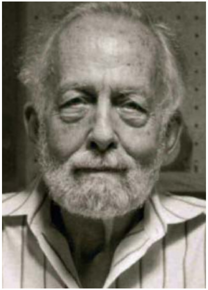
Inexact and Stochastic Generalized Conditional Gradient with Augmented Lagrangian and Proximal Step

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

European Research Council

History and Motivation

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

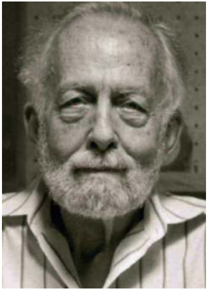


History and Motivation

- 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.



The Frank-Wolfe Algorithm

Algorithm: Frank-Wolfe (Conditional Gradient)

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

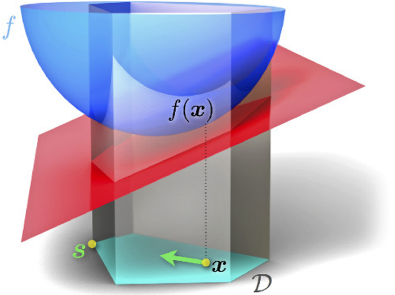
$k = 0$

repeat

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

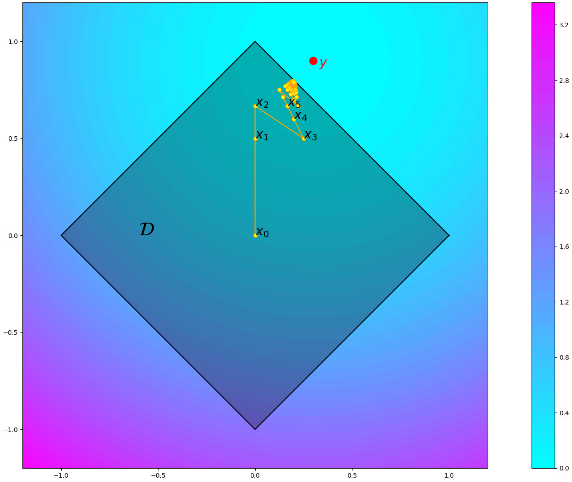
until *convergence*;

Output: x_{k+1}.



(Credit: Stephanie Stutz/Wikipedia)

Frank-Wolfe for Sparse Optimizaiton



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

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.

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.

- 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?

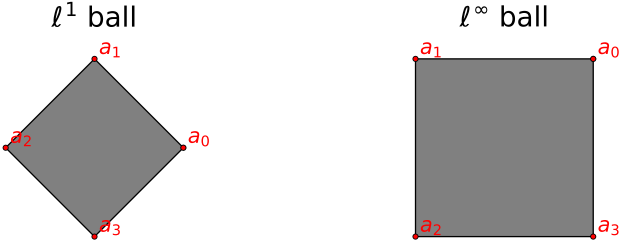
Question: why not just do projected gradient descent?

- 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?

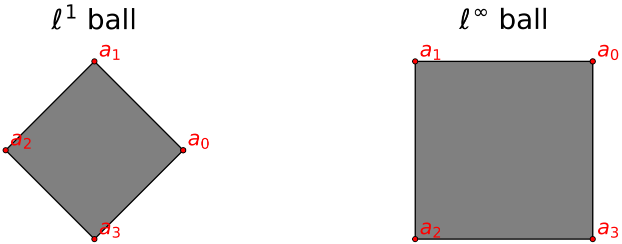
- The set $\mathcal{D}$ might not admit easy projections.
    - Nuclear norm $\|\cdot\|_*$ of a matrix (ℓ^1 norm on singular values).
- The updates of Frank-Wolfe maintain structure.
    - Useful when $\mathcal{D}$ is *atomically generated*, i.e. $\mathcal{D} = \text{conv}\,(a_1, \ldots a_j)$.
    -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



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



Limitations of 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.

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.
- $\mathrm{dom}\, h \ (= \mathcal{D})$ is compact.
- h is Lipschitz-continuous.

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.

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.
- prox_g is accessible.
- $T : \mathcal{H}_p \to \mathcal{H}_v$ and $A : \mathcal{H}_p \to \mathcal{H}_d$ are bounded linear operators.

