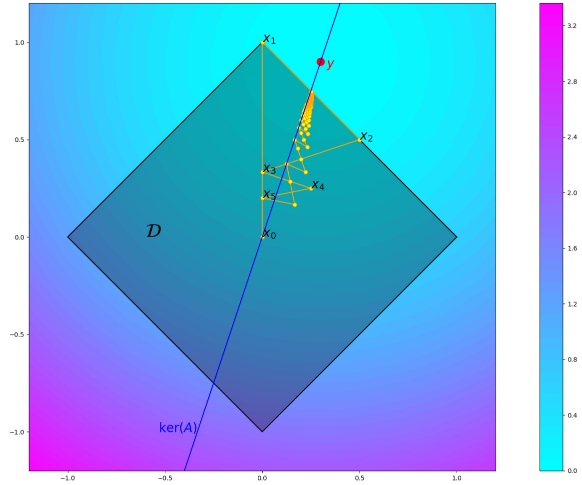
Optimality

We have the ergodic rate:

$$|\Phi(\bar{x}_k) - \Phi(x^\star)| = O\left(\frac{1}{\sqrt{\Gamma_k}}\right) \approx O\left(\frac{1}{(k+2)^{1/6}}\right)$$

GREYC

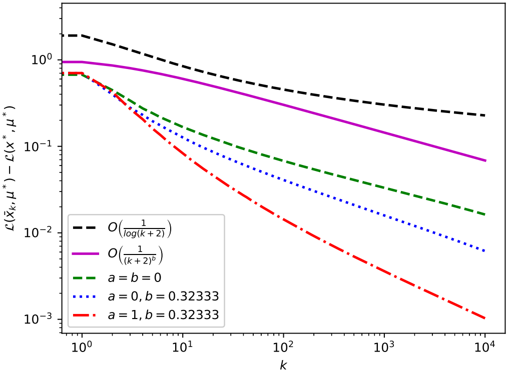
Simple Projection Problem



$$\min_{\substack{\|x\|_1 \leq 1 \\ Ax=0}} \|x - y\|_p \,, \quad p > 1$$

GREYC

Lagrangian Convergence Rate



Ergodic convergence profile for various step size choices,

$$\theta_k = \gamma_k = \frac{(\log(k+2))^a}{(k+1)^{1-b}}$$

GREYC

Matrix Completion Problem

Robust low rank sparse matrix completion problem

$$\min_{X \in \mathbb{R}^{N \times N}} \left\{ \|\Omega X - y\|_1 : \ \|X\|_* \leq \delta_1, \|X\|_1 \leq \delta_2 \right\}$$

GREYC

Matrix Completion Problem

Robust low rank sparse matrix completion problem

$$\min_{X \in \mathbb{R}^{N \times N}} \left\{ \|\Omega X - y\|_1 : \|X\|_* \leq \delta_1, \|X\|_1 \leq \delta_2 \right\}$$

Lift to a product space for CGALP :

$$\min_{\boldsymbol{X} \in \left(\mathbb{R}^{N \times N}\right)^2} \left\{ G\left(\Omega \boldsymbol{X}\right) + H(\boldsymbol{X}) : \Pi_{\mathcal{V}^\perp} \boldsymbol{X} = 0 \right\}$$

with

$$G\left(\Omega \boldsymbol{X}\right) = \frac{1}{2}\left(\left\|\Omega X^{(1)} - y\right\|_1 + \left\|\Omega X^{(2)} - y\right\|_1 \right)$$

and

$$H(\boldsymbol{X}) = \iota_{\mathbb{B}_*^{\delta_1}}\left(X^{(1)}\right) + \iota_{\mathbb{B}_1^{\delta_2}}\left(X^{(2)}\right)$$

Direction Finding Step (2 components)

Linear minimization oracle over $\|\cdot\|_*$ ball

$$S_k^{(1)} \in \underset{S^{(1)} \in \mathbb{B}_{\|\cdot\|_*}^{\delta_1}}{\operatorname{Argmin}} \langle Z_k^{(1)}, S^{(1)} \rangle \qquad \text{(Leading singular vector)}$$

