Non-convex Optimization with Frank-Wolfe Algorithm and Its Variants

Jean Lafond, Hoi-To Wai and Eric Moulines

Abstract

Recently, Frank-Wolfe (a.k.a. conditional gradient) algorithm has become a popular tool for tackling machine learning problems as it avoids the costly projection computation in traditional first-order optimization methods. While the Frank-Wolfe (FW) algorithm has been extensively studied for convex optimization, little is known for the FW algorithm in non-convex optimization. This paper presents a unified convergence analysis for FW algorithm and its variants under the setting of non-convex but smooth objective with a convex, compact constraint set. Our results are based on a novel observation on the so-called Frank-Wolfe gap (FW gap), which measures the closeness of solution to a stationary point. With a diminishing step size, we show that the FW gap decays at a rate of $O\left(\sqrt{\frac{1}{t}}\right)$; and the same rate holds for variants of FW such as the online FW algorithm and decentralized FW algorithm. Numerical experiments are shown to support our findings.

1 Introduction

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable (possibly non-convex) function and $C \subseteq \mathbb{R}^d$ be a closed and bounded convex set, we consider the following optimization problem:

$$\min_{\theta} F(\theta) \text{ s.t. } \theta \in C.$$  

This paper studies the Frank-Wolfe (FW) algorithm that has become popular recently due to its projection-free feature. Comparing to traditional projected gradient algorithms (PGAs), the FW algorithm involves solving a linear optimization (LO) that can be performed much more efficiently than the projection step required by PGA; see [1].

Previous research have focused on convex optimization with FW/FW-based algorithms. For example, [2, 3] studied conditions when FW algorithm converges at a linear rate; [4] studied an Online FW algorithm with a regret bound of $O(1/\sqrt{T})$, where $T$ is the number of rounds played; [5] combined FW algorithm with the popular stochastic variance reduction gradient (SVRG) method to efficiently handle finite-sum optimization problems.

On the other hand, little is known on non-convex optimization with FW/FW-based algorithms. Recent results can be found in several unpublished works, e.g., [6] considered an adaptive step size rule in FW algorithm to yield a convergence rate of $O(1/\sqrt{T})$ which is similar to ours; [7] studied a fixed step size rule but achieved a slightly worse convergence rate than ours; [8] applied the SVRG techniques on FW algorithm with non-convex objectives.

Contributions. This paper presents a unified analysis on the convergence of FW algorithm(s) for non-convex optimization. Under the setting of smooth objective function and a bounded convex constraint set, we show that the limit points of the iterates generated by FW algorithms are stationary points of (1), and they can be found at the fastest rate of $O(1/\sqrt{T})$. We also provide additional conditions when the convergence rate can be accelerated. Lastly, we demonstrate an interesting application on sparse+low rank matrix completion and provide numerical experiments to support our findings.

Notations. For $d \in \mathbb{N}$, we denote the set $\{1, \ldots, d\}$ as $[d]$. The $i$th element of a vector $\theta$ is $[\theta]_i$. The Euclidean norm is denoted by $\| \cdot \|$. A function $f$ is $L$-smooth if $f(\theta) - f(\theta') \leq \langle \nabla f(\theta'), \theta - \theta' \rangle + L\|\theta - \theta'\|^2/2$ for all $\theta, \theta' \in \mathbb{R}^d$. 

*J. Lafond is with Institut Mines-Telecom, Telecom ParisTech, CNRS LTCI, France. H.-T. Wai is with School of ECEE, Arizona State University, USA. The first two authors have contributed equally. E. Moulines is with CMAP, Ecole Polytechnique, France. Emails: lafond.jean@gmail.com, htwai@asu.edu, eric.moulines@polytechnique.edu

Preliminaries. Consider the classical FW algorithm applied to (1). Let \( \theta_t \in \mathcal{C} \) be a feasible solution to (1) and \( \gamma_t \in (0, 1] \) be some step size, at iteration \( t \in \mathbb{N} \), we perform:

\[
\theta_{t+1} = \theta_t + \gamma_t (a_t - \theta_t), \quad \text{where} \quad a_t := \min_{a \in \mathcal{C}} \langle \nabla F(\theta_t), a \rangle.
\] (2)

The latter optimization in (2) is known as the linear optimization (LO) step required by FW algorithm(s). This can be viewed as the projection-free counterpart to the Euclidean projection step required by traditional PGA, i.e., \( \min_{\theta \in \mathbb{R}^d} \| \theta' - \theta \| \). In many interesting cases, the LO step admits a more efficient solution than the projection.

