A Proof of Proposition 1 and Proposition

Since the two propositions can be proved similarly, we only present the proof of Proposition 1.

First, if Θ_* is the optimal dual solution, by replacing $\ell_{y_i}(\cdot)$ in (4) with its conjugate form, the optimal primal solution can be solved by

$$W^* = \arg\min_{W \in \mathbb{R}^{d \times C}} \frac{\lambda}{2}\|W\|_F^2 + \sum_{i=1}^{n} (W^T \mathbf{x}_i)^T [\boldsymbol{\theta}_*]_i \tag{12}$$

Setting the gradient wit respect to $\mathbf{W}$ to zero, we obtain

$$W^* = -\frac{1}{\lambda}\sum_{i=1}^{n} \boldsymbol{x}_i (\boldsymbol{\theta}_i^*)^T = -\frac{1}{\lambda} X \Theta^{*T} \tag{13}$$

Second, let's consider how to obtain the dual solution Θ^* from the primal solution W^*. Note

$$\ell_{y_i}([\boldsymbol{o}^*]_i) = \boldsymbol{o}_i^{*T}\boldsymbol{\theta}_i^* - \ell_{y_i}^*(\boldsymbol{\theta}_i^*) \tag{14}$$

By the Frechel conjugate theory [2,4], we have

$$\nabla L_{y_i}(\boldsymbol{o}_i^*) = \boldsymbol{\theta}_i^* \tag{15}$$

Thus, we get the proposition.

B Proof of the Theorem

We need the following lemma for our proof

Lemma 2 (Corollary 7 [26]). *Let $A \in \mathbb{R}^{r \times m}$ be a standard Gaussian random matrix. For any $0 < \epsilon \leq 1/2$, with a probability at least $1 - \delta$, we have*

$$\left\| \frac{1}{m} AA^T - I \right\|_2 \leq \epsilon, \tag{16}$$

provided

$$m \geq \frac{(r+1)\log(2r/\delta)}{c_0 \epsilon^2}, \tag{17}$$

where $\|\cdot\|_2$ is the spectral norm of matrix and c_0 is a constant whose value is at least $1/4$.

Let the SVD of X be

$$X = U\Sigma V^T = \sum_{i=1}^{r} \lambda_i \mathbf{u}_i \mathbf{v}_i^T, \tag{18}$$

where $\Sigma = \text{diag}(\lambda_1, ..., \lambda_r)$, $U = (\mathbf{u}_1, ..., \mathbf{u}_r)$, $V = (\mathbf{v}_1, ..., \mathbf{v}_r)$, λ_i is the i-th singular value of X, $\mathbf{u}_i \in \mathbb{R}^d$ and $\mathbf{v}_i \in \mathbb{R}^n$ are the corresponding left and right singular vectors of X. We define

$$\Gamma^* = \Sigma V^T \Theta^{*T}, \text{ and } \widetilde{\Gamma} = \Sigma V^T \widehat{\Theta}^{*T} \tag{19}$$

It is straightforward to show that

$$W^* = -\frac{1}{\lambda} U \Sigma V^T \Theta^{*T} = -\frac{1}{\lambda} U \Gamma^*, \text{ and } \widetilde{W} = -\frac{1}{\lambda} U \Sigma V^T \hat{\Theta}^{*T} = -\frac{1}{\lambda} U \widetilde{\Gamma}. \tag{20}$$

Since U is an orthogonal matrix, we have

$$\|W^*\|_F = \frac{1}{\lambda}\|\Gamma^*\|_F, \|\widetilde{W}\|_F = \frac{1}{\lambda}\|\widetilde{\Gamma}\|_F, \text{ and } \|\widetilde{W} - W^*\|_F = \frac{1}{\lambda}\|\widetilde{\Gamma} - \Gamma^*\|_F \tag{21}$$

We define $A = U^T R^T \in \mathbb{R}^{r \times m}$. It is easy to verify that A is a Gaussian matrix of size $r \times m$.

