A Proofs of Deterministic Frank-Wolfe

Lemma A.1. *Consider the proposed zeroth order Frank Wolfe Algorithm. Let Assumptions A1-A5 hold. Then, the sub-optimality* $F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*)$ *satisfies*

$$F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*) \leq (1-\gamma_{t+1})(F(\mathbf{x}_t) - F(\mathbf{x}^*)) + \gamma_{t+1}R\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\| + \frac{LR^2\gamma_{t+1}^2}{2}. \tag{28}$$

Proof. The L-smoothness of the function f yields the following upper bound on $f(\mathbf{x}_{t+1})$:

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^T(\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ &= f(\mathbf{x}_t) + \gamma_{t+1}(\nabla f(\mathbf{x}_t) - \mathbf{d}_t)^T(\mathbf{v}_t - \mathbf{x}_t) + \gamma_{t+1}\mathbf{d}_t^T(\mathbf{v}_t - \mathbf{x}_t) \\ &+ \frac{L\gamma_{t+1}^2}{2}\|\mathbf{v}_t - \mathbf{x}_t\|^2 \end{aligned} \tag{29}$$

Since $\langle \mathbf{x}^*, \mathbf{d}_t \rangle \geq \min_{v \in \mathcal{C}}\{\langle \mathbf{v}, \mathbf{d}_t \rangle\} = \langle \mathbf{v}_t, \mathbf{d}_t \rangle$, we have,

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \gamma_{t+1}(\nabla f(\mathbf{x}_t) - \mathbf{d}_t)^T(\mathbf{v}_t - \mathbf{x}_t) \\ &+ \gamma_{t+1}\mathbf{d}_t^T(\mathbf{x}^* - \mathbf{x}_t) + \frac{L\gamma_{t+1}^2}{2}\|\mathbf{v}_t - \mathbf{x}_t\|^2 \\ &\leq f(\mathbf{x}_t) + \gamma_{t+1}(\nabla f(\mathbf{x}_t) - \mathbf{d}_t)^T(\mathbf{v}_t - \mathbf{x}^*) \\ &+ \gamma_{t+1}\nabla f(\mathbf{x}_t)^T(\mathbf{x}^* - \mathbf{x}_t) + \frac{LR\gamma_{t+1}^2}{2}\|\mathbf{v}_t - \mathbf{x}_t\|^2. \end{aligned} \tag{30}$$

Using Cauchy-Schwarz inequality, we have,

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \gamma_{t+1}\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|\|\mathbf{v}_t - \mathbf{x}^*\| \\ &- \gamma_{t+1}(f(\mathbf{x}_t) - f(\mathbf{x}^*)) + \frac{L\gamma_{t+1}^2}{2}\|\mathbf{v}_t - \mathbf{x}^*\|^2 \\ &\leq f(\mathbf{x}_t) + \gamma_{t+1}R\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\| - \gamma_{t+1}(f(\mathbf{x}_t) - f(\mathbf{x}^*)) \\ &+ \frac{LR^2\gamma_{t+1}^2}{2}, \end{aligned} \tag{31}$$

and subtracting $f(\mathbf{x}^*)$ from both sides of (31), we have,

$$\begin{aligned} f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*) &\leq (1-\gamma_{t+1})(f(\mathbf{x}_t) - f(\mathbf{x}^*)) \\ &+ \gamma_{t+1}R\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\| + \frac{LR^2\gamma_{t+1}^2}{2}. \end{aligned} \tag{32}$$

□

Proof of Theorem 3.1. We have, from Lemma A.1,

$$\begin{aligned} F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*) &\leq (1-\gamma_{t+1})(F(\mathbf{x}_t) - F(\mathbf{x}^*)) \\ &+ \gamma_{t+1}R\|\nabla F(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_t)\| + \frac{LR^2\gamma_{t+1}^2}{2} \\ &\Rightarrow F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*) \leq (1-\gamma_{t+1})(F(\mathbf{x}_t) - F(\mathbf{x}^*)) \\ &+ \frac{c_{t+1}d}{2}\gamma_{t+1}R^2 + \frac{LR^2\gamma_{t+1}^2}{2}. \end{aligned} \tag{33}$$

From, (33), we have,

$$F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*) \leq (1-\gamma_{t+1})(F(\mathbf{x}_t) - F(\mathbf{x}^*)) + LR^2\gamma_{t+1}^2. \tag{34}$$

We use Lemma B.1 to derive the primal gap which then yields,

$$F(\mathbf{x}_t) - F(\mathbf{x}^*) = \frac{Q_{ns}}{t+2}, \tag{35}$$

where $Q_{ns} = \max\{2(F(\mathbf{x}_0) - F(\mathbf{x}^*)), 4LR^2\}$. □

B Proofs of Zeroth Order Stochastic Frank Wolfe: RDSA

Proof of Lemma 3.2 (1). Use the definition $\mathbf{d}_t := (1-\rho_t)\mathbf{d}_{t-1} + \rho_t g(\mathbf{x}_t;\mathbf{y}_t,\mathbf{z}_t)$ to write the difference $\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2$ as

$$\begin{aligned}\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2 &= \|\nabla f(\mathbf{x}_t) - (1-\rho_t)\mathbf{d}_{t-1} \\ &- \rho_t g(\mathbf{x}_t;\mathbf{y}_t,\mathbf{z}_t)\|^2. \end{aligned} \tag{36}$$

Add and subtract the term $(1-\rho_t)\nabla f(\mathbf{x}_{t-1})$ to the right hand side of (36), regroup the terms and expand the squared term to obtain

$$\begin{aligned}&\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2 \\ &= \|\nabla f(\mathbf{x}_t) - (1-\rho_t)\nabla f(\mathbf{x}_{t-1}) + (1-\rho_t)\nabla f(\mathbf{x}_{t-1}) \\ &- (1-\rho_t)\mathbf{d}_{t-1} - \rho_t g(\mathbf{x}_t;\mathbf{y}_t,\mathbf{z}_t)\|^2 \\ &= \rho_t^2\|\nabla f(\mathbf{x}_t) - g(\mathbf{x}_t;\mathbf{y}_t,\mathbf{z}_t)\|^2 + (1-\rho_t)^2\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2 \\ &+ (1-\rho_t)^2\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2 \\ &\quad + 2\rho_t(1-\rho_t)(\nabla f(\mathbf{x}_t) - g(\mathbf{x}_t;\mathbf{y}_t,\mathbf{z}_t))^T(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})) \\ &\quad + 2\rho_t(1-\rho_t)(\nabla f(\mathbf{x}_t) - g(\mathbf{x}_t;\mathbf{y}_t,\mathbf{z}_t))^T(\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}) \\ &\quad + 2(1-\rho_t)^2(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}))^T(\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}). \end{aligned} \tag{37}$$

