Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 11: Constrained convex minimization II

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

EE-556 (Fall 2019)
This work is released under a Creative Commons License with the following terms:

**Attribution**
- The licensor permits others to copy, distribute, display, and perform the work. In return, licensees must give the original authors credit.

**Non-Commercial**
- The licensor permits others to copy, distribute, display, and perform the work. In return, licensees may not use the work for commercial purposes – unless they get the licensor’s permission.

**Share Alike**
- The licensor permits others to distribute derivative works only under a license identical to the one that governs the licensor’s work.

**Full Text of the License**
Outline

This class:

1. Linear minimization oracle
2. Conditional gradient method (CGM)
3. CGM-type methods for problems with affine constraints

Next class

1. Alternating primal-dual methods
Recommended reading material


Motivation

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

In this lecture, we will cover the \textit{conditional gradient}-type methods (a.k.a., Frank-Wolfe algorithm). These methods leverage the so called \textit{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 \( x \in \mathbb{R}^p \). The proximal operator of \( g \) is defined as:

\[
\text{prox}_g(x) \equiv \arg \min_{y \in \mathbb{R}^p} \left\{ g(y) + \frac{1}{2} \|y - x\|_2^2 \right\}.
\] (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 \( \text{prox}_g \) defined by (1) can be computed efficiently.

- "efficiently" = \{closed form solution, low-cost computation, polynomial time\}.
- We denote \( \mathcal{F}_{\text{prox}}(\mathbb{R}^p) \) 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!

<table>
<thead>
<tr>
<th>Name</th>
<th>Function</th>
<th>Proximal operator</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\ell_1)-norm</td>
<td>(f(x) := |x|_1)</td>
<td>(\text{prox}<em>\lambda f(x) = \text{sign}(x) \otimes [|x| - \lambda]</em>+)</td>
<td>(O(p))</td>
</tr>
<tr>
<td>(\ell_2)-norm</td>
<td>(f(x) := |x|_2)</td>
<td>(\text{prox}_\lambda f(x) = \left[1 - \lambda/|x|<em>2\right]</em>+ x)</td>
<td>(O(p))</td>
</tr>
<tr>
<td>Support function</td>
<td>(f(x) := \max_{y \in C} x^T y)</td>
<td>(\text{prox}_\lambda f(x) = x - \lambda \pi_C(x))</td>
<td>(O(p))</td>
</tr>
<tr>
<td>Box indicator</td>
<td>(f(x) := \delta_{[a,b]}(x))</td>
<td>(\text{prox}<em>\lambda f(x) = \pi</em>{[a,b]}(x))</td>
<td>(O(p))</td>
</tr>
<tr>
<td>Positive semidefinite cone indicator</td>
<td>(f(X) := \delta_{\mathbb{S}^p_+}(X))</td>
<td>(\text{prox}<em>\lambda f(X) = U[\Sigma]</em>+ U^T, \text{where } X = U\Sigma U^T)</td>
<td>(O(p^3))</td>
</tr>
<tr>
<td>Hyperplane indicator</td>
<td>(f(x) := \delta_{\mathcal{X}}(x),\mathcal{X} := {x : a^T x = b})</td>
<td>(\text{prox}<em>\lambda f(x) = \pi</em>{\mathcal{X}}(x) = x + \left(\frac{b-a^T x}{|a|_2}\right)a)</td>
<td>(O(p))</td>
</tr>
<tr>
<td>Simplex indicator</td>
<td>(f(x) := \delta_{\mathcal{X}}(x),\mathcal{X} := {x : x \geq 0, 1^T x = 1})</td>
<td>(\text{prox}_\lambda f(x) = (x - \nu 1)) for some (\nu \in \mathbb{R}), which can be efficiently calculated</td>
<td>(\tilde{O}(p))</td>
</tr>
<tr>
<td>Convex quadratic</td>
<td>(f(x) := \frac{1}{2}x^T Q x + q^T x)</td>
<td>(\text{prox}_\lambda f(x) = (\lambda I + Q)^{-1} x)</td>
<td>(O(p \log p) \rightarrow O(p^3))</td>
</tr>
<tr>
<td>Square (\ell_2)-norm</td>
<td>(f(x) := \frac{1}{2}|x|_2^2)</td>
<td>(\text{prox}_\lambda f(x) = \frac{1}{1 + \lambda} x)</td>
<td>(O(p))</td>
</tr>
<tr>
<td>log-function</td>
<td>(f(x) := -\log(x))</td>
<td>(\text{prox}_\lambda f(x) = \frac{(x^2 + 4\lambda)^{1/2} + x}{2})</td>
<td>(O(1))</td>
</tr>
<tr>
<td>log det-function</td>
<td>(f(x) := -\log \det(X))</td>
<td>(\text{prox}_\lambda f(X) = \text{log-function prox applied to the individual eigenvalues of } X)</td>
<td>(O(p^3))</td>
</tr>
</tbody>
</table>

