Structure of This Document

This supplementary document is the appendix section of the AISTATS'19 paper entitled "Distributed Inexact Newton-type Pursuit for Non-Convex Sparse Learning". It is organized as follows: In Section A we present several technical lemmas that will be used for proving the main results. In Section B we give the proofs of main results appeared in Section 3 of the paper. In Section C, we proivde the proof of Theorem 3 in Section 4 of the paper.

A Technical Lemmas

The following lemma shows that the estimation error of the truncated average of estimators is well upper bounded by the average error of those estimators.

Lemma 2. *Let $\bar{w}$ be $\bar{k}$-sparse vector. For a set of k-sparse vectors $\{w_j\}_{j=1}^m$ with $k \geq \bar{k}$, it holds that*

$$\left\| \mathrm{H}_k\left(\frac{1}{m}\sum_{j=1}^m w_j\right) - \bar{w}\right\| \leq \frac{1.62}{m}\sum_{j=1}^m \|w_j - \bar{w}\|.$$

Moreover, if $k > \bar{k}$, then

$$\left\| \mathrm{H}_k\left(\frac{1}{m}\sum_{j=1}^m w_j\right) - \bar{w}\right\| \leq \frac{1}{m}\sqrt{1 + 2\sqrt{\frac{\bar{k}}{k-\bar{k}}}}\sum_{j=1}^m \|w_j - \bar{w}\|.$$

Proof. The first claim follows readily from (Shen & Li, 2017, Theorem 1) and triangle inequality. The second claim is a direct consequence of the result in (Li et al., 2016, Lemma 3.3). □

The following lemma is key to our analysis.

Lemma 3. *Let $\bar{w}$ be a $\bar{k}$-sparse target vector with $\bar{k} \leq k$. Assume that each component $F_j(w)$ is μ_{3k}-strongly-convex and $\eta F(w) - F_j(w) - \frac{\gamma}{2}\|w\|^2$ has α_{3k}-RLG. Then*

$$\|w^{(t)} - \bar{w}\| \leq \frac{3.24\alpha_{3k}}{\gamma + \mu_{3k}}\|w^{(t-1)} - \bar{w}\| + \frac{5.62\eta\sqrt{k}}{\gamma + \mu_{3k}}\|\nabla F(\bar{w})\|_\infty + 2.3\sqrt{\frac{\epsilon}{\gamma+\mu_{3k}}}.$$

Moreover, assume that each $F_j(w)$ has β_{3k}-RLH. Let $\bar{H}_j = \nabla^2 F_j(\bar{w})$ and $\bar{H} = \frac{1}{m}\sum_{j=1}^m \bar{H}_j$. Then

$$\begin{aligned}\|w^{(t)} - \bar{w}\| \leq & \frac{3.24(\gamma + \max_j \|\bar{H}_j - \eta\bar{H}\|)}{\gamma + \mu_{3k}}\|w^{(t-1)} - \bar{w}\| + \frac{1.62(1+\eta)\beta_{3k}}{\gamma + \mu_{3k}}\|w^{(t-1)} - \bar{w}\|^2 \\ & + \frac{5.62\eta\sqrt{k}}{\gamma + \mu_{3k}}\|\nabla F(\bar{w})\|_\infty + 2.3\sqrt{\frac{\epsilon}{\gamma+\mu_{3k}}}.\end{aligned}$$

Proof. For any $j \in [m]$, since $F_j(w)$ is μ_{3k}-strongly-convex, we have that $P_j(w; w^{(t-1)} \mid \eta, \gamma)$ is $(\gamma+\mu_{3k})$-strongly-convex. Let $S_j^{(t)} = \mathrm{supp}(w_j^{(t)})$, $S^{(t-1)} = \mathrm{supp}(w^{(t-1)})$ and $\bar{S} = \mathrm{supp}(\bar{w})$. Consider $S = S_j^{(t)} \cup S^{(t-1)} \cup \bar{S}$. Then

