Generalized Conditional Gradient with Augmented Lagrangian

Cesare Molinari
(joint work with J. Fadili and A. Silveti-Falls)

PostDoc at UniCaen
GREYC Lab, Image Group

Dubrovnik, 13 Febr 2019

Outline

- Conditional gradient (Frank-Wolfe)
  - Problem, algorithm and examples
  - Convergence analysis
- FW + Smoothing
  - Moreau-Yosida Envelope
  - Algorithm
- FW + Smoothing + Augmented Lagrangian

Problem and assumptions

Main purpose: design and analysis of iterative algorithms for optimization problems, as

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

Hypothesis (H):

($\mathbf{H_1}$) $\mathcal{C} \subset \mathcal{H}$ is non-empty, convex and compact ($\mathcal{H}$ is Hilbert);

($\mathbf{H_2}$) $f : \mathcal{C} \to \mathbb{R}$ is convex, $\mathcal{G}$-differentiable and ∇f is Lipschitz continuous (with constant L)

Optimality Condition

Existence (not uniqueness); $\bar{x}$ is solution **iff**, for every $c \in \mathcal{C}$,

$$\langle -\nabla f(\bar{x}), \; c - \bar{x} \rangle \leq 0$$

Classical approach: projected gradient

For a step-size $\lambda \in (0, 2/L)$, iterate

$$x_{k+1} = P_C\left(x_k - \lambda \nabla f(x_k)\right)$$

Problem: the projection can be computational expansive

Quadratic approximation

$$x_k - \lambda \nabla f(x_k) = \text{Argmin}_x \left\{ f(x_k) + \langle \nabla f(x_k),\ x - x_k \rangle + \frac{1}{2\lambda} \|x - x_k\|^2 \right\}$$

(Proof...)

Frank-Wolfe: projection-free algorithm

M. Franke and P. Wolfe, *An algorithm for quadratic programming.* Naval research logistics quarterly, 1956

M. Jaggi, *Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization.* Proceedings on Machine Learning Research, 2013

FW Algorithm:

$$\text{(LMO)} \qquad s_k \in \operatorname{Argmin}_{s \in \mathcal{C}} \langle \nabla f(x_k),\ s \rangle;$$

$$\text{(Update)} \qquad x_{k+1} = x_k + \gamma_k (s_k - x_k)$$

Iterates feasibility

If $\gamma_k \in [0, 1]$, $x_k \in \mathcal{C}$; indeed,

$$x_{k+1} = (1 - \gamma_k) x_k + \gamma_k s_k$$

Convex analysis: a refresher

Let $f: X \to \mathbb{R} \cup \{+\infty\}$.

Definition (Subdifferential)

$\partial f(x) = \{x' \in \mathcal{H} : \quad f(y) \geq f(x) + \langle x',\ y - x \rangle \ \ \forall y \in \mathcal{H}\}$

Definition (Fenchel conjugate)

$f^\star(x) = \max_z \{\langle x,\ z \rangle - f(z)\}$

Theorem

If f is proper convex and l.s.c., then

$$(\partial f)^{-1} = \partial f^\star;$$

moreover, $\partial f^\star(x) = \mathit{Argmax}_z \{\langle x,\ z\rangle - f(z)\}$ **(Proof...)**

Notation: $\mathcal{N}_C := \partial \delta_C$ (*normal cone*) and $\sigma_C := \delta_C^\star$ (*support function*)

Frank-Wolfe: mimicking the opt cond

- **Problem Opt Cond**: $0 \in \nabla f(\bar{x}) + \mathcal{N}_{\mathcal{C}}(\bar{x})$

- **FW Algorithm**: $0 \in \nabla f(x_k) + \mathcal{N}_{\mathcal{C}}(s_k)$

Equivalently, the LMO reads as

$$s_k \in \partial \sigma_C \left(-\nabla f(x_k)\right)$$

(Proof...)

Remark

FW is an **inexact subgradient descent on the dual**

$\implies$ the step-size γ_k has to go to zero

Example: norm constraint

$$\mathcal{C} = \{x \in X : \ \|x\| \leq t\}$$

Given $v \in \mathcal{H}$, the LMO is equivalent to

$$\text{Argmin}_{s \in \mathcal{C}} \ \langle v, \ s \rangle = -t \ \partial \| \cdot \|_* (v)$$

(Proof...)

