Frank-Wolfe Method is Automatically Adaptive to Error Bound Condition

Yi Xu
Tianbao Yang

Department of Computer Science, The University of Iowa, Iowa City, IA 52242

October 10, 2018

Abstract

Error bound condition has recently gained revived interest in optimization. It has been leveraged to derive faster convergence for many popular algorithms, including subgradient methods, proximal gradient method and accelerated proximal gradient method. However, it is still unclear whether the Frank-Wolfe (FW) method can enjoy faster convergence under error bound condition. In this short note, we give an affirmative answer to this question. We show that the FW method (with a line search for the step size) for optimization over a strongly convex set is automatically adaptive to the error bound condition of the problem. In particular, the iteration complexity of FW can be characterized by \(O(\max(1/\epsilon^{1-\theta}, \log(1/\epsilon)))\) where \(\theta \in [0, 1]\) is a constant that characterizes the error bound condition. Our results imply that if the constrained set is characterized by a strongly convex function and the objective function can achieve a smaller value outside the considered domain, then the FW method enjoys a fast rate of \(O(1/t^2)\).

1. Introduction

In this draft, we consider the following constrained convex optimization problem:

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

where \(f(w)\) is a smooth function and \(\Omega \subseteq \mathbb{E}\) is a bounded strongly convex set. We assume that linear optimization over \(\Omega\) is much more cheaper than projection onto \(\Omega\), which makes the FW method more suitable for solving the above problem than gradient methods. The goal of this paper is to show that the FW method is automatically adaptive to an error bound condition of the optimization problem. Below, we will first review the FW method and the error bound condition. In next section, we will prove that the FW method is automatically adaptive to the error bound condition.

The original FW method, introduced by Frank and Wolfe (1956) (a.k.a. Conditional Gradient method (Levitin and Polyak, 1966)), is a projection-free first-order method for minimizing smooth convex objective functions over a convex set. In recent years, the FW method has gained an increasing interest in large-scale optimization and machine learning (e.g., (Garber and Hazan, 2015; Freund and Grigas, 2016; Nesterov, 2018; Narasimhan, 2018)). Many existing works have shown the convergence rate of the standard FW method is \(O(1/t)\) even for strongly convex objectives (Clarkson, 2008; Hazan, 2008; Jaggi, 2013), and in general the rate could not be improved. Under different assumptions or for some
special cases, a series of works tried to get faster rates of the FW method and its variants (Levitin and Polyak, 1966; Demyanov and Rubinov, 1970; Dunn, 1979; Guélat and Marcotte, 1986; Beck and Teboulle, 2004; Garber and Hazan, 2013; Lan, 2013; Lacoste-Julien and Jaggi, 2013; Garber and Hazan, 2015; Lacoste-Julien and Jaggi, 2015; Lan and Zhou, 2016). For example, for minimizing smooth and strongly convex objective functions over a strongly convex set, Garber and Hazan (2015) showed that the FM method enjoyed fast rate of $O(1/t^2)$.

In this paper, we first consider the FW method shown in Algorithm 1, where $L_f$ denotes a smoothness constant of $f(x)$ with respect to $\| \cdot \|$ such that $f(x) \leq f(y) + \nabla f(y)^\top (x - y) + \frac{L_f}{2} \| x - y \|^2$ holds for any $x, y \in \Omega$. Note that both options for selecting the step size have been considered in the literature (Jaggi, 2013; Garber and Hazan, 2015). Option I requires evaluating the objective function but does not need to know the smoothness constant. Option II could be cheaper but requires knowing the Lipschitz constant of the gradient. Our analysis applies to both options. In the sequel, we will focus on option I, with which we have

$$f(x_{t+1}) \leq f(x_t) + \eta(y_t - x_t), \forall \eta \in [0, 1]$$

$$\leq f(x_t) + \eta(y_t - x_t)^\top \nabla f(x_t) + \frac{\eta^2 L_f}{2} \| y_t - x_t \|^2, \forall \eta \in [0, 1] \quad (2)$$

Note that for option II, the second inequality above still holds.

We consider the following definition of error bound condition for the optimization problem (1).