GREYC

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,

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

GREYC

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,

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

$K_{(F,\zeta,\mathcal{D})}$ 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,

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

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

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

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}.

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

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

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

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

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

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

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

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}.

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}.

Example Parameters

General example: take, for $k \in \mathbb{N}$,

$$\rho_k \equiv \rho > 0, \quad \gamma_k = \frac{1}{(k+1)^{1-b}}, \quad \beta_k = \frac{1}{(k+1)^{1-\delta}}, \quad \text{with}$$

$$0 \leq 2b < \delta < 1, \quad \delta < 1 - b, \quad \rho > 2^{2-b}/c, \quad c > 0.$$

Example Parameters

General example: take, for $k \in \mathbb{N}$,

$$\rho_k \equiv \rho > 0, \quad \gamma_k = \frac{1}{(k+1)^{1-b}}, \quad \beta_k = \frac{1}{(k+1)^{1-\delta}}, \quad \text{with}$$
$$0 \leq 2b < \delta < 1, \quad \delta < 1 - b, \quad \rho > 2^{2-b}/c, \quad c > 0.$$

Simple example: take, for $k \in \mathbb{N}$,

$$\rho > 4, \quad \gamma_k = \frac{1}{k+1}, \quad \beta_k = \frac{1}{\sqrt{k+1}}, \quad \theta_k = \gamma_k,$$

i.e., $b = 0$, $\delta = \frac{1}{2}$, $c = 1$.

Asymptotic Feasibility

Theorem

Let $(x_k)_{k\in\mathbb{N}}$ be a sequence of iterates generated by CGALP for a problem which satisfies the previous assumptions on both the functions and the parameters. The the following holds,

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

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

GREYC

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

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 $(x_{k_j})_{j \in \mathbb{N}}$ such that

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

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

- 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 and $(x^\star, \mu^\star)$ a saddle-point pair for the Lagrangian. Assuming the problem satisfies the previous assumptions on both the functions and the parameters, the following holds

- *Convergence of the Lagrangian:*

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

GREYC

Convergence to Optimality

Theorem

Let $(x_k)_{k\in\mathbb{N}}$ *be the sequence of primal iterates generated by CGALP and* $(x^\star, \mu^\star)$ *a saddle-point pair for the Lagrangian. Assuming the problem satisfies the previous assumptions on both the functions and the parameters, 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:

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

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

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

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

A Remark on Subsequential Rates

Our main result shows that

$$\lim_{k\to\infty}\left[\mathcal{L}(x_k,\mu^\star)-\mathcal{L}(x^\star,\mu^\star)+\frac{\rho_k}{2}\|Ax_k-b\|^2\right]=0$$

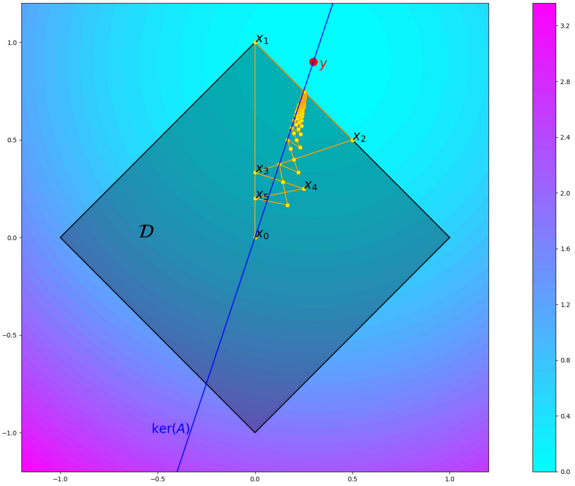
and, subsequentially,

$$\mathcal{L}\left(x_{k_j},\mu^\star\right)-\mathcal{L}(x^\star,\mu^\star)+\frac{\rho_{k_j}}{2}\left\|Ax_{k_j}-b\right\|^2\leq\frac{1}{\Gamma_{k_j}}$$

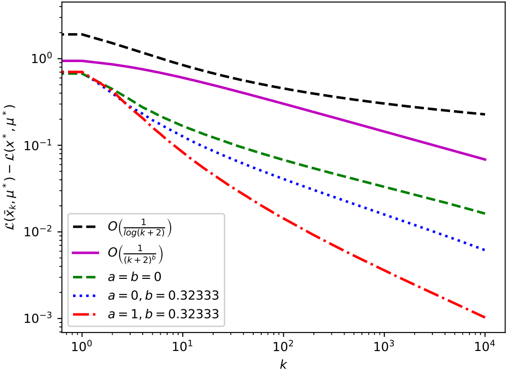
so that our subsequential rates are for the *same* subsequence.