Recall the dual problem for the original problem,

$$\max_{\Theta} -\sum_{i=1}^{n} \ell^*_{y_i}(\boldsymbol{\theta}_i) - \frac{1}{2\lambda}\|X\Theta^{*T}\|_F^2 \tag{22}$$

and the dual problem for the projected problem,

$$\max_{\Theta} -\sum_{i=1}^{n} \ell^*_{y_i}(\boldsymbol{\theta}_i) - \frac{1}{2\lambda}\left\|\frac{1}{\sqrt{m}} R^T X \Theta^{*T}\right\|_F^2 \tag{23}$$

Define $L(\Theta)$ and $\hat{L}(\Theta)$ as

$$\begin{aligned}
L(\Theta) &= -\sum_{i=1}^{n} \ell^*_{y_i}(\boldsymbol{\theta}_i) - \frac{1}{2\lambda}\|X\Theta^T\|_F^2 = -\sum_{i=1}^{n} \ell^*_{y_i}(\boldsymbol{\theta}_i) - \frac{1}{2\lambda} Tr(\Theta X^T X \Theta^T) \\
\hat{L}(\Theta) &= -\sum_{i=1}^{n} \ell^*_{y_i}(\boldsymbol{\theta}_i) - \frac{1}{2\lambda}\left\|\frac{R^T X}{\sqrt{m}}\Theta^T\right\|_F^2 \\
&= -\sum_{i=1}^{n} \ell^*_{y_i}(\boldsymbol{\theta}_i) - \frac{1}{2\lambda} Tr\left(\Theta X^T \frac{RR^T}{m} X \Theta^T\right)
\end{aligned}$$

Define

$$\begin{aligned}
\widehat{H}(t) &= \hat{L}(t(\widehat{\Theta}^* - \Theta^*) + \Theta^*) \\
&= -\sum_{i=1}^{n} L^*_{y_i}(t(\hat{\boldsymbol{\theta}}^*_i - \boldsymbol{\theta}^*_i) + \boldsymbol{\theta}^*_i) - \frac{1}{2\lambda} t^2 Tr\left((\widehat{\Theta}^* - \Theta^*)\frac{X^T RR^T X}{m}(\widehat{\Theta}^* - \Theta^*)^T)\right) \\
&\quad - \frac{1}{\lambda} t \cdot Tr\left((\widehat{\Theta}^* - \Theta^*)\frac{X^T RR^T X}{m}\Theta^{*T}\right) - \frac{1}{2\lambda} Tr\left(\Theta^* \frac{X^T RR^T X}{m}\Theta^{*T}\right)
\end{aligned}$$

where $0 \le t \le 1$ and $t = 1$ maximizes $\widehat{H}(t)$.

Similarly, define $H(t) = L(t(\widehat{\Theta}_* - \Theta_*) + \Theta_*)$, where $0 \leq t \leq 1$ and $t = 0$ maximizes $H(t)$.

We can see that $\widehat{H}(t)$ is $\frac{1}{\lambda} Tr\left((\widehat{\Theta}^* - \Theta^*)\frac{X^T R R^T X}{m}(\widehat{\Theta}^* - \Theta^*)^T\right)$-strongly concave. Thus, it follows

$$\widehat{H}(1) \geq \widehat{H}(0) + \frac{1}{2\lambda} Tr\left((\widehat{\Theta}^* - \Theta^*)\frac{X^T R R^T X}{m}(\widehat{\Theta}^* - \Theta^*)^T\right) \tag{24}$$

Using strongly concave of $\hat{H}(t)$ (strong convex of $-\widehat{H}(t)$) again,

$$-\widehat{H}(1) \geq -\widehat{H}(0) - \frac{1}{\lambda}\nabla\widehat{H}(0) + \frac{1}{2\lambda} Tr\left((\widehat{\Theta}^* - \Theta^*)\frac{X^T R R^T X}{m}(\widehat{\Theta}^* - \Theta^*)^T\right)$$

Rearranging terms, we have

$$\begin{aligned}
&\widehat{H}(1) + \frac{1}{2\lambda} Tr\left((\widehat{\Theta}^* - \Theta^*)\frac{X^T R R^T X}{m}(\widehat{\Theta}^* - \Theta^*)^T\right) \\
&\leq \widehat{H}(0) + \frac{1}{\lambda}(\nabla\widehat{H}(0) - \nabla H(0) + \nabla H(0)) \\
&\leq \widehat{H}(0) + \frac{1}{\lambda}(\nabla\widehat{H}(0) - \nabla H(0)) \\
&= \hat{H}(0) - \frac{1}{\lambda} Tr\left((\widehat{\Theta}^* - \Theta^*)\frac{X^T R R^T X}{m}\Theta^{*T}\right) + \frac{1}{\lambda} Tr\left((\widehat{\Theta}^* - \Theta^*)X^T X \Theta^{*T}\right) \\
&= \widehat{H}(0) + \frac{1}{\lambda} Tr\left((\widehat{\Theta}^* - \Theta^*)\left(X^T X - \frac{X^T R R^T X}{m}\right)\Theta^{*T}\right)
\end{aligned} \tag{25}$$

where the second equality follows from the fact that $\nabla H(0) \leq 0$ since $t = 0$ maximizes $H(t)$ over $0 \leq t \leq 1$.