**Definition 1 (Hölderian error bound (HEB))** A function $f(x)$ is said to satisfy a HEB condition on $\Omega$ if there exist $\theta \in [0, 1]$ and $0 < c < \infty$ such that for any $x \in \Omega$

$$\min_{w \in \Omega_*} \| x - w \| \leq c(f(x) - f_*)^\theta. \quad (3)$$

where $\Omega_*$ denotes the optimal set of $\min_{x \in \Omega} f(x)$ and $f_*$ denotes the optimal objective value.

It is notable that $\theta = 0$ is a trivial condition since it always hold due to that $\Omega$ is a compact set. The above HEB condition has been considered for deriving faster convergence of subgradient methods (Yang and Liu, 2018), proximal gradient method (Liu and Yang, 2017), accelerated gradient method (Xu et al., 2016), and stochastic subgradient methods (Xu et al., 2017a,b; Liu and Yang, 2017; Yang and Lin, 2018). For example, when functions are semialgebraic and “regular” (for instance, continuous), the above inequality is known to hold on any compact set (c.f. (Bolte et al., 2017) and references therein).

The last definition in this section is regarding the strongly convex set.

**Definition 2** A convex set $\Omega$ is a $\alpha$-strongly convex with respect to $\| \cdot \|$ if for any $x, y \in \Omega$, any $\gamma \in [0, 1]$ and any vector $z \in \mathbb{E}$ such that $\| z \| = 1$, it holds that

$$\gamma x + (1 - \gamma)y + \gamma(1 - \gamma)^\alpha \frac{\alpha}{2} \| x - y \|^2 z \in \Omega.$$

**Remark.** Many previous works (e.g., (Levitin and Polyak, 1966; Demyanov and Rubinov, 1970; Dunn, 1979; Garber and Hazan, 2015)) considered this condition of feasible set when studying the FW method.
**Algorithm 1** Frank-Wolfe Method

Initilization: $x_0 \in \Omega$

for $t = 0, \ldots, T$ do

Compute $y_t \in \arg \min_{y \in \Omega} \nabla f(x_t)^\top y$

Option I: Set $\eta_t = \arg \min_{\eta \in [0, 1]} f(x_t + \eta(y_t - x_t))$

Option II: Set $\eta_t = \arg \min_{\eta \in [0, 1]} \eta(y_t - x_t)^\top \nabla f(x_t) + \frac{\eta^2 L f}{2} \|y_t - x_t\|^2$

Compute $x_{t+1} = x_t + \eta_t(y_t - x_t)$

end for

2. Adaptive Convergence of the FW method

In this section, we show that the FW method is automatically adaptive to the HEB condition, enjoying a faster convergence rate than the standard $O(1/t)$ rate without the knowledge of the HEB condition.

We first prove the following lemma.

**Lemma 3** Assume $f(x)$ obeys the HEB condition on $\Omega$ with $\theta \in [0, 1]$, then it holds that

$$\|\nabla f(x)\|_* \geq \frac{1}{c} (f(x) - f_*)^{1-\theta}.$$

**Proof** Let $x_*$ denote the optimal solution in $\Omega_*$ that is closest to $x$ measured in $\| \cdot \|$. By convexity of $f(\cdot)$, we have

$$f(x_*) \geq f(x) + \nabla f(x)^\top (x_* - x).$$

Thus,

$$f(x) - f(x_*) \leq \|\nabla f(x)\|_* \|x_* - x\| \leq c(f(x) - f_*)^\theta \|\nabla f(x)\|_*.$$

As a result,

$$\|\nabla f(x)\|_* \geq \frac{1}{c} (f(x) - f_*)^{1-\theta}.\]

The second lemma is from (Garber and Hazan, 2015).

**Lemma 4** For the FW method given in Algorithm 1, for $t = 0, \ldots, T$ we have

$$f(x_{t+1}) - f_* \leq (f(x_t) - f_*) \max \left\{ \frac{1}{2}, \left( 1 - \frac{\alpha \|\nabla f(x_t)\|_*}{8 L f} \right) \right\}.$$

Finally, we prove the following theorem.

**Theorem 1** For every $t \geq 1$, we have

$$f(x_t) - f_* \leq \begin{cases} \frac{c}{(t+k)^{1/(1-\theta)}} & \text{if } \theta \in [0, 1) \\ \rho^t (f(x_0) - f_*) & \text{otherwise} \end{cases}$$
where $k \geq \max \left\{ \frac{2-2^{1-\theta}}{2^{1-\theta}-1}, C' \right\}$, $C \geq \max \left\{ \frac{L_f D^2 (1+k)^{1-\theta}/2}{2}, 2 \left( \frac{C'}{M'} \right)^{1-\theta} \right\}$, $C' = \frac{1}{1-\theta(2^{1-\theta}-1)}$, and $\rho = \max \left\{ \frac{1}{2}, 1 - \frac{\alpha}{8cL_f} \right\}$.