Example (ℓ^1-norm)

If $\mathcal{C}$ is the ℓ^1-ball, we obtain the **greedy coordinate descent**:

$$i_k \in \text{Argmax}_i \ |\partial_i f(x_k)|$$
$$x_{k+1} = (1 - \gamma_k) x_k - \gamma_k t \ \text{sign}(\partial_{i_k} f(x_k)) \ e_{i_k}$$

(*cheaper that projection*)

Property: affine invariance

Consider the change of variable $x = B\tilde{x}$ and $h(\tilde{x}) = f(B\tilde{x})$; then

$$\text{Argmin}_{x\in\mathcal{C}}\ f(x) = B\ \left[\text{Argmin}_{B\tilde{x}\in\mathcal{C}}\ h(\tilde{x})\right]$$

Starting from $x \in \mathcal{C}$,

- **Gradient method**:

$$x_+^{(1)} = x - \lambda \nabla f(x);$$
$$x_+^{(2)} = x - \lambda B B^* \nabla f(x);$$

- **FW**:

$$x_+^{(1)} = x_+^{(2)}$$

Outline

Conditional gradient (Frank-Wolfe)
- Problem, algorithm and examples
- Convergence analysis

FW + Smoothing
- Moreau-Yosida Envelope
- Algorithm

FW + Smoothing + Augmented Lagrangian

Convergence result (1)

Theorem

Let $\gamma_k \in \ell^2 \setminus \ell^1$*. Then* $\lim_k f(x_k) = f(\bar{x})$*, there exist a subsequence such that*

$$(0 \leq)\ f(x_{k_j}) - f(\bar{x}) \leq \Gamma_{k_j}^{-1}, \qquad \text{where}\ \ \Gamma_n = \sum_{i=1}^{n} \gamma_i$$

and every weak cluster point is a solution.
In particular, if the solution is unique, $x_k \rightharpoonup \bar{x}$*.*

Two lemmas on real sequences

Lemma (Quasi-Fejér monotonicity)

If

$$r_{k+1} - r_k + a_k \leq z_k \in \ell^1,$$

then r_k is convergent and $a_k \in \ell^1$.

Lemma (Subsequencial rate)

If $(\gamma_k w_k) \in \ell^1$ and $\gamma_k \notin \ell^1$, then there exists a subsequence w_{k_j} s.t.

$$w_{k_j} \leq \Gamma_{k_j}^{-1}, \qquad \text{where } \; \Gamma_n = \sum_{i=1}^{n} \gamma_i$$

If moreover $w_k - w_{k+1} \leq \alpha\gamma_k$ for some $\alpha > 0$, then $\lim_k w_k = 0$

Quadratic upper bound

Lemma (Descent Lemma)

Let $f : \mathcal{C} \to \mathbb{R}$ be $\mathcal{G}$-differentiable with L-Lipschitz continuous gradient. Then, for every x and $y \in \mathcal{H}$,

$$f(y) \leq f(x) + \langle \nabla f(x),\ y - x \rangle + \frac{L}{2}\|y - x\|^2 \qquad (*)$$

(Proof...)

Baillon-Haddad Theorem

If f is convexity and differentiable, the following are equivalent:

(i) Descent inequality $(*)$;

(ii) ∇f is L-Lipschitz continuous;

(iii) ∇f is $1/L$ cocoercive

Main estimations

Lemma

For the Frank-Wolfe algorithm,

(i) *denoting* $r_k := f(x_k) - f(\bar{x})$,

$$r_{k+1} - (1 - \gamma_k)\, r_k \le d_{\mathcal{C}}^2 L \gamma_k^2 / 2;$$

(ii) *denoting* $M := d_{\mathcal{C}}\ \max_{x \in \mathcal{C}} \|\nabla f(x)\|$,

$$r_k - r_{k+1} \le M \gamma_k$$

(recall that $\mathcal{C}$ *is compact and* ∇f *is continuous)*

(Proof...)

Convergence result (2)

Theorem

If $\gamma_k = 2/(k+2)$, *then*

$$f(x_k) - f(\bar{x}) \leq \frac{2d_{\mathcal{C}}^2 L}{k+2}$$

(Proof...)

Generalizations

The same proof holds when

(i) **linesearch** for the step-size (closed-loop choice):

$$\gamma_k \in \text{Argmin}_{\gamma \in [0,1]} \;\; f\left(x_k + \gamma\left(s_k - x_k\right)\right);$$

(ii) replace the hypothesis Lipschitz-continuity of ∇f with the boundedness of the **curvature constant**:

$$C_f = \sup\left\{\frac{2}{\gamma^2}\left[f(y) - f(x) - \langle \nabla f(x),\; y - x\rangle\right]\right\},$$

taken on $\gamma \in (0,1]$, $x, s \in \mathcal{C}$, $y = x + \gamma\left(s - x\right)$

Remark

Lipschitz-continuity of ∇f implies $C_f \leq d_{\mathcal{C}}^2 L$ **(Proof...)**

Duality gap

For $x_k \in \mathcal{C}$ (iterate generated by the FW algorithm), define

$$\begin{aligned} \text{gap}(x_k) :&= \langle \nabla f(x_k),\ x_k - s_k \rangle \\ &\geq f(x_k) - f(\bar{x}) \end{aligned}$$

(Proof...)