Compute the expectation $\mathbb{E}\left[(.)\mid\mathcal{F}_t\right]$ for both sides of (37), where $\mathcal{F}_t$ is the σ-algebra given by $\{\{\mathbf{y}_s\}_{s=0}^{t-1}, \{\mathbf{z}_s\}_{s=0}^{t-1}\}$to obtain

$$\begin{aligned}&\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2 \mid \mathcal{F}_t\right] \\ &= \rho_t^2\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - g(\mathbf{x}_t;\mathbf{y}_t,\mathbf{z}_t)\|^2 \mid \mathcal{F}_t\right] \\ &+ (1-\rho_t)^2\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2 \\ &+ (1-\rho_t)^2\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2 \\ &+ 2(1-\rho_t)^2(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}))^T(\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}) \\ &+ 2\rho_t(1-\rho_t)\mathbb{E}\left[(\nabla f(\mathbf{x}_t) - g(\mathbf{x}_t;\mathbf{y}_t,\mathbf{z}_t))^T(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})) \mid \mathcal{F}_t\right] \\ &+ 2\rho_t(1-\rho_t)\mathbb{E}\left[(\nabla f(\mathbf{x}_t) - g(\mathbf{x}_t;\mathbf{y}_t,\mathbf{z}_t))^T(\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}) \mid \mathcal{F}_t\right] \\ &\le \rho_t^2\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - g(\mathbf{x}_t;\mathbf{y}_t,\mathbf{z}_t)\|^2 \mid \mathcal{F}_t\right] \\ &+ (1-\rho_t)^2\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2 \\ &+ (1-\rho_t)^2\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2 \\ &+ (1-\rho_t)^2\beta_t\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2 + \frac{(1-\rho_t)^2}{\beta_t}\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2 \\ &+ 2\rho_t(1-\rho_t)(c_t L\mathbf{v}(\mathbf{x}, c_t))^\top(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})) \\ &+ 2\rho_t(1-\rho_t)(c_t L\mathbf{v}(\mathbf{x}, c_t))^\top(\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}) \\ &\le \rho_t^2\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - g(\mathbf{x}_t;\mathbf{y}_t,\mathbf{z}_t)\|^2 \mid \mathcal{F}_t\right] \\ &+ (1-\rho_t)^2\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2 \\ &+ (1-\rho_t)^2\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2 \\ &+ (1-\rho_t)^2\beta_t\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2 + \frac{(1-\rho_t)^2}{\beta_t}\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2 \\ &+ 2\rho_t(1-\rho_t)c_t^2\left\|L\mathbf{v}(\mathbf{x}, c_t)\right\|^2 + \rho_t(1-\rho_t)\left\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\right\|^2 \\ &+ \rho_t(1-\rho_t)\left\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\right\|^2 \\ &\Rightarrow \mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] \\ &\le \rho_t^2\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \nabla F(\mathbf{x}_t,\mathbf{y}_t) + \nabla F(\mathbf{x}_t,\mathbf{y}_t) - g(\mathbf{x}_t;\mathbf{y}_t,\mathbf{z}_t)\|^2\right] \\ &+ (1-\rho_t)^2\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2\right] \\ &+ (1-\rho_t)^2\|\mathbb{E}\left[\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right] \end{aligned}$$

$$
\begin{aligned}
&+ (1-\rho_t)^2\beta_t\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right] \\
&+ \frac{(1-\rho_t)^2}{\beta_t}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2\right] \\
&+ \frac{\rho_t}{4}(1-\rho_t)c_t^2L^2M(\mu) + \rho_t(1-\rho_t)\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2\right] \\
&+ \rho_t(1-\rho_t)\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right] \\
&\leq 2\rho_t^2\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \nabla F(\mathbf{x}_t, \mathbf{y}_t)\|^2\right] \\
&+ 2\rho_t^2\mathbb{E}\left[\|\nabla F(\mathbf{x}_t, \mathbf{y}_t) - g(\mathbf{x}_t; \mathbf{y}_t, \mathbf{z}_t)\|^2\right] \\
&+ \left(1-\rho_t+\frac{(1-\rho_t)^2}{\beta_t}\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2\right] \\
&+ \left(1-\rho_t+(1-\rho_t)^2\beta_t\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right] \\
&+ \frac{\rho_t}{2}(1-\rho_t)c_t^2L^2M(\mu) \\
&\leq 2\rho_t^2\sigma^2 + 4\rho_t^2\mathbb{E}\left[\|\nabla F(\mathbf{x}_t, \mathbf{y}_t)\|^2\right] + 4\rho_t^2\mathbb{E}\left[\|g(\mathbf{x}_t; \mathbf{y}_t, \mathbf{z}_t)\|^2\right] \\
&+ \left(1-\rho_t+\frac{(1-\rho_t)^2}{\beta_t}\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2\right] \\
&+ \left(1-\rho_t+(1-\rho_t)^2\beta_t\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right] \\
&+ \frac{\rho_t}{2}(1-\rho_t)c_t^2L^2M(\mu) \\
&\leq 2\rho_t^2\sigma^2 + 4\rho_t^2L_1^2 + 8\rho_t^2s(d)L_1^2 + 2\rho_t^2c_t^2L^2M(\mu) \\
&+ \left(1-\rho_t+\frac{(1-\rho_t)^2}{\beta_t}\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2\right] \\
&+ \left(1-\rho_t+(1-\rho_t)^2\beta_t\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right] \\
&+ \frac{\rho_t}{2}c_t^2L^2M(\mu),
\end{aligned} \tag{38}
$$

where we used the gradient approximation bounds as stated in (15) and used Young's inequality to substitute the inner products and in particular substituted $2\langle\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}), \nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\rangle$ by the upper bound $\beta_t\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2 + (1/\beta_t)\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2$ where $\beta_t > 0$ is a free parameter.
By assumption A4, the norm $\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|$ is bounded above by $L\|\mathbf{x}_t - \mathbf{x}_{t-1}\|$. In addition, the condition in Assumption A1 implies that $L\|\mathbf{x}_t - \mathbf{x}_{t-1}\| = L\gamma_t\|\mathbf{v}_t - \mathbf{x}_t\| \leq \gamma_t LR$. Therefore, we can replace $\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|$ by its upper bound $\gamma_t LR$ and since we assume that $\rho_t \leq 1$ we can replace all the terms $(1-\rho_t)^2$. Furthermore, using $\beta_t := \rho_t/2$ we have,

$$
\begin{aligned}
&\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] \\
&\leq 2\rho_t^2\sigma^2 + 4\rho_t^2L_1^2 + 8\rho_t^2s(d)L_1^2 + 2\rho_t^2c_t^2L^2M(\mu) \\
&+ \gamma_t^2(1-\rho_t)\left(1+\frac{2}{\rho_t}\right)L^2R^2 + \frac{\rho_t}{2}c_t^2L^2M(\mu) \\
&+ (1-\rho_t)\left(1+\frac{\rho_t}{2}\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right].
\end{aligned} \tag{39}
$$

Now using the inequalities $(1-\rho_t)(1+(2/\rho_t)) \leq (2/\rho_t)$ and $(1-\rho_t)(1+(\rho_t/2)) \leq (1-\rho/2)$ we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] \leq 2\rho_t^2\sigma^2 + 4\rho_t^2L_1^2 \\
&+ 8\rho_t^2s(d)L_1^2 + 2\rho_t^2c_t^2L^2M(\mu) \\
&+ \frac{2L^2R^2\gamma_t^2}{\rho_t} + \frac{\rho_t}{2}c_t^2L^2M(\mu) \\
&+ \left(1-\frac{\rho_t}{2}\right))\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right].
\end{aligned} \tag{40}
$$

□

Then, we have, from Lemma A.1

$$
\mathbb{E}\left[f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*)\right] \leq (1-\gamma_{t+1})\mathbb{E}\left[(f(\mathbf{x}_t) - f(\mathbf{x}^*))\right]
$$

$$+ \gamma_{t+1} R \mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|\right] + \frac{LR^2\gamma_{t+1}^2}{2}, \tag{41}$$

and then by using Jensen's inequality, we obtain,

$$\begin{aligned} \mathbb{E}\left[f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*)\right] &\leq (1-\gamma_{t+1})\mathbb{E}\left[(f(\mathbf{x}_t) - f(\mathbf{x}^*))\right] \\ &+ \gamma_{t+1} R \sqrt{\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right]} + \frac{LR^2\gamma_{t+1}^2}{2}. \end{aligned} \tag{42}$$

We state a Lemma next which will be crucial for the rest of the paper.

