Model Function Based Conditional Gradient Method
with Armijo-like Line Search

Peter Ochs
Mathematical Optimization Group
Saarland University
— 06.08.2019 —

joint work: Yura Malitsky
### Overview

\[
\min_{x \in C} f(x)
\]

**Classic setting:**
- \( f \) smooth, non-convex
- \( C \) compact, convex

**Oracle:**

\[
y^{(k)} \in \arg\min_{y \in C} \left\langle \nabla f(x^{(k)}), y \right\rangle
\]

**Update (line-search for \( \gamma_k \):)**

\[
x^{(k+1)} = \gamma_k y^{(k)} + (1 - \gamma_k) x^{(k)}
\]

**Convergence condition:**
- Armijo line search
- Descent Lemma (curv. cond.)
### Overview

<table>
<thead>
<tr>
<th>Classic setting:</th>
<th>Generalized setting:</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f$ smooth, non-convex</td>
<td>$f$ non-smooth, non-convex</td>
</tr>
<tr>
<td>$C$ compact, convex</td>
<td>$C$ compact, convex</td>
</tr>
</tbody>
</table>

**Oracle:**

Classical setting:

$$y^{(k)} \in \arg\min_{y \in C} \left\langle \nabla f(x^{(k)}), y \right\rangle$$

Generalized setting:

$$y^{(k)} \in \arg\min_{y \in C} f_{x^{(k)}}(y)$$

**Update (line-search for $\gamma_k$):**

Classical setting:

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

Generalized setting:

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

**Convergence condition:**

Classical setting:

Armijo line search
Descent Lemma (curv. cond.)

Generalized setting:

Armijo-like line search
Generalized Descent Lemma
Generalizing the Descent Lemma

Descent Lemma:

\[ \exists L > 0: \forall x, y: \quad \| \nabla f(x) - \nabla f(y) \| \leq L \| x - y \| \]

\[ \implies | f(x) - f(\bar{x}) - \langle \nabla f(\bar{x}), x - \bar{x} \rangle | \leq \frac{L}{2} \| x - \bar{x} \|^2 \]

provides a measure for the \textbf{linearization error}

\[ \leadsto \text{quadratic growth} \]
Generalizing the Descent Lemma

Generalization of the Descent Lemma:

\[ \exists \psi \text{ continuous}, \psi(0) = 0: \forall x, y: \quad \|\nabla f(x) - \nabla f(y)\| \leq \psi(\|x - y\|) \]

\[ \implies |f(x) - f(\bar{x}) - \langle \nabla f(\bar{x}), x - \bar{x} \rangle| \leq \omega(\|x - \bar{x}\|), \quad \omega(t) = \int_0^1 t\psi(st) \, ds \]

provides a measure for the linearization error

\[ \sim \text{ growth given by } \omega \]
Generalizing the Descent Lemma

Impose Generalization of the Descent Lemma:

Model assumption:

\[ |f(x) - f_{\bar{x}}(x)| \leq \omega(\|x - \bar{x}\|) \]

provides a measure for the approximation error

\[ \sim \quad \text{growth given by “growth function” } \omega \]
Model Assumption \[ |f(x) - f_{\bar{x}}(x)| \leq \omega(\|x - \bar{x}\|) \]
Generalized Conditional Gradient Setting

**Setting:** \[ \min_{x \in C} f(x) \]

- \( C \subset \mathbb{R}^N \) non-empty, compact, convex
- \( f : \mathbb{R}^N \rightarrow (-\infty, \infty] \) proper, lsc with \( \text{dom } f \subset C \) and bounded below

**Growth Function:**
- \( \omega : \mathbb{R}_+ \rightarrow \mathbb{R}_+ \) continuous with \( \omega(0) = 0 \) and \( \omega'_+(0) = 0 \).