$$\begin{aligned}& P_j(w_j^{(t)}; w^{(t-1)} \mid \eta, \gamma) \\ \geq & P_j(\bar{w}; w^{(t-1)} \mid \eta, \gamma) + \langle \nabla P_j(\bar{w}; w^{(t-1)} \mid \eta, \gamma), w_j^{(t)} - \bar{w}\rangle + \frac{\gamma + \mu_{3k}}{2}\|w_j^{(t)} - \bar{w}\|^2 \\ = & P_j(\bar{w}; w^{(t-1)} \mid \eta, \gamma) + \langle \nabla_S P_j(\bar{w}; w^{(t-1)} \mid \eta, \gamma), w_j^{(t)} - \bar{w}\rangle + \frac{\gamma + \mu_{3k}}{2}\|w_j^{(t)} - \bar{w}\|^2 \\ \overset{\xi_1}{\geq} & P_j(w^{(t)}; w^{(t-1)} \mid \eta, \gamma) - \epsilon - \|\nabla_S P_j(\bar{w}; w^{(t-1)} \mid \eta, \gamma)\|\|w_j^{(t)} - \bar{w}\| + \frac{\gamma + \mu_{3k}}{2}\|w_j^{(t)} - \bar{w}\|^2,\end{aligned}$$

where "ξ_1" follows from the definition of $w^{(t)}$ as an ϵ-approximate k-sparse minimizer of $P_j(w; w^{(t-1)} \mid \eta, \gamma)$. By rearranging both sides of the above inequality with proper elementary calculation we get

$$
\begin{aligned}
&\|w_j^{(t)} - \bar{w}\| \\
\leq & \frac{2}{\gamma + \mu_{3k}} \|\nabla_S P_j(\bar{w}; w^{(t-1)} \mid \eta, \gamma)\| + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}} \\
= & \frac{2}{\gamma + \mu_{3k}} \|\eta \nabla_S F(w^{(t-1)}) - \nabla_S F_j(w^{(t-1)}) + \gamma(\bar{w} - w^{(t-1)}) + \nabla_S F_j(\bar{w})\| + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}} \\
= & \frac{2}{\gamma + \mu_{3k}} \|\eta \nabla_S F(w^{(t-1)}) - \eta \nabla_S F(\bar{w}) - (\nabla_S F_j(w^{(t-1)}) - \nabla_S F_j(\bar{w})) + \gamma(\bar{w} - w^{(t-1)}) + \eta \nabla_S F(\bar{w})\| + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}} \\
\overset{\xi_1}{\leq} & \frac{2}{\gamma + \mu_{3k}} \left\| \left(\eta \nabla_S F(w^{(t-1)}) - \nabla_S F_j(w^{(t-1)}) - \gamma w^{(t-1)} \right) - (\eta \nabla_S F(\bar{w}) - \nabla_S F_j(\bar{w}) - \gamma \bar{w}) \right\| + \frac{2\eta}{\gamma + \mu_{3k}} \|\nabla_S F(\bar{w})\| \\
& + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}} \\
\leq & \frac{2\alpha_{3k}}{\gamma + \mu_{3k}} \|w^{(t-1)} - \bar{w}\| + \frac{2\eta\sqrt{3k}}{\gamma + \mu_{3k}} \|\nabla F(\bar{w})\|_\infty + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}},
\end{aligned}
$$

where ζ_1 is according to the assumption that $\eta F(w) - F_j(w) - \frac{\gamma}{2}\|w\|^2$ has α_{3k}-RLG. Since $w^{(t)} = \mathrm{H}_k\left(\frac{1}{m}\sum_{j=1}^m w_j^{(t)}\right)$, by applying the first claim in Lemma 2 we obtain

$$
\begin{aligned}
\|w^{(t)} - \bar{w}\| = & 1.62 \left\| \frac{1}{m} \sum_{j=1}^m w_j^{(t)} - \bar{w} \right\| \\
& \leq \frac{1.62}{m} \sum_{j=1}^m \left\| w_j^{(t)} - \bar{w} \right\| \leq \frac{3.24\alpha_{3k}}{\gamma + \mu_{3k}} \|w^{(t-1)} - \bar{w}\| + \frac{5.62\eta\sqrt{k}}{\gamma + \mu_{3k}} \|\nabla F(\bar{w})\|_\infty + 2.3\sqrt{\frac{\epsilon}{\gamma + \mu_{3k}}}.
\end{aligned}
$$

This shows the validity of the first part.