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,

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

GREYC

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)

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

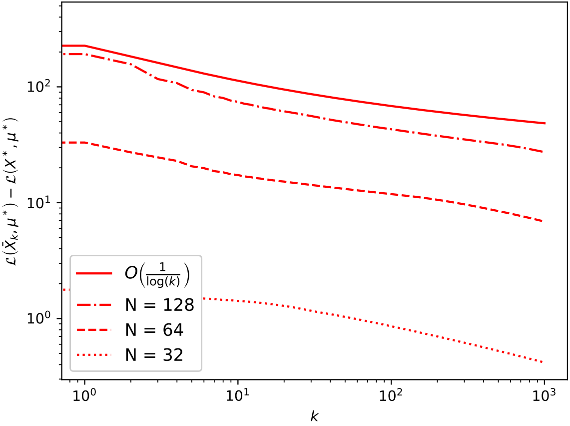
$$\left. + \frac{1}{2} \left(\mu_k^{(1)} - \mu_k^{(2)} + \rho_k \left(X_k^{(1)} - X_k^{(2)} \right) \right), S^{(1)} \right\rangle$$

Direction Finding Step (2 components)

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

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

CGALP Ergodic Convergence Rate



Ergodic convergence profiles for CGALP.

Can We Extend the Algorithm?

What if we have noise?

- On the computation of

$$\nabla f(x_k) + \frac{T^*\left(Tx_k - \mathrm{prox}_{\beta_k g}(Tx_k)\right)}{\beta_k} + \rho_k A^*(Ax_k - b)? \ (\lambda_k^z)$$

Can We Extend the Algorithm?

What if we have noise?

- On the computation of
  $$\nabla f\left(x_k\right) + \frac{T^*\left(Tx_k - \text{prox}_{\beta_k g}\left(Tx_k\right)\right)}{\beta_k} + \rho_k A^*\left(Ax_k - b\right)? \ (\lambda_k^z)$$
- On the linear minimization oracle itself? (λ_k^s)

Inexact CGALP

Algorithm: ICGALP

Input: $x_0 \in \mathcal{D} \stackrel{\text{def}}{=} \operatorname{dom}(h)$; $\mu_0 \in \operatorname{ran}(A)$; $(\gamma_k)_{k\in\mathbb{N}}$, $(\beta_k)_{k\in\mathbb{N}}$, $(\theta_k)_{k\in\mathbb{N}}, (\rho_k)_{k\in\mathbb{N}} \in \ell_+$, $k = 0$.

repeat

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

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

until *convergence*;

Algorithm: ICGALP

Input: $x_0 \in \mathcal{D} \stackrel{\text{def}}{=} \mathrm{dom}\,(h)$; $\mu_0 \in \mathrm{ran}(A)$; $(\gamma_k)_{k\in\mathbb{N}}$, $(\beta_k)_{k\in\mathbb{N}}$, $(\theta_k)_{k\in\mathbb{N}}, (\rho_k)_{k\in\mathbb{N}} \in \ell_+$, $k = 0$.

repeat

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

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

$$s_k \in \mathrm{Argmin}_{s\in\mathcal{H}_p}\{h(s) + \langle z_k, s\rangle\}$$

$$\widehat{s}_k \in \{s : \langle s, z_k\rangle + h(s) \le \langle s_k, z_k\rangle + h(s_k) + \lambda_k^s\}$$

until *convergence*;

Inexact CGALP

Algorithm: ICGALP

Input: $x_0 \in \mathcal{D} \stackrel{\text{def}}{=} \mathrm{dom}\,(h)$; $\mu_0 \in \mathrm{ran}(A)$; $(\gamma_k)_{k\in\mathbb{N}}$, $(\beta_k)_{k\in\mathbb{N}}$, $(\theta_k)_{k\in\mathbb{N}}, (\rho_k)_{k\in\mathbb{N}} \in \ell_+$, $k = 0$.

