Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 12: Constrained convex minimization II

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

EE-556 (Fall 2018)

lions@epfl

License Information for Mathematics of Data Slides

Outline

- This class:
  1. Linear minimization oracle
  2. Conditional gradient method (CGM)
  3. CGM-type methods for problems with affine constraints
- Next class
  1. Primal-dual subgradient methods

Recommended reading material

- M. Jaggi, *Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization* In Proc. 30th International Conference on Machine Learning, 2013.
- A. Yurtsever, O. Fercoq, F. Locatello and V. Cevher, *A Conditional Gradient Framework for Composite Convex Minimization with Applications to Semidefinite Programming* In Proc. 35th International Conference on Machine Learning, 2018.

Motivation

Motivation

In previous class, we learned optimization techniques for solving constrained convex minimization problems, based on the powerful proximal gradient framework. Unfortunately, the *proximal operator* can impose an undesirable *computational burden* and even intractability in many applications.

In this lecture, we will cover the *conditional gradient*-type methods (*a.k.a.*, Frank-Wolfe algorithm). These methods leverage the so called *linear minimization oracle*, which is arguably cheaper to evaluate than proximal operator.

Recall the proximal operator

Definition (Proximal operator)

Let $g \in \mathcal{F}(\mathbb{R}^p)$ and $\mathbf{x} \in \mathbb{R}^p$. The proximal operator of g is defined as:

$$\boxed{\mathrm{prox}_g(\mathbf{x}) \equiv \arg\min_{\mathbf{y} \in \mathbb{R}^p} \left\{ g(\mathbf{y}) + \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 \right\}.} \tag{1}$$

Proximal operator helps us processing nonsmooth terms.

Definition (Tractable proximity)

Given $g \in \mathcal{F}(\mathbb{R}^p)$. We say that g is proximally tractable if prox_g defined by (1) can be computed **efficiently**.

- "**efficiently**" = {closed form solution, low-cost computation, polynomial time}.
- We denote $\mathcal{F}_{\mathrm{prox}}(\mathbb{R}^p)$ the class of proximally tractable convex functions.

Not all non-smooth functions are prox-friendly

Surprisingly, proximal operator can be intractable, e.g., for dual of structural SVMs [5].

Even some tractable proximal operators can impose
undesirable computational burden!

Name	Function	Proximal operator	Complexity
ℓ_1-norm	$f(\mathbf{x}) := \|\mathbf{x}\|_1$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = \mathrm{sign}(\mathbf{x}) \otimes [\lvert\mathbf{x}\rvert - \lambda]_+$	$\mathcal{O}(p)$
ℓ_2-norm	$f(\mathbf{x}) := \|\mathbf{x}\|_2$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = [1 - \lambda/\|\mathbf{x}\|_2]_+\mathbf{x}$	$\mathcal{O}(p)$
Support function	$f(\mathbf{x}) := \max_{\mathbf{y}\in\mathcal{C}} \mathbf{x}^T\mathbf{y}$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = \mathbf{x} - \lambda\pi_{\mathcal{C}}(\mathbf{x})$	
Box indicator	$f(\mathbf{x}) := \delta_{[\mathbf{a},\mathbf{b}]}(\mathbf{x})$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = \pi_{[\mathbf{a},\mathbf{b}]}(\mathbf{x})$	$\mathcal{O}(p)$
Positive semidefinite cone indicator	$f(\mathbf{X}) := \delta_{\mathbb{S}_+^p}(\mathbf{X})$	$\mathrm{prox}_{\lambda f}(\mathbf{X}) = \mathbf{U}[\Sigma]_+\mathbf{U}^T$, where $\mathbf{X} = \mathbf{U}\Sigma\mathbf{U}^T$	$\mathcal{O}(p^3)$
Hyperplane indicator	$f(\mathbf{x}) := \delta_{\mathcal{X}}(\mathbf{x}),\ \mathcal{X} := \{\mathbf{x} : \mathbf{a}^T\mathbf{x} = b\}$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = \pi_{\mathcal{X}}(\mathbf{x}) = \mathbf{x} + \left(\frac{b-\mathbf{a}^T\mathbf{x}}{\|\mathbf{a}\|_2}\right)\mathbf{a}$	$\mathcal{O}(p)$
Simplex indicator	$f(\mathbf{x}) = \delta_{\mathcal{X}}(\mathbf{x}), \mathcal{X} := \{\mathbf{x} : \mathbf{x} \geq 0,\ \mathbf{1}^T\mathbf{x} = 1\}$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = (\mathbf{x} - \nu\mathbf{1})$ for some $\nu \in \mathbb{R}$, which can be efficiently calculated	$\tilde{\mathcal{O}}(p)$
Convex quadratic	$f(\mathbf{x}) := (1/2)\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{q}^T\mathbf{x}$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = (\lambda\mathbb{I} + \mathbf{Q})^{-1}\mathbf{x}$	$\mathcal{O}(p\log p) \to \mathcal{O}(p^3)$
Square ℓ_2-norm	$f(\mathbf{x}) := (1/2)\|\mathbf{x}\|_2^2$	$\mathrm{prox}_{\lambda f}(\mathbf{x}) = (1/(1+\lambda))\mathbf{x}$	$\mathcal{O}(p)$
log-function	$f(\mathbf{x}) := -\log(x)$	$\mathrm{prox}_{\lambda f}(x) = ((x^2 + 4\lambda)^{1/2} + x)/2$	$\mathcal{O}(1)$
$\log\det$-function	$f(\mathbf{x}) := -\log\det(\mathbf{X})$	$\mathrm{prox}_{\lambda f}(\mathbf{X})$ is the $\log$-function prox applied to the individual eigenvalues of $\mathbf{X}$	$\mathcal{O}(p^3)$

Here: $[\mathbf{x}]_+ := \max\{0, \mathbf{x}\}$ and $\delta_{\mathcal{X}}$ is the indicator function of the convex set $\mathcal{X}$, sign is the sign function, $\mathbb{S}_+^p$ is the cone of symmetric positive semidefinite matrices.