**Model Function:** For each \( \bar{x} \):
- proper, lsc, convex function \( f_{\bar{x}} : \mathbb{R}^N \rightarrow (-\infty, \infty] \) (model function)
- \( \text{dom } f = \text{dom } f_{\bar{x}} \)
- \( |f(x) - f_{\bar{x}}(x)| \leq \omega(\|x - \bar{x}\|), \quad \forall x \in C \)
### Overview

#### Classic setting:
- $f$ smooth, non-convex
- $C$ compact, convex

#### Generalized setting:
- $f$ non-smooth, non-convex
- $C$ compact, convex

#### Oracle:

<table>
<thead>
<tr>
<th>$y^{(k)} \in \text{argmin}_{y \in C} \langle \nabla f(x^{(k)}), y \rangle$</th>
<th>$y^{(k)} \in \text{argmin}<em>{y \in C} f</em>{x^{(k)}}(y)$</th>
</tr>
</thead>
</table>

#### Update (line-search for $\gamma_k$):

<table>
<thead>
<tr>
<th>$x^{(k+1)} = \gamma_k y^{(k)} + (1 - \gamma_k) x^{(k)}$</th>
<th>$x^{(k+1)} = \gamma_k y^{(k)} + (1 - \gamma_k) x^{(k)}$</th>
</tr>
</thead>
</table>

#### Convergence condition:

- Armijo line search
- Descent Lemma (curv. cond.)

- Armijo-like line search
- Generalized Descent Lemma
Examples: Generalizing the Descent Lemma

Model Assumption: \[ |f(x) - f_{\bar{x}}(x)| \leq \omega(\|x - \bar{x}\|) \]

▶ Example: Additive composite problem:

\[
\min_{x \in C} f(x), \quad f(x) = g(x) + \begin{array}{c}
\text{non-smooth convex} \\
\psi\text{-uniform smooth}
\end{array} + h(x)
\]

▶ model function:

\[
f_{\bar{x}}(x) = g(x) + h(\bar{x}) + \langle \nabla h(\bar{x}), x - \bar{x} \rangle
\]

▶ generalized Conditional Gradient oracle:

\[
\arg\min_{y \in C} g(y) + \left\langle \nabla h(x^{(k)}), y \right\rangle
\]
Examples: Generalizing the Descent Lemma

Model Assumption:  \[ |f(x) - f_{\bar{x}}(x)| \leq \omega(\|x - \bar{x}\|) \]

Example: Proximal Gradient Descent:

\[
\min_{x \in C} f(x), \quad f(x) = \begin{aligned}
g(x) + h(x) & \text{ non-smooth convex} \\ ψ\text{-uniform smooth}
\end{aligned}
\]

Model Function:

\[
f_{\bar{x}}(x) = g(x) + h(\bar{x}) + \langle \nabla h(\bar{x}), x - \bar{x} \rangle + \frac{1}{2\lambda} \|x - \bar{x}\|^2
\]

generalized Conditional Gradient Oracle:

\[
\arg\min_{y \in C} g(y) + \langle \nabla h(x^{(k)}), y \rangle + \frac{1}{2\lambda} \|y - x^{(k)}\|^2
\]
Examples: Generalizing the Descent Lemma

Model Assumption: \[ |f(x) - f_{\bar{x}}(x)| \leq \omega(\|x - \bar{x}\|) \]

▶ Example: Proximal Gradient Descent:

\[
\min_{x \in C} f(x), \quad f(x) = g(x) + h(x)
\]

non-smooth convex \psi-uniform smooth

▶ model function:

\[
f_{\bar{x}}(x) = g(x) + h(\bar{x}) + \langle \nabla h(\bar{x}), x - \bar{x} \rangle + \frac{1}{2\lambda} \|x - \bar{x}\|^2
\]

▶ generalized Conditional Gradient oracle:

\[
\arg\min_{y \in C} g(y) + \langle \nabla h(x^{(k)}), y \rangle + \frac{1}{2\lambda} \|y - x^{(k)}\|^2
\]