Here: \([x]_+ := \max\{0,x\}\) and \(\delta_{\mathcal{X}}\) is the indicator function of the convex set \(\mathcal{X}\), \(\text{sign}\) is the sign function, \(\mathbb{S}^p_+\) is the cone of symmetric positive semidefinite matrices.
Example: \( \text{prox} \) for the indicator of a nuclear-norm ball

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

**Proximal operator of \( \delta_{\mathcal{X}}(X) \)**

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

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

This can be computed as follows:

- Compute SVD of \( X \) \( \implies U\Sigma V^T = X \).
- Form a vector \( s \in \mathbb{R}^p \) by the diagonal entries of \( \Sigma \) \( \implies s = \text{diag} \left( \Sigma \right) \).
- Project \( s \) onto \( \ell_1 \) norm ball \( \implies \hat{s} = \arg \min_x \{ \|s - x\| : \|x\|_1 \leq \alpha \} \).
- Form a diagonal matrix with entries \( \hat{s} \) \( \implies \hat{\Sigma} = \text{diag}^*(\hat{s}) \).
- Form the output \( \implies \text{proj}_{\mathcal{X}}(X) = U\hat{\Sigma}V^T \)

Finding SVD is costly in when \( p \) is big!
A basic constrained problem setting

Problem setting

\[
f^* := \min_{x \in \mathbb{R}^p} \left\{ f(x) : x \in \mathcal{X} \right\}, \tag{2}
\]

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

<table>
<thead>
<tr>
<th>Basic proximal-gradient scheme (ISTA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Choose ( x^0 \in \text{dom}(F) ) arbitrarily as a starting point.</td>
</tr>
<tr>
<td>2. For ( k = 0, 1, \ldots ), generate a sequence ( {x^k}_{k \geq 0} ) as:</td>
</tr>
</tbody>
</table>
| \[
x^{k+1} := \text{prox}_{\alpha g} \left( x^k - \alpha \nabla f(x^k) \right)
\]
| where \( \alpha := \frac{1}{L} \). |

- Prox-operator of indicator of \( \mathcal{X} \) is projection onto \( \mathcal{X} \) \( \implies \) ensures feasibility

How else can we ensure feasibility?
Frank-Wolfe’s approach - 1

\[ f^* := \min_{x \in \mathbb{R}^p} \left\{ f(x) : x \in \mathcal{X} \right\}, \]

**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 \( x^k \)
   \[ \phi_k(x) := f(x^k) + \nabla f(x^k)^T (x - x^k) \]

2. Minimize this approximation within constraint set
   \[ \hat{x}^k \in \min_{x \in \mathcal{X}} \phi_k(x) = \min_{x \in \mathcal{X}} \nabla f(x^k)^T x \]

3. Take a step towards \( \hat{x}^k \) with step-size \( \gamma_k \in [0, 1] \)
   \[ x^{k+1} = x^k + \gamma_k (\hat{x}^k - x^k) \]