Remark
An upper-bound for optimality is available at each iteration (and similar convergence-rates for $\text{gap}(x_k)$ hold)

Outline

Problem 2

Now we consider the following problem:

$$\min_{x \in \mathcal{C}} \{f(x) + g(x)\}$$

Hypothesis (H):

- $(\mathbf{H_1})$ $\mathcal{C} \subset \mathcal{H}$ is non-empty, convex and compact;
- $(\mathbf{H_2})$ $f : \mathcal{C} \to \mathbb{R}$ is convex, $\mathcal{G}$-differentiable and ∇f is Lipschitz continuous (with constant L);
- $(\mathbf{H_3})$ $g : \mathcal{C} \to \mathbb{R} \cup \{+\infty\}$ is proper, convex and l.s.c. (non diff)

Outline

Moreau-Yosida envelope

Definition

$$g_\lambda(x) := \inf_y \left\{ g(y) + \frac{1}{2\lambda} \|y - x\|^2 \right\}$$

Motivation: convexification of the dual

$$g_\lambda = g \,\Box \left(\frac{1}{2\lambda} \| \cdot \|^2 \right) = g^{\star\star} \,\Box \left(\frac{\lambda}{2} \| \cdot \|^2 \right)^\star = \left(g^\star + \frac{\lambda}{2} \| \cdot \|^2 \right)^\star$$

(Proof...)

Proximal-point operator

Definition

$$\text{prox}_{\lambda g}(x) := \text{Argmin}_y \left\{ g(y) + \frac{1}{2\lambda}\|y - x\|^2 \right\}$$

- Projection generalization: for $g = \delta_C$, $\text{prox}_{\lambda g} = P_C$

Moreau identity

$$x = \text{prox}_{\lambda g^\star}(x) + \lambda \text{prox}_{\lambda^{-1} g}(\lambda^{-1} x)$$

(Proof...)

Differentiability

Theorem

If $g^\star$ is strongly-convex, then ∇g is Lipschitz-continuous (in particular, g is differentiable)

Theorem

∇g_λ is $(1/\lambda)$-Lipschitz continuous with

$$\nabla g_\lambda(x) = \frac{x - \mathit{prox}_{\lambda g}(x)}{\lambda}$$

(Proof...)

Other properties

For every x in $\mathcal{H}$, denote

$$[\partial g(x)]^0 = \text{Argmin}_{y \in \partial g(x)} \|y\|$$

$i)$ $\inf\ g_\lambda = \inf\ g$ & $\text{Argmin}\ g_\lambda = \text{Argmin}\ g$;

$ii)$ For $\lambda \searrow 0^+$, $g_\lambda(x) \nearrow g(x)$ with

$$g(x) - g_\lambda(x) \leq \frac{\lambda}{2} \| [\partial g(x)]^0 \|^2;$$
$$g_{\lambda'}(x) - g_\lambda(x) \leq \frac{1}{2} \left(\lambda - \lambda'\right) \|\nabla g_{\lambda'}(x)\|^2.$$

$iii)$ For $\lambda \searrow 0^+$, $\nabla g_\lambda(x) \to [\partial g(x)]^0$ with

$$\|\nabla g_\lambda(x)\| \nearrow \| [\partial g(x)]^0 \|$$

Lax-Hopf formula

iv)

$$\left[\frac{\partial}{\partial\lambda} g_\lambda(x)\right]_{\lambda=\lambda'} = -\frac{1}{2}\|\nabla g_{\lambda'}(x)\|^2 \qquad (*)$$

Hamilton-Jacobi equation

For $H: \mathcal{H} \to \mathbb{R}$ convex and 1-coercive and $g_0: \mathcal{H} \to \mathbb{R}$, consider

$$\begin{cases} \frac{\partial}{\partial\lambda} g + H\left(\nabla_x g\right) = 0 & (x,\lambda) \in \mathcal{H} \times (0,+\infty) \\ g\left(x,0\right) = g_0\left(x\right) & x \in \mathcal{H} \end{cases}$$

The (*viscosity*) solution is given by the **Lax-Hopf formula**:

$$g\left(x,\lambda\right) := \inf_{y\in\mathcal{H}} \left\{ g_0(y) + \lambda H^\star\left(\frac{y-x}{\lambda}\right)\right\}$$

For $H(p) = \frac{1}{2}\|p\|^2$, then $H^\star(p) = \frac{1}{2}\|p\|^2$ and we recover $(*)$