GREYC

Direction Finding Step (2 components)

Linear minimization oracle over $\|\cdot\|_*$ ball

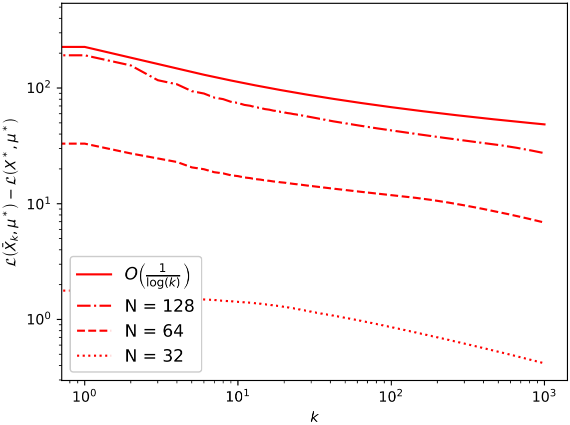
$$S_k^{(1)} \in \underset{S^{(1)} \in \mathbb{B}_{\|\cdot\|_*}^{\delta_1}}{\text{Argmin}} \langle Z_k^{(1)}, S^{(1)} \rangle \qquad \text{(Leading singular vector)}$$

Linear minimization oracle over $\|\cdot\|_1$ ball

$$S_k^{(2)} \in \underset{S^{(2)} \in \mathbb{B}_{\|\cdot\|_1}^{\delta_2}}{\text{Argmin}} \langle Z_k^{(2)}, S^{(2)} \rangle \qquad \text{(Largest entry in magnitude)}$$

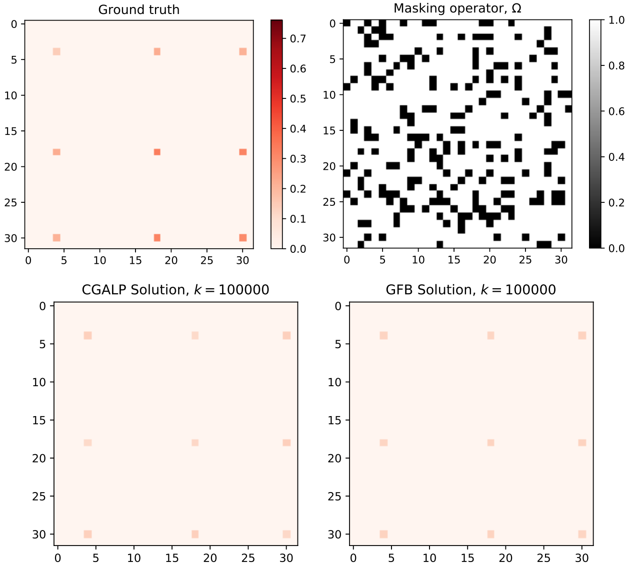
GREYC

CGALP Ergodic Convergence Rate



Ergodic convergence profiles for CGALP.

CGALP Recovered Matrix



Compared to Generalized Forward-Backward [Raguet et al. 2013]

Let's Recap

Part I - SBPD algorithm

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

Part II - ICGALP algorithm

GREYC

Let's Recap

Part I - SBPD algorithm

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

Part II - ICGALP algorithm

GREYC

Let's Recap

Part I - SBPD algorithm

- 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.

Part II - ICGALP algorithm

GREYC

Let's Recap

Part I - SBPD algorithm

- 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.

Part II - ICGALP algorithm

- Allow for nonsmooth functions.

GREYC

Let's Recap

Part I - SBPD algorithm

- 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.

Part II - ICGALP algorithm

- Allow for nonsmooth functions.
- Affine constraint for $\bigcap_i \mathcal{D}_i$.

GREYC

Let's Recap

Part I - SBPD algorithm

- 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.

Part II - ICGALP algorithm

- Allow for nonsmooth functions.
- Affine constraint for $\bigcap_i \mathcal{D}_i$.
- Hybridize proximal and conditional gradient methods: best of both worlds.