Adding (24) and (25), we get

$$Tr\left((\widehat{\Theta}^* - \Theta^*)\frac{X^T R R^T X}{m}(\widehat{\Theta}^* - \Theta^*)^T\right) \leq Tr\left((\widehat{\Theta}^* - \Theta^*)\left(X^T X - \frac{X^T R R^T X}{m}\right)\Theta^{*T}\right)$$

Let $A = U^T R \in \mathbb{R}^{r \times m}$. It can be verified that A is a Gaussian matrix. By Lemma 2, with a probability of at least $1 - \delta$, we have $\|I - \frac{1}{m}AA^T\|_2 \leq \epsilon$, under the given condition on m. Hence, the largest eigenvalue of $I - \frac{1}{m}AA^T$ is no greater than ϵ. Therefore, we have

$$\begin{aligned}
&Tr\left((\Theta^* - \widehat{\Theta}^*)\left(X^T X - \frac{X^T R R^T X}{m}\right)\Theta^{*T}\right) \\
&= Tr\left((\Gamma^* - \widetilde{\Gamma}^*)\left(I - \frac{AA^T}{m}\right)\Gamma^{*T}\right) \\
&\leq \epsilon\|\Gamma^* - \widetilde{\Gamma}^*\|_F\|\Gamma^*\|_F
\end{aligned} \tag{26}$$

If $\|I - \frac{1}{m}AA^T\|_2 \leq \epsilon$, we have

$$\max_{e:e^T e=1} e^T(I - AA^T)e \leq \epsilon, \tag{27}$$

thus, $\min_{e:e^Te=1} e^TAA^Te \geq 1-\epsilon$. This is to say, the minimal eigenvalue of AA^T is no less than $1-\epsilon$.

$$\begin{aligned} &Tr\left((\widehat{\Theta}^* - \Theta^*)\frac{X^TRR^TX}{m}(\widehat{\Theta}^* - \Theta^*)^T\right) \\ &= Tr\left((\widetilde{\Gamma} - \Gamma^*)^T\frac{U^TRR^TU}{m}(\widetilde{\Gamma} - \Gamma^*)\right) \\ &\geq (1-\epsilon)\|\Gamma^* - \widetilde{\Gamma}\|_F^2 \end{aligned}$$

Thus, we have

$$\|\Gamma^* - \widetilde{\Gamma}\|_F^2 \leq \frac{\epsilon}{1-\epsilon}\|\Gamma^*\|_F \tag{28}$$

Applying this to (21), we get the proposition.

C Proof of Lemma 1

Let $\mathbf{z}_\psi^*$ denote the optimum of $\psi(\mathbf{z})$, i.e., $\mathbf{z}_\psi^* = \arg\min_{\mathbf{z}} \psi(\mathbf{z})$. $\psi(\mathbf{z}) = \sum_{k=1}^{K} \psi_k(\mathbf{z})$, where $\psi_k(\mathbf{z})$ is the objective on k-th machine. ξ_t^k denotes a random sample drawn on k-th machine at t-th iteration. And $\nabla\psi_k(\mathbf{z};\xi_t^k)$ denotes a stochastic gradient.

Let $\bar{\mathbf{z}}_t = \frac{1}{K}\sum_{k=1}^{K}\mathbf{z}_t^k$ denote the average of the model across K machines at iteration t. Note that this is a virtual sequence used in the analysis. In the algorithm, it is only available for the iterations where communication has been done to synchronize machines.