Remark. In order to find an $\epsilon$-approximate solution $x_t$ such that $f(x_t) - f_* \leq \epsilon$, the iteration complexity of FW method is $O(\max(1/\epsilon^{1-\theta}, \log(1/\epsilon)))$ with $\theta \in [0, 1]$.

Proof When $\theta = 1$, the conclusion is trivial, which follows directly from Lemma 4. Next, we prove for $\theta \in [0, 1)$. Let $\beta = 1 - \theta$, $h_t = f(x_t) - f_*$. Combining Lemma 3 and Lemma 4, we have

$$h_{t+1} \leq h_t \max \left\{ \frac{1}{2}, 1 - \frac{\alpha}{8cL_f} h_t^\beta \right\} = h_t \max \left\{ \frac{1}{2}, 1 - M h_t^\beta \right\} \tag{4}$$

We prove by induction that $h_t \leq \frac{C}{(t+k)^{1/\beta}}$.

For the case $t = 1$, following (2) we have

$$h_1 \leq h_0(1 - \eta) + \frac{L_f \eta D^2}{2} \leq \max \left\{ \frac{L_f D^2}{2}, h_0 \right\} \leq \frac{L_f D^2}{2}, \forall \eta \in [0, 1],$$

where we use the fact that $h_0 = f(x_0) - f(x_*) \leq \nabla f(x_*)^T (x_0 - x_*) + \frac{L_f}{2} \|x_0 - x_*\|^2 = \frac{L_f}{2} \|x_0 - x_*\|^2 \leq \frac{L_f D^2}{2}$. As long as $C/(1+k)^{1/\beta} \geq L_f D^2/2$, we have the conclusion holds for $t = 1$.

Next, we consider $t \geq 1$. First assume that the max operation in (4) gives $1/2$, i.e.,

$$h_{t+1} \leq \frac{h_t}{2} \leq \frac{C}{2(t+k)^{1/\beta}} \leq \frac{C}{(t+1+k)^{1/\beta}} \leq \frac{C}{(t+1+k)^{1/\beta}},$$

where the last inequality holds as long as

$$\frac{(t+1+k)^{1/\beta}}{(t+k)^{1/\beta}} \leq 1, \forall t \geq 1, \text{ i.e., } 1 + \frac{1}{t+k} \leq 2^\beta, \forall t \geq 1, \text{ i.e., } k \geq \frac{2-2^\beta}{2^\beta-1}.$$ 

Next, consider the case that the max operation is the second argument. In this case, if $h_t \leq \frac{C}{2(t+k)^{1/\beta}}$, the same conclusion holds under the above condition of $k$. Otherwise, $h_t \geq \frac{C}{2(t+k)^{1/\beta}}$. We have

$$h_{t+1} \leq h_t (1 - M h_t^\beta) \leq \frac{C}{(t+k)^{1/\beta}} \left( 1 - M \left( \frac{C}{2} \right)^{1/\beta} \frac{1}{t+k} \right) \leq \frac{C}{(t+k+1)^{1/\beta}} \left( \frac{(t+k+1)^{1/\beta}}{(t+k)^{1/\beta}} \left( 1 - M \left( \frac{C}{2} \right)^{1/\beta} \frac{1}{t+k} \right) \right) \leq \frac{C}{(t+k+1)^{1/\beta}} \left( 1 + \frac{C'}{t+k} \right) \left( 1 - \frac{C'}{t+k} \right)$$

To show the last inequality holds, we can set $C' = \frac{1}{\beta - (1-\beta)(2^{1-\beta}-1)} > 1$ and $C \geq 2(C'/M)^{1/\beta}$.