For convex problems, a well known fact is that the objective value of the FW algorithm converges to the minimum at a rate of at least \( O(1/t) \) [1], i.e., \( f(\theta_t) - f(\theta^*) = O(1/t) \), where \( \theta^* \) is an optimum solution to (1). This rate can be accelerated under some conditions, e.g., see [2, 3]. For non-convex problems, it is not possible to take the objective values’ differences as a benchmark. Instead, we focus on the following FW/Duality gap:

\[
g_t := \max_{\theta \in \mathbb{C}} \langle \nabla f(\theta_t), \theta - \theta_t \rangle = \langle \nabla f(\theta_t), \theta_t - a_t \rangle.
\] (3)

Importantly, if \( g_t = 0 \), then \( \langle \nabla f(\theta_t), \theta_t - \theta \rangle \leq 0 \) for all \( \theta \in \mathcal{C} \), i.e., \( \theta^t \) is a stationary point to Problem (1). Like [6, 7, 8], we can take \( g_t \) as a measure of the stationarity of the iterate \( \theta_t \). As mentioned, the convergence of the FW algorithm for non-convex objective functions has only been considered recently by a few authors [6, 7, 8], and they studied the convergence rate of FW algorithm in terms of the FW gap bound can be improved to \( O(1/T^\alpha) \) for the epoch considered. (iii) Finally, when \( C_0 = 0 \), i.e., \( F_t(\theta) = F(\theta) \) for all \( t \geq 1, \eta + \alpha > 1 \) and \( \alpha > 0.5 \). We further assume that \( F(\theta) \) takes a finite number of values for the stationary points \( \theta \), then the sequence \( \{\theta_t\}_{t \geq 1} \) has limit points and each limit point \( \bar{\theta} \) satisfies

\[
\max_{\theta \in \mathbb{C}} \langle \nabla F(\theta), \theta - \theta_t \rangle = 0.
\] (7)

We relegate the proof of Theorem 1 to Section 3 and Appendix A. Notice that if \( \eta \geq 1, \alpha \geq 0.5 \), then a convergence rate of \( O(1/\sqrt{T}) \) is achieved. This matches the rate for PGA on non-convex problems in [9]. Moreover, in our numerical experiment, we observe that the FW gap often decays at \( O(1/t^\alpha) \) for \( \alpha > 0.5 \), this can be accounted for using (6). Below we list a few cases that satisfy H1, H2, and thus can be analyzed using our results.
2.1 Online FW (O-FW) algorithm

Like [4], we consider a fully informational setting for the online FW (O-FW) algorithm. At round/iteration $t$, an online learner plays $\theta_t$ and receives full information about the instantaneous loss function $f_t(\theta)$. For example, $f_t(\theta)$ may correspond to the data observed at the current round. To account for the loss functions from the past, we design the time varying objective function as the temporal average $F_t(\theta) := t^{-1} \sum_{s=1}^{t} f_s(\theta)$.

Now, as $F_t(\theta)$ is fully known, its gradient $\nabla F_t(\theta)$ can be exactly evaluated at round $t$. As such, the FW algorithm [10] can be directly applied with $H_2$ automatically satisfied with $C_g = 0$. If $f_t(\theta)$ is bounded by $B$ for all $\theta \in C$, then

$$|F_t(\theta) - F_{t-1}(\theta)| = \frac{1}{t} f_t(\theta) + \sum_{s=1}^{t-1} \left( \frac{1}{t} - \frac{1}{t-1} \right) f_s(\theta) = \frac{1}{t} (f_t(\theta) - F_{t-1}(\theta)) \leq \frac{2B}{t}, \quad \forall t \geq 1,$$

i.e., $H_1$ is satisfied with $\beta = 1$, $C_b = 2B$. Consequently, the results from Theorem 1 apply directly.