Since $\mathcal{L}(\mathbf{z})$ is L-weakly convex, we know that $\psi(\mathbf{z})$ is L-strong convexity. Thus, we get

$$\psi(\mathbf{z}_\psi^*) - \psi(\bar{\mathbf{z}}_{t-1}) \geq \langle\nabla\psi(\bar{\mathbf{z}}_{t-1}), \mathbf{z}_\psi^* - \bar{\mathbf{z}}_{t-1}\rangle + \frac{L}{2}\|\bar{\mathbf{z}}_{t-1} - \mathbf{z}_\psi^*\|^2 \tag{29}$$

By the $3L$-smooth of $\psi(\mathbf{z})$,

$$\frac{3L}{2}\|\bar{\mathbf{z}}_{t-1} - \bar{\mathbf{z}}_t\|^2 + \langle\nabla\psi(\bar{\mathbf{z}}_{t-1}), \bar{\mathbf{z}}_t - \bar{\mathbf{z}}_{t-1}\rangle + \psi(\bar{\mathbf{z}}_{t-1}) \geq \psi(\bar{\mathbf{z}}_t) \tag{30}$$

Therefore, by (29) and (30), we get

$$\psi(\bar{\mathbf{z}}_t) - \psi(\mathbf{z}_\psi^*) + \frac{L}{2}\|\bar{\mathbf{z}}_{t-1} - \mathbf{z}_*\|^2 \leq \frac{3L}{2}\|\bar{\mathbf{z}}_{t-1} - \bar{\mathbf{z}}_t\|^2 + \langle\nabla\psi(\bar{\mathbf{z}}_{t-1}), \bar{\mathbf{z}}_t - \mathbf{z}_\psi^*\rangle \tag{31}$$

Let's consider the last term,

$$\begin{aligned}
&\langle \nabla\psi(\bar{\mathbf{z}}_{t-1}), \bar{\mathbf{z}}_t - \mathbf{z}_\psi^* \rangle \\
&= \left\langle \frac{1}{K}\sum_{k=1}^{K} \nabla\psi_k(\bar{\mathbf{z}}_{t-1}), \bar{\mathbf{z}}_t - \mathbf{z}_\psi^* \right\rangle \\
&= \left\langle \frac{1}{K}\sum_{k=1}^{K} [\nabla\psi_k(\bar{\mathbf{z}}_{t-1}) - \nabla\psi_k(\mathbf{z}_{t-1}^k)], \bar{\mathbf{z}}_t - \mathbf{z}_\psi^* \right\rangle \qquad ① \\
&\quad + \left\langle \frac{1}{K}\sum_{k=1}^{K} [\nabla\psi_k(\mathbf{z}_{t-1}^k) - \nabla\psi_k(\mathbf{z}_{t-1}^k; \xi_{t-1}^k)], \bar{\mathbf{z}}_t - \mathbf{z}_\psi^* \right\rangle \qquad ② \\
&\quad + \left\langle \frac{1}{K}\sum_{k=1}^{K} [\nabla\psi_k(\mathbf{z}_{t-1}^k; \xi_{t-1}^k)], \bar{\mathbf{z}}_t - \mathbf{z}_\psi^* \right\rangle \qquad ③.
\end{aligned} \tag{32}$$

First, the term ① can be bounded by using Young's inequality,

$$\begin{aligned}
&\left\langle \frac{1}{K}\sum_{k=1}^{K} [\nabla\psi_k(\bar{\mathbf{z}}_{t-1}) - \nabla\psi_k(\mathbf{z}_{t-1}^k)], \bar{\mathbf{z}}_t - \mathbf{z}_\psi^* \right\rangle \\
&\leq \frac{1}{2L} \left\| \frac{1}{K}\sum_{k=1}^{K} [\nabla\psi_k(\bar{\mathbf{z}}_{t-1})] - \nabla\psi_k(\mathbf{z}_{t-1}^k) \right\|^2 + \frac{L}{2}\|\bar{\mathbf{z}}_t - \mathbf{z}_\psi^*\|^2.
\end{aligned} \tag{33}$$

Second, let $\hat{\mathbf{z}}_t = \bar{\mathbf{z}}_{t-1} - \frac{\eta}{K}\sum_{k=1}^{K} \nabla\psi_k(\mathbf{z}_{t-1}^k)$, the term ② will be