Next we prove the second part. Similar to the above argument we can derive the following:

$$
\begin{aligned}
&\|w_j^{(t)} - \bar{w}\| \\
\leq & \frac{2}{\gamma + \mu_{3k}} \|\nabla_S P_j(\bar{w}; w^{(t-1)} \mid \eta, \gamma)\| + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}} \\
\leq & \frac{2}{\gamma + \mu_{3k}} \left\| \gamma(w^{(t-1)} - \bar{w}) + \eta \nabla_S F(w^{(t-1)}) - \eta \nabla_S F(\bar{w}) - (\nabla_S F_j(w^{(t-1)}) - \nabla_S F_j(\bar{w})) \right\| \\
& + \frac{2\eta}{\gamma + \mu_{3k}} \|\nabla_S F(\bar{w})\| + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}} \\
\leq & \frac{2}{\gamma + \mu_{3k}} \left\| \gamma(w^{(t-1)} - \bar{w}) + \eta \nabla_{SS}^2 F(\bar{w})(w^{(t-1)} - \bar{w}) - \nabla_{SS}^2 F_j(\bar{w})(w^{(t-1)} - \bar{w}) \right\| \\
& + \frac{2\eta}{\gamma + \mu_{3k}} \|\nabla_S F(\bar{w})\| + \frac{2\eta}{\gamma + \mu_{3k}} \left\| \nabla_S F(w^{(t-1)}) - \nabla_S F(\bar{w}) - \nabla_{SS}^2 F(\bar{w})(w^{(t-1)} - \bar{w}) \right\| \\
& + \frac{2}{\gamma + \mu_{3k}} \left\| \nabla_S F_j(w^{(t-1)}) - \nabla_S F_j(\bar{w}) - \nabla_{SS}^2 F_j(\bar{w})(w^{(t-1)} - \bar{w}) \right\| + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}} \\
\leq & \frac{2}{\gamma + \mu_{3k}} \left(\gamma + \|\eta \nabla_{SS}^2 F(\bar{w}) - \nabla_{SS}^2 F_j(\bar{w})\| \right) \|w^{(t-1)} - \bar{w}\| + \frac{2\eta}{\gamma + \mu_{3k}} \|\nabla_S F(\bar{w})\| \\
& + \frac{(1+\eta)\beta_{3k}}{\gamma + \mu_{3k}} \|w^{(t-1)} - \bar{w}\|^2 + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}} \\
\leq & \frac{2(\gamma + \max_{j'} \|\bar{H}_{j'} - \eta \bar{H}\|)}{\gamma + \mu_{3k}} \|w^{(t-1)} - \bar{w}\| + \frac{(1+\eta)\beta_{3k}}{\gamma + \mu_{3k}} \|w^{(t-1)} - \bar{w}\|^2 + \frac{2\eta\sqrt{3k}}{\gamma + \mu_{3k}} \|\nabla F(\bar{w})\|_\infty + \sqrt{\frac{2\epsilon}{\gamma + \mu_{3k}}}.
\end{aligned}
$$

Again, from the definition of $w^{(t)}$ and by applying the first claim in Lemma 2 we have

$$\begin{aligned}&\left\|w^{(t)} - \bar{w}\right\| \\ \le & \frac{3.24(\gamma + \max_j \|\bar{H}_j - \eta\bar{H}\|)}{\gamma + \mu_{3k}}\|w^{(t-1)} - \bar{w}\| + \frac{1.62(1+\eta)\beta_{3k}}{\gamma + \mu_{3k}}\|w^{(t-1)} - \bar{w}\|^2 + \frac{5.62\eta\sqrt{k}}{\gamma + \mu_{3k}}\|\nabla F(\bar{w})\|_\infty \\ & + 2.3\sqrt{\frac{\epsilon}{\gamma + \mu_{3k}}}.\end{aligned}$$