2.2 Decentralized FW (DeFW) algorithm

In the decentralized FW (DeFW) algorithm [10], we consider solving an optimization problem of the form:

$$\min_{\theta} \left( \frac{1}{N} \sum_{i=1}^{N} f_i(\theta) \right) \text{ s.t. } \theta \in C. \quad (9)$$

The problem is to be solved distributively by a network of $N$ agents, each of them holding a private objective function $f_i(\theta)$. Our goal is for the agents to cooperatively find a stationary point to the above problem through exchanging information over network.

Following standard set-ups in distributed optimization [11], we assume that the network is described by an undirected graph $G = (V, E)$, where $V = [N]$ and $E \subseteq V \times V$. The graph is associated with a doubly stochastic weight matrix $W \in \mathbb{R}^{N \times N}$ such that $[W]_{ij} = 0$ if and only if $(i, j) \notin E$. To describe the DeFW algorithm, we denote $\theta^i_t$ as the local copy of $\theta_t$ kept by the $i$th agent at iteration $t$. We perform the following updates in order — for each $i \in [N]$,

$$\theta^i_t = \sum_{j=1}^{N} W_{ij} \cdot \theta^j_t, \quad \overline{\nabla F}_t = \sum_{j=1}^{N} W_{ij} \cdot (\nabla f_{j}(\theta^j_{t-1}) + \nabla f_{j}(\theta^j_{t})), \quad (10a)$$

$$\theta^i_{t+1} = \theta^i_t + \gamma_i (\alpha^i_t - \theta^i_t), \quad \text{where } \alpha^i_t = \arg\min_{\alpha \in C} \langle \overline{\nabla F}_t, \alpha \rangle. \quad (10b)$$

Notice that (10a) represents the gossip-based average consensus (GAC) updates [12] (executed for one round) for averaging the parameter variables and the gradient vectors; while (10b) is the standard FW update.

We analyze the convergence of the DeFW algorithm [10] by studying the average iterate $\bar{\theta}_t := N^{-1} \sum_{j=1}^{N} \theta^j_t$. Using [10], we see that $\bar{\theta}_{t+1} = \bar{\theta}_t + \gamma_i (N^{-1} \sum_{j=1}^{N} a^i_t - \bar{\theta}_t)$ and the DeFW algorithm can be analyzed under the framework of [4]. Now, $H_1$ is satisfied with $C_b = 0$ since the objective function is not time varying. Secondly, if $\|\theta\| \leq \rho$ for all $\theta \in C$, the following inequality holds for all $i \in [N]$:

$$\langle \nabla F(\bar{\theta}_t), \alpha^i_t \rangle \leq \langle \overline{\nabla F}_t, \alpha^i_t \rangle + \rho \|\overline{\nabla F}_t - \nabla F(\bar{\theta}_t)\| \leq \langle \overline{\nabla F}_t, \alpha^i_t \rangle + 2\rho \|\overline{\nabla F}_t - \nabla F(\bar{\theta}_t)\|, \quad (11)$$

where the first and last inequalities are due to Cauchy-Schwarz, and the second inequality is due to the optimality of $\alpha^i_t$. If each of $f_i$ is $L$-smooth, it can be proven that $\|\nabla F(\bar{\theta}_t)\| \leq \gamma_i C'_g$ for some $C'_g < \infty$; see [10]. Therefore, when we set $\gamma_i = t^{-\alpha}$, then $H_2$ is satisfied with $\eta = \alpha$, $C_g = 2\rho C'_g$ and thus Theorem 1 applies. Lastly, we remark that $|\theta^i_t - \bar{\theta}_t|$ also decays at the order of $O(\gamma_i)$ [10], therefore the results in Theorem 1 applies to each local variable.

