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:

$$y^{(k)} \in \operatorname*{argmin}_{y\in C} \left\langle \nabla f(x^{(k)}), y \right\rangle$$

Update (line-search for γ_k):

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

Convergence condition:

Armijo line search
Descent Lemma (curv. cond.)

Overview

$$\min_{x\in C} f(x)$$

Classic setting:	Generalized setting:
f smooth, non-convex	f non-smooth, non-convex
C compact, convex	C compact, convex

Oracle:

$$y^{(k)} \in \operatorname*{argmin}_{y\in C} \left\langle \nabla f(x^{(k)}), y \right\rangle \qquad \Big| \qquad y^{(k)} \in \operatorname*{argmin}_{y\in C} f_{x^{(k)}}(y)$$

Update (line-search for γ_k):

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

Convergence condition:

Classic setting	Generalized setting
Armijo line search	Armijo-like line search
Descent Lemma (curv. cond.)	Generalized Descent Lemma

Generalizing the Descent Lemma

Descent Lemma:

$$\exists L > 0\colon \forall x, y\colon \quad \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

$$\Longrightarrow |f(x) - f(\bar{x}) - \langle \nabla f(\bar{x}), x - \bar{x}\rangle| \leq \frac{L}{2}\|x - \bar{x}\|^2$$

provides a measure for the **linearization error**

$\rightsquigarrow$ quadratic growth

Generalizing the Descent Lemma

Generalization of the Descent Lemma:

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

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

provides a measure for the **linearization error**

$\rightsquigarrow$ growth given by ω

Generalizing the Descent Lemma

Impose Generalization of the Descent Lemma:

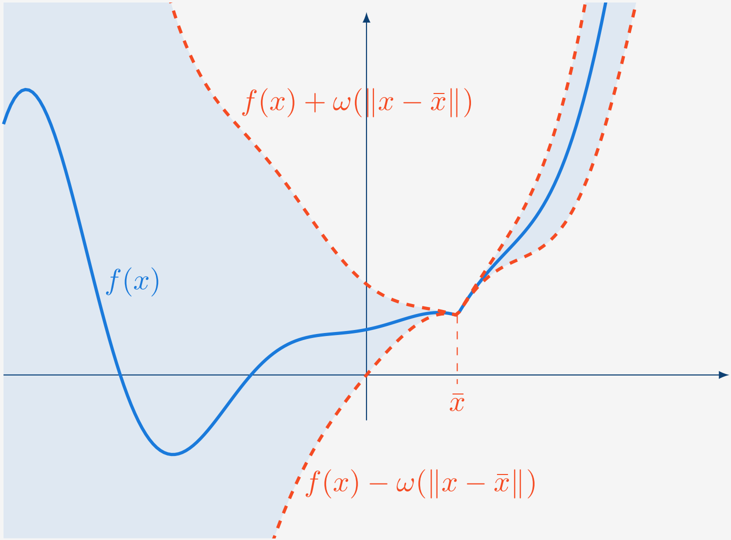
Model assumption:

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

provides a measure for the **approximation error**

$\rightsquigarrow$ growth given by "growth function" ω

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
- $f\colon \mathbb{R}^N \to (-\infty, \infty]$ proper, lsc with $\operatorname{dom} f \subset C$ and bounded below

Growth Function:

- $\omega\colon \mathbb{R}_+ \to \mathbb{R}_+$ continuous with $\omega(0) = 0$ and $\omega'_+(0) = 0$.

Model Function: For each $\bar{x}$:

- proper, lsc, convex function $f_{\bar{x}}\colon \mathbb{R}^N \to (-\infty, \infty]$ (**model function**)
- $\operatorname{dom} f = \operatorname{dom} f_{\bar{x}}$
- $|f(x) - f_{\bar{x}}(x)| \leq \omega(\|x - \bar{x}\|)\,, \quad \forall x \in C$

Overview

$$\min_{x \in C} f(x)$$

Classic setting:	**Generalized setting:**
f smooth, non-convex	f non-smooth, non-convex
C compact, convex	C compact, convex

Oracle:

$y^{(k)} \in \operatorname*{argmin}_{y \in C} \left\langle \nabla f(x^{(k)}), y \right\rangle$	$y^{(k)} \in \operatorname*{argmin}_{y \in C} f_{x^{(k)}}(y)$

Update (line-search for γ_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	Armijo-like line search
Descent Lemma (curv. cond.)	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:**

$$\min_{x\in C} f(x)\,, \quad f(x) = \underbrace{g(x)}_{\text{non-smooth convex}} + \underbrace{h(x)}_{\psi\text{-uniform smooth}}$$