This proves desired bound in the second part. □

B Proofs for the Main Results in Section 3

B.1 Proof of Proposition 1

Proof. Consider an index set S with cardinality $|S| \le s$ and all w, w' with $\text{supp}(w) \cup \text{supp}(w') \subseteq S$. Since $\sigma(z)$ is Lipschitz continuous with constant 1, we have that

$$|\sigma(2y_i w^\top x_i) - \sigma(2y_i w'^\top x_i)| \le |2(w - w')^\top y_i x_i| \le 2\|[x_i]_S\|\|w - w'\| \le 2r_s\|w - w'\|.$$

Using this above inequality and the fact that $\sigma(z) \le 1$ we obtain

$$\begin{aligned}&|\sigma(2v_i w^\top u_i)(1 - \sigma(2v_i w^\top u_i)) - \sigma(2v_i w'^\top u_i)(1 - \sigma(2v_i w'^\top u_i))| \\ \le & |\sigma(2v_i w^\top u_i) - \sigma(2v_i w'^\top u_i)|(1 + \sigma(2v_i w^\top u_i) + \sigma(2v_i w'^\top u_i)) \\ \le & 3|\sigma(2v_i w^\top u_i) - \sigma(2v_i w'^\top u_i)| \le 6r_s\|w - w'\|.\end{aligned}$$

This yields $\|\Lambda(w) - \Lambda(w')\| \le 24r_s\|w - w'\|$. Therefore,

$$\left\|\nabla^2_{SS} f(w) - \nabla^2_{SS} f(w')\right\| \le \frac{1}{n}\|X^n_S\|^2\|\Lambda(w) - \Lambda(w')\| \overset{\zeta_1}{\le} 24r_s\left\|\frac{1}{n}X^n_S(X^n_S)^\top\right\|\|w - w'\| \le 24r_s\rho^{\max}_s(\Sigma_n)\|w - w'\|,$$

where the "ζ_1" follows from the standard matrix norm equality $\|A\|^2 = \|AA^\top\|$. This proves the desired result. □

B.2 Proof of Theorem 1 and Corollary 1

Proof of Theorem 1. Since the local objective functions F_j are quadratic, we can simply set $\beta_s = 0$ for all cardinality s. By assumption $F_j(w)$ is μ_{3k}-strongly-convex. Then by invoking the second part of Lemma 3 with $\beta_{3k} = 0$, $\gamma = 0$ and $\epsilon \le \frac{k\eta^2\|\nabla F(\bar{w})\|^2_\infty}{5.29\mu_{3k}}$ we get

$$\|w^{(t)} - \bar{w}\| \le \frac{3.24\max_j \|\bar{H}_j - \eta\bar{H}\|}{\mu_{3k}}\|w^{(t-1)} - \bar{w}\| + \frac{6.62\eta\sqrt{k}}{\mu_{3k}}\|\nabla F(\bar{w})\|_\infty.$$

It can be readily verified that the factor $\frac{3.24\max_j \|\bar{H}_j - \eta\bar{H}\|}{\mu_{3k}} \le \theta < 1$. Based on the above recursion inequality we can show

$$\|w^{(t)} - \bar{w}\| \le \theta^t\|w^{(0)} - \bar{w}\| + \frac{6.62\eta\sqrt{k}\|\nabla F(\bar{w})\|_\infty}{(1-\theta)\mu_{3k}}.$$

Based on the inequality $1 - x \le \exp(-x)$ we need

$$t \ge \frac{1}{1-\theta}\log\frac{(1-\theta)\mu_{3k}\|w^{(0)} - \bar{w}\|}{\eta\sqrt{k}\|\nabla F(\bar{w})\|_\infty}$$

rounds of iteration/communication to achieve the precision of $\|w^{(t)} - \bar{w}\| \le \frac{7.62\eta\sqrt{k}\|\nabla F(\bar{w})\|_\infty}{(1-\theta)\mu_{3k}}$. This proves the desired complexity bound. □

Based on the results in Theorem 1 and Lemma 1 we can straightforwardly prove Corollary 1.