3 Sketch of Proof for Theorem 1

We only sketch the proof of [5] and [6] in Theorem 1, the proof for [7] will be postponed to the appendix. Let us define \( \rho := \max_{\theta \in C} \|\theta - \theta^i_t\| \) as the diameter of $C$. Now as $F_t$ is $L$-smooth and using $H_2$ we have

$$F_t(\theta_{t+1}) \leq F_t(\theta_t) + \gamma_t \langle \nabla F_t(\theta_t), \hat{\alpha}_t - \theta_t \rangle + \gamma_t \frac{L\rho^2}{2} \leq F_t(\theta_t) + \gamma_t \langle \nabla F_t(\theta_t), \alpha_t - \theta_t \rangle + \gamma_t C_g \cdot t^{-\eta} - \frac{L\rho^2}{2} \max_{\theta \in C} \langle \nabla F_t(\theta_t), \theta_t - \theta \rangle \quad (12)$$
To obtain (5), we sum up the both sides of (12) from $t = T/2 + 1$ to $t = T$ to yield,

$$
\sum_{t=T/2+1}^{T} t^{-\alpha} \langle \nabla F_t(\theta), \theta - a_t \rangle \leq \sum_{t=T/2+1}^{T} t^{-\alpha} \left( C_g \cdot t^{-\eta} + (L\beta^2/2) \cdot t^{-\alpha} \right) + \sum_{t=T/2+1}^{T} (F_t(\theta) - F_t(\theta_{t+1}))
$$

(13)

As $\langle \nabla F_t(\theta), \theta - a_t \rangle$ is non-negative for all $t$, the left hand side can be lower bounded by $\Omega(T^{1-\alpha}) \cdot \min_{t \in [T/2+1, T]} \max_{\theta \in C} \langle \nabla F_t(\theta), \theta - \theta \rangle$. Meanwhile, the first summation on the right can be upper bounded by $O(T^{1-\min\{2\alpha, \alpha + \eta\}})$ and the summation on the right can be bounded as:

$$
F_{T/2+1}(\theta_{T/2+1}) - F_{T/2+1}(\theta_{T/2+2}) + F_{T/2+2}(\theta_{T/2+2}) - \cdots - F_{T-1}(\theta_T) + F_T(\theta_T) - F_T(\theta_{T+1})
$$

$$
\leq F_{T/2+1}(\theta_{T/2+1}) - F_{T/2+1}(\theta_{T+1}) + \sum_{t=T/2+1}^{T-1} C_b \cdot t^{-\beta} = O(T^{1-\min\{1, \beta\}})
$$

(14)

We notice that the above results hold as the summation is taken from $t = T/2 + 1$ to $t = T$ instead of $t = 1$ to $t = T$. Thus, the right hand side of (13) can be bounded by $O(T^{1-\min\{1, \beta, 2\alpha, \alpha + \eta\}})$, dividing with $\Omega(T^{1-\alpha})$ yields (5).

The results in (6) can be observed directly from (12). In particular, when statement (a) in (6) is violated, i.e., $C_t t^{-\eta} + (L^2/2) t^{-\alpha} < \max_{\theta \in C} \langle \nabla F_t(\theta), \theta - \theta \rangle$ for all $t \in [T/2 + 1, T]$, then $F_t(\theta_{t+1}) < F_t(\theta_t)$ and statement (b) holds. Otherwise, statement (a) holds whenever statement (b) is violated.

## 4 Numerical Experiment

To illustrate our analytical findings, we consider a non-convex matrix completion problem where the observations are contaminated with sparse noise. This is related to the so-called sparse+low rank matrix completion formulation [13].

Let $\sigma > 0, R > 0$ be a fixed parameter, we consider:

$$
\min_{\theta \in \mathbb{R}^{m_1 \times m_2}} \sum_{(k,l) \in \Omega} \left( 1 - \exp\left( - (Y_{k,l} - [\theta]_{k,l})^2 / \sigma \right) \right) \text{ s.t. } \| \theta \|_* \leq R,
$$

(15)

where $Y_{k,l}$ is the noisy observations on the $(k,l)$th entry of the matrix to be estimated, $\Omega \subseteq [m_1] \times [m_2]$ is the set of observed entries’ locations, $\| \cdot \|_*$ is the nuclear norm and we promote a low rank solution by the nuclear norm constraint in (15). The negated Gaussian loss gives better tolerance to outlier entries over the standard square loss.