Lemma B.1. *Let $z(k)$ be a non-negative (deterministic) sequence satisfying:*

$$z(k+1) \leq (1 - r_1(k))\, z_1(k) + r_2(k),$$

where $\{r_1(k)\}$ and $\{r_2(k)\}$ are deterministic sequences with

$$\frac{a_1}{(k+1)^{\delta_1}} \leq r_1(k) \leq 1 \text{ and } r_2(k) \leq \frac{a_2}{(k+1)^{2\delta_1}},$$

with $a_1 > 0$, $a_2 > 0$, $1 > \delta_1 > 1/2$ and $k_0 \geq 1$. Then,

$$z(k+1) \leq \exp\left(-\frac{a_1\delta_1(k+k_0)^{1-\delta_1}}{4(1-\delta_1)}\right)\left(z(0) + \frac{a_2}{k_0^{\delta_1}(2\delta_1 - 1)}\right) + \frac{a_2 2^{\delta_1}}{a_1\left(k+k_0\right)^{\delta_1}}.$$

Proof of Lemma B.1. We have,

$$\begin{aligned} z(k+1) &\leq \prod_{l=0}^{k}\left(1 - \frac{a_1}{(l+k_0)^{\delta_1}}\right) z(0) \sum_{l=0}^{\lfloor \frac{k}{2} \rfloor - 1} \prod_{m=l+1}^{k}\left(1 - \frac{a_1}{(m+k_0)^{\delta_1}}\right)\frac{a_2}{(k+k_0)^{2\delta_1}} \\ &+ \sum_{l=\lfloor \frac{k}{2} \rfloor}^{k} \prod_{m=l+1}^{k}\left(1 - \frac{a_1}{(m+k_0)^{\delta_1}}\right)\frac{a_2}{(k+k_0)^{2\delta_1}} \\ &\leq \exp\left(\sum_{l=0}^{k}\left(1 - \frac{a_1}{(l+k_0)^{\delta_1}}\right)\right) z(0) + \prod_{m=l+1}^{k}\left(1 - \frac{a_1}{(m+k_0)^{\delta_1}}\right)\sum_{l=0}^{\lfloor \frac{k}{2} \rfloor - 1}\frac{a_2}{(k+k_0)^{2\delta_1}} \\ &+ \frac{a_2 2^{\delta_1}}{a_1\left(k+k_0\right)^{\delta_1}} \sum_{l=\lfloor \frac{k}{2} \rfloor}^{k} \prod_{m=l+1}^{k}\left(1 - \frac{a_1}{(m+k_0)^{\delta_1}}\right)\frac{a_1}{(k+k_0)^{\delta_1}} \\ &\leq \exp\left(-\sum_{l=0}^{k}\frac{a_1}{(l+k_0)^{\delta_1}}\right) z(0) + \frac{a_2}{a_1 k_0^{\delta_1}} \exp\left(-\sum_{m=\lfloor \frac{k}{2} \rfloor}^{k}\frac{a_1}{(m+k_0)^{\delta_1}}\right)\sum_{l=0}^{\lfloor \frac{k}{2} \rfloor - 1}\frac{a_1}{(k+k_0)^{2\delta_1}} \\ &+ \frac{a_2 2^{\delta_1}}{a_1\left(k+k_0\right)^{\delta_1}} \sum_{l=\lfloor \frac{k}{2} \rfloor}^{k}\left(\prod_{m=l+1}^{k}\left(1 - \frac{a_1}{(m+k_0)^{\delta_1}}\right) - \prod_{m=l}^{k}\left(1 - \frac{a_1}{(m+k_0)^{\delta_1}}\right)\right) \\ &\leq \exp\left(-\sum_{l=0}^{k}\frac{a_1}{(l+k_0)^{\delta_1}}\right) z(0) + \frac{a_2 2^{\delta_1}}{a_1\left(k+k_0\right)^{\delta_1}} + \frac{a_2}{a_1 k_0^{\delta_1}} \exp\left(-\sum_{m=\lfloor \frac{k}{2} \rfloor}^{k}\frac{a_1}{(m+k_0)^{\delta_1}}\right)\sum_{l=0}^{\lfloor \frac{k}{2} \rfloor - 1}\frac{a_1}{(k+k_0)^{2\delta_1}} \\ &\leq \exp\left(-\sum_{l=0}^{k}\frac{a_1}{(l+k_0)^{\delta_1}}\right) z(0) + \frac{a_2 2^{\delta_1}}{a_1\left(k+k_0\right)^{\delta_1}} + \frac{a_2}{k_0^{\delta_1}} \exp\left(-\frac{a_1\delta_1}{4(1-\delta_1)}(k+k_0)^{1-\delta_1}\right)\frac{1}{2\delta_1 - 1}, \end{aligned} \tag{43}$$

where we used the inequality that,

$$\sum_{m=\lfloor \frac{k}{2} \rfloor}^{k}\frac{1}{(m+k_0)^{\delta_1}} \geq \frac{1}{2(1-\delta_1)}(k+k_0)^{1-\delta_1} - \frac{1}{2(1-\delta_1)}\left(\frac{k}{2} + k_0\right)^{1-\delta_1}$$

$$\geq \frac{1}{2^{1+\delta_1}(1-\delta_1)}(k+k_0)^{1-\delta_1}\left(2^{1-\delta_1}-1-\frac{(1-\delta_1)k_0}{k+k_0}\right) \geq \frac{\delta_1}{4(1-\delta_1)}(k+k_0)^{1-\delta_1}$$

Following up with (43), we have,

$$z(k+1) \leq \exp\left(-\sum_{l=0}^{k} -\frac{a_1}{(l+k_0)^{\delta_1}}\right) z(0) + \frac{a_2 2^{\delta_1}}{a_1\,(k+k_0)^{\delta_1}} + \frac{a_2}{k_0^{\delta_1}} \exp\left(-\frac{a_1\delta_1}{4(1-\delta_1)}(k+k_0)^{1-\delta_1}\right) \frac{1}{2\delta_1-1}$$
$$\leq \exp\left(-\frac{a_1\delta_1(k+k_0)^{1-\delta_1}}{4(1-\delta_1)}\right)\left(z(0)+\frac{a_2}{k_0^{\delta_1}(2\delta_1-1)}\right) + \frac{a_2 2^{\delta_1}}{a_1\,(k+k_0)^{\delta_1}}. \tag{44}$$

For $\delta = 2/3$, we have,

$$z(k+1) \leq \exp\left(-\frac{a_1(k+k_0)^{1/3}}{2}\right)\left(z(0)+\frac{3a_2}{k_0^{2/3}}\right) + \frac{a_2 2^{2/3}}{a_1\,(k+k_0)^{2/3}}.$$

□

Proof of Theorem 3.4 (1). Now using the result in Lemma B.1 we can characterize the convergence of the sequence of expected errors $\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right]$ to zero. To be more precise, using the result in Lemma 3.2 and setting $\gamma_t = 2/(t+8)$, $\rho_t = 4/d^{1/3}(t+8)^{2/3}$ and $c_t = 2/\sqrt{M(\mu)}(t+8)^{1/3}$ for any $\epsilon > 0$ to obtain

