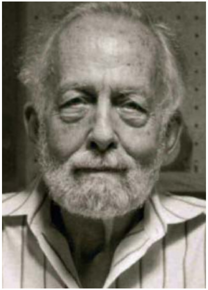
Generalized Conditional Gradient with Augmented Lagrangian for Composite Minimization

Antonio Silveti-Falls
(Joint work with Cesare Molinari and Jalal Fadili)
SPARS 2019

History and Motivation

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

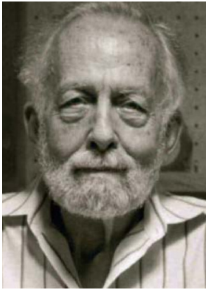


History and Motivation

- 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.



GREYC

The Frank-Wolfe Algorithm

Algorithm: Frank-Wolfe (Conditional Gradient)

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

$k = 0$

repeat

$\gamma_k = \frac{1}{k+2}$

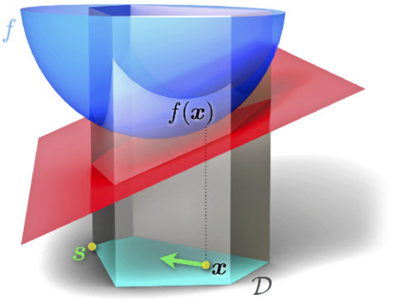
$s_k \in \underset{s \in \mathcal{D}}{\text{Argmin}} \left\langle \nabla f\left(x_k\right), s\right\rangle$

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

$k \leftarrow k + 1$

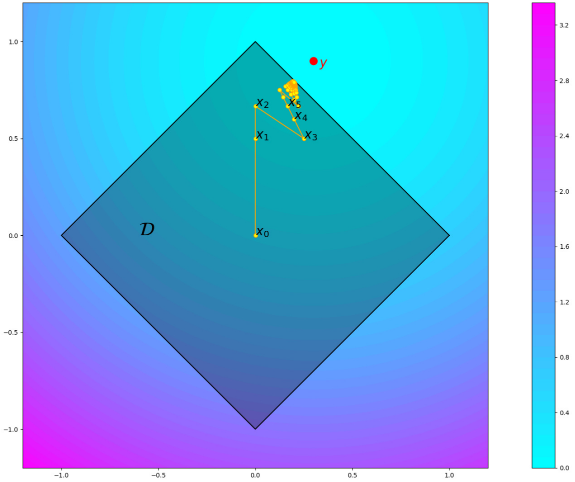
until *convergence*;

Output: x_{k+1}.



(Credit: Stephanie Stutz/Wikipedia)

Frank-Wolfe for sparse optimizaiton



$$\min_{\|x\|_1 \leq 1} \|x - y\|^2$$

Assumptions for Frank-Wolfe

2011 Martin Jaggi PhD Thesis: *Sparse Convex Optimization Methods for Machine Learning*

- Curvature constant:

$$C_f = \sup_{\substack{x,z\in\mathcal{D} \\ \gamma\in[0,1] \\ y=\gamma z+(1-\gamma)x}} \frac{2}{\gamma^2}\left(f(y) - f(x) - \langle y-x, \nabla f(x)\rangle\right)$$

  We call $D_f(y,x) = f(y) - f(x) - \langle y-x, \nabla f(x)\rangle$ the Bregman divergance associated to f.

2011 Martin Jaggi PhD Thesis: *Sparse Convex Optimization Methods for Machine Learning*

- Curvature constant:

$$C_f = \sup_{\substack{x,z\in\mathcal{D}\\ \gamma\in[0,1]\\ y=\gamma z+(1-\gamma)x}} \frac{2}{\gamma^2}\left(f(y) - f(x) - \langle y - x, \nabla f(x)\rangle\right)$$

  We call $D_f(y,x) = f(y) - f(x) - \langle y - x, \nabla f(x)\rangle$ the Bregman divergance associated to f.

- Bounded by the Lipschitz constant L_f of ∇f on D:

$$\forall x, y \in \mathcal{D}, \quad \|\nabla f(x) - \nabla f(y)\| \leq L_f \|x - y\|$$

Advantages of Frank-Wolfe

Question: why not just do projected gradient descent?