$$\begin{aligned}
&\left\langle \frac{1}{K}\sum_{k=1}^{K} [\nabla\psi_k(\mathbf{z}_{t-1}^k) - \nabla\psi_k(\mathbf{z}_{t-1}^k; \xi_{t-1}^k)], \bar{\mathbf{z}}_t - \mathbf{z}_\psi^* \right\rangle \\
&= \left\langle \frac{1}{K}\sum_{k=1}^{K} [\nabla\psi_k(\mathbf{z}_{k,t-1}) - \nabla\psi_k(\mathbf{z}_{t-1}^k; \xi_{t-1}^k)], \bar{\mathbf{z}}_t - \hat{\mathbf{z}}_t \right\rangle \\
&\quad + \left\langle \frac{1}{K}\sum_{k=1}^{K} [\nabla\psi_k(\mathbf{z}_{t-1}^k) - \nabla\psi_k(\mathbf{w}_{k,t-1}^f; \xi_{t-1}^k)], \hat{\mathbf{z}}_t - \mathbf{z}_\psi^* \right\rangle \\
&= \eta \left\| \frac{1}{K}\sum_{k=1}^{K} [\nabla\psi_k(\mathbf{z}_{t-1}^k) - \nabla\psi_k(\mathbf{z}_{t-1}^k; \xi_{t-1}^k)] \right\|^2 \\
&\quad + \left\langle \frac{1}{K}\sum_{k=1}^{K} [\nabla\psi_k(\mathbf{z}_{t-1}^k) - \nabla\psi_k(\mathbf{z}_{t-1}^k; \xi_{t-1}^k)], \hat{\mathbf{z}}_t - \mathbf{z}_\psi^* \right\rangle.
\end{aligned} \tag{34}$$

Third, the term ③ becomes

$$\begin{aligned}&\left\langle \frac{1}{K}\sum_{k=1}^{K}[\nabla\psi_k(\mathbf{z}_{t-1}^k;\xi_{t-1}^k)], \bar{\mathbf{z}}_t - \mathbf{z}_\psi^* \right\rangle \\ &= \frac{1}{\eta}\left\langle \bar{\mathbf{z}}_{t-1} - \bar{\mathbf{z}}_t, \bar{\mathbf{z}}_t - \mathbf{z}_\psi^* \right\rangle \\ &= \frac{\|\bar{\mathbf{z}}_{t-1} - \mathbf{z}_\psi^*\|^2 - \|\bar{\mathbf{z}}_{t-1} - \bar{\mathbf{z}}_t\|^2 - \|\bar{\mathbf{z}}_t - \mathbf{z}_\psi^*\|^2}{2\eta}.\end{aligned} \tag{35}$$

Finally, plug (33), (34) and (35) into (31), we get

$$\begin{aligned}\psi(\bar{\mathbf{z}}_t) - \psi(\mathbf{z}_\psi^*) \leq& \left(\frac{3L}{2} - \frac{1}{2\eta}\right)\|\bar{\mathbf{z}}_{t-1} - \bar{\mathbf{z}}_t\|^2 \\ &+ \left(\frac{1}{2\eta} - \frac{L}{2}\right)\|\bar{\mathbf{z}}_{t-1} - \mathbf{z}_*\|^2 - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\|\bar{\mathbf{z}}_t - \mathbf{z}_*\|^2 \\ &+ \frac{1}{2L}\left\|\frac{1}{K}\sum_{k=1}^{K}[\nabla\psi_k(\bar{\mathbf{z}}_{t-1})] - \nabla\psi_k(\mathbf{z}_{t-1}^k)\right\|^2 \\ &+ \eta\left\|\frac{1}{K}\sum_{k=1}^{K}[\nabla\psi_k(\mathbf{z}_{t-1}^k) - \nabla\psi_k(\mathbf{z}_{t-1}^k;\xi_{t-1}^k)]\right\|^2 \\ &+ \left\langle \frac{1}{K}\sum_{k=1}^{K}[\nabla\psi_k(\mathbf{z}_{t-1}^k) - \nabla\psi_k(\mathbf{z}_{t-1}^k;\xi_{t-1}^k)], \hat{\mathbf{z}}_t - \mathbf{z}_\psi^* \right\rangle.\end{aligned} \tag{36}$$

By the setting of $\eta \leq \frac{1}{3L}$ and the unbiased stochastic gradient $E[\nabla\psi_k(\bar{\mathbf{z}}_{t-1}^k;\xi_{t-1}^k)] = \nabla\psi_k(\bar{\mathbf{z}}_{t-1}^k)$. We sum up over $t = 1, ..., T$ and take the expectation,