Proof of Corollary 1. From the definition of θ and Lemma 1 we get $\max_j \|H_j - H\| \le \frac{\theta\mu_{3k}}{3.24}$ holds with probability at least $1-\delta$. Since $n > \frac{336L^2\log(mp/\delta)}{\mu_{3k}^2}$, we have $\theta \in (0,1)$. The desired bound is then directly implied by Theorem 1. □

Implications for distributed sparse linear regression. Given a $\bar{k}$-sparse parameter vector $\bar{w}$, assume the samples are generated according to the linear model $y = \bar{w}^\top x + \varepsilon$ where ε is a zero-mean Gaussian random noise variable with parameter σ. Assume the data samples $\{D_j = \{x_{ji}, y_{ji}\}_{i=1}^n\}_{j=1}^m$ are distributed over m machines and let $F_j(w) = \frac{1}{2n}\sum_{i=1}^n \|y_{ji} - w^\top x_{ji}\|^2$, $j \in [m]$ be the least square loss over D_j and $F(w) = \frac{1}{m}\sum_{j=1}^m F_j(w)$ be the average of local loss. This example belongs to the quadratic case for which the performance of DINPS is analyzed in Section 3.2. Suppose x_{ji} are drawn from Gaussian distribution with covariance Σ. Then it holds with high probability that $F_j(w)$ has restricted strong-convexity constant $\mu_{3k} \ge \lambda_{\min}(\Sigma) - \mathcal{O}(k\log p/n)$ and smoothness constant $L \le \max_{j,i}\|x_{ji}\|$; and $\|\nabla F(\bar{w})\|_\infty = \mathcal{O}\left(\sigma\sqrt{\log p/(mn)}\right)$ and $\|\nabla F_j(\bar{w})\|_\infty = \mathcal{O}\left(\sigma\sqrt{\log p/n}\right)$. Consider the local initialization strategy of $w^{(0)} \approx \arg\min_{\|w\|_0\le k} F_1(w)$. Then according to the bound in (6), if the sample size $n = \mathcal{O}\left(\frac{L^2\log(mp)}{\mu_{3k}^2}\right)$ is sufficiently large, DINPS needs $\mathcal{O}(\log m)$ rounds of iteration/communication to reach the statistical error level $\mathcal{O}\left(\sigma\sqrt{k\log p/(mn)}\right)$.

B.3 Proof of Theorem 2 and Corollary 2

Proof of Theorem 2. We first claim that $\|w^{(t)} - \bar{w}\| \le \frac{\mu_{3k}\theta}{3.24(1+\eta)\beta_{3k}}$ holds for all $t \ge 0$. This can be shown by induction. Based on the theorem assumptions the claim holds for $t=0$. Now suppose that $\|w^{(t-1)} - \bar{w}\| \le \frac{\mu_{3k}\theta}{3.24(1+\eta)\beta_{3k}}$ for some $t \ge 1$. Since $\gamma = 0$, according to Lemma 3 we have