To see this, we need to show that

$$\log(1 + C' x) - \frac{1}{\beta} \log(1 + x) \geq 0, \forall 0 \leq x \leq 2^\beta - 1.$$
In fact, due to $\frac{C'}{1 + C'} - \frac{1}{\beta(1 + x)} \geq 0, \forall 0 \leq x \leq 2^\beta - 1$, it gives $1 + C'x \geq (1 + x)^{1/\beta}$ holds for all $0 \leq x \leq 2^\beta - 1$. Plugging $x = 1/(t + k) \leq 2^\beta - 1$ into this inequality, we get what we want $1 + C'/(t + k) \geq (1 + \frac{1}{t + k})^{1/\beta}$.

\[ \boxed{1 + C'/(t + k) \geq (1 + \frac{1}{t + k})^{1/\beta}.} \]

### 3. Examples

Lastly, we give examples exhibiting the HEB condition with $\theta = 1/2$. In particular, let us consider

\[ \min_{g(x) \leq r} f(x) \tag{5} \]

where $g(x)$ is a non-negative, strongly and smooth function. It is shown that $\Omega = \{x : g(x) \leq r\}$ is a strongly convex set (Garber and Hazan, 2015).

**Lemma 5** Assume that $\min_x f(x) < \min_{g(x) \leq r} f(x)$ and there exists a $x_0$ such that $g(x_0) < r$, then the above problem satisfies HEB with $\theta = 1/2$.

**Proof** We set $\Omega = \{x : g(x) \leq r\}$ and $\Omega_* = \arg \min_{g(x) \leq r} f(x)$, and we define an indicator function as follows,

\[ I_{\Omega}(x) = \begin{cases} 0 & \text{if } x \in \Omega, \\ +\infty & \text{if } x \notin \Omega. \end{cases} \]

Then the problem of (5) can be written as

\[ \min_x \hat{f}(x) := f(x) + I_{\Omega}(x), \]

and thus we also have $\Omega_* = \arg \min_x \hat{f}(x)$. We only need to consider any fixed $x_* \in \Omega_*$. By the condition of $g(x_0) < r$ and Corollary 28.2.1 of (Rockafellar, 1970), there exists $\lambda^* \geq 0$ such that

\[ \hat{f}(x_*) = \min_x \hat{f}(x) = f(x_*) = \min_{x \in \Omega} f(x) = \min_x \{f(x) + \lambda^*(g(x) - r)\} \]

\[ \leq f(x_*) + \lambda^*(g(x_*) - r) \leq f(x_*), \tag{6} \]

where the first inequality is due to $x_* \in \Omega_*$; the second inequality uses the fact that $x_* \in \Omega_*$ hence $g(x_*) - r \leq 0$. Then, equality holds for (6), which implies $f(x_*) + \lambda^*(g(x_*) - r) = f(x_*)$, that is,

\[ \lambda^*(g(x_*) - r) = 0. \tag{7} \]

On the other hand, let $u_* \in \arg \min_x f(x)$, then based on the assumption of $\min_x f(x) < \min_{g(x) \leq r} f(x)$ we know $u_* \notin \Omega_*$ hence $u_* \notin \Omega$. By (6), we also know

\[ f(u_*) < \min_{x \in \Omega} f(x) = \min_x \{f(x) + \lambda^*(g(x) - r)\} \leq f(u_*) + \lambda^*(g(u_*) - r), \]

\[ \boxed{f(u_*) < \min_{x \in \Omega} f(x) = \min_x \{f(x) + \lambda^*(g(x) - r)\} \leq f(u_*) + \lambda^*(g(u_*) - r),} \]
which implies
\[ \lambda^*(g(u^*) - r) > 0. \] (8)

Since \( u^* \notin \Omega \), then \( g(u^*) - r > 0 \). In order to have (8), we need \( \lambda^* > 0 \). Thus, by (7) we have
\[ g(x^*) - r = 0. \] (9)

For any such \( \lambda^* > 0 \), then by Theorem 28.1 of (Rockafellar, 1970), we also have
\[
\Omega_* = \{ x : g(x) = r \} \cap \arg \min_x \{ f(x) + \lambda^*(g(x) - r) \}. \tag{10}
\]

Since \( g(x) \) is strongly convex, \( f(x) \) is convex and \( \lambda^* > 0 \), then \( f(x) + \lambda^*(g(x) - r) \) is also strongly convex, implying that \( v_* = \arg \min_x \{ f(x) + \lambda^*(g(x) - r) \} \) is a unique constant. Due to \( \lambda^* > 0 \), \( g(v_*) \) is also a constant (Li and Pong, 2017). By (10) we have \( g(v_*) = g(x_*) = r \). Therefore,
\[
\Omega_* = \arg \min_x \{ f(x) + \lambda^*(g(x) - r) \}. \tag{11}
\]