repeat

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

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

$$s_k \in \mathrm{Argmin}_{s\in\mathcal{H}_p}\{h(s) + \langle z_k, s\rangle\}$$

$$\widehat{s}_k \in \{s : \langle s, z_k\rangle + h(s) \le \langle s_k, z_k\rangle + h(s_k) + \lambda_k^s\}$$

$$x_{k+1} = x_k - \gamma_k(x_k - \widehat{s}_k)$$

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

$$k \leftarrow k + 1$$

until *convergence*;

Technical Setup

Let λ_k^z and λ_k^s be random variables from $(\Omega, \mathcal{F}, \mathbb{P})$ to $\mathcal{H}_p$ and $\mathbb{R}_+$ respectively.
Define the filtration $\mathcal{S} \stackrel{\text{def}}{=} (\mathcal{S}_k)_{k\in\mathbb{N}}$ where $\mathcal{S}_k \stackrel{\text{def}}{=} \sigma\left(x_0, \mu_0, \widehat{s}_0, \ldots, \widehat{s}_k\right)$ is the σ-algebra generated by the random variables $x_0, \mu_0, \widehat{s}_0, \ldots, \widehat{s}_k$.

Technical Setup

Let λ_k^z and λ_k^s be random variables from $(\Omega, \mathcal{F}, \mathbb{P})$ to $\mathcal{H}_p$ and $\mathbb{R}_+$ respectively.
Define the filtration $\mathcal{S} \stackrel{\text{def}}{=} (\mathcal{S}_k)_{k \in \mathbb{N}}$ where $\mathcal{S}_k \stackrel{\text{def}}{=} \sigma\left(x_0, \mu_0, \widehat{s}_0, \ldots, \widehat{s}_k\right)$ is the σ-algebra generated by the random variables $x_0, \mu_0, \widehat{s}_0, \ldots, \widehat{s}_k$.
We will assume:

- $\left(\gamma_{k+1} \mathbb{E}\left[\left\|\lambda_{k+1}^z\right\| \mid \mathcal{S}_k\right]\right)_{k \in \mathbb{N}} \in \ell_+^1(\mathfrak{S})$
- $\left(\gamma_{k+1} \mathbb{E}\left[\lambda_{k+1}^s \mid \mathcal{S}_k\right]\right)_{k \in \mathbb{N}} \in \ell_+^1(\mathfrak{S})$

GREYC

Technical Setup

Let λ_k^z and λ_k^s be random variables from $(\Omega, \mathcal{F}, \mathbb{P})$ to $\mathcal{H}_p$ and $\mathbb{R}_+$ respectively.

Define the filtration $\mathcal{S} \stackrel{\text{def}}{=} (\mathcal{S}_k)_{k\in\mathbb{N}}$ where $\mathcal{S}_k \stackrel{\text{def}}{=} \sigma\left(x_0, \mu_0, \widehat{s}_0, \ldots, \widehat{s}_k\right)$ is the σ-algebra generated by the random variables $x_0, \mu_0, \widehat{s}_0, \ldots, \widehat{s}_k$.

We will assume:

- $\left(\gamma_{k+1}\mathbb{E}\left[\left\|\lambda_{k+1}^z\right\| \mid \mathcal{S}_k\right]\right)_{k\in\mathbb{N}} \in \ell_+^1\left(\mathfrak{S}\right)$
- $\left(\gamma_{k+1}\mathbb{E}\left[\lambda_{k+1}^s \mid \mathcal{S}_k\right]\right)_{k\in\mathbb{N}} \in \ell_+^1\left(\mathfrak{S}\right)$

We can further refine these assumptions by decomposing λ_{k+1}^z depending on the structure of the noise, e.g.
$\lambda_{k+1}^z = \lambda_{k+1}^f - T^*\lambda_{k+1}^g/\beta_{k+1} + \rho_k\lambda_{k+1}^A$ where λ_{k+1}^f, λ_{k+1}^g, and λ_{k+1}^A represent the error in computing $\nabla f\left(x_{k+1}\right)$, $\mathrm{prox}_{\beta_{k+1}g}\left(Tx_{k+1}\right)$ and $A^*\left(Ax_k - b\right)$ respectively.