Example: prox for the indicator of a nuclear-norm ball

Consider $\delta_{\mathcal{X}}$, the indicator of nuclear-norm ball $\mathcal{X} := \left\{\mathbf{X} : \mathbf{X} \in \mathbb{R}^{p\times p}, \ \|\mathbf{X}\|_* \leq \alpha\right\}$

Proximal operator of $\delta_{\mathcal{X}}(\mathbf{X})$

$$\mathrm{prox}_{\delta_{\mathcal{X}}}(\mathbf{X}) \equiv \arg\min_{\mathbf{Y}\in\mathbb{R}^{p\times p}} \left\{\delta_{\mathcal{X}}(\mathbf{Y}) + \frac{1}{2}\|\mathbf{Y}-\mathbf{X}\|_F^2\right\} \equiv \mathrm{proj}_{\mathcal{X}}(\mathbf{X})$$

prox of the indicator nuclear-norm ball is equivalent to proj onto nuclear norm-ball.

This can be computed as follows:

- Compute SVD of $\mathbf{X}$ $\Longrightarrow$ $\boldsymbol{U\Sigma V}^T = \mathbf{X}$.
- Form a vector $\mathbf{s} \in \mathbb{R}^p$ by the diagonal entries of $\boldsymbol{\Sigma}$ $\Longrightarrow$ $\mathbf{s} = \mathrm{diag}(\boldsymbol{\Sigma})$.
- Project $\mathbf{s}$ onto ℓ_1 norm ball $\Longrightarrow$ $\hat{\mathbf{s}} = \arg\min_{\mathbf{x}}\{\|\mathbf{s}-\mathbf{x}\| : \|\mathbf{x}\|_1 \leq \alpha\}$
- Form a diagonal matrix with entries $\hat{\mathbf{s}}$ $\Longrightarrow$ $\hat{\boldsymbol{\Sigma}} = \mathrm{diag}^*(\hat{\mathbf{s}})$
- Form the output $\Longrightarrow$ $\mathrm{proj}_{\mathcal{X}}(\mathbf{X}) = \boldsymbol{U\hat{\Sigma}V}^T$

Finding SVD is costly in when p is big!

A basic constrained problem setting

Problem setting

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \Big\{ f(\mathbf{x}) : \mathbf{x} \in \mathcal{X} \Big\}, \tag{2}$$

Assumptions

- $\mathcal{X}$ is nonempty, convex, closed and bounded.
- $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^p)$ (i.e., convex with Lipschitz gradient).

Recall proximal gradient algorithm

Basic proximal-gradient scheme (ISTA)

1. Choose $\mathbf{x}^0 \in \mathrm{dom}(F)$ arbitrarily as a starting point.
2. For $k = 0, 1, \cdots,$ generate a sequence $\{\mathbf{x}^k\}_{k\geq 0}$ as:

$$\mathbf{x}^{k+1} := \mathrm{prox}_{\alpha g}\left(\mathbf{x}^k - \alpha\nabla f(\mathbf{x}^k)\right)$$

where $\alpha := \frac{1}{L}$.

- Prox-operator of indicator of $\mathcal{X}$ is projection onto $\mathcal{X}$ $\Longrightarrow$ ensures feasibility

How else can we ensure feasibility?

Frank-Wolfe's approach - I

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) : \mathbf{x} \in \mathcal{X} \right\},$$

Conditional gradient method (CGM, see [4] for review)

A plausible strategy which dates back to 1956 [2]. At iteration k:

1. Consider the linear approximation of f at $\mathbf{x}^k$

$$\phi_k(\mathbf{x}) := f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^T(\mathbf{x} - \mathbf{x}^k)$$

2. Minimize this approximation within constraint set

$$\hat{\mathbf{x}}^k \in \min_{x\in\mathcal{X}} \phi_k(\mathbf{x}) = \min_{x\in\mathcal{X}} \nabla f(\mathbf{x}^k)^T\mathbf{x}$$

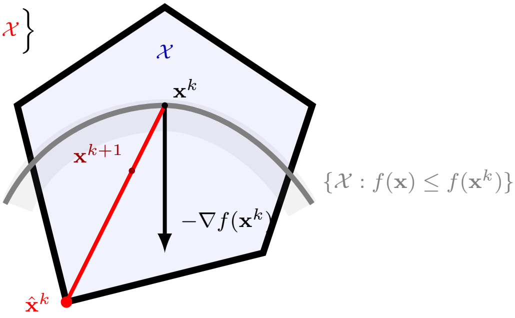
3. Take a step towards $\hat{\mathbf{x}}^k$ with step-size $\gamma_k \in [0, 1]$

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \gamma_k(\hat{\mathbf{x}}^k - \mathbf{x}^k)$$

- $\mathbf{x}^{k+1}$ is feasible since it is convex combination of two other feasible points.

Frank-Wolfe's approach - II

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) : \mathbf{x} \in \mathcal{X} \right\}$$



Conditional gradient method (CGM)

1. Choose $\mathbf{x}^0 \in \mathcal{X}$.

2. For $k = 0, 1, \ldots$ perform:

$$\begin{cases} \hat{\mathbf{x}}^k & := \arg\min_{\mathbf{x}\in\mathcal{X}} \nabla f(\mathbf{x}^k)^T \mathbf{x} \\ \mathbf{x}^{k+1} & := (1-\gamma_k)\mathbf{x}^k + \gamma_k \hat{\mathbf{x}}^k, \end{cases}$$

where $\gamma_k := \frac{2}{k+2}$.

On the linear minimization oracle

Definition (Linear minimization oracle)

Let $\mathcal{X}$ be a convex, closed and bounded set. Then, the linear minimization oracle of $\mathcal{X}$ ($\mathrm{lmo}_{\mathcal{X}}$) returns a vector $\hat{\mathbf{x}}$ such that

$$\boxed{\mathrm{lmo}_{\mathcal{X}}(\mathbf{x}) := \hat{\mathbf{x}} \in \arg\min_{\mathbf{y} \in \mathcal{X}} \ \mathbf{x}^T \mathbf{y}} \tag{3}$$

- $\mathrm{lmo}_{\mathcal{X}}$ returns an extreme point of $\mathcal{X}$.
- $\mathrm{lmo}_{\mathcal{X}}$ is arguably cheaper than projection.
- $\mathrm{lmo}_{\mathcal{X}}$ is not single valued, note $\in$ in the definition.