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right]$$
$$\leq \left(1-\frac{2}{d^{1/3}(t+8)^{2/3}}\right)\mathbb{E}\left[\|\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right]$$
$$+ \frac{32d^{-1/3}\sigma^2 + 64d^{-1/3}L_1^2 + 128d^{2/3}L_1^2 + 2L^2R^2d^{2/3} + 416d^{2/3}L^2}{(t+8)^{4/3}}. \tag{45}$$

According to the result in Lemma B.1, the inequality in (45) implies that

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] \leq \overline{Q} + \frac{Q}{(t+8)^{2/3}} \leq \frac{2Q}{(t+8)^{2/3}}, \tag{46}$$

where $Q = 32d^{-1/3}\sigma^2 + 64d^{-1/3}L_1^2 + 128d^{2/3}L_1^2 + 2L^2R^2d^{2/3} + 416d^{2/3}L^2$, where $\overline{Q}$ is a function of $\mathbb{E}\left[\|\nabla f(\mathbf{x}_0) - \mathbf{d}_0\|^2\right]$ and decays exponentially. Now we proceed by replacing the term $\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right]$ in (42) by its upper bound in (46) and γ_{t+1} by $2/(t+9)$ to write

$$\mathbb{E}\left[f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*)\right] \leq \left(1-\frac{2}{t+9}\right)\mathbb{E}\left[(f(\mathbf{x}_t) - f(\mathbf{x}^*))\right]$$
$$+ \frac{R\sqrt{Q}}{(t+9)^{4/3}} + \frac{2LR^2}{(t+9)^2}. \tag{47}$$

Note that we can write $(t+9)^2 = (t+9)^{4/3}(t+9)^{2/3} \geq (t+9)^{4/3}9^{2/3} \geq 4(t+9)^{4/3}$.Therefore,

$$\mathbb{E}\left[f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*)\right] \leq \left(1-\frac{2}{t+9}\right)\mathbb{E}\left[(f(\mathbf{x}_t) - f(\mathbf{x}^*))\right]$$
$$+ \frac{2R\sqrt{Q} + LD^2/2}{(t+9)^{4/3}}. \tag{48}$$

We use induction to prove for $t \geq 0$,

$$\mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \leq \frac{Q'}{(t+9)^{1/3}},$$

where $Q' = \max\{9^{1/3}(f(\mathbf{x}_0) - f(\mathbf{x}^*)), 2R\sqrt{2Q} + LR^2/2\}$. For $t = 0$, we have that $\mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \leq \frac{Q'}{9^{1/3}}$, which is turn follows from the definition of Q'. Assume for the induction hypothesis holds for $t = k$. Then, for $t = k+1$, we have,

$$\mathbb{E}\left[f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)\right] \leq \left(1-\frac{2}{k+9}\right)\mathbb{E}\left[(f(\mathbf{x}_k) - f(\mathbf{x}^*))\right]$$

$$+ \frac{2R\sqrt{2Q} + LD^2/2}{(k+9)^{4/3}}$$
$$\leq \left(1 - \frac{2}{k+9}\right)\frac{Q'}{(t+9)^{1/3}} + \frac{Q'}{(t+9)^{4/3}} \leq \frac{Q'}{(t+10)^{1/3}}.$$

Thus, for $t \geq 0$ from Lemma B.1 we have that,

$$\mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \leq \frac{Q'}{(t+9)^{1/3}} = O\left(\frac{d^{1/3}}{(t+9)^{1/3}}\right). \tag{49}$$

where $Q' = \max\{2(f(\mathbf{x}_0) - f(\mathbf{x}^*)), 2R\sqrt{2Q} + LR^2/2\}$. □

Proof of Theorem 3.5(1). Then, we have,

$$F(\mathbf{x}_{t+1}) \leq F(\mathbf{x}_t) + \gamma_t\langle \mathbf{g}(\mathbf{x}_t), \mathbf{v}_t - \mathbf{x}_t\rangle$$
$$+ \gamma_t\langle \nabla F(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_t), \mathbf{v}_t - \mathbf{x}_t\rangle + \frac{LR^2\gamma_t^2}{2}$$
$$\Rightarrow F(\mathbf{x}_{t+1}) \leq F(\mathbf{x}_t) + \gamma_t\langle \mathbf{g}(\mathbf{x}_t), \operatorname*{argmin}_{\mathbf{v}\in\mathcal{C}}\langle \mathbf{v}, \nabla F(\mathbf{x}_t)\rangle - \mathbf{x}_t\rangle$$
$$+ \gamma_t\langle \nabla F(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_t), \mathbf{v}_t - \mathbf{x}_t\rangle + \frac{LR^2\gamma_t^2}{2}$$
$$\Rightarrow F(\mathbf{x}_{t+1}) \leq F(\mathbf{x}_t) + \gamma_t\langle \nabla F(\mathbf{x}_t), \operatorname*{argmin}_{\mathbf{v}\in\mathcal{C}}\langle \mathbf{v}, \nabla F(\mathbf{x}_t)\rangle - \mathbf{x}_t\rangle$$
$$+ \gamma_t\langle \nabla F(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_t), \mathbf{v}_t - \operatorname*{argmin}_{\mathbf{v}\in\mathcal{C}}\langle \mathbf{v}, \nabla F(\mathbf{x}_t)\rangle\rangle + \frac{LR^2\gamma_t^2}{2}$$
$$\Rightarrow F(\mathbf{x}_{t+1}) \leq F(\mathbf{x}_t) - \gamma_t\mathcal{G}\left(\mathbf{x}_t\right)$$
$$+ \gamma_t\langle \nabla F(\mathbf{x}_t) - \mathbf{g}(\mathbf{x}_t), \mathbf{v}_t - \operatorname*{argmin}_{\mathbf{v}\in\mathcal{C}}\langle \mathbf{v}, \nabla F(\mathbf{x}_t)\rangle\rangle + \frac{LR^2\gamma_t^2}{2}$$
$$\Rightarrow \gamma_t\mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right] \leq \mathbb{E}\left[F(\mathbf{x}_t) - F(\mathbf{x}_{t+1})\right] + \gamma_t R\frac{\sqrt{2Q}}{(t+8)^{1/3}} + \frac{LR^2\gamma_t^2}{2} +$$
$$\Rightarrow \mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right] \leq \mathbb{E}\left[\frac{t+7}{2}F(\mathbf{x}_t) - \frac{t+8}{2}F(\mathbf{x}_{t+1}) + \frac{1}{2}F(\mathbf{x}_t)\right] + R\frac{\sqrt{2Q}}{(t+8)^{1/3}} + \frac{LR^2\gamma_t}{2}$$
$$\Rightarrow \sum_{t=0}^{T-1}\mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right] \leq \mathbb{E}\left[\frac{7}{2}F(\mathbf{x}_0) - \frac{T+7}{2}F(\mathbf{x}_T) + \sum_{t=0}^{T-1}\left(\frac{1}{2}F(\mathbf{x}_t)\right] + R\frac{\sqrt{2Q}}{(t+8)^{1/3}} + \frac{LR^2\gamma_t}{2}\right)$$
$$\Rightarrow \sum_{t=0}^{T-1}\mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right] \leq \mathbb{E}\left[\frac{7}{2}F(\mathbf{x}_0) - \frac{7}{2}F(\mathbf{x}^*)\right] + \sum_{t=0}^{T-1}\left(\frac{1}{2}\left(F(\mathbf{x}_t) - F(\mathbf{x}^*)\right) + R\frac{\sqrt{2Q}}{(t+8)^{1/3}} + \frac{LR^2\gamma_t}{2}\right)$$
$$\Rightarrow \sum_{t=0}^{T-1}\mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right] \leq \frac{7}{2}F(\mathbf{x}_0) - \frac{7}{2}F(\mathbf{x}^*) + \sum_{t=0}^{T-1}\left(\frac{Q' + R\sqrt{2Q}}{2(t+8)^{1/3}} + \frac{LR^2}{(t+8)}\right)$$
$$\Rightarrow T\mathbb{E}\left[\min_{t=0,\cdots,T-1}\mathcal{G}\left(\mathbf{x}_t\right)\right] \leq \frac{7}{2}F(\mathbf{x}_0) - \frac{7}{2}F(\mathbf{x}^*) + LR^2\ln(T+7) + \frac{Q' + R\sqrt{2Q}}{2}(T+7)^{2/3}$$
$$\Rightarrow \mathbb{E}\left[\min_{t=0,\cdots,T-1}\mathcal{G}\left(\mathbf{x}_t\right)\right] \leq \frac{7(F(\mathbf{x}_0) - F(\mathbf{x}^*))}{2T} + \frac{LR^2\ln(T+7)}{T} + \frac{Q' + R\sqrt{2Q}}{2T}(T+7)^{2/3}. \tag{50}$$