Asymptotic Feasibility

Theorem (Feasibility)

Let $(x_k)_{k\in\mathbb{N}}$ be a sequence of iterates generated by ICGALP for a problem which satisfies the previous assumptions on both the functions, the parameters, and the noise. The the following holds,

- *Asymptotic feasbility:* $\lim_{k\to\infty} \|Ax_k - b\| = 0 \quad (\mathbb{P}\text{-a.s.})$.

GREYC

Asymptotic Feasibility Rate

- Pointwise rate:

$$\inf_{0\leq i\leq k} \|Ax_i - b\| = O\left(\frac{1}{\sqrt{\Gamma_k}}\right) \;(\mathbb{P}\text{-a.s.}) \;.$$

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

$$\|Ax_{k_j} - b\| \leq \frac{1}{\sqrt{\Gamma_{k_j}}} \;(\mathbb{P}\text{-a.s.}) \;,$$

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

Asymptotic Feasibility Rate

- Pointwise rate:

$$\inf_{0\le i\le k} \|Ax_i - b\| = O\left(\frac{1}{\sqrt{\Gamma_k}}\right) \; (\mathbb{P}\text{-a.s.}) \; .$$

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

$$\|Ax_{k_j} - b\| \le \frac{1}{\sqrt{\Gamma_{k_j}}} \; (\mathbb{P}\text{-a.s.}) \; ,$$

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

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

$$\|A\bar{x}_k - b\| = O\left(\frac{1}{\sqrt{\Gamma_k}}\right) \; (\mathbb{P}\text{-a.s.}) \; .$$

Convergence to Optimality

Theorem (Optimality)

Let $(x_k)_{k\in\mathbb{N}}$ be the sequence of primal iterates generated by ICGALP and $(x^\star, \mu^\star)$ a saddle-point pair for the Lagrangian. Assuming the problem satisfies the previous assumptions on both the functions, the parameters, and the noise, the following holds

- *Convergence of the Lagrangian:*

$$\lim_{k\to\infty} \mathcal{L}(x_k, \mu^\star) = \mathcal{L}(x^\star, \mu^\star) \;\; (\mathbb{P}\text{-a.s.}) \; . \quad (1)$$

GREYC

Convergence to Optimality

Theorem (Optimality)

Let $(x_k)_{k\in\mathbb{N}}$ be the sequence of primal iterates generated by ICGALP and $(x^\star, \mu^\star)$ a saddle-point pair for the Lagrangian. Assuming the problem satisfies the previous assumptions on both the functions, the parameters, and the noise, the following holds

- *Convergence of the Lagrangian:*

$$\lim_{k\to\infty} \mathcal{L}(x_k, \mu^\star) = \mathcal{L}(x^\star, \mu^\star) \quad (\mathbb{P}\text{-a.s.}) . \tag{1}$$

- *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 $(\mathbb{P}\text{-a.s.})$.*

GREYC

Lagrangian Convergence Rate

- Pointwise rate:

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

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

$$\mathcal{L}\left(x_{k_j+1}, \mu^\star\right) - \mathcal{L}(x^\star, \mu^\star) \leq \frac{1}{\Gamma_{k_j}} \ (\mathbb{P}\text{-a.s.}) \ .$$

GREYC

Lagrangian Convergence Rate

- Pointwise rate:

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

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

$$\mathcal{L}\left(x_{k_j+1},\mu^\star\right) - \mathcal{L}(x^\star,\mu^\star) \leq \frac{1}{\Gamma_{k_j}} \ (\mathbb{P}\text{-a.s.}) \ .$$

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

$$\mathcal{L}(\bar{x}_k,\mu^\star) - \mathcal{L}(x^\star,\mu^\star) = O\left(\frac{1}{\Gamma_k}\right) \ (\mathbb{P}\text{-a.s.}) \ .$$

GREYC

Model Problem

Consider the following risk minimization problem,

$$\min_{\substack{x \in \mathcal{C} \subset \mathcal{H} \\ Ax = b}} f(x) \left[\stackrel{\text{def}}{=} \mathbb{E}\left[L(x, \eta)\right] \right]$$

assuming that

- ∇f is Hölder-continuous with constant C_f and exponent τ_f.
- $\nabla_x L(\cdot, \eta)$ is Hölder-continuous for every η with constant C_f and exponent τ_f, η being a random variable.
- $\nabla f(x) = \mathbb{E}\left[\nabla_x L(x, \eta)\right] \quad (\mathbb{P}\text{-a.e.})$.

