PUC- RIO
DEPARTAMENTO DE INFORÁTICA
PROF.: MARCO MOLINARO

Frank-Wolfe Method and Convergence Rates

Andrew W. Rosemberg[1]

[1]Pontifical Catholic University of Rio de Janeiro
[†]Work supported by CAPES Foundation

December 14, 2018

Agenda I

Theme

The focus of the current presentation:

- Overview of **The Frank-Wolfe method**.
- Prove the well-known and *tight* convergence rate on the order of $O(\frac{1}{t})$.
- Discuss the faster convergence rates for specific cases on the order of $O(\frac{1}{t^2})$.

Based on the article [Garber et al., 2014] and the chapter 7 of the book [Hazan et al., 2016].

Introduction

- **The Frank-Wolfe method** (or **Conditional Gradient method**) is a first order method for the minimization of a smooth convex function over a convex set.
- Introduced by Frank and Wolfe in the 1950's [Frank and Wolfe, 1956].
- A first-order and projection-free method.

The Frank-Wolfe Algorithm

Preliminaries

- Convexity: function and set.
- Optimization.
- Definitions:

For the following definition let E be a finite vector space and $||\cdot||$ be a norm over E.

Definition 1 (smooth function). *We say that a function $f : E \to \mathbb{R}$ is β smooth over a convex set $\mathcal{K} \subset E$ with respect to $||\cdot||$ if it holds that:*
$\forall x, y \in \mathcal{K}$,

$$f(y) \leq f(x) + \nabla f(x) \cdot (y - x) + \frac{\beta}{2}||x - y||^2 \tag{1}$$

The Algorithm

- **Objective**: Minimization of a convex function $f : E \to \mathbb{R}$ which is β_f-smooth with respect to a norm $||\cdot||$.
- **Feasible region**: A convex and compact set $\mathcal{K} \subset E$.
- $\mathcal{K}$ is given terms of a LP oracle $\mathcal{O}_\mathcal{K} : E \to \mathcal{K}$, such that

$$x = \mathcal{O}_\mathcal{K}(c), \quad x \in \arg\min_{y \in \mathcal{K}} y \cdot c$$

Algorithm 1: Frank-Wolfe Algorithm

Let x_0 be an arbitrary point in $\mathcal{K}$.
for $t = 0, 1, \ldots$ **do**
    $p_t \leftarrow \mathcal{O}_\mathcal{K}(\nabla f(x_t))$;
    $\eta_t \leftarrow \arg\min_{\eta \in [0,1]} \eta(p_t - x_t) \cdot \nabla f(x_t) + \eta^2 \frac{\beta_f}{2} ||p_t - x_t||^2$;
    $x_{t+1} \leftarrow x_t + \eta_t(p_t - x_t)$;
end

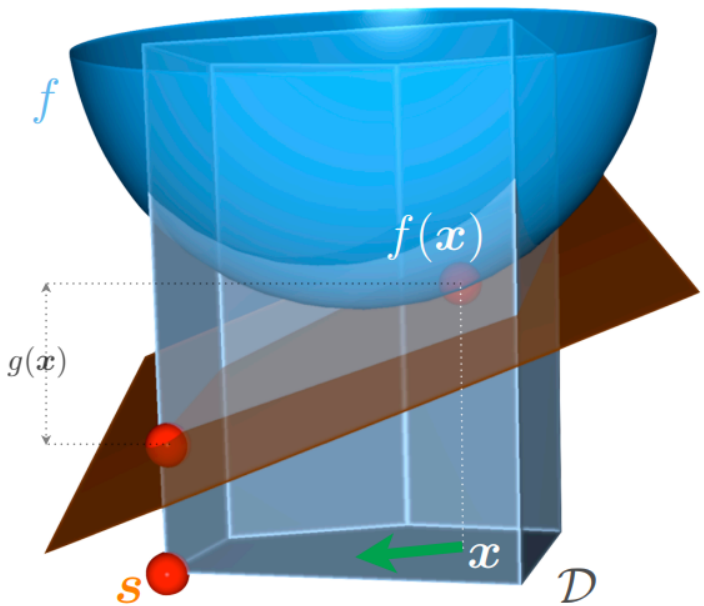

Figure: [Jaggi, 2013]