Advantages of Frank-Wolfe

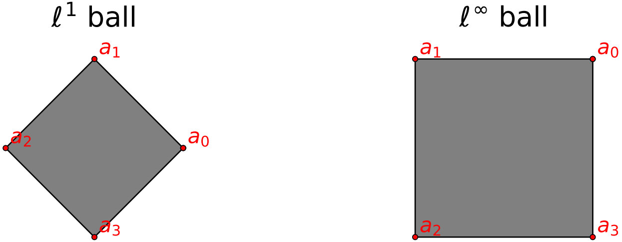
Question: why not just do projected gradient descent?

- 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

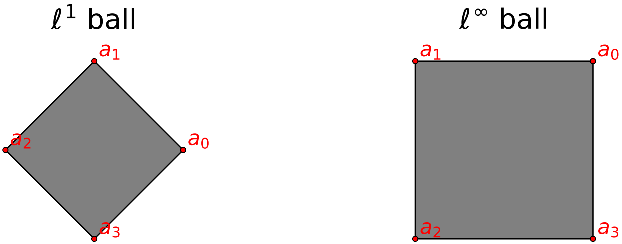
Question: why not just do projected gradient descent?

- 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. $\mathcal{D} = \text{conv}(a_1, \dots a_j)$.
  - Sparsity, low-rank, etc.



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. $\mathcal{D} = \operatorname{conv}(a_1, \dots a_j)$.
  - Sparsity, low-rank, etc.
- The iterates are always feasible, i.e. $x_k \in \mathcal{D}$ for all $k \in \mathbb{N}$.



Limitations of classical Frank-Wolfe/Conditional Gradient

- Lipschitz-smoothness can be a strong assumption.

Limitations of classical Frank-Wolfe/Conditional Gradient

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

Limitations of classical Frank-Wolfe/Conditional Gradient

- 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.

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.

GREYC

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):

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

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):

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

- f is *relatively* smooth.

GREYC

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):

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

- f is *relatively* smooth.
- $\operatorname{dom} h \ (= \mathcal{D})$ is compact.
- h is Lipschitz-continuous.

GREYC

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):

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

- f is *relatively* smooth.
- $\mathrm{dom}\, h \; (= \mathcal{D})$ is compact.
- h is Lipschitz-continuous.
- 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):

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

- f is *relatively* smooth.
- $\mathrm{dom}\, h \; (= \mathcal{D})$ is compact.
- h is Lipschitz-continuous.
- prox_g is accessible.
- T and A are bounded linear operators.

Relative Smoothness

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,

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

$K_{(F,\zeta,\mathcal{C})}$ is a far-reaching generalization of the standard curvature constant.

Moreau-Yosida Regularization

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

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

Moreau-Yosida Regularization

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

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

- 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,

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

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

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

The proximal operator associated to g with parameter β is given by,

$$\mathrm{prox}_{\beta g}(x) = \operatorname*{Argmin}_{y} g(y) + \frac{1}{2\beta}\|x - y\|^2$$

GREYC

What about the affine constraint $Ax = b$?

- Constrained optimization problems can be reformulated as a Lagrangian saddle point problem,

$$\min_{Ax=b} f(x) = \min_{x} \max_{\mu} f(x) + \langle \mu, Ax - b \rangle$$

which admits a so-called dual problem,

$$\max_{\mu} \min_{x} f(x) + \langle \mu, Ax - b \rangle$$

GREYC

What about the affine constraint $Ax = b$?

- Constrained optimization problems can be reformulated as a Lagrangian saddle point problem,

$$\min_{Ax=b} f(x) = \min_{x} \max_{\mu} f(x) + \langle \mu, Ax - b \rangle$$

which admits a so-called dual problem,

$$\max_{\mu} \min_{x} f(x) + \langle \mu, Ax - b \rangle$$

- *Augmented* Lagrangian problem,

$$\min_{Ax=b} f(x) = \min_{x} \max_{\mu} f(x) + \langle \mu, Ax - b \rangle + \frac{\rho}{2} \|Ax - b\|^2$$

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}.