Example: lmo of nuclear-norm bal

Consider $\delta_{\mathcal{X}}$, the indicator of nuclear-norm ball $\mathcal{X} := \left\{\mathbf{X} : \mathbf{X} \in \mathbb{R}^{p\times p}, \ \ \|\mathbf{X}\|_* \leq \alpha\right\}$

lmo of nuclear-norm ball

$$\mathrm{lmo}_{\mathcal{X}}(\mathbf{X}) := \hat{\mathbf{X}} \in \arg\min_{\mathbf{Y}\in\mathcal{X}} \ \langle \mathbf{Y}, \mathbf{X}\rangle$$

This can be computed as follows:

- Compute top singular vectors of $\mathbf{X}$ $\Longrightarrow$ $(\mathbf{u}_1, \sigma_1, \mathbf{v}_1) = \mathtt{svds}(\mathbf{X}, 1)$.
- Form the rank-1 output $\Longrightarrow$ $\mathbf{X} = -\mathbf{u}_1 \alpha \mathbf{v}_1^T$

We can efficiently approximate top singular vectors by power method!

Convergence guarantees of CGM

Problem setting

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) : \mathbf{x} \in \mathcal{X} \right\},$$

Assumptions

- $\mathcal{X}$ is nonempty, convex, closed and bounded.
- $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^p)$ (i.e., convex with Lipschitz gradient).

Theorem

Under ***assumptions*** *listed above, CGM with step size* $\gamma_k = \frac{2}{k+2}$ *satisfies*

$$\boxed{f(\mathbf{x}^k) - f(\mathbf{x}^\star) \leq \frac{4LD_{\mathcal{X}}}{k+1}} \tag{4}$$

where $D_{\mathcal{X}} := \max_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \|\mathbf{x}-\mathbf{y}\|_2$ *is diameter of constraint set.*

Proof of convergence rate of CGM - part I (self study)

Proof

First, recall the following result about Lipschitz gradient functions $f \in \mathcal{F}_L^{1,1}$

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2.$$

Remark that $\mathbf{x}^{k+1} - \mathbf{x}^k = \gamma_k(\hat{\mathbf{x}}^k - \mathbf{x}^k)$

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \gamma_k \langle \nabla f(\mathbf{x}^k), \hat{\mathbf{x}}^k - \mathbf{x}^k \rangle + \gamma_k^2 \frac{L}{2}\|\hat{\mathbf{x}}^k - \mathbf{x}^k\|_2^2. \tag{5}$$

Since $\mathbf{x}^k$, $\hat{\mathbf{x}}^k$ and $\mathbf{x}^\star$ are all in $\mathcal{X}$, we have

$$\begin{cases} \langle \nabla f(\mathbf{x}^k), \hat{\mathbf{x}}^k - \mathbf{x}^k \rangle = \min_{\mathbf{x}\in\mathcal{X}} \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle \leq \underbrace{\langle \nabla f(\mathbf{x}^k), \mathbf{x}^\star - \mathbf{x}^k \rangle \leq f^\star - f(\mathbf{x}^k)}_{\text{since } f \text{ is convex}} \\ \|\hat{\mathbf{x}}^k - \mathbf{x}^k\|_2 \leq \max_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \|\mathbf{x} - \mathbf{y}\|_2 = D_{\mathcal{X}} \end{cases}$$

Substituting into (5) and substracting $f^\star$ we get

$$\boxed{f(\mathbf{x}^{k+1}) - f^\star \leq (1-\gamma_k)(f(\mathbf{x}^k) - f^\star) + \gamma_k^2 \frac{L}{2} D_{\mathcal{X}}^2}$$

Proof of convergence rate of CGM - part II (self study)

$$f(\mathbf{x}^{k+1}) - f^\star \leq (1-\gamma_k)(f(\mathbf{x}^k) - f^\star) + \gamma_k^2 \frac{L}{2} D_{\mathcal{X}}^2$$

Proof (Continued)

We will use induction technique: First note

$$\gamma_0 = 1 \quad \Longrightarrow \quad f(\mathbf{x}^1) - f^\star \leq \tfrac{1}{2} L D_{\mathcal{X}}^2$$

Now, suppose (4) holds, then

$$\begin{aligned} f(\mathbf{x}^{k+1}) - f^\star &\leq (1-\gamma_k)\frac{4LD_{\mathcal{X}}}{k+1} + \gamma_k^2 \frac{L}{2} D_{\mathcal{X}}^2 \\ &= \frac{k}{k+2}\frac{4LD_{\mathcal{X}}}{k+1} + \frac{4}{(k+2)^2}\frac{L}{2}D_{\mathcal{X}}^2 \leq \frac{4LD_{\mathcal{X}}}{k+2} \end{aligned}$$

which completes the proof by induction.

*Example: Phase retrieval

Phase retrieval

Aim: Recover signal $\mathbf{x}^\natural \in \mathbb{C}^p$ from the measurements $\mathbf{b} \in \mathbb{R}^n$:

$$b_i = \left|\langle \mathbf{a}_i, \mathbf{x}^\natural \rangle\right|^2 + \omega_i.$$

($\mathbf{a}_i \in \mathbb{C}^p$ are known measurement vectors, ω_i models noise).
• Non-linear measurements → **non-convex** maximum likelihood estimators.

PhaseLift [1]

Phase retrieval can be solved as a convex matrix completion problem, following a combination of

- semidefinite relaxation $(\mathbf{x}^\natural {\mathbf{x}^\natural}^H = \mathbf{X}^\natural)$
- convex relaxation $(\mathtt{rank} \to \|\cdot\|_*)$

albeit in terms of the lifted variable $\mathbf{X} \in \mathbb{C}^{p\times p}$.