$$\begin{aligned}&E\left[\frac{1}{T}\sum_{t=1}^{T}\psi(\bar{\mathbf{z}}_t) - \psi(\mathbf{z}_\psi^*)\right] \\ &\leq \frac{1}{T}\left(\frac{1}{2\eta} - \frac{L}{2}\right)\|\bar{\mathbf{z}}_0 - \mathbf{z}_\psi^*\|^2 + \frac{1}{2L}\frac{1}{T}\sum_{t=1}^{T}\left\|\frac{1}{K}\sum_{k=1}^{K}[\nabla\psi_k(\bar{\mathbf{z}}_{t-1})] - \nabla\psi_k(\mathbf{z}_{t-1}^k)\right\|^2 \\ &+ \frac{1}{T}\sum_{t=1}^{T}\eta\left\|\frac{1}{K}\sum_{k=1}^{K}[\nabla\psi_k(\mathbf{z}_{t-1}^k) - \nabla\psi_k(\mathbf{z}_{t-1}^k;\xi_{t-1}^k)]\right\|^2 \\ &\leq \frac{1}{2\eta T}\|\bar{\mathbf{z}}_{t-1} - \mathbf{z}_\psi^*\|^2 + \frac{1}{2L}\frac{1}{T}\sum_{t=1}^{T}\left\|\frac{1}{K}\sum_{k=1}^{K}[\nabla\psi_k(\bar{\mathbf{z}}_{t-1})] - \nabla\psi_k(\mathbf{z}_{t-1}^k)\right\|^2 \\ &+ \eta\frac{1}{K^2}\sum_{k=1}^{K}\|[\nabla\psi_k(\mathbf{z}_{t-1}^k) - \nabla\psi_k(\mathbf{z}_{t-1}^k;\xi_{t-1}^k)]\|^2 \\ &\leq \frac{1}{2\eta T}\|\bar{\mathbf{z}}_{t-1} - \mathbf{z}_\psi^*\|^2 + \frac{\eta\sigma^2}{K} + \frac{1}{2L}\frac{1}{T}\sum_{t=1}^{T}\left\|\frac{1}{K}\sum_{k=1}^{K}[\nabla\psi_k(\bar{\mathbf{z}}_{t-1})] - \nabla\psi_k(\mathbf{z}_{t-1}^k)\right\|^2.\end{aligned}$$

$$\begin{aligned}
&\frac{1}{2L}\frac{1}{T}\sum_{t=1}^{T}\left\|\frac{1}{K}\sum_{k=1}^{K}[\nabla\psi_k(\bar{\mathbf{z}}_{t-1})]-\nabla\psi_k(\mathbf{z}_{t-1}^k)\right\|^2\\
&\leq\frac{1}{2LT}\sum_{t=1}^{T}\frac{1}{K}\sum_{k=1}^{K}\|\nabla\psi_k(\bar{\mathbf{z}}_{t-1})-\nabla\psi_k(\mathbf{z}_{t-1}^k)\|^2\\
&\leq\frac{1}{2L}\frac{1}{K}\sum_{k=1}^{K}L^2\|\bar{\mathbf{z}}_{t-1}-\mathbf{z}_{t-1}^k\|^2\\
&\leq 2L\eta^2I^2B^2.
\end{aligned}\tag{37}$$

Thus,

$$E\left[\frac{1}{T}\sum_{t=1}^{T}\psi(\bar{\mathbf{z}}_t)-\psi(\mathbf{z}_\psi^*)\right]\leq\frac{1}{2\eta T}\|\bar{\mathbf{z}}_0-\mathbf{z}_\psi^*\|^2+\frac{\eta\sigma^2}{K}+2L\eta^2I^2B^2.\tag{38}$$

D Proof of Theorem 2

Proof. Define $\mathcal{L}_s(\mathbf{z})=\mathcal{L}(\mathbf{z})+\frac{1}{2\gamma}\|\mathbf{z}-\mathbf{z}_{s-1}\|^2$. We can see that $\mathcal{L}_s(\mathbf{v})$ is convex and smooth since $\gamma\leq 1/L$. The smooth coefficient of $\mathcal{L}_s$ is $\hat{L}=L+1/\gamma=3L$. According to Theorem 2.1.5 of [21], we have

$$\|\nabla\mathcal{L}_s(\mathbf{z}_s)\|^2\leq 2\hat{L}(\mathcal{L}_s(\mathbf{z}_s)-\mathcal{L}_s(\mathbf{z}_{\mathcal{L}_s}^*)).\tag{39}$$

Applying Lemma 1, we have

$$E_{s-1}[\mathcal{L}_s(\mathbf{z}_s)-\mathcal{L}_s(\mathbf{z}_{\mathcal{L}_s}^*)]\leq\frac{1}{2\eta_sT_s}\|\mathbf{z}_{s-1}-\mathbf{z}_{\mathcal{L}_s}^*\|^2+2L\eta_s^2I_s^2B^2\mathbb{I}_{I_s>1}+\frac{\eta_s\sigma^2}{K}.\tag{40}$$