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

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

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

$$s_k \in \mathrm{Argmin}_s\left\{h\left(s\right) + \langle z_k, s\rangle\right\}$$

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

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

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

$$s_k \in \mathrm{Argmin}_s\left\{h\left(s\right) + \langle z_k, s\rangle\right\}$$

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

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

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

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

$$s_k \in \mathrm{Argmin}_s \left\{h\left(s\right) + \langle z_k, s\rangle\right\}$$

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

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

$$k \leftarrow k+1$$

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.,*

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

GREYC

Asymptotic Feasibility Rate

- Pointwise rate:

$$\inf_{0 \le i \le k} \|Ax_i - b\| = O\left(\frac{1}{\sqrt{\Gamma_k}}\right)$$

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

$$\|Ax_{k_j} - b\| \le \frac{1}{\sqrt{\Gamma_{k_j}}},$$

where $\Gamma_k = \sum_{i=0}^{k} \gamma_i$.

Asymptotic Feasibility Rate

- Pointwise rate:

$$\inf_{0\leq i\leq k} \|Ax_i - b\| = O\left(\frac{1}{\sqrt{\Gamma_k}}\right)$$

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

$$\|Ax_{k_j} - b\| \leq \frac{1}{\sqrt{\Gamma_{k_j}}},$$

where $\Gamma_k = \sum_{i=0}^{k} \gamma_i$.

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

$$\|A\bar{x}_k - b\| = O\left(\frac{1}{\sqrt{\Gamma_k}}\right)$$

Convergence to Optimality

Theorem

Let $(x_k)_{k\in\mathbb{N}}$ *be the sequence of primal iterates generated by CGALP,* $(\mu_k)_{k\in\mathbb{N}}$ *the sequence of dual iterates, and* $(x^\star,\mu^\star)$ *a saddle-point pair for the Lagrangian. Then the following holds,*

Convergence to Optimality

Theorem

Let $(x_k)_{k\in\mathbb{N}}$ be the sequence of primal iterates generated by CGALP, $(\mu_k)_{k\in\mathbb{N}}$ the sequence of dual iterates, and $(x^\star, \mu^\star)$ a saddle-point pair for the Lagrangian. Then the following 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 $(x_k)_{k\in\mathbb{N}}$ be the sequence of primal iterates generated by CGALP, $(\mu_k)_{k\in\mathbb{N}}$ the sequence of dual iterates, and $(x^\star, \mu^\star)$ a saddle-point pair for the Lagrangian. Then the following 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 $(\mu_k)_{k\in\mathbb{N}}$ is bounded.*

Lagrangian Convergence Rate

- Pointwise rate:

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

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

- Pointwise rate:

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

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

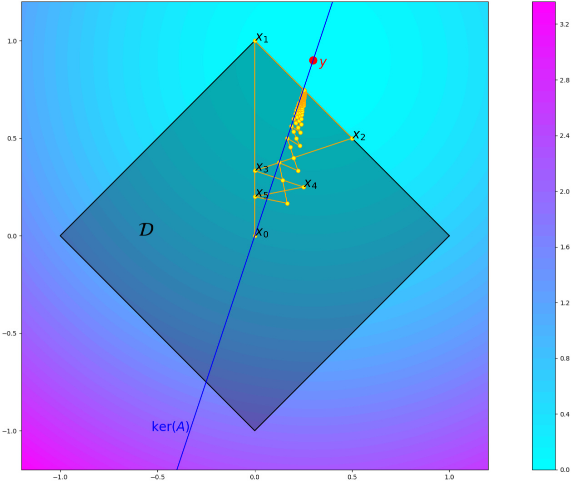
$$\mathcal{L}\left(x_{k_j+1}, \mu^\star\right) - \mathcal{L}(x^\star, \mu^\star) \leq \frac{1}{\Gamma_{k_j}}$$

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

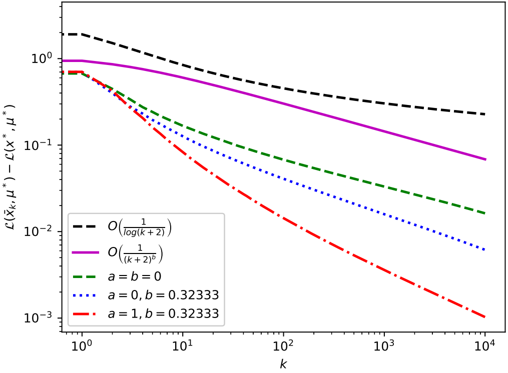
$$\mathcal{L}(\bar{x}_k, \mu^\star) - \mathcal{L}(x^\star, \mu^\star) = O\left(\frac{1}{\Gamma_k}\right)$$