Example: Phase retrieval - II

Problem formulation

We solve the following PhaseLift variant:

$$f^\star := \min_{\mathbf{X} \in \mathbb{C}^{p \times p}} \left\{ \frac{1}{2} \|\mathcal{A}(\mathbf{X}) - \mathbf{b}\|_2^2 : \quad \|\mathbf{X}\|_* \leq \kappa, \quad \mathbf{X} \geq 0 \right\}. \tag{6}$$

Experimental setup [12]

Coded diffraction pattern measurements, $\mathbf{b} = [\mathbf{b}_1, \ldots, \mathbf{b}_L]$ with $L = 20$ different masks

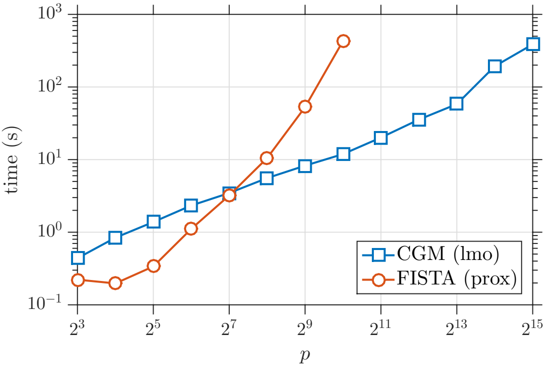
$$\mathbf{b}_\ell = |\texttt{fft}(\mathbf{d}_\ell^H \odot \mathbf{x}^\natural)|^2$$

→ $\odot$ denotes Hadamard product; $|\cdot|^2$ applies element-wise

→ $\mathbf{d}_\ell$ are randomly generated octonary masks (distributions as proposed in [1])

→ Parametric choices: $\lambda^0 = \mathbf{0}^n$; $\quad \epsilon = 10^{-2}$; $\quad \kappa = \texttt{mean}(\mathbf{b})$.

Example: Phase retrieval - III



Test with synthetic data: Prox vs sharp

→ Synthetic data: $\mathbf{x}^\natural = \texttt{randn}(p,1) + i \cdot \texttt{randn}(p,1)$.

→ Stopping criteria: $\frac{\|\mathbf{x}^\natural - \mathbf{x}^k\|_2}{\|\mathbf{x}^\natural\|_2} \leq 10^{-2}$.

→ Averaged over 10 Monte-Carlo iterations.

Note that the problem is $p \times p$ dimensional!

Recall the prototype problem

A **primal problem** prototype

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} - \mathbf{b} \in \mathcal{K},\ \mathbf{x} \in \mathcal{X} \right\}, \tag{7}$$

- f is a proper, closed and convex function
- $\mathcal{X}$ and $\mathcal{K}$ are nonempty, closed convex sets
- $\mathbf{A} \in \mathbb{R}^{n\times p}$ and $\mathbf{b} \in \mathbb{R}^n$ are known
- An optimal solution $\mathbf{x}^\star$ to (7) satisfies $f(\mathbf{x}^\star) = f^\star$, $\mathbf{A}\mathbf{x}^\star = \mathbf{b}$ and $\mathbf{x}^\star \in \mathcal{X}$
- *We further assume $\mathcal{X}$ is a bounded set!*

Classical CGM does not apply to (7)

- lmo of the intersection of $\{\mathbf{x} : \mathbf{A}\mathbf{x} - \mathbf{b} \in \mathcal{K}\}$ and $\mathcal{X}$ is difficult to compute.

CGM with quadratic penalty

Quadratic penalty strategy for $\min\{f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \in \mathcal{X}\}$

A quadratic penalty formulation:

$$\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) + \frac{1}{2\beta}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 : \mathbf{x} \in \mathcal{X} \right\}$$

- $\beta > 0$ is the penalty parameter.
- $f_\beta(\mathbf{x}) := f(\mathbf{x}) + \frac{1}{2\beta}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$ is the penalized objective function.
- Note that $f_\beta(\mathbf{x})$ is smooth with parameter $L + \beta^{-1}\|\mathbf{A}\|^2$.

Our strategy [13] $\Rightarrow$ Take a CGM step on f_β and decrease β progressively to 0

Homotopy conditional gradient method (HCGM)
1. Choose $\mathbf{x}^0 \in \mathcal{X}$, and $\beta_0 > 0$.   **2.** For $k = 0, 1, \ldots$ perform: $\begin{cases} \hat{\mathbf{x}}^k & := \mathrm{lmo}_{\mathcal{X}}(\nabla f(\mathbf{x}^k) + \beta_k^{-1}\mathbf{A}^T(\mathbf{A}\mathbf{x}^k - \mathbf{b})) \\ \mathbf{x}^{k+1} & := (1-\gamma_k)\mathbf{x}^k + \gamma_k\hat{\mathbf{x}}^k, \end{cases}$ where $\gamma_k := \frac{2}{k+2}$ and $\beta_k = \frac{\beta_0}{\sqrt{k+2}}$.

Convergence guarantees of HCGM

Recall Lagrange duality

$$\mathcal{L}(\mathbf{x}, \lambda) := f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle$$

$$\underbrace{\max_{\lambda} \min_{\mathbf{x} \in \mathcal{X}} \mathcal{L}(\mathbf{x}, \lambda)}_{\text{dual problem}} \quad \leq \quad \underbrace{\min_{\mathbf{x} \in \mathcal{X}} \max_{\lambda} \mathcal{L}(\mathbf{x}, \lambda)}_{\text{primal problem}} \qquad \text{(Duality)}$$

- λ is called the Lagrange multiplier.
- The function $d(\lambda)$ is called the dual function, and it is concave!
- The optimal dual objective value is $d^\star = d(\lambda^\star)$.

(Duality) holds with equality under vague assumptions $\Rightarrow$ (Strong duality).