\( x^{k+1} \) is feasible since it is convex combination of two other feasible points.
Frank-Wolfe’s approach - II

\[ f^* := \min_{x \in \mathbb{R}^p} \left\{ f(x) : x \in \mathcal{X} \right\} \]

\[ \mathcal{X} : f(x) \leq f(x^k) \]

\[ x^k \]

\[ -\nabla f(x^k) \]

\[ \hat{x}^k \]

\[ x^{k+1} \]

\[ \{ \mathcal{X} : f(x) \leq f(x^k) \} \]

---

**Conditional gradient method (CGM)**

1. Choose \( x^0 \in \mathcal{X} \).
2. For \( k = 0, 1, \ldots \) perform:

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

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 $\mathcal{X}$ ($\text{lmo}_\mathcal{X}$) returns a vector $\hat{x}$ such that

$$\text{lmo}_\mathcal{X}(x) := \hat{x} \in \arg\min_{y \in \mathcal{X}} x^T y$$

(3)

- $\text{lmo}_\mathcal{X}$ returns an extreme point of $\mathcal{X}$.
- $\text{lmo}_\mathcal{X}$ is arguably cheaper than projection.
- $\text{lmo}_\mathcal{X}$ is not single valued, note $\in$ in the definition.
Consider $\delta_\mathcal{X}$, the indicator of nuclear-norm ball $\mathcal{X} := \{ \mathbf{X} : \mathbf{X} \in \mathbb{R}^{p \times p}, \|\mathbf{X}\|_* \leq \alpha \}$

**lmo of nuclear-norm ball**

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

This can be computed as follows:

- Compute top singular vectors of $\mathbf{X}$ $\implies (u_1, \sigma_1, v_1) = \text{svds}(\mathbf{X}, 1)$.
- Form the rank-1 output $\implies \mathbf{X} = -u_1 \alpha v_1^T$

We can efficiently approximate top singular vectors by power method!
Convergence guarantees of CGM

Problem setting

\[ f^* := \min_{x \in \mathbb{R}^p} \left\{ f(x) : x \in X \right\}, \]

Assumptions

- \( X \) 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

\[
f(x^k) - f(x^*) \leq \frac{4LD_X}{k+1} \tag{4}\]

where \( D_X := \max_{x,y \in X} \|x - y\|_2 \) 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 $f \in \mathcal{F}_{L}^{1,1}$

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

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

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

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

$$\begin{aligned}
&\langle \nabla f(x^k), \hat{x}^k - x^k \rangle = \min_{x \in \mathcal{X}} \nabla f(x^k), x - x^k \rangle \\
&\|\hat{x}^k - x^k\|_2 \leq \max_{x, y \in \mathcal{X}} \|x - y\|_2 = D_\mathcal{X}
\end{aligned}$$

Substituting into (5) and subtracting $f^*$ we get

$$f(x^{k+1}) - f^* \leq (1 - \gamma_k)(f(x^k) - f^*) + \gamma_k^2 \frac{L}{2} D_\mathcal{X}^2.$$
Proof of convergence rate of CGM - part II (self study)

\[
f(x^{k+1}) - f^* \leq (1 - \gamma_k)(f(x^k) - f^*) + \gamma_k^2 \frac{L}{2} D^2 \chi
\]

Proof (Continued)

We will use induction technique: First note
\[
\gamma_0 = 1 \quad \implies \quad f(x^1) - f^* \leq \frac{1}{2} LD^2 \chi
\]

Now, suppose (4) holds, then
\[
f(x^{k+1}) - f^* \leq (1 - \gamma_k) \frac{4LD \chi}{k + 1} + \gamma_k^2 \frac{L}{2} D^2 \chi
\]

\[
= \frac{k}{k + 2} \frac{4LD \chi}{k + 1} + \frac{4}{(k + 2)^2} \frac{L}{2} D^2 \chi \leq \frac{4LD \chi}{k + 2}
\]

which completes the proof by induction.
Example: Phase retrieval

Phase retrieval

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

$$b_i = \left| \langle a_i, x^\dagger \rangle \right|^2 + \omega_i.$$  

($a_i \in \mathbb{C}^p$ are known measurement vectors, $\omega_i$ models noise).