- **model function:**

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

- **generalized Conditional Gradient oracle:**

$$\operatorname*{argmin}_{y\in C} g(y) + \left\langle \nabla h(x^{(k)}), y \right\rangle$$

Examples: Generalizing the Descent Lemma

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

- **Example: Proximal Gradient Descent:**

$$\min_{x \in C} f(x)\,, \quad f(x) = \underbrace{g(x)}_{\text{non-smooth convex}} + \underbrace{h(x)}_{\psi\text{-uniform smooth}}$$

- **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:**

$$\operatorname*{argmin}_{y \in C} g(y) + \left\langle \nabla h(x^{(k)}), y \right\rangle + \frac{1}{2\lambda}\|y - x^{(k)}\|^2$$

Examples: Generalizing the Descent Lemma

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

- **Example: Proximal Gradient Descent:**

$$\min_{x \in C} f(x)\,, \quad f(x) \;=\; \underbrace{g(x)}_{\text{non-smooth convex}} \;+\; \underbrace{h(x)}_{\psi\text{-uniform smooth}}$$

- **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:**

$$\operatorname*{argmin}_{y \in C} g(y) + \left\langle \nabla h(x^{(k)}), y \right\rangle + \frac{1}{2\lambda}\|y - x^{(k)}\|^2$$

- 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:**

$$\min_{x\in C} f(x)\,, \quad f(x) = \underbrace{g(x)}_{\text{non-smooth convex}} + \underbrace{h(x)}_{\text{twice diff. } \psi\text{-uniform smooth}}$$

- **model function:**

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

- **generalized Conditional Gradient oracle:**

$$\operatorname*{argmin}_{y\in C} g(y) + \left\langle \nabla h(x^{(k)}), y\right\rangle + \frac{1}{2}\left\langle y - x^{(k)}, [\nabla^2 h(x^{(k)})]_+(y - x^{(k)})\right\rangle$$

Examples: Generalizing the Descent Lemma

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

- **Example: Hybrid Proximal–Conditional Gradient:**

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

- **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:**

$$\begin{cases} \operatorname*{argmin}_{y_1 \in C_1} & g(y_1) + \frac{1}{2\lambda}\|y_1 - \big(x_1^{(k)} + \lambda \nabla_1 h(x_1^{(k)}, x_2^{(k)})\big)\|^2 \\ \operatorname*{argmin}_{y_2 \in C_2} & \Big\langle \nabla_2 h(x_1^{(k)}, x_2^{(k)}), y_2 \Big\rangle \end{cases}$$

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):**

$$\min_{x\in C} f(x)\,, \quad f(x) = \underbrace{g}_{\substack{\text{non-smooth}\\ \text{convex}\\ \text{Lipschitz}}} \Big(\underbrace{F(x)}_{\substack{\psi\text{-uniform}\\ \text{smooth}}} \Big)$$

- **model function:**

$$f_{\bar{x}}(x) = g\big(F(\bar{x}) + DF(\bar{x})(x - \bar{x})\big)$$

- **generalized Conditional Gradient oracle:**

$$\underset{y\in C}{\operatorname{argmin}}\, g\big(F(x^{(k)}) + DF(x^{(k)})(y - x^{(k)})\big)$$

Examples: Generalizing the Descent Lemma

Design model functions for your problem

such that the oracle is easy to evaluate !

Overview

$$\min_{x \in C} f(x)$$

Classic setting:	**Generalized setting:**
f smooth, non-convex	f non-smooth, non-convex
C compact, convex	C compact, convex

Oracle:

$$y^{(k)} \in \operatorname*{argmin}_{y \in C} \left\langle \nabla f(x^{(k)}), y \right\rangle \qquad \Big| \qquad y^{(k)} \in \operatorname*{argmin}_{y \in C} f_{x^{(k)}}(y)$$

Update (line-search for γ_k):

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

Convergence condition:

Classic	Generalized
Armijo line search	Armijo-like line search
Descent Lemma (curv. cond.)	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
$$y^{(k)} \in \operatorname*{argmin}_{y \in C} f_{x^{(k)}}(y)$$
$$x^{(k+1)} = x^{(k)} + \gamma_k (y^{(k)} - x^{(k)})$$

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

$$f(x^{(k+1)}) \leq f(x^{(k)}) - \rho\gamma_k \underbrace{\left(f_{x^{(k)}}(x^{(k)}) - f_{x^{(k)}}(y^{(k)}) \right)}_{\text{model improvement}} .$$

Algorithm and Convergence

Convergence results:

- Algorithm and line-search are well-defined.
- **If** ω 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