□

C Proofs for Improvised RDSA

Proof of Lemma 3.2(2). Following as in the proof of RDSA, we have,

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2 \mid \mathcal{F}_t\right]$$
$$\leq \rho_t^2\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - g(\mathbf{x}_t; \mathbf{y}_t, \mathbf{z}_t)\|^2 \mid \mathcal{F}_t\right]$$
$$+ (1-\rho_t)^2\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2$$
$$+ (1-\rho_t)^2\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2$$
$$+ (1-\rho_t)^2\beta_t\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2$$

$$
\begin{aligned}
&+ \frac{(1-\rho_t)^2}{\beta_t}\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2 \\
&+ 2\rho_t(1-\rho_t)\frac{c_t^2}{m^2}\left\|L\mathbf{v}\left(\mathbf{x}, c_t\right)\right\|^2 + \rho_t(1-\rho_t)\left\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\right\|^2 \\
&+ \rho_t(1-\rho_t)\left\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\right\|^2 \\
&\Rightarrow \mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] \leq 2\rho_t^2\sigma^2 + 4\rho_t^2\mathbb{E}\left[\|\nabla F(\mathbf{x}_t, \mathbf{y}_t)\|^2\right] \\
&+ 4\rho_t^2\mathbb{E}\left[\|g(\mathbf{x}_t; \mathbf{y}_t, \mathbf{z}_t)\|^2\right] \\
&+ \left(1 - \rho_t + \frac{(1-\rho_t)^2}{\beta_t}\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2\right] \\
&+ \left(1 - \rho_t + (1-\rho_t)^2\beta_t\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right] \\
&+ \frac{\rho_t}{2}(1-\rho_t)c_t^2L^2M(\mu) \\
&\leq 2\rho_t^2\sigma^2 + 4\rho_t^2L_1^2 + 8\rho_t^2\left(1 + \frac{s(d)}{m}\right)L_1^2 + \left(\frac{1+m}{2m}\right)\rho_t^2c_t^2L^2M(\mu) \\
&+ \left(1 - \rho_t + \frac{(1-\rho_t)^2}{\beta_t}\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2\right] \\
&+ \left(1 - \rho_t + (1-\rho_t)^2\beta_t\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right] \\
&+ \frac{\rho_t}{2m^2}c_t^2L^2M(\mu), \qquad (51)
\end{aligned}
$$

where we used the gradient approximation bounds as stated in (15) and used Young's inequality to substitute the inner products and in particular substituted $2\langle\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}), \nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\rangle$ by the upper bound $\beta_t\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2 + (1/\beta_t)\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2$ where $\beta_t > 0$ is a free parameter.
According to Assumption A4, the norm $\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|$ is bounded above by $L\|\mathbf{x}_t - \mathbf{x}_{t-1}\|$. In addition, the condition in Assumption A1 implies that $L\|\mathbf{x}_t - \mathbf{x}_{t-1}\| = L\gamma_t\|\mathbf{v}_t - \mathbf{x}_t\| \leq \gamma_t LR$. Therefore, we can replace $\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|$ by its upper bound $\gamma_t LR$ and since we assume that $\rho_t \leq 1$ we can replace all the terms $(1-\rho_t)^2$. Furthermore, using $\beta_t := \rho_t/2$ we have,

$$
\begin{aligned}
&\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] \\
&\leq 2\rho_t^2\sigma^2 + 4\rho_t^2L_1^2 + 8\rho_t^2\left(1 + \frac{s(d)}{m}\right)L_1^2 + \frac{\rho_t}{2m^2}c_t^2L^2M(\mu) \\
&+ \gamma_t^2(1-\rho_t)\left(1 + \frac{2}{\rho_t}\right)L^2R^2 + \left(\frac{1+m}{2m}\right)\rho_t^2c_t^2L^2M(\mu) \\
&+ (1-\rho_t)\left(1 + \frac{\rho_t}{2}\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right]. \qquad (52)
\end{aligned}
$$

Now using the inequalities $(1-\rho_t)(1 + (2/\rho_t)) \leq (2/\rho_t)$ and $(1-\rho_t)(1 + (\rho_t/2)) \leq (1 - \rho/2)$ we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] \leq 2\rho_t^2\sigma^2 + 4\rho_t^2L_1^2 + 8\rho_t^2\left(1 + \frac{s(d)}{m}\right)L_1^2 \\
&+ \left(\frac{1+m}{2m}\right)\rho_t^2c_t^2L^2M(\mu) + \frac{2L^2R^2\gamma_t^2}{\rho_t} + \frac{\rho_t}{2m^2}c_t^2L^2M(\mu) \\
&+ \left(1 - \frac{\rho_t}{2}\right))\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right]. \qquad (53)
\end{aligned}
$$

□

Proof of Theorem 3.4(2). Now using the result in Lemma B.1 we can characterize the convergence of the sequence of expected errors $\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right]$ to zero. To be more precise, using the result in Lemma 3.2 and setting $\gamma_t = 2/(t+8)$, $\rho_t = 4/\left(1 + \frac{d}{m}\right)^{1/3}(t+8)^{2/3}$ and $c_t = 2\sqrt{m}/\sqrt{M(\mu)}(t+8)^{1/3}$, we have,

$$
\begin{aligned}
&\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] \\
&\leq \left(1 - \frac{2}{\left(1 + \frac{d}{m}\right)^{1/3}(t+8)^{2/3}}\right)\mathbb{E}\left[\|\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right] \\
&+ \frac{32\left(1 + \frac{d}{m}\right)^{-1/3}\sigma^2 + 64L_1^2\left(1 + \frac{d}{m}\right)^{-1/3} + 128\left(1 + \frac{d}{m}\right)^{2/3}L_1^2}{(t+8)^{4/3}}
\end{aligned}
$$

$$+\frac{2L^2R^2\left(1+\frac{d}{m}\right)^{2/3}+416\left(1+\frac{d}{m}\right)^{2/3}L^2}{(t+8)^{4/3}}. \tag{54}$$