By the strong convexity of \( f(x) + \lambda^*(g(x) - r) \) we know for any \( x \in \Omega \) and \( x_* \in \Omega_* \subseteq \Omega \),
\[
\frac{1}{c^2} \| x - x_* \|^2 \leq f(x) + \lambda^*(g(x) - r) - [f(x_*) + \lambda^*(g(x_*) - r)],
\]
where \( c > 0 \). Since \( \lambda^* > 0 \), \( g(x) - r \leq 0 \) and \( g(x_*) - r = 0 \), we get
\[
\frac{1}{c^2} \| x - x_* \|^2 \leq f(x) - f(x_*).
\]

Therefore, for any \( x \in \Omega \)
\[
\min_{w \in \Omega_*} \| x - w \| \leq c(f(x) - f_*)^{1/2},
\]
which implies \( \theta = 1/2 \).

References

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

Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W Suter. From error bounds to the complexity of first-order descent methods for convex functions. Mathematical Programming, 165(2):471–507, 2017.

Kenneth L Clarkson. Coresets, sparse greedy approximation, and the frank-wolfe algorithm. In Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms (SODA), pages 922–931. Society for Industrial and Applied Mathematics, 2008.
Vladimir Fedorovich Demyanov and Aleksandr Moiseevich Rubinov. *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.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

Robert M Freund and Paul Grigas. New analysis and results for the frank–wolfe method. *Mathematical Programming*, 155(1-2):199–230, 2016.

Dan Garber and Elad Hazan. Playing non-linear games with linear oracles. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 420–428, 2013.

Dan Garber and Elad Hazan. Faster rates for the frank-wolfe method over strongly-convex sets. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 541–549, 2015.

Jacques Guélat and Patrice Marcotte. Some comments on wolfe’s away step. *Mathematical Programming*, 35(1):110–119, 1986.

Elad Hazan. Sparse approximate solutions to semidefinite programs. In *Latin American symposium on theoretical informatics*, pages 306–316. Springer, 2008.

Martin Jaggi. Revisiting frank-wolfe: projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 427–435, 2013.

Simon Lacoste-Julien and Martin Jaggi. An affine invariant linear convergence analysis for frank-wolfe algorithms. *arXiv preprint arXiv:1312.7864*, 2013.

Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of frank-wolfe optimization variants. In *Advances in Neural Information Processing Systems (NIPS)*, pages 496–504, 2015.

Guanghui Lan. The complexity of large-scale convex programming under a linear optimization oracle. *arXiv preprint arXiv:1309.5550*, 2013.

Guanghui Lan and Yi Zhou. Conditional gradient sliding for convex optimization. *SIAM Journal on Optimization*, 26(2):1379–1409, 2016.

ES Levitin and BT Polyak. Constrained minimization methods. *USSR Computational Mathematics and Mathematical Physics*, 6(5):1–50, 1966.

Guoyin Li and Ting Kei Pong. Calculus of the exponent of kurdyka–lojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of Computational Mathematics*, pages 1–34, 2017.
Mingrui Liu and Tianbao Yang. Adaptive accelerated gradient converging method under hölderian error bound condition. In *Advances in Neural Information Processing Systems*, pages 3104–3114, 2017.

Harikrishna Narasimhan. Learning with complex loss functions and constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 1646–1654, 2018.

Yu Nesterov. Complexity bounds for primal-dual methods minimizing the model of objective function. *Mathematical Programming*, 171(1-2):311–330, 2018.

R Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Yi Xu, Yan Yan, Qihang Lin, and Tianbao Yang. Homotopy smoothing for non-smooth problems with lower complexity than $O(1/\epsilon)$. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1208–1216, 2016.

Yi Xu, Qihang Lin, and Tianbao Yang. Stochastic convex optimization: Faster local growth implies faster global convergence. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 3821–3830, 2017a.

Yi Xu, Mingrui Liu, Qihang Lin, and Tianbao Yang. ADMM without a fixed penalty parameter: Faster convergence with new adaptive penalization. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 1267–1277, 2017b.

Tianbao Yang and Qihang Lin. Rsg: Beating subgradient method without smoothness and strong convexity. *Journal of Machine Learning Research*, 19(6), 2018.