$$
\begin{aligned}
&\|w^{(t)} - \bar{w}\| \\
\le& \frac{3.24\max_j\|\bar{H}_j - \eta\bar{H}\|}{\mu_{3k}}\|w^{(t-1)} - \bar{w}\| + \frac{1.62(1+\eta)\beta_{3k}}{\mu_{3k}}\|w^{(t-1)} - \bar{w}\|^2 + \frac{5.62\eta\sqrt{k}}{\mu_{3k}}\|\nabla F(\bar{w})\|_\infty + 2.3\sqrt{\frac{\epsilon}{\mu_{3k}}} \\
\overset{\zeta_1}{\le}& \frac{\theta}{2}\|w^{(t-1)} - \bar{w}\| + \frac{1.62(1+\eta)\beta_{3k}}{\mu_{3k}}\|w^{(t-1)} - \bar{w}\|^2 + \frac{6.62\eta\sqrt{k}}{\mu_{3k}}\|\nabla F(\bar{w})\|_\infty \\
\le& \theta\|w^{(t-1)} - \bar{w}\| + \frac{6.62\eta\sqrt{k}}{\mu_{3k}}\|\nabla F(\bar{w})\|_\infty \\
\overset{\zeta_2}{\le}& \frac{\mu_{3k}\theta^2}{3.24(1+\eta)\beta_{3k}} + \frac{\mu_{3k}\theta(1-\theta)}{3.24(1+\eta)\beta_{3k}} = \frac{\mu_{3k}\theta}{3.24(1+\eta)\beta_{3k}},
\end{aligned}
$$

where "ζ_1" follows from the assumptions on $\max_j\|\bar{H}_j - \eta\bar{H}\|$ and ϵ, and "ζ_2" follows from the condition of $\|\nabla F(\bar{w})\|_\infty \le \frac{(1-\theta)\theta\mu_{3k}^2}{21.45\eta(1+\eta)\beta_{3k}\sqrt{k}}$. Thus by induction $\|w^{(t)} - \bar{w}\| \le \frac{\mu_{3k}\theta}{3.24(1+\eta)\beta_{3k}}$ holds for all $t \ge 1$. Then it follows from the inequality below "ζ_1" we get that for all $t \ge 0$,

$$\|w^{(t)} - \bar{w}\| \le \theta\|w^{(t-1)} - \bar{w}\| + \frac{6.62\eta\sqrt{k}}{\mu_{3k}}\|\nabla F(\bar{w})\|_\infty.$$

By recursively applying the above inequality we get

$$\|w^{(t)} - \bar{w}\| \le \theta^t\|w^{(0)} - \bar{w}\| + \frac{6.62\eta\sqrt{k}}{(1-\theta)\mu_{3k}}\|\nabla F(\bar{w})\|_\infty.$$

Based on the inequality $1 - x \le \exp(-x)$ we need

$$t \ge \frac{1}{1-\theta}\log\left(\frac{(1-\theta)\mu_{3k}\|w^{(0)} - \bar{w}\|}{\eta\sqrt{k}\|\nabla F(\bar{w})\|_\infty}\right)$$

rounds of iteration/communication to achieve $\|w^{(t)} - \bar{w}\| \le \frac{7.62\eta\sqrt{k}\|\nabla F(\bar{w})\|_\infty}{(1-\theta)\mu_{3k}}$. This proves the desired complexity bound. □

Corollary 2 can be readily proved by applying Lemma 1 to Theorem 2.

Proof of Corollary 2. From the definition of θ and Lemma 1 we get that $\max_j \|H_j - H\| \le \frac{\theta \mu_{3k}}{6.48}$ holds with probability at least $1-\delta$. Since $n > \frac{1344 L^2 \log(mp/\delta)}{\mu_{3k}^2}$, we have $\theta \in (0,1)$. By invoking Theorem 2 we get the desired result. $\square$

C Proof of Theorem 3 in Section 4

Proof. Recall that we update $w^{(t)} = w_1^{(t)}$ in this non-convex setting. Then the assumption $\|\nabla P_1(w_1^{(t)}; w^{(t-1)} \mid \eta, \gamma)\| \le \epsilon$ implies

$$\|\nabla F_1(w^{(t)}) + \eta \nabla F(w^{(t-1)}) - \nabla F_1(w^{(t-1)}) + \gamma(w^{(t)} - w^{(t-1)})\| \le \epsilon. \tag{C.1}$$