According to the result in Lemma B.1, the inequality in (45) implies that

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\mathbf{d}_t\|^2\right]\leq\overline{Q}_{ir}+\frac{Q_{ir}}{(t+8)^{2/3}}\leq\frac{Q_{ir}}{(t+8)^{2/3}}, \tag{55}$$

where $Q_{ir}=32\left(1+\frac{d}{m}\right)^{-1/3}\sigma^2+128\left(1+\frac{d}{m}\right)^{2/3}L_1^2+64\left(1+\frac{d}{m}\right)^{-1/3}L_1^2+2L^2R^2\left(1+\frac{d}{m}\right)^{2/3}+416\left(1+\frac{d}{m}\right)^{2/3}L^2$ and $\overline{Q}_{ir}$ is a function of $\mathbb{E}\left[\|\nabla f(\mathbf{x}_0)-\mathbf{d}_0\|^2\right]$ and decays exponentially. Now we proceed by replacing the term $\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\mathbf{d}_t\|^2\right]$ in (42) by its upper bound in (55) and γ_{t+1} by $2/(t+9)$ to write

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_{t+1})-f(\mathbf{x}^*)\right]&\leq\left(1-\frac{2}{t+9}\right)\mathbb{E}\left[(f(\mathbf{x}_t)-f(\mathbf{x}^*))\right]\\&+\frac{R\sqrt{2Q_{ir}}}{(t+9)^{4/3}}+\frac{2LR^2}{(t+9)^2}.\end{aligned} \tag{56}$$

Note that we can write $(t+9)^2=(t+9)^{4/3}(t+9)^{2/3}\geq(t+9)^{4/3}9^{2/3}\geq4(t+9)^{4/3}$.Therefore,

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_{t+1})-f(\mathbf{x}^*)\right]&\leq\left(1-\frac{2}{t+9}\right)\mathbb{E}\left[(f(\mathbf{x}_t)-f(\mathbf{x}^*))\right]\\&+\frac{2R\sqrt{Q}+LD^2/2}{(t+9)^{4/3}}.\end{aligned} \tag{57}$$

Following the induction steps as in (49), we have,

$$\mathbb{E}\left[f(\mathbf{x}_t)-f(\mathbf{x}^*)\right]\leq\frac{Q_{ir}^{'}}{(t+8)^{1/3}}=O\left(\frac{(d/m)^{1/3}}{(t+9)^{1/3}}\right). \tag{58}$$

where $Q_{ir}'=\max\{2(f(\mathbf{x}_0)-f(\mathbf{x}^*)),2R\sqrt{2Q_{ir}}+LR^2/2\}$. □

Proof of Theorem 3.5(2). Following as in (50), we have,

$$\begin{aligned}&\gamma_t\mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right]\leq\mathbb{E}\left[F(\mathbf{x}_t)-F(\mathbf{x}_{t+1})\right]+\gamma_tR\frac{\sqrt{2Q_{ir}}}{(t+8)^{1/3}}+\frac{LR^2\gamma_t^2}{2}+\\&\Rightarrow\mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right]\leq\mathbb{E}\left[\frac{t+7}{2}F(\mathbf{x}_t)-\frac{t+8}{2}F(\mathbf{x}_{t+1})+\frac{1}{2}F(\mathbf{x}_t)\right]+R\frac{\sqrt{2Q_{ir}}}{(t+8)^{1/3}}+\frac{LR^2\gamma_t}{2}\\&\Rightarrow\sum_{t=0}^{T-1}\mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right]\leq\mathbb{E}\left[\frac{7}{2}F(\mathbf{x}_0)-\frac{T+7}{2}F(\mathbf{x}_T)+\sum_{t=0}^{T-1}\left(\frac{1}{2}F(\mathbf{x}_t)\right]+R\frac{\sqrt{2Q_{ir}}}{(t+8)^{1/3}}+\frac{LR^2\gamma_t}{2}\right)\\&\Rightarrow\sum_{t=0}^{T-1}\mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right]\leq\mathbb{E}\left[\frac{7}{2}F(\mathbf{x}_0)-\frac{7}{2}F(\mathbf{x}^*)\right]+\sum_{t=0}^{T-1}\left(\frac{1}{2}\left(F(\mathbf{x}_t)-F(\mathbf{x}^*)\right)+R\frac{\sqrt{2Q_{ir}}}{(t+8)^{1/3}}+\frac{LR^2\gamma_t}{2}\right)\\&\Rightarrow\sum_{t=0}^{T-1}\mathbb{E}\left[\mathcal{G}\left(\mathbf{x}_t\right)\right]\leq\frac{7}{2}F(\mathbf{x}_0)-\frac{7}{2}F(\mathbf{x}^*)+\sum_{t=0}^{T-1}\left(\frac{Q_{ir}^{'}+R\sqrt{2Q_{ir}}}{2(t+8)^{1/3}}+\frac{LR^2}{(t+8)}\right)\\&\Rightarrow T\mathbb{E}\left[\min_{t=0,\cdots,T-1}\mathcal{G}\left(\mathbf{x}_t\right)\right]\leq\frac{7}{2}F(\mathbf{x}_0)-\frac{7}{2}F(\mathbf{x}^*)+LR^2\ln(T+7)+\frac{Q_{ir}^{'}+R\sqrt{2Q_{ir}}}{2}(T+7)^{2/3}\\&\Rightarrow\mathbb{E}\left[\min_{t=0,\cdots,T-1}\mathcal{G}\left(\mathbf{x}_t\right)\right]\leq\frac{7(F(\mathbf{x}_0)-F(\mathbf{x}^*))}{2T}+\frac{LR^2\ln(T+7)}{T}+\frac{Q_{ir}^{'}+R\sqrt{2Q_{ir}}}{2T}(T+7)^{2/3}\end{aligned} \tag{59}$$

□

D Proofs for KWSA

Proof of Lemma 3.2(3). Following as in the proof of Lemma 3.2, we have,

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\mathbf{d}_t\|^2\right]$$

$$
\begin{aligned}
&\leq (1-\rho_t)^2\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_{t-1})\|^2\right] \\
&+ (1-\rho_t)^2\|\mathbb{E}\left[\nabla f(\mathbf{x}_{t-1})-\mathbf{d}_{t-1}\|^2\right] \\
&+ (1-\rho_t)^2\beta_t\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1})-\mathbf{d}_{t-1}\|^2\right] \\
&+ \frac{(1-\rho_t)^2}{\beta_t}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_{t-1})\|^2\right] \\
&+ \frac{\rho_t}{2}(1-\rho_t)c_t^2L^2d \\
&+ \rho_t(1-\rho_t)\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_{t-1})\|^2\right] \\
&+ \rho_t(1-\rho_t)\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1})-\mathbf{d}_{t-1}\|^2\right] \\
&\leq 2\rho_t^2\sigma^2+2\rho_t^2c_t^2dL^2 \\
&+ \left(1-\rho_t+\frac{(1-\rho_t)^2}{\beta_t}\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_{t-1})\|^2\right] \\
&+ \left(1-\rho_t+(1-\rho_t)^2\beta_t\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1})-\mathbf{d}_{t-1}\|^2\right] \\
&+ \frac{\rho_t}{2}(1-\rho_t)c_t^2L^2d \\
&\leq 2\rho_t^2\sigma^2+2\rho_tc_t^2dL^2 \\
&+ \left(1-\rho_t+\frac{(1-\rho_t)^2}{\beta_t}\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_{t-1})\|^2\right] \\
&+ \left(1-\rho_t+(1-\rho_t)^2\beta_t\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1})-\mathbf{d}_{t-1}\|^2\right], \qquad (60)
\end{aligned}
$$