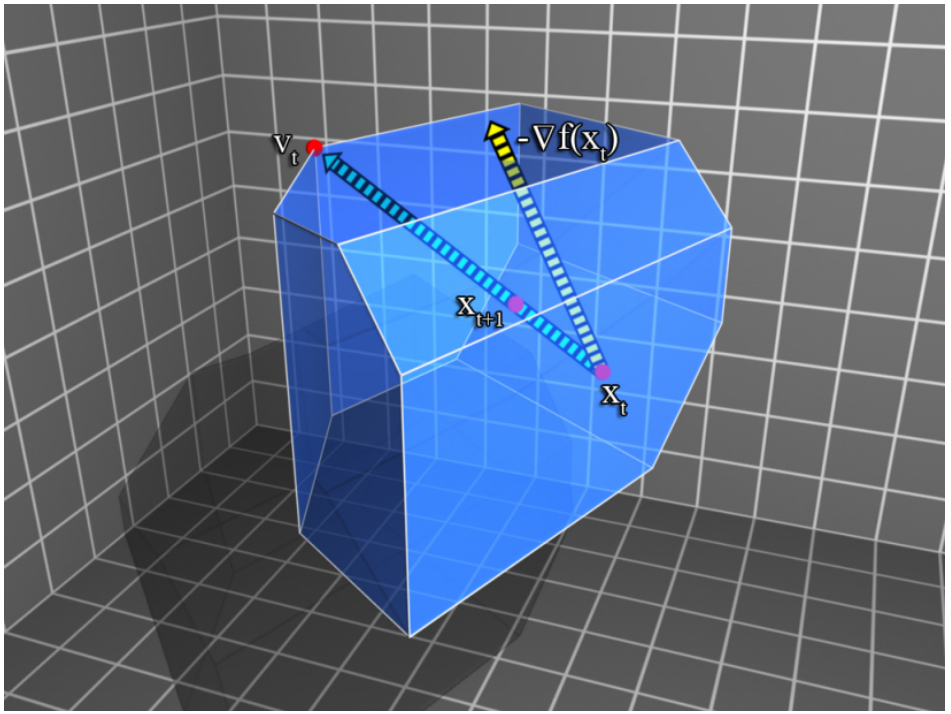

Figure: [Hazan et al., 2016]

Convergence Rates

General Case

Theorem 1 *Let* $x^* \in \arg\min_{x \in \mathcal{K}} f(x)$ *and denote* $D_{\mathcal{K}} = \max_{x,y \in \mathcal{K}} ||x - y||$ *(the diameter of the set with respect to* $||\cdot||$*). For every* $t \geq 1$ *the iterate* x_t *of Algorithm 1 satisfies:*

$$f(x_t) - f(x^*) \leq \frac{8\beta_f D_{\mathcal{K}}^2}{t} = O\left(\frac{1}{t}\right) \tag{2}$$

Step Variants with Same Rate

Let $x^* \in \arg\min_{x \in \mathcal{K}} f(x)$, $h_t = f(x_t) - f(x^*)$ and $d_t = p_t - x_t$

- $\eta_t = min\{\frac{h_t}{\beta_f ||d_t||^2}, 1\}$
- $\eta_t = \frac{2 \cdot H}{t}$, where $H \geq max\{h_1, 1\}$
- $\eta_t = \frac{2 \cdot t}{t+2}$

Preliminaries 2

For the following definitions let E be a finite vector space and $||\cdot||$, $||\cdot||_*$ be a pair of dual norms over E.