We define $I_s'=1/\sqrt{K\eta_s}=\frac{1}{K\sqrt{\eta_0}}\exp(\frac{d(s-1)}{2})$. Applying this to (39), we get

$$E[\|\nabla\mathcal{L}_s(\mathbf{z}_s)\|^2]\leq 2\hat{L}\left[\frac{\|\mathbf{z}_{s-1}-\mathbf{z}_{\mathcal{L}_s}^*\|^2}{2\eta_sT_s}+2L\eta_s^2{I'}_s^2B^2+\frac{\eta_s\sigma^2}{K}\right].\tag{41}$$

Note that $\mathcal{L}_s(\mathbf{v})$ is L-strongly convex, we have

$$\mathcal{L}_s(\mathbf{z}_{s-1})\geq\mathcal{L}_s(\mathbf{z}_{\mathcal{L}_s}^*)+\frac{L}{2}\|\mathbf{z}_{s-1}-\mathbf{z}_{\mathcal{L}_s}^*\|^2.\tag{42}$$

Plug (42) into (40), we get

$$\begin{aligned}
&E_{s-1}[\mathcal{L}(\mathbf{z}_s)+L_s\|\mathbf{z}_s-\mathbf{z}_{s-1}\|^2]\\
&\leq\mathcal{L}_s(\mathbf{z}_{s-1})-\frac{L}{2}\|\mathbf{z}_{s-1}-\mathbf{z}_{\mathcal{L}_s}^*\|^2+\frac{\|\mathbf{z}_{s-1}-\mathbf{z}_{\mathcal{L}_s}^*\|^2}{2\eta_sT_s}+2L\eta_s^2{I'}_s^2B^2+\frac{\eta_s\sigma^2}{K}.
\end{aligned}\tag{43}$$

Noting $\eta_s T_s L = 2$ and $\mathcal{L}_s(\mathbf{z}_{s-1}) = f(\mathbf{z}_{s-1})$, we rearrange terms and get

$$\frac{\|\mathbf{z}_{s-1} - \mathbf{z}^*_{\mathcal{L}_s}\|^2}{2\eta_s T_s} \leq \mathcal{L}(\mathbf{z}_{s-1}) - E_{s-1}[\mathcal{L}(\mathbf{z}_s)] + 2L\eta_s^2 {I'}_s^2 B^2 + \frac{\eta_s \sigma^2}{K}. \tag{44}$$

Combining (41) and (44), we get

$$\begin{aligned} E_{s-1}\|\nabla \mathcal{L}_s(\mathbf{z}_s)\|^2 &\leq 2\hat{L}\left[\mathcal{L}(\mathbf{z}_{s-1}) - E_{s-1}[\mathcal{L}(\mathbf{z}_s)] + 4L\eta_s^2 {I'}_s^2 B^2 + \frac{2\eta_s\sigma^2}{K}\right] \\ &= 6L\left[\mathcal{L}(\mathbf{z}_{s-1}) - E_{s-1}[\mathcal{L}(\mathbf{z}_s)] + 4L\eta_s^2 {I'}_s^2 B^2 + \frac{2\eta_s\sigma^2}{K}\right]. \end{aligned} \tag{45}$$

Taking expectation on both sides over all randomness until $\mathbf{z}_{s-1}$ is generated and by tower property, we have

$$E\|\nabla \mathcal{L}_s(\mathbf{z}_s)\|^2 \leq 6L\left(E[\mathcal{L}(\mathbf{z}_{s-1}) - f(\mathbf{z}^*_{\mathcal{L}})] - E[f(\mathbf{z}_s) - f(\mathbf{z}^*_{\mathcal{L}})] + 4L\eta_s^2 {I'}_s^2 B^2 + \frac{2\eta_s\sigma^2}{K}\right). \tag{46}$$