GREYC

Simple Projection Problem



$$\min_{\substack{\|x\|_1 \leq 1 \\ Ax = 0}} \|x - y\|^2$$



Ergodic convergence profile for various step size choices,

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

Matrix Completion Problem

Consider the following matrix completion problem,

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

Matrix Completion Problem

Consider the following 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

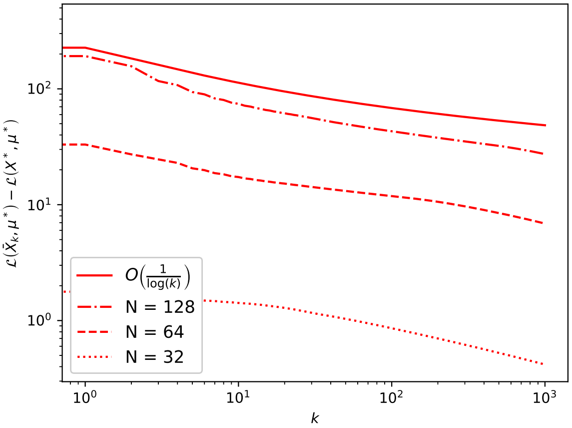
$$H(\boldsymbol{X}) = \iota_{\mathbb{B}_*^{\delta_1}} \left(X^{(1)} \right) + \iota_{\mathbb{B}_1^{\delta_2}} \left(X^{(2)} \right)$$

Direction Finding Step (2 components)

$$S_k^{(1)} \in \operatorname*{Argmin}_{S^{(1)} \in \mathbb{B}^{\delta_1}_{\|\cdot\|_*}} \left\langle \frac{\Omega^* \left(\Omega X_k^{(1)} - y - \operatorname{prox}_{\frac{\beta_k}{2}\|\cdot\|_1}\left(\Omega X_k^{(1)} - y\right)\right)}{\beta_k} + \frac{1}{2}\left(\mu_k^{(1)} - \mu_k^{(2)} + \rho_k \left(X_k^{(1)} - X_k^{(2)}\right)\right), S^{(1)} \right\rangle$$

$$S_k^{(2)} \in \operatorname*{Argmin}_{S^{(2)} \in \mathbb{B}^{\delta_2}_{\|\cdot\|_1}} \left\langle \frac{\Omega^* \left(\Omega X_k^{(2)} - y - \operatorname{prox}_{\frac{\beta_k}{2}\|\cdot\|_1}\left(\Omega X_k^{(2)} - y\right)\right)}{\beta_k} + \frac{1}{2}\left(\mu_k^{(2)} - \mu_k^{(1)} + \rho_k \left(X_k^{(2)} - X_k^{(1)}\right)\right), S^{(2)} \right\rangle$$

CGALP Ergodic Convergence Rate



Ergodic convergence profiles for CGALP.

Future Work

- Stochastic setting: noise on ∇f, noise on $\mathrm{prox}_{\beta g}$, noise on linear minimization oracle.

Future Work

- Stochastic setting: noise on ∇f, noise on $\mathrm{prox}_{\beta g}$, noise on linear minimization oracle.
  - Almost sure feasibility, almost sure convergence of Lagrangian to optimal value, almost sure weak convergence of $(\mu_k)_{k\in\mathbb{N}}$ to solution of the dual problem, almost sure rates, etc.

GREYC

Future Work

- Stochastic setting: noise on ∇f, noise on $\text{prox}_{\beta g}$, noise on linear minimization oracle.
  - Almost sure feasibility, almost sure convergence of Lagrangian to optimal value, almost sure weak convergence of $(\mu_k)_{k\in\mathbb{N}}$ to solution of the dual problem, almost sure rates, etc.
- (Reflexive) Banach space setting: applicable to more general problems.

Thanks for listening

Thanks for listening.

Full paper available on arxiv: https://arxiv.org/abs/1901.01287

"Generalized Conditional Gradient with Augmented Lagrangian for Composite Minimization" - Antonio Silveti-Falls, Cesare Molinari, Jalal Fadili.