Definition 2 (strongly convex function). *We say that a function $f : E \to \mathbb{R}$ is α-strongly convex over a convex set $\mathcal{K} \subset E$ with respect to $||\cdot||$ if it satisfies the following two equivalent conditions:*

1.$\forall x, y \in \mathcal{K}$:

$$f(y) \geq f(x) + \nabla f(x) \cdot (x - y) + \frac{\alpha}{2}||x - y||^2 \tag{3}$$

2.$\forall x, y \in \mathcal{K}, \lambda \in [0, 1]$:

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) - \frac{\alpha}{2}\lambda(1 - \lambda)||x - y||^2 \tag{4}$$

Definition 3 (strongly convex set). *We say that a convex set $\mathcal{K} \subset E$ is is α-strongly convex with respect to $||\cdot||$ if $\forall x, y \in \mathcal{K}, \lambda \in [0,1]$: and any vector $z \in E$ such that $||z|| = 1$ it holds that:*

$$\lambda x + (1-\lambda)y + \lambda(1-\lambda)\frac{\alpha}{2}||x-y||^2 z \in \mathcal{K} \tag{5}$$

That is, $\mathcal{K}$ contains a ball of of radius $\lambda(1-\lambda)\frac{\alpha}{2}||x-y||^2$ induced by the norm $||\cdot||$ centered at $\lambda x + (1-\lambda)y$.

Strongly Convex Case

Contribution of the article [Garber et al., 2014].
Consider:

- f β_f-smooth with respect to $||\cdot||$.
- f α_f-strongly convex with respect to $||\cdot||$ (further relaxed on the article).
- $\mathcal{K}$ $\alpha_{\mathcal{K}}$-strongly convex with respect to $||\cdot||$.

Theorem 2 *Let* $x^* \in \arg\min_{x\in\mathcal{K}} f(x)$ *and* $M = \frac{\sqrt{\alpha_f}\alpha_{\mathcal{K}}}{8\sqrt{2}\beta_f}$. *Denote* $D_{\mathcal{K}} = \max_{x,y\in\mathcal{K}} ||x-y||$ *(the diameter of the set with respect to* $||\cdot||$*). For every* $t \geq 1$ *the iterate* x_t *of Algorithm 1 satisfies:*

$$f(x_t) - f(x^*) \leq \frac{max\{\frac{9}{2}\beta_f D_{\mathcal{K}}^2, 18M^{-2}\}}{(t+2)^2} = O\left(\frac{1}{t^2}\right) \tag{6}$$

Other Cases

Reference	$\mathcal{K}$	f	Location of x^*	Conv. rate
[Jaggi, 2013]	convex	convex	unrestricted	$1/t$
[Guélat et al., 1986]	polytope	strongly convex	interior	$exp(-\Theta(t))$
[Beck et al., 2004]	convex	$f(x) = \|\|Ax - b\|\|_2^2$	interior	$exp(-\Theta(t))$
[Levitin et al., 1966] [Demyanov et al., 1970] [Dunn, 1979]	strongly convex	$\|\|\nabla f(x)\|\| \geq c > 0$	unrestricted	$exp(-\Theta(t))$
[Garber et al., 2014]	strongly convex	strongly convex	unrestricted	$1/t^2$

Bibliography

Garber et al. Faster rates for the frank-wolfe method over strongly-convex sets. *arXiv preprint arXiv:1406.1305*, 2014.

Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956. doi: 10.1002/nav.3800030109. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800030109`.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 427–435, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL `http://proceedings.mlr.press/v28/jaggi13.html`.

Guélat et al. Some comments on wolfe's 'away step'. *Mathematical Programming*, 35(1):110–119, 1986.

Beck et al. A conditional gradient method with linear rate of convergence for solving convex linear systems. *Mathematical Methods of Operations Research*, 59(2):235–247, 2004.

Levitin et al. Constrained minimization methods. *USSR Computational mathematics and mathematical physics*, 6(5):1–50, 1966.

Demyanov et al. *Approximate methods in optimization problems*, volume 32. Elsevier Publishing Company, 1970.

Joseph C Dunn. Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals. *SIAM Journal on Control and Optimization*, 17(2):187–211, 1979.