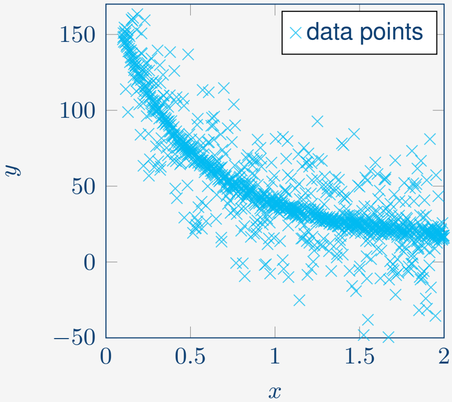
    $$\min_{x\in C} f(x)\,,$$

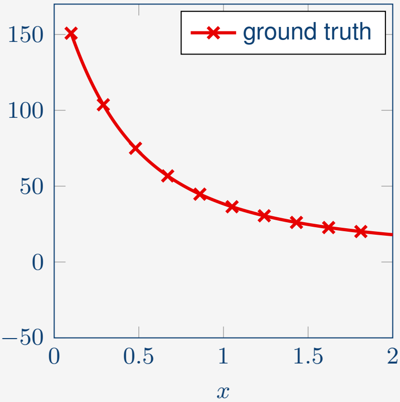
  - there exists at least one limit point, and
  - $(f(x^{(k)}))_{k\in\mathbb{N}}$ converges to the value of f at the limit point.

Application: Robust Sparse Non-linear Regression

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-CompLinLS`)

$$\min_{u=(a,b)\in C} \sum_{i=1}^{M} \|\mathcal{K}_i u - y_i^\diamond\|_1 + \mu\|a\|_1\,, \quad \mathcal{K}_i := DF_i(u^{(k)})\,.$$

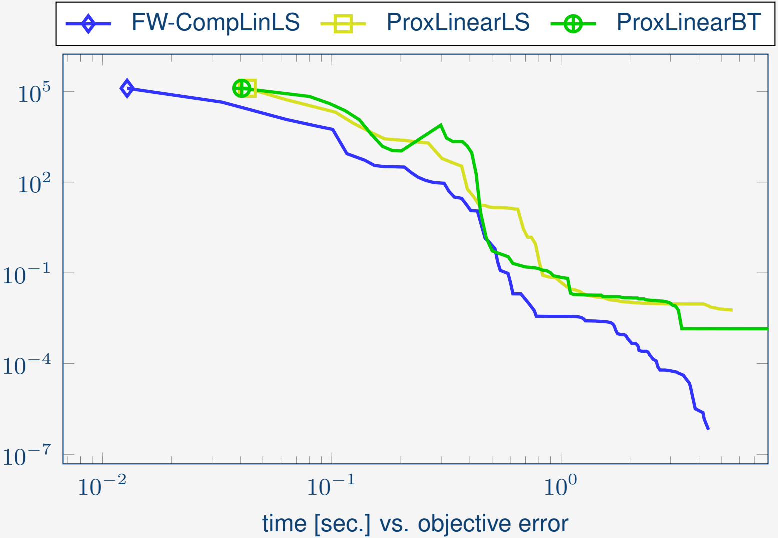
- ProxLinear oracle [Lewis and Wright 2016]:

$$\min_{u=(a,b)\in C} \sum_{i=1}^{M} \|\mathcal{K}_i u - y_i^\diamond\|_1 + \mu\|a\|_1 + \frac{1}{2\tau}\|u - u^{(k)}\|^2\,.$$

  - `ProxLinearLS`: Armijo-like line search (as above).
  - `ProxLinearBT`: Backtracking for τ.

Solve subproblems by PDHG [Pock and Chambolle 2011].

Application: Robust Sparse Non-linear Regression



Structured Matrix Factorization

Applications:

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

Optimization Problem:

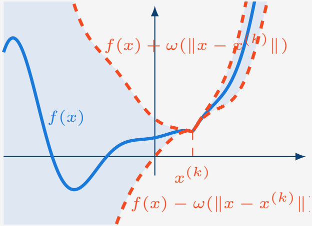
$$\min_{X,Y} \underbrace{\frac{1}{2}\|A - XY\|_F^2}_{=:h(X,Y)} + g(X) \quad s.t.\ X \in \mathcal{X},\ Y \in \mathcal{Y}\,,$$

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

$$|f(x) - f_{\bar{x}}(x)| \leq \omega(\|x - \bar{x}\|)\,.$$

- Flexible design of subproblems

$$\operatorname*{argmin}_{y \in C}\, f_{x^{(k)}}(y)\,.$$

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

- Subsequences converge to stationary points.