▶ works also with Bregman proximal term

▶ setting without constraint set: [O., Fadili, Brox 18]
Examples: Generalizing the Descent Lemma

Model Assumption: \[ |f(x) - f_{\bar{x}}(x)| \leq \omega(\|x - \bar{x}\|) \]

Example: Newton–Conditional Gradient:

\[
\min_{x \in C} f(x), \quad f(x) = g(x) + h(x)
\]

- non-smooth convex
- twice diff. \(\psi\)-uniform smooth

Model function:

\[
f_{\bar{x}}(x) = g(x) + h(\bar{x}) + \langle \nabla h(\bar{x}), x - \bar{x} \rangle + \frac{1}{2} \langle x - \bar{x}, [\nabla^2 h(\bar{x})]_+(x - \bar{x}) \rangle
\]

generalized Conditional Gradient oracle:

\[
\arg\min_{y \in C} g(y) + \langle \nabla h(x^{(k)}), y \rangle + \frac{1}{2} \langle y - x^{(k)}, [\nabla^2 h(x^{(k)})]_+(y - x^{(k)}) \rangle
\]
Examples: Generalizing the Descent Lemma

Model Assumption: \[ |f(x) - f_{\bar{x}}(x)| \leq \omega(\|x - \bar{x}\|) \]

- Example: Hybrid Proximal–Conditional Gradient:

\[
\min_{x_1 \in C_1} \min_{x_2 \in C_2} f(x_1, x_2), \quad f(x_1, x_2) = g(x_1) + h(x_1, x_2) \\
\text{non-smooth convex} \quad \psi\text{-uniform smooth}
\]

- Model function:

\[
f_{\bar{x}}(x_1, x_2) = h(\bar{x}) + \langle \nabla h(\bar{x}), x - \bar{x} \rangle + g(x_1) + \frac{1}{2\lambda} \|x_1 - \bar{x}_1\|^2, \quad x = (x_1, x_2)
\]

- Generalized Conditional Gradient oracle:

\[
\begin{cases}
\text{argmin}_{y_1 \in C_1} g(y_1) + \frac{1}{2\lambda} \|y_1 - (x_1^{(k)} + \lambda \nabla_1 h(x_1^{(k)}, x_2^{(k)}))\|^2 \\
\text{argmin}_{y_2 \in C_2} \langle \nabla_2 h(x_1^{(k)}, x_2^{(k)}), y_2 \rangle
\end{cases}
\]
Examples: Generalizing the Descent Lemma

Model Assumption: \[ |f(x) - f_{\bar{x}}(x)| \leq \omega(\|x - \bar{x}\|) \]

▶ Example: Non-linear composite problems (Gauss–Newton):

\[
\min_{x \in C} f(x), \quad f(x) = \begin{cases} g_\text{non-smooth} & F(x) \\ \text{convex} & F(x) \\ \psi\text{-uniform} & F(x) \\ \text{smooth} & F(x) \end{cases}
\]

▶ model function:

\[ f_{\bar{x}}(x) = g(F(\bar{x}) + DF(\bar{x})(x - \bar{x})) \]

▶ generalized Conditional Gradient oracle:

\[ \arg\min_{y \in C} g(F(x^{(k)}) + DF(x^{(k)})(y - x^{(k)})) \]
Examples: Generalizing the Descent Lemma

Design model functions for your problem
such that the oracle is easy to evaluate!
## Overview

### Classic setting:

- $f$ smooth, non-convex
- $C$ compact, convex

### Generalized setting:

- $f$ non-smooth, non-convex
- $C$ compact, convex

### Oracle:

$y^{(k)} \in \arg\min_{y \in C} \langle \nabla f(x^{(k)}), y \rangle$

$y^{(k)} \in \arg\min_{y \in C} f_{x^{(k)}}(y)$

### Update (line-search for $\gamma_k$):

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

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

### Convergence condition:

- Armijo line search
- Descent Lemma (curv. cond.)