GREYC

Growing Batch Size

At each iteration $k \in \mathbb{N}$, we compute the average of a batch of $n(k)$ samples of the gradient,

$$\widehat{\nabla f}_k \stackrel{\text{def}}{=} \frac{1}{n(k)} \sum_{i=1}^{n(k)} \nabla_x L(x_k, \eta_i)$$

Growing Batch Size

At each iteration $k \in \mathbb{N}$, we compute the average of a batch of $n(k)$ samples of the gradient,

$$\widehat{\nabla f}_k \stackrel{\text{def}}{=} \frac{1}{n(k)} \sum_{i=1}^{n(k)} \nabla_x L(x_k, \eta_i)$$

We make the assumption each η_i is i.i.d. according to a fixed distribution and that the number of samples in each batch k can vary with k (growing).

Growing Batch Size

At each iteration $k \in \mathbb{N}$, we compute the average of a batch of $n(k)$ samples of the gradient,

$$\widehat{\nabla f}_k \stackrel{\text{def}}{=} \frac{1}{n(k)} \sum_{i=1}^{n(k)} \nabla_x L(x_k, \eta_i)$$

We make the assumption each η_i is i.i.d. according to a fixed distribution and that the number of samples in each batch k can vary with k (growing).
If $n(k)$ grows sufficiently fast, i.e. like $\gamma_k^{-2\tau_f}$, then the summability condition for the error is met,

$$\left(\gamma_{k+1}\mathbb{E}\left[\left\|\lambda_{k+1}^z\right\| \mid \mathcal{S}_k\right]\right)_{k\in\mathbb{N}} \in \ell_+^1(\mathfrak{S})$$

GREYC

Variance Reduction

Fix $\gamma_k = \frac{1}{(k+1)^{1-b}}$ and introduce a weight $\nu_k = \gamma_k^{\frac{2}{3}\tau_f}$. Recursively define,

$$\widehat{\nabla f}_k \stackrel{\text{def}}{=} (1-\nu_k)\,\widehat{\nabla f}_{k-1} + \nu_k \nabla_x L(x_k, \eta_k); \quad \widehat{\nabla f}_{-1} = 0$$

Variance Reduction

Fix $\gamma_k = \frac{1}{(k+1)^{1-b}}$ and introduce a weight $\nu_k = \gamma_k^{\frac{2}{3}\tau_f}$. Recursively define,

$$\widehat{\nabla f}_k \stackrel{\text{def}}{=} (1-\nu_k)\,\widehat{\nabla f}_{k-1} + \nu_k \nabla_x L(x_k, \eta_k)\,; \quad \widehat{\nabla f}_{-1} = 0$$

Here the batch size need not grow, it may even be 1 for all k. The choice of b is more restricted in order to meet summability conditions, we must take $b < 1 - \left(1 + \frac{\tau_f}{3}\right)^{-1}$ to fulfill

$$\left(\gamma_{k+1}\mathbb{E}\left[\left\|\lambda_{k+1}^z\right\| \mid \mathcal{S}_k\right]\right)_{k\in\mathbb{N}} \in \ell_+^1(\mathfrak{S})$$

Deterministic Sweeping for Finite Sum Minimization

For finite sum minimization problems of the form

$$\min_{\substack{x \in \mathcal{C} \subset \mathcal{H} \\ Ax = b}} \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

with $n > 1$ fixed and each f_i Hölder-smooth with constant C_f and exponent τ_f.

For finite sum minimization problems of the form

$$\min_{\substack{x \in \mathcal{C} \subset \mathcal{H} \\ Ax=b}} \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

with $n > 1$ fixed and each f_i Hölder-smooth with constant C_f and exponent τ_f.
Requires computing the gradient of a single f_i at each iteration and keeping a running average of past n sampled gradients.