where we used the gradient approximation bounds as stated in (15) and used Young's inequality to substitute the inner products and in particular substituted $2\langle\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_{t-1}),\nabla f(\mathbf{x}_{t-1})-\mathbf{d}_{t-1}\rangle$ by the upper bound $\beta_t\|\nabla f(\mathbf{x}_{t-1})-\mathbf{d}_{t-1}\|^2+(1/\beta_t)\|\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_{t-1})\|^2$ where $\beta_t>0$ is a free parameter.
According to Assumption A4, the norm $\|\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_{t-1})\|$ is bounded above by $L\|\mathbf{x}_t-\mathbf{x}_{t-1}\|$. In addition, the condition in Assumption A1 implies that $L\|\mathbf{x}_t-\mathbf{x}_{t-1}\|=L\gamma_t\|\mathbf{v}_t-\mathbf{x}_t\|\leq\gamma_tLR$. Therefore, we can replace $\|\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_{t-1})\|$ by its upper bound γ_tLR and since we assume that $\rho_t\leq 1$ we can replace all the terms $(1-\rho_t)^2$. Furthermore, using $\beta_t:=\rho_t/2$ we have,

$$
\begin{aligned}
&\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\mathbf{d}_t\|^2\right] \\
&\leq 2\rho_t^2\sigma^2+2\rho_tc_t^2dL^2+\gamma_t^2(1-\rho_t)\left(1+\frac{2}{\rho_t}\right)L^2R^2 \\
&+ (1-\rho_t)\left(1+\frac{\rho_t}{2}\right)\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1})-\mathbf{d}_{t-1}\|^2\right]. \qquad (61)
\end{aligned}
$$

Now using the inequalities $(1-\rho_t)(1+(2/\rho_t))\leq(2/\rho_t)$ and $(1-\rho_t)(1+(\rho_t/2))\leq(1-\rho/2)$ we obtain

$$
\begin{aligned}
\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\mathbf{d}_t\|^2\right] &\leq 2\rho_t^2\sigma^2+2\rho_tc_t^2dL^2+\frac{2L^2R^2\gamma_t^2}{\rho_t} \\
&+ \left(1-\frac{\rho_t}{2}\right))\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1})-\mathbf{d}_{t-1}\|^2\right]. \qquad (62)
\end{aligned}
$$

□

Proof of Theorem 3.4(3). Now using the result in Lemma B.1 we can characterize the convergence of the sequence of expected errors $\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\mathbf{d}_t\|^2\right]$ to zero. To be more precise, using the result in Lemma 3.2 and setting $\gamma_t=2/(t+8)$, $\rho_t=4/(t+8)^{2/3}$ and $c_t=2/\sqrt{d}(t+8)^{1/3}$ for any $\epsilon>0$ to obtain

$$
\begin{aligned}
&\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\mathbf{d}_t\|^2\right]\leq \\
&\left(1-\frac{2}{(t+8)^{2/3}}\right)\mathbb{E}\left[\|\nabla F(\mathbf{x}_{t-1})-\mathbf{d}_{t-1}\|^2\right] \\
&+ \frac{32\sigma^2+32L^2+2L^2R^2}{(t+8)^{4/3}}. \qquad (63)
\end{aligned}
$$

According to the result in Lemma B.1, the inequality in (45) implies that

$$
\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)-\mathbf{d}_t\|^2\right]\leq\frac{Q_{kw}}{(t+8)^{2/3}}, \qquad (64)
$$

where

$$Q = \max\left\{4\|\nabla f(\mathbf{x}_0) - \mathbf{d}_0\|^2, 32\sigma^2 + 32L^2 + 2L^2R^2\right\}$$

Now we proceed by replacing the term $\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right]$ in (42) by its upper bound in (55) and γ_{t+1} by $2/(t+9)$ to write

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*)\right] &\leq \left(1 - \frac{2}{t+9}\right)\mathbb{E}\left[(f(\mathbf{x}_t) - f(\mathbf{x}^*))\right] \\ &+ \frac{R\sqrt{Q_{kw}}}{(t+9)^{4/3}} + \frac{2LR^2}{(t+9)^2}.\end{aligned} \tag{65}$$

Note that we can write $(t+9)^2 = (t+9)^{4/3}(t+9)^{2/3} \geq (t+9)^{4/3}9^{2/3} \geq 4(t+9)^{4/3}$.Therefore,

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*)\right] &\leq \left(1 - \frac{2}{t+9}\right)\mathbb{E}\left[(f(\mathbf{x}_t) - f(\mathbf{x}^*))\right] \\ &+ \frac{2R\sqrt{Q_{kw}} + LD^2/2}{(t+9)^{4/3}}.\end{aligned} \tag{66}$$

Thus, for $t \geq 0$ by induction we have,

$$\mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^*)\right] \leq \frac{Q'}{(t+9)^{1/3}} = O\left(\frac{d^0}{(t+9)^{1/3}}\right). \tag{67}$$

where $Q' = \max\{2(f(\mathbf{x}_0) - f(\mathbf{x}^*)), 2R\sqrt{Q_{kw}} + LR^2/2\}$. □

Proof of Theorem 3.5(3). Following as in (50), we have,

$$\begin{aligned}
&\gamma_t\mathbb{E}\left[\mathcal{G}(\mathbf{x}_t)\right] \leq \mathbb{E}\left[F(\mathbf{x}_t) - F(\mathbf{x}_{t+1})\right] + \gamma_t R\frac{\sqrt{2Q_{kw}}}{(t+8)^{1/3}} + \frac{LR^2\gamma_t^2}{2} + \\
&\Rightarrow \mathbb{E}\left[\mathcal{G}(\mathbf{x}_t)\right] \leq \mathbb{E}\left[\frac{t+7}{2}F(\mathbf{x}_t) - \frac{t+8}{2}F(\mathbf{x}_{t+1}) + \frac{1}{2}F(\mathbf{x}_t)\right] + R\frac{\sqrt{2Q_{kw}}}{(t+8)^{1/3}} + \frac{LR^2\gamma_t}{2} \\
&\Rightarrow \sum_{t=0}^{T-1}\mathbb{E}\left[\mathcal{G}(\mathbf{x}_t)\right] \leq \mathbb{E}\left[\frac{7}{2}F(\mathbf{x}_0) - \frac{T+7}{2}F(\mathbf{x}_T) + \sum_{t=0}^{T-1}\left(\frac{1}{2}F(\mathbf{x}_t)\right] + R\frac{\sqrt{2Q_{kw}}}{(t+8)^{1/3}} + \frac{LR^2\gamma_t}{2}\right) \\
&\Rightarrow \sum_{t=0}^{T-1}\mathbb{E}\left[\mathcal{G}(\mathbf{x}_t)\right] \leq \mathbb{E}\left[\frac{7}{2}F(\mathbf{x}_0) - \frac{7}{2}F(\mathbf{x}^*)\right] + \sum_{t=0}^{T-1}\left(\frac{1}{2}\left(F(\mathbf{x}_t) - F(\mathbf{x}^*)\right) + R\frac{\sqrt{2Q_{kw}}}{(t+8)^{1/3}} + \frac{LR^2\gamma_t}{2}\right) \\
&\Rightarrow \sum_{t=0}^{T-1}\mathbb{E}\left[\mathcal{G}(\mathbf{x}_t)\right] \leq \frac{7}{2}F(\mathbf{x}_0) - \frac{7}{2}F(\mathbf{x}^*) + \sum_{t=0}^{T-1}\left(\frac{Q'_{kw} + R\sqrt{2Q_{kw}}}{2(t+8)^{1/3}} + \frac{LR^2}{(t+8)}\right) \\
&\Rightarrow T\mathbb{E}\left[\min_{t=0,\cdots,T-1}\mathcal{G}(\mathbf{x}_t)\right] \leq \frac{7}{2}F(\mathbf{x}_0) - \frac{7}{2}F(\mathbf{x}^*) + LR^2\ln(T+7) + \frac{Q'_{kw} + R\sqrt{2Q_{kw}}}{2}(T+7)^{2/3} \\
&\Rightarrow \mathbb{E}\left[\min_{t=0,\cdots,T-1}\mathcal{G}(\mathbf{x}_t)\right] \leq \frac{7(F(\mathbf{x}_0) - F(\mathbf{x}^*))}{2T} + \frac{LR^2\ln(T+7)}{T} + \frac{Q'_{kw} + R\sqrt{2Q_{kw}}}{2T}(T+7)^{2/3}
\end{aligned} \tag{68}$$