We consider the movielens100k dataset which contains $10^4$ records of movie ratings from $m_1 = 943$ users on $m_2 = 1692$ movies. We assign $8 \times 10^3$ (resp. $2 \times 10^3$) records for the training (resp. testing) purpose. The standard FW, O-FW, DeFW algorithms are tested in the experiment. Specifically, the O-FW algorithm takes a batch of $B = 20$ new records at each round, and the DeFW algorithm is simulated on a network with $N = 50$ agents, connected via an Erdos-Renyi graph with connectivity 0.1. The weight matrix $W$ for the DeFW algorithm is designed using the Metropolis-Hastings rule. We consider the case when $\ell = 1.3$ GAC information exchange rounds are performed per iteration in the DeFW algorithm. In addition, we also test the implementations with a standard square loss. We set the step size to $\gamma_t = 2/(t+1)$ for the convex square loss; $\gamma_t = t^{-\alpha}$ with $\alpha = 0.75$ for the non-convex negated Gaussian loss; and we set $R = 10^4$.

![Fig. 1. Convergence of FW algorithms on the movielens100k dataset. (Left) noiseless observations. (Right) sparse-noise contaminated observations (20% sparsity). Same legend are used for the right figure.](image)

The numerical results are presented in Fig. 1, where we show the test mean square error (MSE) and the FW gap versus the iteration/round number. Notice that the O-FW algorithm is terminated after 4000 rounds since we only have $8 \times 10^4$ training data. As seen, for the non-convex loss, the FW gaps of the tested algorithms decreases with the iteration number at an order of $\sim O(1/\sqrt{t})$ and achieves better MSE with the sparse noise data than the standard square loss formulation. The observed convergence rate corroborates with our analysis in Theorem 1.

**Conclusions.** This paper presents a unified analysis for FW/FW'-like algorithms on non-convex optimizations. Open problem includes a more in-depth investigation to the accelerated convergence rate observed in [1]. We also remark that with a slight modification, our analysis can be applied to the away-step algorithm in [3] and its online variant.
References


We supply details to the proof sketch in Section 3 by identifying the non-asymptotic constants. Let us begin from (13).

As \( \gamma_t = t^{-\alpha} \), we can lower bound the left hand side by:

\[
\sum_{t=T/2+1}^{T} t^{-\alpha} \langle \nabla F_t(\theta_t), \theta_t - a_i \rangle \geq \left( \sum_{t=T/2+1}^{T} t^{-\alpha} \right) \cdot \min_{t \in [T/2+1, T]} \langle \nabla F_t(\theta_t), \theta_t - a_i \rangle \\
\geq \frac{T^{1-\alpha}}{1-\alpha} \left( 1 - \left( \frac{2}{3} \right)^{1-\alpha} \right) \cdot \min_{t \in [T/2+1, T]} \langle \nabla F_t(\theta_t), \theta_t - a_i \rangle,
\]

for all \( T \geq 6 \). On the other hand, the right hand side of (13) can be bounded as

\[
\sum_{t=T/2+1}^{T} t^{-\alpha} \left( C_g \cdot t^{-\eta} + (L\rho^2/2) \cdot t^{-\alpha} \right) \leq (C_g + (L\rho^2/2)) \cdot \sum_{t=T/2+1}^{T} t^{-\min(2\alpha, 2+\alpha)},
\]

and using (14),

\[
\sum_{t=T/2+1}^{T} (F_t(\theta_t) - F_t(\theta_{t+1})) \leq F_{T/2+1}(\theta_{T/2+1}) - F_T(\theta_{T+1}) + \sum_{t=T/2+1}^{T-1} C_b \cdot t^{-\beta}.
\]

For any \( 1 > \delta > 0 \), the latter summations in the above can be bounded as

\[
\sum_{t=T/2+1}^{T} t^{-\delta} \leq \int_{T/2}^{T} t^{-\delta} dt \leq \frac{T^{1-\delta}}{1-\delta} \left( 1 - \left( \frac{1}{2} \right)^{1-\delta} \right).
\]