Theorem

Assume that strong duality holds. Then, the iterates of HCGM satisfies

$$\begin{cases} -\|\mathbf{A}\mathbf{x}^k - \mathbf{b}\| \|\lambda^\star\| \leq & f(\mathbf{x}^k) - f^\star & \leq 2D_{\mathcal{X}} \left(\frac{L}{k+1} + \frac{\|\mathbf{A}\|^2}{\beta_0 \sqrt{k+1}} \right) \\ & \|\mathbf{A}\mathbf{x}^k - \mathbf{b}\| & \leq \frac{2\beta_0}{\sqrt{k+1}} \left(\|\lambda^\star\| + D_{\mathcal{X}} \sqrt{\frac{L}{\beta_0} + \frac{\|\mathbf{A}\|^2}{\beta_0^2}} \right). \end{cases}$$

Augmented Lagrangian CGM: CGAL

Quadratic penalty strategy for $\min\{f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \in \mathcal{X}\}$

Augmented problem formulation:

$$\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) + \frac{1}{2\beta}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 : \mathbf{A}\mathbf{x} = \mathbf{b},\ \ \mathbf{x} \in \mathcal{X} \right\}$$

- Write down the Lagrangian:

$$\mathcal{L}_{1/\beta}(\mathbf{x}, \lambda) = f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle + \frac{1/\beta}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$$

- Note that $\mathcal{L}_{1/\beta}(\cdot\,\lambda)$ is smooth with parameter $L + \beta^{-1}\|\mathbf{A}\|^2$.

Our strategy [11] $\Rightarrow$
1. Take a CGM step wrt $\mathcal{L}_{1/\beta}(\cdot\,, \lambda)$
2. Take a gradient step wrt $\mathcal{L}_{1/\beta}(\mathbf{x}, \cdot\,)$
3. Decrease β progressively to 0

Challenge: Step size in dual (step 2.)

Convergence guarantees of CGAL

Conditional gradient augmented Lagrangian method (CGAL)

1. Choose $\mathbf{x}^0 \in \mathcal{X}$, $\lambda^0 \in \mathbb{R}^n$, and $\beta_0 > 0$.
2. For $k = 0, 1, \ldots$ perform:

$$\begin{cases} \hat{\mathbf{x}}^k & := \mathrm{lmo}_{\mathcal{X}}(\nabla f(\mathbf{x}^k) + \mathbf{A}^T\lambda^k + \beta_k^{-1}\mathbf{A}^T(\mathbf{A}\mathbf{x}^k - \mathbf{b})) \\ \mathbf{x}^{k+1} & := (1-\gamma_k)\mathbf{x}^k + \gamma_k \hat{\mathbf{x}}^k \\ \lambda^{k+1} & := \lambda^k + \omega_k(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{b}) \end{cases}$$

where $\gamma_k := \frac{2}{k+2}$ and $\beta_k = \frac{\beta_0}{\sqrt{k+2}}$.

Theorem

Assume that strong duality holds. Let us choose dual step size ω_k by the following rule

$$\omega_k = \alpha_k := \min\{\frac{1}{\beta_0}, \frac{\eta_k^2(L_f + \lambda_{k+1})D_{\mathcal{X}}^2}{2\|\mathbf{A}\mathbf{x}^{k+1} - \mathbf{b}\|^2}\} \quad \textit{if} \quad \|\lambda^k + \alpha_k(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{b})\| \le D_{\mathcal{Y}}$$

and $\omega_k = 0$ otherwise, for some $D_{\mathcal{Y}} \ge 0$. Then, the iterates of CGAL satisfies

$$\begin{cases} -\|\mathbf{A}\mathbf{x}^k - \mathbf{b}\|\|\lambda^\star\| \le & f(\mathbf{x}^k) - f^\star & \le 4D_{\mathcal{X}}\left(\frac{L}{k+1} + \frac{\|\mathbf{A}\|^2}{\beta_0\sqrt{k+1}}\right) + \frac{\beta_0 D_{\mathcal{Y}}}{2\sqrt{k+1}} \\ & \|\mathbf{A}\mathbf{x}^k - \mathbf{b}\| & \le \frac{2\beta_0}{\sqrt{k+1}}\left(\frac{3D_{\mathcal{Y}}}{2} + \|\lambda^\star\| + \frac{D_{\mathcal{X}}}{\beta_0}\sqrt{L\beta_0 + \|\mathbf{A}\|^2}\right) \end{cases}$$

*Generalization of HCGM for $\mathbf{A}\mathbf{x} - \mathbf{b} \in \mathcal{K}$

Quadratic penalty strategy for $\min\{f(\mathbf{x}) : \mathbf{A}\mathbf{x} - \mathbf{b} \in \mathcal{K}, \mathbf{x} \in \mathcal{X}\}$

Define the distance function

$$\mathrm{dist}(\mathbf{y}, \mathcal{K}) := \min_{\mathbf{z} \in \mathcal{K}} \|\mathbf{y} - \mathbf{z}\|.$$

Quadratic penalty takes the form

$$\min_{\mathbf{x} \in \mathbb{R}^p} \left\{ f(\mathbf{x}) + \frac{1}{2\beta} \mathrm{dist}^2(\mathbf{A}\mathbf{x} - \mathbf{b}, \mathcal{K}) : \mathbf{x} \in \mathcal{X} \right\}$$

Gradient of $\mathrm{dist}^2(\mathbf{z}, \mathcal{K})$ is

$$\nabla \mathrm{dist}^2(\mathbf{y}, \mathcal{K}) = 2(\mathbf{y} - \mathrm{proj}_{\mathcal{K}}(\mathbf{y})).$$

Hence, HCGM can be generalized by changing lmo step as

$$\hat{\mathbf{x}}^k := \mathrm{lmo}_{\mathcal{X}}(\nabla f(\mathbf{x}^k) + \beta_k^{-1} \mathbf{A}^T(\mathbf{A}\mathbf{x}^k - \mathbf{b} - \mathrm{proj}_{\mathcal{K}}(\mathbf{A}\mathbf{x}^k - \mathbf{b}))).$$

Same guarantees hold, by replacing $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|$ by $\mathrm{dist}(\mathbf{A}\mathbf{x} - \mathbf{b}, \mathcal{K})$.