□

E Proofs for Non Convex Stochastic Frank Wolfe

Proof of Theorem 3.6. We reuse the following characterization derived earlier:

Lemma E.1. *Let Assumptions A3-A6 hold. Given the recursion in* (12)*, we have that* $\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2$ *satisfies*

$$\begin{aligned}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] &\leq 2\rho_t^2\sigma^2 + 4\rho_t^2 L_1^2 \\ &+ 8\rho_t^2\left(1 + \frac{s(d)}{m}\right)L_1^2 + \left(\frac{1+m}{2m}\right)\rho_t^2 c_t^2 L^2 M(\mu) \\ &+ \frac{2L^2R^2\gamma^2}{\rho_t} + \frac{\rho_t}{2m^2}c_t^2 L^2 M(\mu) \\ &+ \left(1 - \frac{\rho_t}{2}\right))\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right].\end{aligned} \tag{69}$$

Now using the result in Lemma B.1 we can characterize the convergence of the sequence of expected errors $\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right]$ to zero. To be more precise, using the result in Lemma 3.2 and setting $\gamma = T^{-3/4}$, $\rho_t = 4/\left(1+\frac{d}{m}\right)^{1/3}(t+8)^{1/2}$ and $c_t = 2\sqrt{m}/\sqrt{M(\mu)}(t+8)^{1/4}$ to obtain for all $t = 0, \cdots, T-1$,

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] \\
&\leq \left(1 - \frac{2}{\left(1+\frac{d}{m}\right)^{1/3}(t+8)^{1/2}}\right)\mathbb{E}\left[\|\nabla F(\mathbf{x}_{t-1}) - \mathbf{d}_{t-1}\|^2\right] \\
&+ \frac{32\sigma^2 + 64L_1^2 + 128\left(1+\frac{d}{m}\right)^{1/3}L_1^2}{(t+8)} \\
&+ \frac{8L^2R^2\left(1+\frac{d}{m}\right)^{1/3} + 416L^2}{(t+8)}.
\end{aligned} \tag{70}$$

Using Lemma B.1, we then have,

$$\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \mathbf{d}_t\|^2\right] = O\left(\frac{(d/m)^{2/3}}{(t+9)^{1/2}}\right), \forall\, t = 0, \cdots, T-1 \tag{71}$$

Finally, we have,

$$\begin{aligned}
&F(\mathbf{x}_{t+1}) \leq F(\mathbf{x}_t) + \gamma_t\langle \mathbf{d}_t, \mathbf{v}_t - \mathbf{x}_t\rangle \\
&+ \gamma\langle \nabla F(\mathbf{x}_t) - \mathbf{d}_t, \mathbf{v}_t - \mathbf{x}_t\rangle + \frac{LR^2\gamma^2}{2} \\
&\leq F(\mathbf{x}_t) + \gamma\langle \mathbf{d}_t, \operatorname*{argmin}_{\mathbf{v}\in\mathcal{C}}\langle \mathbf{v}, \nabla F(\mathbf{x}_t)\rangle - \mathbf{x}_t\rangle \\
&+ \gamma\langle \nabla F(\mathbf{x}_t) - \mathbf{d}_t, \mathbf{v}_t - \mathbf{x}_t\rangle + \frac{LR^2\gamma^2}{2} \\
&\leq F(\mathbf{x}_t) + \gamma\langle \nabla F(\mathbf{x}_t), \operatorname*{argmin}_{\mathbf{v}\in\mathcal{C}}\langle \mathbf{v}, \nabla F(\mathbf{x}_t)\rangle - \mathbf{x}_t\rangle \\
&+ \gamma\langle \nabla F(\mathbf{x}_t) - \mathbf{d}_t, \mathbf{v}_t - \operatorname*{argmin}_{\mathbf{v}\in\mathcal{C}}\langle \mathbf{v}, \nabla F(\mathbf{x}_t)\rangle\rangle + \frac{LR^2\gamma^2}{2} \\
&\leq F(\mathbf{x}_t) - \gamma\mathcal{G}(\mathbf{x}_t) + \frac{LR^2\gamma^2}{2} \\
&+ \gamma\langle \nabla F(\mathbf{x}_t) - \mathbf{d}_t, \mathbf{v}_t - \operatorname*{argmin}_{\mathbf{v}\in\mathcal{C}}\langle \mathbf{v}, \nabla F(\mathbf{x}_t)\rangle\rangle \\
&\Rightarrow \gamma\mathbb{E}\left[\mathcal{G}(\mathbf{x}_t)\right] \leq \mathbb{E}\left[F(\mathbf{x}_t)\right] - \mathbb{E}\left[F(\mathbf{x}_{t+1})\right] \\
&+ \gamma R\mathbb{E}\left[\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|\right] + \frac{LR^2\gamma^2}{2} \\
&\leq \mathbb{E}\left[F(\mathbf{x}_t)\right] - \mathbb{E}\left[F(\mathbf{x}_{t+1})\right] + \gamma_t R\sqrt{\mathbb{E}\left[\|\nabla F(\mathbf{x}_t) - \mathbf{d}_t\|^2\right]} + \frac{LR^2\gamma^2}{2} \\
&\leq \mathbb{E}\left[F(\mathbf{x}_t)\right] - \mathbb{E}\left[F(\mathbf{x}_{t+1})\right] + Q_{nc}\gamma\rho_t^{1/2}R(d/m)^{1/3} + \frac{LR^2\gamma^2}{2} \\
&\Rightarrow \mathbb{E}\left[\mathcal{G}_{min}\right]T\gamma \leq \mathbb{E}\left[F(\mathbf{x}_0)\right] - \mathbb{E}\left[F(\mathbf{x}_{t+1})\right] \\
&+ Q_{nc}\gamma R(d/m)^{1/3}\sum_{t=0}^{T-1}\rho_t^{1/2} + \frac{LR^2T\gamma^2}{2} \\
&\Rightarrow \mathbb{E}\left[\mathcal{G}_{min}\right] \leq \frac{\mathbb{E}\left[F(\mathbf{x}_0)\right] - \mathbb{E}\left[F(\mathbf{x}^*)\right]}{T\gamma} \\
&+ \gamma Q_{nc}R(d/m)^{1/3}\frac{\sum_{t=0}^{T-1}\rho_t^{1/2}}{T\gamma} + \frac{LR^2T\gamma^2}{2T\gamma} \\
&\Rightarrow \mathbb{E}\left[\mathcal{G}_{min}\right] \leq \frac{\mathbb{E}\left[F(\mathbf{x}_0)\right] - \mathbb{E}\left[F(\mathbf{x}^*)\right]}{T^{1/4}} \\
&+ \frac{Q_{nc}Rd^{1/3}}{T^{1/4}m^{1/3}} + \frac{LR^2}{2T},
\end{aligned} \tag{72}$$

where $\mathcal{G}_{min} = \min_{t=0,\cdots,T-1}\mathcal{G}(\mathbf{x}_t)$. □