- Non-linear measurements $\rightarrow$ non-convex maximum likelihood estimators.

PhaseLift [1]

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

- semidefinite relaxation ($x^\dagger x^\dagger^H = X^\dagger$)  
- convex relaxation ($\text{rank} \rightarrow \| \cdot \|_*$)

albeit in terms of the lifted variable $X \in \mathbb{C}^{p \times p}$.
Example: Phase retrieval - II

Problem formulation
We solve the following PhaseLift variant:

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

Experimental setup [13]
Coded diffraction pattern measurements, \( b = [b_1, \ldots, b_L] \) with \( L = 20 \) different masks

\[
 b_\ell = |\text{fft}((d_\ell^H \odot x^h))|^2
\]

→ \( \odot \) denotes Hadamard product; \( | \cdot |^2 \) applies element-wise
→ \( d_\ell \) are randomly generated octonary masks (distributions as proposed in [1])
→ Parametric choices: \( \lambda^0 = 0^n; \ \epsilon = 10^{-2}; \ \kappa = \text{mean}(b) \).
Test with synthetic data: Prox vs sharp

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

→ 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

\[ f^* := \min_{x \in \mathbb{R}^p} \left\{ f(x) : Ax = b, \ x \in X \right\}, \]  

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

Classical CGM does not apply to (7)

- \( \text{Imo} \) of the intersection of \( \{x : Ax = b\} \) and \( X \) is difficult to compute.
CGM with quadratic penalty

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

A quadratic penalty formulation:

\[
\min_{x \in \mathbb{R}^p} \left\{ f(x) + \frac{1}{2\beta} \|Ax - b\|^2 : x \in \mathcal{X} \right\}
\]

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

Our strategy \([14]\) \( \Rightarrow \) Take a CGM step on \( f_\beta \) and decrease \( \beta \) progressively to 0

<table>
<thead>
<tr>
<th>Homotopy conditional gradient method (HCGM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Choose ( x^0 \in \mathcal{X} ), and ( \beta_0 &gt; 0 ).</td>
</tr>
<tr>
<td>2. For ( k = 0, 1, \ldots ) perform:</td>
</tr>
</tbody>
</table>
| \[
\begin{align*}
\hat{x}^k &:= \text{lmo}_{\mathcal{X}}(\nabla f(x^k) + \beta_k^{-1} A^T(Ax^k - b)) \\
x^{k+1} &:= (1 - \gamma_k)x^k + \gamma_k \hat{x}^k,
\end{align*}
\]
| where \( \gamma_k := \frac{2}{k+2} \) and \( \beta_k = \frac{\beta_0}{\sqrt{k+2}} \). |
Convergence guarantees of HCGM

Recall Lagrange duality

\[ \mathcal{L}(x, \lambda) := f(x) + \langle \lambda, Ax - b \rangle \]

\[
\max_{\lambda} \min_{x \in \mathcal{X}} \mathcal{L}(x, \lambda) \leq \min_{x \in \mathcal{X}} \max_{\lambda} \mathcal{L}(x, \lambda) \quad \text{(Duality)}
\]

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

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

Theorem

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

\[
\begin{cases}
- \|Ax^k - b\| \|\lambda^*\| \leq f(x^k) - f^* \leq 2D_x \left( \frac{L}{k+1} + \frac{\|A\|^2}{\beta_0 \sqrt{k+1}} \right) \\
\|Ax^k - b\| \leq \frac{2\beta_0}{\sqrt{k+1}} \left( \|\lambda^*\| + D_x \sqrt{\frac{L}{\beta_0} + \frac{\|A\|^2}{\beta_0^2}} \right)
\end{cases}
\]
Augmented Lagrangian CGM: CGAL

Quadratic penalty strategy for \( \min \{ f(x) : Ax = b, x \in X \} \)

Augmented problem formulation:

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

- Write down the Lagrangian:

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

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

Our strategy [12] \(\Rightarrow\) \[
\begin{align*}
1. & \text{ Take a CGM step wrt } \mathcal{L}_{1/\beta}(\cdot, \lambda) \\
2. & \text{ Take a gradient step wrt } \mathcal{L}_{1/\beta}(x, \cdot) \\
3. & \text{ Decrease } \beta \text{ progressively to } 0
\end{align*}
\]