GREYC

Let's Recap

Part I - SBPD algorithm

- 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.

Part II - ICGALP algorithm

- Allow for nonsmooth functions.
- Affine constraint for $\bigcap_i \mathcal{D}_i$.
- Hybridize proximal and conditional gradient methods: best of both worlds.
- Avoid projecting: full SVD vs leading singular vector.

GREYC

Let's Recap

Part I - SBPD algorithm

- 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.

Part II - ICGALP algorithm

- Allow for nonsmooth functions.
- Affine constraint for $\bigcap_i \mathcal{D}_i$.
- Hybridize proximal and conditional gradient methods: best of both worlds.
- Avoid projecting: full SVD vs leading singular vector.

Note

Code (NumPy) is available on github.

Perspectives

Future work

- More general spaces: CGALP for Banach space, Riemannian CGALP, etc.

Perspectives

Future work

- More general spaces: CGALP for Banach space, Riemannian CGALP, etc.
- CGALP beyond bounded sets.

Perspectives

Future work

- More general spaces: CGALP for Banach space, Riemannian CGALP, etc.
- CGALP beyond bounded sets.
- Nonconvex settings.

GREYC

Perspectives

Future work

- More general spaces: CGALP for Banach space, Riemannian CGALP, etc.
- CGALP beyond bounded sets.
- Nonconvex settings.
- Acceleration under stricter assumptions.

Perspectives

Future work

- More general spaces: CGALP for Banach space, Riemannian CGALP, etc.
- CGALP beyond bounded sets.
- Nonconvex settings.
- Acceleration under stricter assumptions.
- Optimal complexity analysis/performance estimation problems.

Perspectives

Future work

- More general spaces: CGALP for Banach space, Riemannian CGALP, etc.
- CGALP beyond bounded sets.
- Nonconvex settings.
- Acceleration under stricter assumptions.
- Optimal complexity analysis/performance estimation problems.
- Fancier stochastic gradients.

GREYC

Perspectives

Future work

- More general spaces: CGALP for Banach space, Riemannian CGALP, etc.
- CGALP beyond bounded sets.
- Nonconvex settings.
- Acceleration under stricter assumptions.
- Optimal complexity analysis/performance estimation problems.
- Fancier stochastic gradients.

The end

Thanks for listening.

GREYC

Relative Strong Convexity

Recall that f is L_p relatively smooth with respect to ϕ_p if

$$D_f\left(x_1, x_2\right) \leq L_p D_{\phi_p}\left(x_1, x_2\right).$$

We can similarly define *relative strong convexity*,

$$D_f\left(x_1, x_2\right) \geq m_p D_{\phi_p}\left(x_1, x_2\right).$$

Relative Strong Convexity

Recall that f is L_p relatively smooth with respect to ϕ_p if

$$D_f(x_1, x_2) \leq L_p D_{\phi_p}(x_1, x_2).$$

We can similarly define *relative strong convexity*,

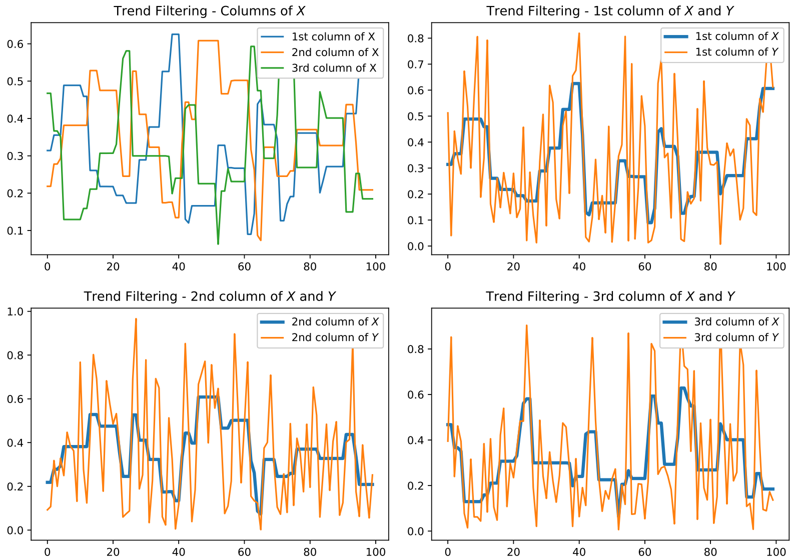
$$D_f(x_1, x_2) \geq m_p D_{\phi_p}(x_1, x_2).$$