For \( \delta \geq 1 \), the summation is bounded by \( \sum_{t=T/2+1}^{T} t^{-\delta} \leq \int_{T/2}^{T} t^{-\delta} dt \leq \log 2 \). Therefore, we can write the upper bound as

\[
\sum_{t=T/2+1}^{T} t^{-\delta} \leq T^{\max(0,1-\delta)} C(\delta), \text{ where } C(\delta) := \begin{cases} (1 - (1/2)^{1-\delta})/(1-\delta) & \text{if } 0 < \delta < 1, \\ \log 2 & \text{if } \delta \geq 1. \end{cases}
\]

Notice that the bound is decreasing with \( \delta \). Consequently, let \( F_{T/2+1}(\theta_{T/2+1}) - F_T(\theta_{T+1}) \leq 2B \) since the objective values are bounded and we have

\[
\min_{t \in [T/2+1, T]} \max_{\theta \in C} \langle \nabla F_t(\theta_t), \theta_t - \theta \rangle \leq \left( 1 - \left( \frac{2}{3} \right)^{1-\alpha} \right)^{-1} \cdot \left( 2B + \left( C_g + C_b + \frac{L\rho^2}{2} \right) \cdot C(\min(2\alpha, \eta + \alpha, \beta)) \right) \cdot T^{-\min(1-\alpha, \eta, \beta - \alpha)}.
\]

We now prove the third statement in the Theorem. Define the following set of stationary points to (1):

\[
C^* := \{ \theta \in C : \max_{\theta \in C} \langle \nabla F(\theta), \theta - \theta \rangle = 0 \}.
\]

**Theorem 2.** [14] *Theorem 1* For an arbitrary convergent subsequence \( \{\theta_{s_t}\}_{t \geq 1} \) in \( C \) with the limit point \( \theta \). If the following conditions hold:

A1. it holds that \( \lim_{t \to \infty} \|\theta_{t+1} - \theta_t\| = 0 \),

A2. if \( \theta \notin C^* \), there exists \( \epsilon_0 > 0 \) such that for all \( 0 < \epsilon \leq \epsilon_0 \), the integer quantity \( \tau_t \) is finite with

\[
\tau_t := \min_{s > t} \text{ s.t. } \|\theta_s - \theta_{s_t}\| > \epsilon,
\]

A3. taking the same \( \tau_t \) defined above, there exists a continuous function \( W(\theta) \) that takes a finite number of values in \( C^* \) such that

\[
\limsup_{t \to \infty} W(\theta_{\tau_t}) < \lim_{t \to \infty} W(\theta_{s_t}),
\]

then the sequence \( \{W(\theta_t)\}_{t \geq 1} \) converges and the limit points of the sequence \( \{\theta_t\}_{t \geq 1} \) belong to the set \( C^* \).
Our plan is to apply the theorem above to prove \( \mathcal{C} \). We first observe that as \( C \) is closed and bounded, by Bolzano-Weierstrass there exists a convergent subsequence \( \{ \theta_{s_k} \}_{k=1} \) of the sequence of iterates generated by the generalized FW algorithm \( \mathcal{C} \). Moreover, condition A1 can be easily verified since

\[
\| \theta_{t+1} - \theta_t \| \leq \gamma_t \| \hat{a}_t - \theta_t \| \leq \gamma_t \bar{t} ,
\]

and \( \gamma_t \to 0 \) as \( t \to \infty \).

Now, let \( \theta \) be the limit of the subsequence \( \{ \theta_{s_k} \}_{k=1} \) and \( \theta \notin C^* \). We shall verify condition A2 in Theorem 2 by contradiction. In particular, we assume that the following holds for all \( 0 < \epsilon \leq \epsilon_0 \):

\[
\| \theta_s - \theta_{s_k} \| \leq \epsilon, \forall \ s > s_k .
\]