Deterministic Sweeping for Finite Sum Minimization

$$\widehat{\nabla f}_0 = \frac{1}{n}(\quad 0 + \quad 0 + \ldots \quad + 0)$$

Deterministic Sweeping for Finite Sum Minimization

$$\widehat{\nabla f}_0 = \frac{1}{n}(\quad 0 + \quad 0 + \ldots \quad + 0)$$

$$\widehat{\nabla f}_1 = \frac{1}{n}(\quad \nabla f_1(x_1) + \quad 0 + \ldots \quad + 0)$$

Deterministic Sweeping for Finite Sum Minimization

$$\widehat{\nabla f}_0 = \frac{1}{n}(\quad 0 + \quad 0 + \ldots \quad +0)$$

$$\widehat{\nabla f}_1 = \frac{1}{n}(\quad \nabla f_1(x_1) + \quad 0 + \ldots \quad +0)$$

$$\widehat{\nabla f}_2 = \frac{1}{n}(\quad \nabla f_1(x_1) + \quad \nabla f_2(x_2) + \ldots \quad +0)$$

Deterministic Sweeping for Finite Sum Minimization

$$
\begin{aligned}
\widehat{\nabla f}_0 &= \frac{1}{n}(&& 0 + && 0 + \ldots && +0) \\
\widehat{\nabla f}_1 &= \frac{1}{n}(&& \nabla f_1(x_1) + && 0 + \ldots && +0) \\
\widehat{\nabla f}_2 &= \frac{1}{n}(&& \nabla f_1(x_1) + && \nabla f_2(x_2) + \ldots && +0) \\
& && \vdots && && \\
\widehat{\nabla f}_{n+1} &= \frac{1}{n}(&& \nabla f_1(x_{n+1}) + && \nabla f_2(x_2) + \ldots && +\nabla f_n(x_n))
\end{aligned}
$$

GREYC

Deterministic Sweeping for Finite Sum Minimization

$$\widehat{\nabla f}_0 = \frac{1}{n}(\quad 0 + \quad 0 + \dots \quad +0)$$

$$\widehat{\nabla f}_1 = \frac{1}{n}(\quad \nabla f_1(x_1) + \quad 0 + \dots \quad +0)$$

$$\widehat{\nabla f}_2 = \frac{1}{n}(\quad \nabla f_1(x_1) + \quad \nabla f_2(x_2) + \dots \quad +0)$$

$$\vdots$$

$$\widehat{\nabla f}_{n+1} = \frac{1}{n}(\quad \nabla f_1(x_{n+1}) + \quad \nabla f_2(x_2) + \dots \quad + \nabla f_n(x_n))$$

$$\widehat{\nabla f}_{n+2} = \frac{1}{n}(\quad \nabla f_1(x_{n+1}) + \quad \nabla f_2(x_{n+2}) + \dots \quad + \nabla f_n(x_n))$$

$$\vdots$$

GREYC

Projection Problem with Sampling

We apply the variance reduction method and the sweeping method to the projection problem,

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

by letting η take value in $\{1, \ldots, n\}$ with $L(x, \eta) = \frac{1}{2}(x_\eta - y_\eta)$ and $f_i(x) = \frac{1}{2}(x_i - y_i)^2$ respectively.

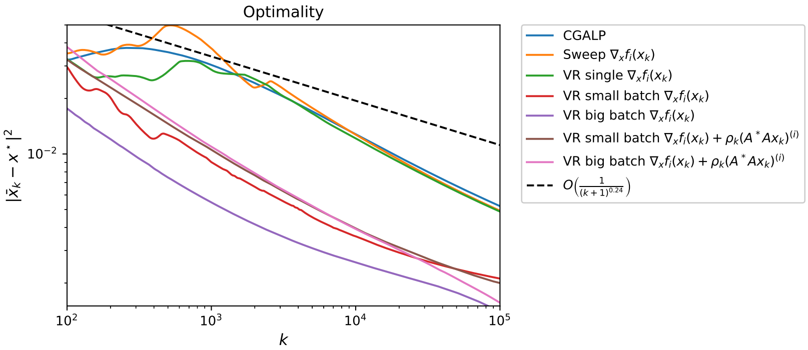
Projection Problem with Sampling

We apply the variance reduction method and the sweeping method to the projection problem,