Theorem

Assume additionally that $f + g$ is relatively strongly convex with respect to ϕ_p and ϕ_p is totally convex. Then $(x_k)_{k \in \mathbb{N}}$ converges strongly to the solution of the primal problem $x^\star$.

GREYC

Results - Recovered Trends



Moreau-Yosida Regularization

Given a closed, convex, proper function g, the Moreau envelope (Moreau-Yosida regularization) of g is,

$$g^{\beta}(x) = \min_{y} \left\{ g(y) + \frac{1}{2\beta} \|x - y\|^2 \right\}$$

GREYC

Moreau-Yosida Regularization

Given a closed, convex, proper function g, the Moreau envelope (Moreau-Yosida regularization) of g is,

$$g^{\beta}(x) = \min_{y} \left\{ g(y) + \frac{1}{2\beta} \|x - y\|^2 \right\}$$

- The Moreau envelope is always Lipschitz-smooth.

Moreau-Yosida Regularization

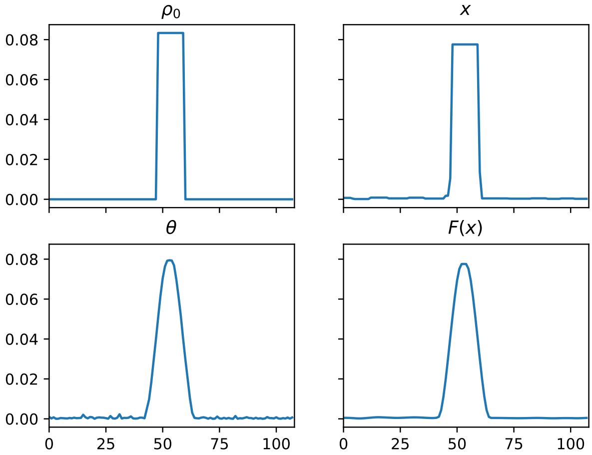
Given a closed, convex, proper function g, the Moreau envelope (Moreau-Yosida regularization) of g is,

$$g^{\beta}(x) = \min_{y} \left\{ g(y) + \frac{1}{2\beta} \|x - y\|^2 \right\}$$

- The Moreau envelope is always Lipschitz-smooth.
- Gradient is given by,

$$\nabla g^{\beta}(x) = \frac{x - \mathrm{prox}_{\beta g}(x)}{\beta}$$

Results - Recovered Probability Measure



Relative Smoothness Condition

Let $F : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ and $\zeta :]0,1] \to \mathbb{R}_+$. The pair $(f, \mathcal{D})$, where $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ and $\mathcal{D} \subset \mathrm{dom}(f)$, is said to be (F,ζ)-smooth if there exists an open set $\mathcal{D}_0$ such that $\mathcal{D} \subset \mathcal{D}_0 \subset \mathrm{int}\,(\mathrm{dom}\,(F))$ and

- F and f are differentiable on $\mathcal{D}_0$;
- $F - f$ is convex on $\mathcal{D}_0$;
- The following holds,

$$K_{(F,\zeta,\mathcal{D})} = \sup_{\substack{x,s\in\mathcal{D};\ \gamma\in]0,1] \\ z=x+\gamma(s-x)}} \frac{D_F(z,x)}{\zeta(\gamma)} < +\infty.$$

$K_{(F,\zeta,\mathcal{C})}$ measures the "curvature" of F on $\mathcal{D}$.