For some sufficiently large \( k \) and \( s > s_k \), since \( \{ \theta_{s_k} \}_{k=1} \) converges to \( \theta \) as \( t \to \infty \), we have \( \theta_s \in B_{2\epsilon}(\theta) \), i.e., the ball of radius \( 2\epsilon \) centered at \( \theta \). Furthermore, as \( \theta \notin C^* \) and \( C^* \) is closed, the following holds,

\[
\langle \nabla F(\theta_s), \theta - \theta_s \rangle \leq -\delta < 0, \forall \ \theta \in C, \forall \ s > s_k ,
\]

for some \( \delta > 0 \). In particular, we have \( \langle \nabla F(\theta_s), \alpha_s - \theta_s \rangle \leq -\delta \). From (12), H1 and H2 it holds true for all \( t \geq 1 \) that:

\[
F(\theta_{t+1}) - F(\theta_t) \leq \gamma_t \cdot \langle \nabla F(\theta_t), \alpha_t - \theta_t \rangle + \gamma_t t^{-\eta} + \frac{1}{2} \gamma_t^2 L \bar{\rho}^2 .
\]

To arrive at a contradiction, we let \( s > s_k \) and sum up the both side of the above from \( t = s_k \) to \( t = s \). Consider the following chain of inequality:

\[
F(\theta_s) - F(\theta_{s_k}) \leq \gamma_{s_k} \cdot \langle \nabla F(\theta_{s_k}), \alpha_{s_k} - \theta_{s_k} \rangle + C_g \cdot t^{-\eta} + (L \bar{\rho}^2 / 2) \cdot \ell^{-\alpha}
\]

\[
\leq -\delta \sum_{t=s_k}^{s} \gamma_t + \sum_{t=s_k}^{s} \ell^{-\alpha} (C_g \cdot \ell^{-\eta} + (L \bar{\rho}^2 / 2) \cdot \ell^{-\alpha}) ,
\]

where the second inequality is due to (27). Letting \( s \to \infty \) and observe that \( \sum_{t=s_k}^{s} \gamma_t \to +\infty \) implies

\[
\lim_{s \to \infty} F(\theta_s) - F(\theta_{s_k}) < -\infty ,
\]

since \( \lim_{s \to \infty} \sum_{t=s_k}^{s} \ell^{-\alpha} (C_g \cdot \ell^{-\eta} + (L \bar{\rho}^2 / 2) \cdot \ell^{-\alpha}) < \infty \), which is due to \( \eta + \alpha > 1 \) and \( 2\alpha > 1 \). This leads to a contradiction since \( \hat{F}(\theta) \) is bounded over \( C \). We conclude that condition A2 holds for the FW algorithm.

The remaining task is to verify condition A3. We shall take \( W(\cdot) = F(\cdot) \). By the definition of \( \tau_t \), we have \( \theta_{s_k} \in B_{\epsilon}(\theta_{s_k}) \) for all \( s_k \leq s \leq \tau_t - 1 \). Again for some sufficiently large \( t \), we have \( \theta_s \in B_t(\theta_{s_k}) \subseteq B_{2\epsilon}(\theta) \) and the inequality (29) holds for \( s = \tau_t - 1 \). This gives:

\[
F(\theta_{\tau_t}) - F(\theta_{s_k}) \leq \sum_{t=s_k}^{\tau_t-1} \gamma_t \cdot (-\delta + C_g \cdot \ell^{-\eta} + (L \bar{\rho}^2 / 2) \cdot \ell^{-\alpha}) .
\]

On the other hand, we have \( \theta_{\tau_t} \notin B_t(\theta_{s_k}) \) and thus

\[
\epsilon < \| \theta_{\tau_t} - \theta_{s_k} \| \leq \sum_{t=s_k}^{\tau_t-1} \gamma_t \| \hat{a}_t - \theta_t \| \leq \bar{\rho} \sum_{t=s_k}^{\tau_t-1} \gamma_t .
\]

The above implies that \( \sum_{t=s_k}^{\tau_t-1} \gamma_t > \epsilon / \bar{\rho} > 0 \). Considering (31) again, observe that the latter two terms decay to zero, for some sufficiently large \( t \), we have \( -\delta + O(\ell^{-\min(\eta,\alpha)}) \leq -\delta' < 0 \) if \( \ell \geq s_k \). Therefore, (31) leads to

\[
F(\theta_{\tau_t}) - F(\theta_{s_k}) \leq -\delta' \sum_{t=s_k}^{\tau_t-1} \gamma_t < -\frac{\delta' \epsilon}{\bar{\rho}} < 0 .
\]

Taking the limit \( t \to \infty \) on both sides lead to (24) and completes the proof.