*Generalization of CGAL for $\mathbf{A}\mathbf{x}-\mathbf{b}\in\mathcal{K}$

Augmented Lagrangian for $\min\{f(\mathbf{x}):\mathbf{A}\mathbf{x}-\mathbf{b}\in\mathcal{K},\mathbf{x}\in\mathcal{X}\}$

Similarly, CGAL can be extended for $\mathbf{A}\mathbf{x}-\mathbf{b}\in\mathcal{K}$ constraint, by replacing

- lmo step as

$$\hat{\mathbf{x}}^k := \mathrm{lmo}_{\mathcal{X}}\left(\nabla f(\mathbf{x}^k) + \mathbf{A}^T\lambda^k + \beta_k^{-1}\mathbf{A}^T\left(\mathbf{A}\mathbf{x}^k - \mathbf{b} - \mathrm{proj}_{\mathcal{K}}(\mathbf{A}\mathbf{x}^k - \mathbf{b} + \beta_k\lambda^k)\right)\right)$$

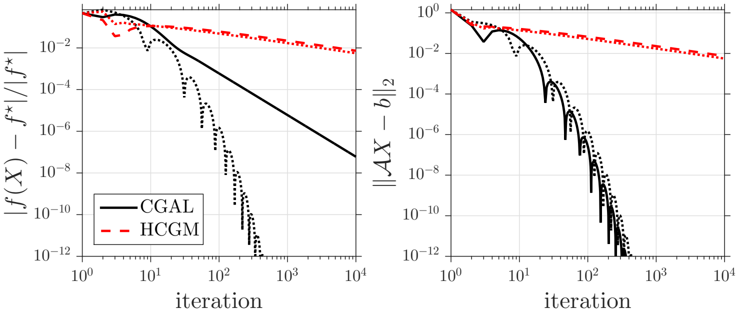
- and dual update step as

$$\lambda^{k+1} := \lambda^k + \omega_k\left(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{b} + \mathrm{proj}_{\mathcal{K}}(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{b} + \beta_{k+1}\lambda^k)\right)$$

Same guarantees hold, by replacing $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|$ by $\mathrm{dist}(\mathbf{A}\mathbf{x}-\mathbf{b},\mathcal{K})$.

Example: Generalized eigenvalue problem

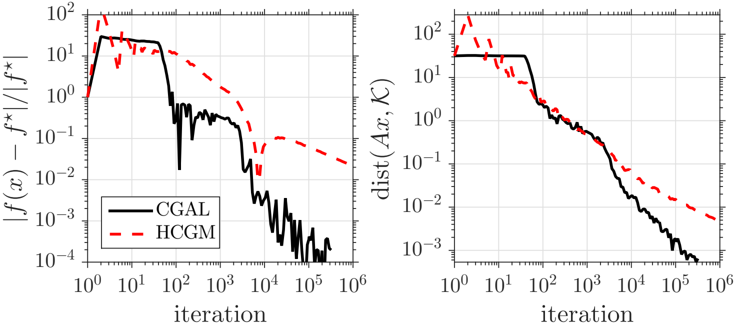
$$\max_{\mathbf{X}\in\mathbb{R}^{p\times p}} \left\{ \mathrm{Tr}\left(\mathbf{BX}\right) : \mathrm{Tr}\left(\mathbf{AX}\right) = 1,\ \mathbf{X} \in \mathcal{S}_{+}^{p},\ \mathrm{Tr}\left(\mathbf{X}\right) \leq \alpha \right\}$$



- $\mathbf{A}$ and $\mathbf{B}$ generated synthetically with iid Gaussian entries.
- $p = 1000$
- $\alpha > 0$ is a model parameter
- Dotted lines represent $\hat{\mathbf{X}}^k$ (output of lmo)

Example: k-means clustering

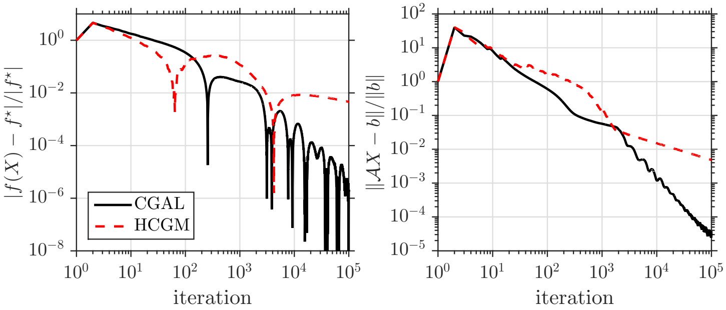
$$\min_{\mathbf{X}\in\mathbb{R}^{p\times p}} \left\{ \mathrm{Tr}\left(\mathbf{X}\right) : \mathbf{X1} = \mathbf{1},\ \mathbf{X} \geq 0,\ \mathbf{X} \in \mathcal{S}_{+}^{p},\ \mathrm{Tr}\left(\mathbf{X}\right) = \alpha \right\}$$



- Test setup with preprocessed MNIST dataset [13]
- $p = 1000$
- $\alpha = 10$ is the number of clusters

Example: Max-cut SDP

$$\max_{\mathbf{X}\in\mathbb{R}^{p\times p}} \left\{ \frac{1}{4}\mathrm{Tr}\left(\mathbf{L}\mathbf{X}\right) : \mathrm{diag}(\mathbf{X}) = \mathbf{1},\ \mathbf{X}\in\mathcal{S}_+^p,\ \mathrm{Tr}\left(\mathbf{X}\right) = p \right\}$$



- UF Sparse graphs: GSet collection, G40 dataset $p = 2000$
- $\mathbf{L}$ is graph Laplacian matrix.

*CGM as approximation method for subsolvers

Recall projection oracle

Projection (of $\mathbf{z}$ onto $\mathcal{X}$) oracle returns the solution of the following problem:

$$\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\mathbf{x}-\mathbf{z}\|_2^2 : \quad \mathbf{x}\in\mathcal{X} \right\}$$

CGM applies to this problem.

Conditional gradient sliding [6]

- Consider ISTA or FISTA for solving (8).
- Replace projection step with approximate projection oracle.
- Approximate projection using CGM.

Inexact augmented Lagrangian method (with CGM) [7]

Similar ideas works for more general templates.

- Consider augmented Lagrangian (AL) method for solving (7).
- Replace solution AL subproblem with approximate solution of AL subproblem.
- Approximate solution of AL subproblem using CGM.

A basic constrained stochastic problem

Problem setting (Stochastic)

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \mathbb{E}[f(\mathbf{x},\theta)] : \mathbf{x} \in \mathcal{X} \right\}, \tag{8}$$

Assumptions

- θ is a random vector whose probability distribution is supported on set Θ
- $\mathcal{X}$ is nonempty, convex, closed and bounded.
- $f(\cdot,\theta) \in \mathcal{F}_L^{1,1}(\mathbb{R}^p)$ for all θ (i.e., convex with Lipschitz gradient).