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

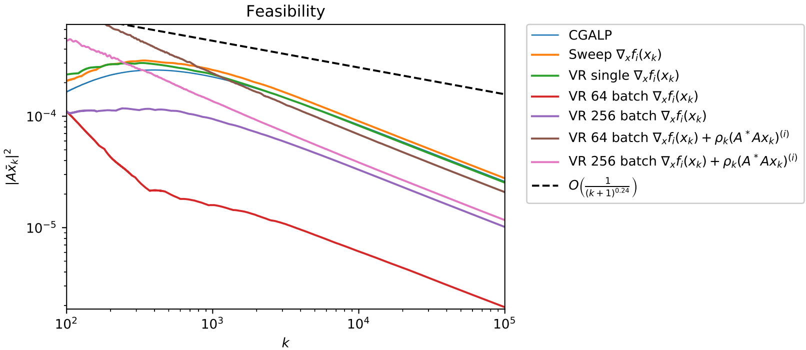
by letting η take value in $\{1, \ldots, n\}$ with $L(x, \eta) = \frac{1}{2}(x_\eta - y_\eta)$ and $f_i(x) = \frac{1}{2}(x_i - y_i)^2$ respectively.
Since the objective is Lipschitz-smooth we have $\tau_f = 1$ and $\alpha = \frac{2}{3}$.
We take $\gamma_k = \frac{1}{(k+1)^{1-b}}, \rho_k \equiv \rho = 2^{2-b} + 1, \theta_k = \gamma_k$.

Optimality - Big Step Size



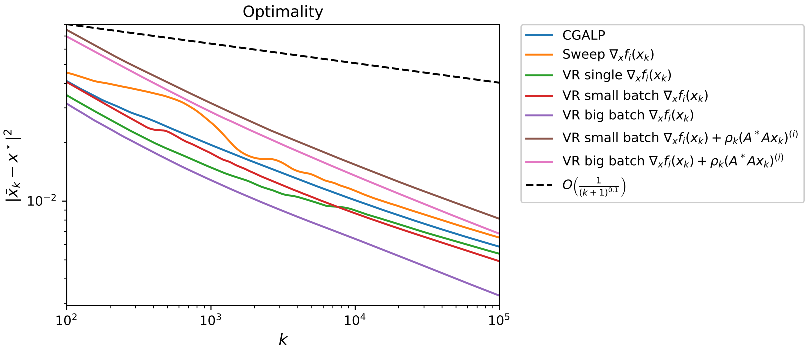
The step size is $\gamma_k = (k+1)^{-\left(1-\frac{1}{4}+0.01\right)}$ and the weight for variance reduction is $\nu_k = \gamma_k^{2/3}$.

Feasibility - Big Step Size



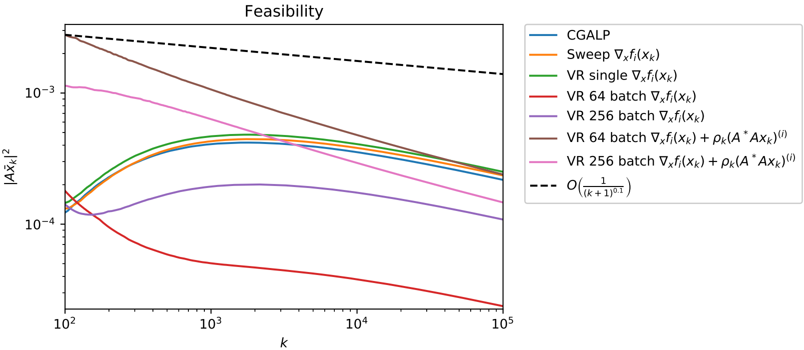
The step size is $\gamma_k = (k+1)^{-\left(1-\frac{1}{4}+0.01\right)}$ and the weight for variance reduction is $\nu_k = \gamma_k^{2/3}$.

Optimality - Small Step Size



The step size is $\gamma_k = (k+1)^{-\left(1-\frac{1}{4}+0.15\right)}$ and the weight for variance reduction is $\nu_k = \gamma_k^{2/3}$.

Feasibility - Small Step Size



The step size is $\gamma_k = (k+1)^{-\left(1-\frac{1}{4}+0.15\right)}$ and the weight for variance reduction is $\nu_k = \gamma_k^{2/3}$.

Thanks for Listening

Thanks for listening.

Full paper available on arxiv: https://arxiv.org/abs/ 2005.05158

"Inexact and Stochastic Generalized Conditional Gradient with Augmented Lagrangian and Proximal Step" - Antonio Silveti-Falls, Cesare Molinari, Jalal Fadili.

Special thanks to Cesare Molinari for the invitation to give this talk.

GREYC