Since $F(w)$ is L_{2k}-smooth,

$$\begin{aligned}
&F(w^{(t)}) \\
\le& F(w^{(t-1)}) + \langle \nabla F(w^{(t-1)}), w^{(t)} - w^{(t-1)} \rangle + \frac{L_{2k}}{2}\|w^{(t)} - w^{(t-1)}\|^2 \\
=& F(w^{(t-1)}) - \frac{1}{\eta}\langle \nabla F_1(w^{(t)}) - \nabla F_1(w^{(t-1)}) + \gamma(w^{(t)} - w^{(t-1)}), w^{(t)} - w^{(t-1)} \rangle + \frac{L_{2k}}{2}\|w^{(t)} - w^{(t-1)}\|^2 \\
&+ \frac{1}{\eta}\langle \nabla F_1(w^{(t)}) + \eta \nabla F(w^{(t-1)}) - \nabla F_1(w^{(t-1)}) + \gamma(w^{(t)} - w^{(t-1)}), w^{(t)} - w^{(t-1)} \rangle \\
\le& F(w^{(t-1)}) - \frac{2\gamma - (\eta+1)L_{2k}}{2\eta}\|w^{(t)} - w^{(t-1)}\|^2 + \frac{\epsilon}{\eta}\|w^{(t)} - w^{(t-1)}\|.
\end{aligned}$$

By rearranging the terms on both sides of the above we get

$$\frac{2\gamma - (\eta+1)L_{2k}}{2\eta}\|w^{(t)} - w^{(t-1)}\|^2 - \frac{\epsilon}{\eta}\|w^{(t)} - w^{(t-1)}\| \le F(w^{(t-1)}) - F(w^{(t)}).$$

By adding both sides of the above from index 1 to t we obtain

$$\begin{aligned}
&\min_{\tau=1,\ldots,t} \frac{2\gamma - (\eta+1)L_{2k}}{2\eta}\|w^{(\tau)} - w^{(\tau-1)}\|^2 - \frac{\epsilon}{\eta}\|w^{(\tau)} - w^{(\tau-1)}\| \\
\le& \frac{1}{t}\sum_{\tau=1}^{t} \frac{2\gamma - (\eta+1)L_{2k}}{2\eta}\|w^{(\tau)} - w^{(\tau-1)}\|^2 - \frac{\epsilon}{\eta}\|w^{(\tau)} - w^{(\tau-1)}\| \\
\le& \frac{1}{t}(F(w^{(0)}) - F(w^{(t)})) \le \frac{1}{t}(F(w^{(0)}) - F(w^*)).
\end{aligned}$$

From the above and the basic fact that $ax^2 - bx - c < 0$ implies $x^2 \le \frac{2b^2}{a^2} + \frac{2c}{a}$ for $a, b, c > 0$, we can verify

$$\min_{\tau=1,\ldots,t} \|w^{(\tau)} - w^{(\tau-1)}\|^2 \le \frac{8\epsilon^2}{(\gamma - (\eta+1)L_{2k})^2} + \frac{4\eta(F(w^{(0)}) - F(w^*))}{(\gamma - (\eta+1)L_{2k})t}.$$

From (C.1) and triangle inequality we can derive that

$$\begin{aligned}
\|\nabla F(w^{(t-1)})\|^2 &\le \left(\frac{1}{\eta}\|\nabla F_1(w^{(t)}) - \nabla F_1(w^{(t-1)}) + \gamma(w^{(t)} - w^{(t-1)})\| + \epsilon\right)^2 \\
&\le \frac{2(L_{2k}+\gamma)^2}{\eta^2}\|w^{(t)} - w^{(t-1)}\|^2 + 2\epsilon^2.
\end{aligned}$$

By combining the preceding two inequalities we get

$$\min_{\tau=1,\ldots,t} \|\nabla F(w^{(\tau)})\|^2 \le \left(\frac{16(L_{2k}+\gamma)^2}{\eta^2(\gamma - (\eta+1)L_{2k})^2} + 2\right)\epsilon^2 + \left(\frac{8(L_{2k}+\gamma)^2(F(w^{(0)}) - F(w^*))}{\eta(\gamma - (\eta+1)L_{2k})}\right)\frac{1}{t}.$$

The desired bound then follows from the setting of $\gamma = (\eta+2)L_{2k}$. This completes the proof. $\square$