Challenge: Step size in dual (step 2.)
Convergence guarantees of CGAL

<table>
<thead>
<tr>
<th>Conditional gradient augmented Lagrangian method (CGAL)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1.</strong> Choose $x^0 \in X$, $\lambda^0 \in \mathbb{R}^n$, and $\beta_0 &gt; 0$.</td>
</tr>
<tr>
<td><strong>2.</strong> For $k = 0, 1, \ldots$ perform:</td>
</tr>
</tbody>
</table>
| \[
\begin{align*}
\hat{x}^k & := \text{lmo}_X(\nabla f(x^k) + A^T \lambda^k + \beta_k^{-1} A^T (Ax^k - b)) \\
x^{k+1} & := (1 - \gamma_k)x^k + \gamma_k \hat{x}^k \\
\lambda^{k+1} & := \lambda^k + \omega_k (Ax^{k+1} - b)
\end{align*}
\]
| 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 $\omega_k$ by the following rule

\[
\omega_k = \alpha_k := \min\left\{ \frac{1}{\beta_0}, \frac{\eta^2_k (L_f + \lambda_{k+1}) D_X^2}{2 \|Ax^{k+1} - b\|^2} \right\} \quad \text{if} \quad \|\lambda^k + \alpha_k (Ax^{k+1} - b)\| \leq D_Y
\]

and $\omega_k = 0$ otherwise, for some $D_Y \geq 0$. Then, the iterates of CGAL satisfies

\[
\begin{align*}
-\|Ax^k - b\| \|\lambda^*\| & \leq f(x^k) - f^* \leq 4D_X \left( \frac{L}{k+1} + \frac{\|A\|^2}{\beta_0 \sqrt{k+1}} \right) + \frac{\beta_0 D_Y}{2 \sqrt{k+1}} \\
\|Ax^k - b\| & \leq \frac{2\beta_0}{\sqrt{k+1}} \left( \frac{3D_Y}{2} + \|\lambda^*\| + \frac{D_X \sqrt{L\beta_0 + \|A\|^2}}{\beta_0} \right)
\end{align*}
\]
*Generalization of HCGM for $Ax - b \in \mathcal{K}$

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

Define the distance function

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

Quadratic penalty takes the form

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

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

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

Hence, HCGM can be generalized by changing lmo step as

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

Same guarantees hold, by replacing $\|Ax - b\|$ by $\text{dist}(Ax - b, \mathcal{K})$. 
Generalization of CGAL for $Ax - b \in \mathcal{K}$

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

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

- lmo step as

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

- and dual update step as

$$\lambda^{k+1} := \lambda^k + \omega_k \left( Ax^{k+1} - b + \text{proj}_\mathcal{K}(Ax^{k+1} - b + \beta_{k+1} \lambda^k) \right)$$

Same guarantees hold, by replacing $\|Ax - b\|$ by $\text{dist}(Ax - b, \mathcal{K})$. 
Example: Generalized eigenvalue problem

\[
\max_{X \in \mathbb{R}^{p \times p}} \left\{ \text{Tr}(BX) : \text{Tr}(AX) = 1, \ X \in S^p_+, \ \text{Tr}(X) \leq \alpha \right\}
\]

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

\[ \min_{\mathbf{X} \in \mathbb{R}^{p \times p}} \left\{ \text{Tr} (\mathbf{X}) : \mathbf{X} \mathbf{1} = 1, \, \mathbf{X} \geq 0, \, \mathbf{X} \in S_+^p, \, \text{Tr} (\mathbf{X}) = \alpha \right\} \]

- Test setup with preprocessed MNIST dataset [14]
- \( p = 1000 \)
- \( \alpha = 10 \) is the number of clusters
Example: Max-cut SDP

\[
\max_{\mathbf{X} \in \mathbb{R}^{p \times p}} \left\{ \frac{1}{4} \text{Tr}(\mathbf{LX}) : \text{diag}(\mathbf{X}) = 1, \quad \mathbf{X} \in S^p_+, \quad \text{Tr}(\mathbf{X}) = p \right\}
\]