Outline

FW + Smoothing

(Smoothing) $y_k = \text{Argmin}_x \left\{ g(x) + \frac{1}{2\lambda_k} \|x - x_k\|^2 \right\}$

(Gradient) $v_k = \nabla f(x_k) + (x_k - y_k) / \lambda_k$

(LMO) $s_k \in \text{Argmin}_{s \in \mathcal{C}} \langle v_k,\ s \rangle$

(Update) $x_{k+1} = x_k + \gamma_k (s_k - x_k)$

A. Yurtsever, O. Fercoq, F. Locatello and V. Cevher, *A Conditional Gradient Framework for Composite Convex Minimization.* Proceedings of the 35th Internat Conf on Machine Learning, 2018

Outline

Problem 3

Now we consider the following problem:

$$\min_{x \in \mathcal{C}} \; \{ \; f(x) + g(x) : \quad Ax = 0 \; \}$$

Hypothesis (H):

($\mathbf{H_1}$) $\mathcal{C} \subset \mathcal{H}$ is non-empty, convex and compact;

($\mathbf{H_2}$) $f : \mathcal{C} \to \mathbb{R}$ is convex, $\mathcal{G}$-differentiable and ∇f is Lipschitz continuous (with constant L);

($\mathbf{H_3}$) $g : \mathcal{C} \to \mathbb{R} \cup \{+\infty\}$ is proper, convex and l.s.c. (non diff)

($\mathbf{H_4}$) A is linear and continuous

Product-space trick

The linear constraint allows to threat the sum of non-differentiable functions by **separate proximal-point operators**

Augmented Lagrangian

To deal with the linear constraint, 3 **techniques** are available

- **Penalization**: $f(x) + g(x) + \rho_k \|Ax\|^2$ with $\rho_k \to +\infty$;
- **Lagrangian duality**: looking at the saddle-points of

$$\mathcal{L}(x, \mu) = f(x) + g(x) + \langle \mu,\ Ax \rangle;$$

- **Augmented Lagrangian**: for fixed $\rho > 0$,

$$\mathcal{E}_k(x, \mu) = f(x) + g_{\lambda_k}(x) + \langle \mu,\ Ax \rangle + \frac{\rho}{2}\|Ax\|^2$$

Algorithm

(Smoothing) $$y_k = \operatorname{Argmin}_x \left\{ g(x) + \frac{1}{2\lambda_k} \|x - x_k\|^2 \right\}$$

(Gradient) $$v_k = \nabla f(x_k) + (x_k - y_k)/\lambda_k + A^* \mu_k + \rho A^* A x_k$$

(LMO) $$s_k \in \operatorname{Argmin}_{s \in \mathcal{C}} \langle v_k,\ s \rangle$$

(Primal update) $$x_{k+1} = x_k + \gamma_k (s_k - x_k)$$

(Dual update) $$\mu_{k+1} = \mu_k + \theta_k A x_k$$

G. Gidel, F. Pedregosa and S. Lacoste-Julien, *Frank-Wolfe Splitting via Augmented Lagrangian Method.*
10th NIPS Workshop on Optimization for Machine Learning, 2018

Conclusions: what we have done...

(i) **Asymptotic feasibility**: $Ax_k \to 0$ *(strongly)*;

(ii) **Lagrangian multiplier boundedness**;

(iii) **Optimality rates**: every weak cluster point of x_k is a solution and μ_k weakly converges to an optimal dual variable with

$$\lim_{k\to\infty} \left[\mathcal{L}\left(x_k, \bar{\mu}\right) - \mathcal{L}\left(\bar{x}, \bar{\mu}\right)\right] = 0$$

and, subsequentially,

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

A. Silveti-Falls, C. M., J. Fadili, *Generalized Conditional Gradient with Augmented Lagrangian for Composite Minimization.* arxiv.org/abs/1901.01287, 2018

Some bibliography

- M. Franke and P. Wolfe, *An algorithm for quadratic programming.* Naval research logistics quarterly, 1956
- M. Jaggi, *Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization.* Proceedings of Machine Learning, 2013
- G. Gidel, F. Pedregosa and S. Lacoste-Julien, *Frank-Wolfe Splitting via Augmented Lagrangian Method.* 10th NIPS Workshop on Optimization for Machine Learning, 2018
- A. Yurtsever, O. Fercoq, F. Locatello and V. Cevher, *A Conditional Gradient Framework for Composite Convex Minimization.* Proceedings of the 35th Internat Conf on Machine Learning, 2018
- A. Silveti-Falls, C. Molinari, J. Fadili, *Generalized Conditional Gradient with Augmented Lagrangian for Composite Minimization.* arxiv.org/abs/1901.01287, 2018