- Armijo-like line search
- Generalized Descent Lemma
Model Based Conditional Gradient Method with Line Search:

- **Initialization**: \( x^{(0)} \in \mathbb{R}^N \) and set \( \rho \in (0, 1) \).

- **Update** (\( k \geq 0 \)):
  
  Compute
  \[
  y^{(k)} \in \arg\min_{y \in C} f_{x^{(k)}}(y)
  \]

  \[
  x^{(k+1)} = x^{(k)} + \gamma_k (y^{(k)} - x^{(k)})
  \]

  with \( \gamma_k \in [0, 1] \) determined by backtracking line search such that

  \[
  f(x^{(k+1)}) \leq f(x^{(k)}) - \rho \gamma_k \left( f_{x^{(k)}}(x^{(k)}) - f_{x^{(k)}}(y^{(k)}) \right) \tag{model improvement}
  \]
Algorithm and Convergence

Convergence results:

- Algorithm and line-search are well-defined.
- If $\omega$ is a growth function and the model assumption holds, then
  
  - every limit point of $(x^{(k)})_{k \in \mathbb{N}}$ is a stationary point of
    \[ \min_{x \in C} f(x), \]
  - there exists at least one limit point, and
  - $(f(x^{(k)}))_{k \in \mathbb{N}}$ converges to the value of $f$ at the limit point.
**Assumptions:**

\[ F_i(a, b) := \sum_{j=1}^{P} a_j \exp(-b_j x_i) \]

- \( y_i = F_i(a, b) + n_i \) where \( n_i \) are iid errors (Laplacian distribution)
- large percentage of coefficients \( a_j \) are zero
Application: Robust Sparse Non-linear Regression

\[
\min_{(a,b) \in C} \sum_{i=1}^{M} \| F_i(a, b) - y_i \|_1 + \mu \| a \|_1
\]

- **Our Generalized Conditional Gradient oracle: \((FW-\text{CompLinLS})\)**

\[
\min_{u=(a,b) \in C} \sum_{i=1}^{M} \| K_i u - y_i^\dagger \|_1 + \mu \| a \|_1 , \quad K_i := DF_i(u^{(k)}).
\]

- **ProxLinear oracle [Lewis and Wright 2016]:**

\[
\min_{u=(a,b) \in C} \sum_{i=1}^{M} \| K_i u - y_i^\dagger \|_1 + \mu \| a \|_1 + \frac{1}{2\tau} \| u - u^{(k)} \|_2^2.
\]

- **ProxLinearLS**: Armijo-like line search (as above).
- **ProxLinearBT**: Backtracking for \(\tau\).

Solve subproblems by PDHG [Pock and Chambolle 2011].
Application: Robust Sparse Non-linear Regression

![Graph showing time vs. objective error for different methods: FW-CompLinLS, ProxLinearLS, and ProxLinearBT. The graph plots time in seconds on a logarithmic scale against objective error also on a logarithmic scale. The methods show varying rates of convergence.]
Structured Matrix Factorization

Applications:
- blind image deblurring, clustering and principal component analysis, source separation, signal processing, dictionary learning, ...

Optimization Problem:
\[
\min_{X,Y} \frac{1}{2} \| A - XY \|_F^2 + g(X) \quad \text{s.t. } X \in \mathcal{X}, \ Y \in \mathcal{Y},
\]
\[=: h(X,Y) \]

Examples:
- constraints on: norm balls, non-negativity, stochasticity, rank, ...
- regularization: (block) sparsity, \(l_2\)-norm, low rank, ...

Model function: Linearization of \(h\), proximal linearization, block-proximal linearization of \(h\), ...
Conclusion

▶ Model function in Conditional Gradient

\[ |f(x) - f(x^*)| \leq \omega(\|x - x^*\|) \].

▶ Flexible design of subproblems

\[ \arg\min_{y \in C} f_{x^{(k)}}(y) \].

▶ Subsequences converge to stationary points.

Design model functions for your problem such that the oracle is easy to evaluate!