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

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

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 work 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)

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

(8)

Assumptions

- \( \theta \) is a random vector whose probability distribution is supported on set \( \Theta \)
- \( \mathcal{X} \) is nonempty, convex, closed and bounded.
- \( f(\cdot, \theta) \in \mathcal{F}_{1,1}^1(\mathbb{R}^p) \) for all \( \theta \) (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 \( \Theta = \{1, 2, \ldots, n\} \)
- \( f_j \in \mathcal{F}_{1,1}^1(\mathbb{R}^p) \) for all \( j \) (i.e., convex with Lipschitz gradient).
Stochastic conditional gradient method - I

<table>
<thead>
<tr>
<th>Stochastic conditional gradient method (SFW1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Choose $x^0 \in \mathcal{X}$.</td>
</tr>
<tr>
<td>2. For $k = 0, 1, \ldots$ perform:</td>
</tr>
</tbody>
</table>
|   $\left\{ \begin{array}{l}
   \hat{x}^k \equiv \text{im}o_{\mathcal{X}}(\tilde{\nabla} f(x^k, \theta_k)) \\
   x^{k+1} \equiv (1 - \gamma_k)x^k + \gamma_k \hat{x}^k,
   \end{array} \right.$ |
| where $\gamma_k := \frac{2}{k+2}$, and $\tilde{\nabla} f$ is an unbiased estimator of $\nabla f$. |

Theorem [3]
Assume that the following variance condition holds

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

(\star)

Then, the iterates of SFW satisfies

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

(\star) $\rightarrow$ SFW requires decreasing variance!
Stochastic conditional gradient method - I

<table>
<thead>
<tr>
<th>Stochastic conditional gradient method (SFW1)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1.</strong> Choose $x^0 \in \mathcal{X}$.</td>
</tr>
<tr>
<td><strong>2.</strong> For $k = 0, 1, \ldots$ perform:</td>
</tr>
</tbody>
</table>
| \[
\begin{align*}
\hat{x}^k & := \underset{x \in \mathcal{X}}{\text{lmo}}(\tilde{\nabla} f(x^k, \theta_k)) \\
x^{k+1} & := (1 - \gamma_k)x^k + \gamma_k \hat{x}^k,
\end{align*}
\]
where $\gamma_k := \frac{2}{k+2}$, and $\tilde{\nabla} f$ is an unbiased estimator of $\nabla f$.

Example (Finite-sum model)

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

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

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

Hence, by choosing $|S_k| = \left( \frac{G(k+1)}{LD} \right)^2$ we satisfy the variance condition for SFW.
Stochastic conditional gradient method (SFW2)

1. Choose $x^0 \in X$ and set $z^0 = 0$.
2. For $k = 0, 1, \ldots$, perform:
   \[
   \begin{align*}
   z^{k+1} &:= (1 - \rho_k) z^k + \rho_k \tilde{\nabla} f(x^k, \theta_k) \\
   \hat{x}^k &:= \text{im}o_X(z^{k+1}) \\
   x^{k+1} &:= (1 - \gamma_k) x^k + \gamma_k \hat{x}^k,
   \end{align*}
   \]
   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.,

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

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

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

Slower rate than SFW1, but requires a single datapoint each iteration in finite-sum!
Stochastic CGM with quadratic penalty