Example (Finite-sum model)

$$\mathbb{E}[f(\mathbf{x},\theta)] = \frac{1}{n}\sum_{j=1}^{n} f_j(\mathbf{x})$$

- $j = \theta$ is a drawn uniformly from $\Theta = \{1, 2, \ldots, n\}$
- $f_j \in \mathcal{F}_L^{1,1}(\mathbb{R}^p)$ for all j (i.e., convex with Lipschitz gradient).

Stochastic conditional gradient method - I

Stochastic conditional gradient method (SFW1)
1. Choose $\mathbf{x}^0 \in \mathcal{X}$. **2.** For $k = 0, 1, \ldots$ perform: $$\begin{cases} \hat{\mathbf{x}}^k & := \mathrm{lmo}_{\mathcal{X}}(\tilde{\nabla} f(\mathbf{x}^k, \theta_k)) \\ \mathbf{x}^{k+1} & := (1-\gamma_k)\mathbf{x}^k + \gamma_k \hat{\mathbf{x}}^k, \end{cases}$$ where $\gamma_k := \frac{2}{k+2}$, and $\tilde{\nabla} f$ is an unbiased estimator of ∇f.

Theorem [3]

Assume that the following variance condition holds

$$\mathbb{E}\left\|\nabla f(\mathbf{x}^k) - \tilde{\nabla} f(\mathbf{x}^k, \theta_k)\right\|^2 \leq \left(\frac{LD}{k+1}\right)^2. \qquad (\star)$$

Then, the iterates of SFW satisfies

$$\mathbb{E}[f(\mathbf{x}^k, \theta)] - f^\star \leq \frac{4LD^2}{k+1}.$$

$(\star)$ $\rightarrow$ SFW requires decreasing variance!

Stochastic conditional gradient method - I

Stochastic conditional gradient method (SFW1)
1. Choose $\mathbf{x}^0 \in \mathcal{X}$. **2.** For $k = 0, 1, \ldots$ perform: $\begin{cases} \hat{\mathbf{x}}^k & := \mathrm{lmo}_{\mathcal{X}}(\tilde{\nabla} f(\mathbf{x}^k, \theta_k)) \\ \mathbf{x}^{k+1} & := (1-\gamma_k)\mathbf{x}^k + \gamma_k \hat{\mathbf{x}}^k, \end{cases}$ where $\gamma_k := \frac{2}{k+2}$, and $\tilde{\nabla} f$ is an unbiased estimator of ∇f.

Example (Finite-sum model)

$$\mathbb{E}[f(\mathbf{x}, \theta)] = \frac{1}{n}\sum_{j=1}^{n} f_j(\mathbf{x})$$

Assume f_j is G-Lipschitz continuous for all j. Suppose that $\mathcal{S}_k$ is a random sampling (with replacement) from $\Theta = \{1, 2, \ldots, n\}$. Then,

$$\tilde{\nabla} f(\mathbf{x}^k, \theta_k) := \frac{1}{|\mathcal{S}_k|}\sum_{j \in \mathcal{S}_k} f_j(\mathbf{x}^k) \quad \Longrightarrow \quad \mathbb{E}\left\|\nabla f(\mathbf{x}) - \tilde{\nabla} f(\mathbf{x}, \theta_k)\right\|^2 \leq \frac{G^2}{|\mathcal{S}_k|}.$$

Hence, by choosing $|\mathcal{S}_k| = (\frac{G(k+1)}{LD})^2$ we satisfy the variance condition for SFW.

Stochastic conditional gradient method - II

Stochastic conditional gradient method (SFW2)

1. Choose $\mathbf{x}^0 \in \mathcal{X}$ and set $\mathbf{z}^0 = \mathbf{0}$.
2. For $k = 0, 1, \ldots$ perform:

$$\begin{cases} \mathbf{z}^{k+1} & := (1-\rho_k)\mathbf{z}^k + \rho_k \tilde{\nabla} f(\mathbf{x}^k, \theta_k) \\ \hat{\mathbf{x}}^k & := \mathrm{lmo}_{\mathcal{X}}(\mathbf{z}^{k+1}) \\ \mathbf{x}^{k+1} & := (1-\gamma_k)\mathbf{x}^k + \gamma_k \hat{\mathbf{x}}^k, \end{cases}$$

where $\gamma_k := \frac{9}{k+8}$, and $\rho_k = \frac{4}{(k+8)^{2/3}}$.

Theorem [9]

Assume that the unbiased estimator $\tilde{\nabla} f$ has a bounded variance, i.e.,

$$\mathbb{E}\left\|\nabla f(\mathbf{x}^k) - \tilde{\nabla} f(\mathbf{x}^k, \theta_k)\right\|^2 \leq \sigma^2 \quad \text{for some } \sigma < \infty.$$

Then, the iterates of SFW2 satisfies $\mathbb{E}[f(\mathbf{x}^k, \theta)] - f^\star \leq \frac{Q}{(k+9)^{1/3}}$,

where $Q := \max\left\{9^{1/3}(f(\mathbf{x}^0) - f^\star), \frac{LD^2}{2} + 2D\max\left\{2\left\|\nabla f(\mathbf{x}^0)\right\|, \sqrt{16\sigma^2 + 2L^2D^2}\right\}\right\}$.

Slower rate than SFW1, but requires a single datapoint each iteration in finite-sum!

Stochastic CGM with quadratic penalty

Stochastic homotopy conditional gradient method (SHCGM)

1. Choose $\mathbf{x}^0 \in \mathcal{X}$, $\beta_0 > 0$, and set $\mathbf{z}^0 = \mathbf{0}$.
2. For $k = 0, 1, \ldots$ perform:

$$\begin{cases} \mathbf{z}^{k+1} & := (1-\rho_k)\mathbf{z}^k + \rho_k \tilde{\nabla} f(\mathbf{x}^k, \theta_k) \\ \hat{\mathbf{x}}^k & := \mathrm{lmo}_{\mathcal{X}}(\mathbf{z}^{k+1} + \beta_k^{-1}\mathbf{A}^T(\mathbf{A}\mathbf{x}^k - \mathbf{b})) \\ \mathbf{x}^{k+1} & := (1-\gamma_k)\mathbf{x}^k + \gamma_k \hat{\mathbf{x}}^k, \end{cases}$$

where $\gamma_k := \frac{9}{k+8}$, $\rho_k = \frac{4}{(k+8)^{2/3}}$, and $\beta_k = \frac{\beta_0}{(k+8)^{1/2}}$.