<table>
<thead>
<tr>
<th>Stochastic homotopy conditional gradient method (SHCGM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Choose ( \mathbf{x}^0 \in \mathcal{X}, \beta_0 &gt; 0 ), and set ( \mathbf{z}^0 = 0 ).</td>
</tr>
<tr>
<td>2. For ( k = 0, 1, \ldots ) perform:</td>
</tr>
</tbody>
</table>
| \[
\begin{aligned}
\mathbf{z}^{k+1} &:= (1 - \rho_k)\mathbf{z}^k + \rho_k \tilde{\nabla} f(\mathbf{x}^k, \theta_k) \\
\hat{\mathbf{x}}^k &:= \text{lmo}_\mathcal{X}(\mathbf{z}^{k+1} + \beta_k^{-1} \mathbf{A}^T (\mathbf{A}\mathbf{x}^k - \mathbf{b})) \\
\mathbf{x}^{k+1} &:= (1 - \gamma_k)\mathbf{x}^k + \gamma_k \hat{\mathbf{x}}^k,
\end{aligned}
\]
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]

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

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

\[
\begin{aligned}
\mathbb{E} f(\mathbf{x}^k, \theta) - f^* &\geq - \|y^*\| \cdot \mathbb{E} \|\mathbf{A}\mathbf{x} - \mathbf{b}\| \\
\mathbb{E} f(\mathbf{x}^k, \theta) - f^* &\in \mathcal{O}\left(\frac{1}{k^{1/3}}\right) \\
\mathbb{E} \|\mathbf{A}\mathbf{x} - \mathbf{b}\| &\in \mathcal{O}\left(\frac{1}{k^{5/12}}\right)
\end{aligned}
\]
Example: Stochastic matrix completion

\[
\max_{X} \left\{ \sum_{(i,j) \in \Omega} (X_{i,j} - Y_{i,j})^2 : \|X\|_* \leq \beta_1, 1 \leq X \leq 5 \right\}
\]

- $\Omega$ is the set of observed entries from the true matrix $Y$
- $1 \leq X \leq 5$ is the hard threshold on the estimated ratings
- $\beta_1 = 7000$, MovieLens100k dataset, 100,000 ratings from 1682 users for 943 movies
Example: Online covariance matrix estimation

\[
\max_{X \in \mathbb{R}^{p \times p}} \{ \mathbb{E} \| X - \omega \omega^\top \|_F^2 : \| X \|_1 \leq \beta, X \in S_p^+, \text{Tr}(X) \leq \alpha \}
\]

- \( p = 1000, \alpha, \beta \) are model parameters
- Entries of vector \( \phi \) are selected uniformly at random from \([-1, 1]\)
- Covariance matrix \( \Sigma \) is block diagonal with 10 blocks of \( \phi \phi^\top \)
- Streaming observations \( \omega \in \mathcal{N}(0, \Sigma) \)
# A basic constrained non-convex problem

## Problem setting

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

### 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, $\|\nabla f(x^*)\| = 0$ may not hold!

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

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

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

<table>
<thead>
<tr>
<th>CGM for non-convex problems</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1.</strong> Choose $x^0 \in \mathcal{X}$, $K &gt; 0$ total number of iterations.</td>
</tr>
<tr>
<td><strong>2.</strong> For $k = 0, 1, \ldots, K - 1$ perform:</td>
</tr>
</tbody>
</table>
| \[
\begin{align*}
\hat{x}^k & := \text{lmo}_\mathcal{X}(\nabla f(x^k)) \\
x^{k+1} & := (1 - \gamma_k)x^k + \gamma_k \hat{x}^k,
\end{align*}
\] |
| where $\gamma_k := \frac{1}{\sqrt{K+1}}$. |

**Theorem**

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

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

Non-convex approach: Burer-Monteiro factorization

- SDP template:

$$\min_{x \in \mathbb{R}^p \times p} \left\{ \text{tr}(cx) : Ax = b, x \succeq 0, x^* = x, \text{tr}(x) = \rho \right\}$$

- Burer-Monteiro splitting

$$\min_{u \in \mathbb{R}^p \times r} \left\{ \text{tr}(cuu^*) : Auu^* = b, u \in \mathcal{U} := \{u : \|u\|_F \leq \sqrt{\rho}\} \right\}$$

- Nonlinear and non-convex problem ($Lu = Auu^* = b, \text{tr}(cuu^*)$)

- Local minima vs. saddle points issues

- Local minima vs. global minimum: $r = \Omega(\sqrt{p})$, due to Pataki and Barvinok

Barvinok, Problems of distance geometry and convex properties of quadratic maps, Disc Comp Geo, 1995.
Inexact augmented Lagrangian framework

• Our idea [11]: Solve primal subproblems with stricter tolerance, i.e., $\epsilon \to 0$

iALM:

$\begin{align*}
\text{Obtain } & u^+ \text{ such that } \\
& \text{dist}(-\nabla u \mathcal{L}_\beta(u^+, y), \partial g(u^+)) \leq \epsilon_f, \text{ or} \\
& \lambda_{\min}(\nabla_{uu} \mathcal{L}_\beta(u^+, y)) \geq -\epsilon_s
\end{align*}$

[1st order stationarity] [2nd order stationarity]

$\begin{align*}
y^+ &= y + \sigma \left( L(u^+) - b \right) \\
\text{Pick } & \beta^+ < \beta \text{ and } \epsilon^+ = \beta^+ \\
\text{Update } & \sigma^+ = \sigma_0 \min \left( \frac{1}{||L(u)-b||k log^2(k+1)}, 1 \right) \quad \implies \quad \text{Bounded dual}
\end{align*}$

$L(u) = Auu^* \& g(u) = \text{tr}(c uu^*)$

• Our result: FOS with $\mathcal{O} \left( \frac{1}{\epsilon^3} \right)$ & SOS $\tilde{\mathcal{O}} \left( \frac{1}{\epsilon^3} \right)$ total complexity

*Numerical experiment: Clustering

- Model free k-means clustering SDP:

\[
\min \left\{ \text{tr}(cx) : \ x1 = 1, \ x \succeq 0, \ x^* = x, \ \text{tr}(x) = \rho \right\},
\]

- Nonconvex formulation:

\[
\min \left\{ \text{tr}(cuu^*) : \ uu^* 1 = 1, \ u \succeq 0, \ \|u\|_F \leq \sqrt{\rho} \right\},
\]

- Preprocessing & setup & rounding as in (Mixon et. al., 2017)

D. Mixon, S. Villar and R. Ward, Clustering subgaussian mixtures by semidefinite programming, 2017
DARN with GANs - Numerical Results (MNIST)

- De-adversarial-noise with generative adversarial networks:

\[
\begin{align*}
\text{minimize} & \quad \|w - (w_0 + \eta)\|_*, \\
\text{subject to} & \quad w = G(z),
\end{align*}
\]

Figure: \(\ell_\infty\) error per iteration

Figure: misclassification error per iteration
*Numerical experiment: Basis Pursuit*

- Convex formulation:
  \[
  \min \left\{ ||x||_1 : Ax = b \right\}
  \]

- Non-convex formulation:
  \[
  \begin{align*}
  \text{change of variables} \quad & \begin{cases} 
  x & := x^+ - x^- \\
  x^+ & := u_1^{o2}, \quad x^- := u_2^{o2} \text{ and } u := [u_1^\top, u_2^\top]^\top \\
  \bar{A} & := [A, -A]
  \end{cases} \\
  \min \left\{ ||u||_2^2 : \bar{A}u^{o2} = b \right\}
  \end{align*}
  \]

- Potential with more structured norms (e.g., latent group lasso norm)
References I

Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming.

An algorithm for quadratic programming.

Variance-reduced and projection-free stochastic optimization.

Revisiting Frank-Wolfe: Projection-free sparse convex optimization.

Block-coordinate Frank-Wolfe optimization for structural SVMs.
Conditional gradient sliding for convex optimization. 

On the non-ergodic convergence rate of an inexact augmented lagrangian framework for composite convex programming. 

Stochastic conditional gradient method for composite convex minimization. 

Stochastic conditional gradient methods: From convex minimization to submodular maximization. 

Stochastic frank-wolfe methods for nonconvex optimization. 
References III

An inexact augmented lagrangian framework for nonconvex optimization with nonlinear constraints.

A conditional gradient-based augmented lagrangian framework.

Scalable convex methods for phase retrieval.

A conditional gradient framework for composite convex minimization with applications to semidefinite programming.