SHCGM template and convergence rates [8]

$$f^\star := \min_{\mathbf{x} \in \mathbb{R}^p} \left\{ \mathbb{E}[f(\mathbf{x}, \theta)] : \mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \in \mathcal{X} \right\},$$

SHCGM is the combination of HCGM and SFW2. Iterates converges with

$$\begin{cases} \mathbb{E} f(\mathbf{x}^k, \theta) - f^\star & \geq & -\|y^\star\| \cdot \mathbb{E}\|\mathbf{A}\mathbf{x} - \mathbf{b}\| \\ \mathbb{E} f(\mathbf{x}^k, \theta) - f^\star & \in & \mathcal{O}\left(\frac{1}{k^{1/3}}\right) \\ \mathbb{E}\|\mathbf{A}\mathbf{x} - \mathbf{b}\| & \in & \mathcal{O}\left(\frac{1}{k^{5/12}}\right) \end{cases}$$

A basic constrained non-convex problem

Problem setting

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) : \mathbf{x} \in \mathcal{X} \right\},$$

Assumptions

- $\mathcal{X}$ is nonempty, convex, closed and bounded.
- f has L-Lipschitz continuous gradients, but it is non-convex.

Stationary point

Due to constraints, $\|\nabla f(\mathbf{x}^\star)\| = 0$ may not hold!

Frank-Wolfe gap: Following measure, known as FW-gap, generalizes the definition of stationary point for constrained problems:

$$g_{FW}(\mathbf{x}) := \max_{\mathbf{y}\in\mathcal{X}} \ (\mathbf{x} - \mathbf{y})^T \nabla f(\mathbf{x})$$

- $g_{FW}(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathcal{X}$.
- $\mathbf{x} \in \mathcal{X}$ is a stationary point if and only if $g_{FW}(\mathbf{x}) = 0$.

CGM for non-convex problems

CGM for non-convex problems

1. Choose $\mathbf{x}^0 \in \mathcal{X}$, $K > 0$ total number of iterations.
2. For $k = 0, 1, \ldots, K-1$ perform:

$$\begin{cases} \hat{\mathbf{x}}^k & := \mathrm{lmo}_{\mathcal{X}}(\nabla f(\mathbf{x}^k)) \\ \mathbf{x}^{k+1} & := (1-\gamma_k)\mathbf{x}^k + \gamma_k \hat{\mathbf{x}}^k, \end{cases}$$

where $\gamma_k := \frac{1}{\sqrt{K+1}}$.

Theorem

Denote $\bar{\mathbf{x}}$ chosen uniformly random from $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^K\}$. Then, CGM satisfies

$$\min_{k=1,2,\ldots,K} g_{FW}(\mathbf{x}^k) \ \leq \ \mathbb{E}[g_{FW}(\bar{\mathbf{x}})] \ \leq \ \frac{1}{\sqrt{K}} \left(f(\mathbf{x}^0) - f^\star + \frac{LD^2}{2} \right).$$

$\star$ There exist stochastic CGM methods for non-convex problems. See [10] for details.

References I

[1] Emmanuel J Candes, T. Strohmer, and V. Voroninski.
Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming.
IEEE Trans. Signal Processing, 60(5):2422–2432, 2012.

[2] Marguerite Frank and Philip Wolfe.
An algorithm for quadratic programming.
Naval Res. Logis. Quart., 3:95–110, 1956.

[3] E. Hazan and H. Luo.
Variance-reduced and projection-free stochastic optimization.
In *Proc. 33rd Int. Conf. Machine Learning*, 2016.

[4] Martin Jaggi.
Revisiting Frank-Wolfe: Projection-free sparse convex optimization.
In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, 2013.

[5] Simon Lacoste-Julien, Martin Jaggi, Mark Schmidt, and Patrick Pletscher.
Block-coordinate Frank-Wolfe optimization for structural SVMs.
In *Proc. 30th International Conference on Machine Learning*, 2013.

References II

[6] Guanghui Lan and Yi Zhou.
Conditional gradient sliding for convex optimization.
SIAM J. Optim., 26(2):1379–1409, 2016.

[7] Ya-Feng Liu, Xin Liu, and Shiqian Ma.
On the non-ergodic convergence rate of an inexact augmented lagrangian framework for composite convex programming.
arXiv preprint arXiv: 1603.05738v3, 2018.

[8] F. Locatello, A. Yurtsever, O. Fercoq, and V. Cevher.
Stochastic conditional gradient method with ane constraints.
Technical report, 2018.

[9] Aryan Mokhtari, Hamed Hassani, and Amin Karbasi.
Stochastic conditional gradient methods: From convex minimization to submodular maximization.
arXiv preprint arXiv:1804.09554, 2018.

[10] Sashank J Reddi, Suvrit Sra, Barnabás Póczos, and Alex Smola.
Stochastic frank-wolfe methods for nonconvex optimization.
arXiv preprint arXiv:1607.08254, 2016.

References III

[11] Alp Yurtsever, Olivier Fercoq, and Volkan Cevher.
A conditional gradient-based augmented lagrangian framework.
Technical report, 2018.

[12] Alp Yurtsever, Ya-Ping Hsieh, and Volkan Cevher.
Scalable convex methods for phase retrieval.
In *$6th$ IEEE Intl. Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, 2015.

[13] Alp Yurtsever, Fercoq Olivier, Locatello Francesco, and Volkan Cevher.
A conditional gradient framework for composite convex minimization with applications to semidefinite programming.
In *Proceedings of the 35th International Conference on International Conference on Machine Learning - Volume 28*, ICML'18, 2018.