Since $\mathcal{L}(\mathbf{z})$ is L-smooth and hence is L-weakly convex, we have

$$\begin{aligned} &\mathcal{L}(\mathbf{z}_{s-1}) \geq \mathcal{L}(\mathbf{z}_s) + \langle \nabla \mathcal{L}(\mathbf{z}_s), \mathbf{z}_{s-1} - \mathbf{z}_s\rangle - \frac{L}{2}\|\mathbf{z}_{s-1} - \mathbf{z}_s\|^2 \\ &= \mathcal{L}(\mathbf{z}_s) + \langle \nabla \mathcal{L}(\mathbf{z}_s) + 2L(\mathbf{z}_s - \mathbf{z}_{s-1}), \mathbf{z}_{s-1} - \mathbf{z}_s\rangle + \frac{3}{2}L\|\mathbf{z}_{s-1} - \mathbf{z}_s\|^2 \\ &= f(\mathbf{z}_s) + \langle \nabla \mathcal{L}_s(\mathbf{z}_s), \mathbf{z}_{s-1} - \mathbf{z}_s\rangle + \frac{3}{2}L\|\mathbf{z}_{s-1} - \mathbf{z}_s\|^2 \\ &= \mathcal{L}(\mathbf{z}_s) - \frac{1}{2L}\langle \nabla \mathcal{L}_s(\mathbf{z}_s), \nabla \mathcal{L}_s(\mathbf{z}_s) - \nabla \mathcal{L}(\mathbf{z}_s)\rangle + \frac{3}{8L}\|\nabla \mathcal{L}_s(\mathbf{z}_s) - \nabla \mathcal{L}(\mathbf{z}_s)\|^2 \\ &= \mathcal{L}(\mathbf{z}_s) - \frac{1}{8L}\|\nabla \mathcal{L}_s(\mathbf{z}_s)\|^2 - \frac{1}{4L}\langle \nabla \mathcal{L}_s(\mathbf{z}_s), \nabla \mathcal{L}(\mathbf{z}_s)\rangle + \frac{3}{8L}\|\nabla \mathcal{L}(\mathbf{z}_s)\|^2. \end{aligned} \tag{47}$$

Rearranging terms, it yields

$$\begin{aligned} &\mathcal{L}(\mathbf{z}_s) - \mathcal{L}(\mathbf{z}_{s-1}) \leq \frac{1}{8L}\|\nabla \mathcal{L}_s(\mathbf{z}_s)\|^2 + \frac{1}{4L}\langle \nabla \mathcal{L}_s(\mathbf{z}_s), \nabla \mathcal{L}(\mathbf{z}_s)\rangle - \frac{3}{8L}\|\nabla \mathcal{L}(\mathbf{z}_s)\|^2 \\ &\leq \frac{1}{8L}\|\nabla \mathcal{L}_s(\mathbf{z}_s)\|^2 + \frac{1}{8L}(\|\nabla \mathcal{L}_s(\mathbf{z}_s)\|^2 + \|\nabla \mathcal{L}(\mathbf{z}_s)\|^2) - \frac{3}{8L}\|\nabla \mathcal{L}(\mathbf{z}_s)\|^2. \end{aligned} \tag{48}$$

Define $\Delta_s = \mathcal{L}(\mathbf{z}_s) - \mathcal{L}(\mathbf{z}^*_{\mathcal{L}})$. Combining (46) and (48), we get

$$E[\Delta_s - \Delta_{s-1}] \leq \frac{3}{2}E(\Delta_{s-1} - \Delta_s) + 6L\eta_s^2 {I'}_s^2 B^2 + \frac{3\eta_s\sigma^2}{K} - \frac{\mu}{2L}E[\Delta_s]. \tag{49}$$

Therefore,

$$\left(\frac{5}{2} + \frac{\mu}{2L}\right)E[\Delta_s] \leq \frac{5}{2}E[\Delta_{s-1}] + 6L\eta_s^2 {I'}_s^2 B^2 + \frac{3\eta_s\sigma^2}{K}. \tag{50}$$

Using $c_1 = \frac{\mu/L}{5+\mu/L}$ as defined in the theorem,

$$\begin{aligned}
E[\Delta_S] &\leq \frac{5L}{5L+\mu} E[\Delta_{S-1}] + \frac{2L}{5L+\mu}\left[6L\eta_S^2 {I'}_S^2 B^2 + \frac{3\eta_S\sigma^2}{K}\right] \\
&= (1-c_1)\left[E[\Delta_{S-1}] + \frac{2}{5}\left(6L\eta_S^2 {I'}_S^2 B^2 + \frac{3\eta_S\sigma^2}{K}\right)\right] \\
&\leq (1-c_1)^S E[\Delta_0] + \frac{12LB^2}{5}\sum_{j=1}^{S}\eta_j^2 {I'}_j^2 (1-c_1)^{S+1-j} + \frac{6\sigma^2}{5K}\sum_{j=1}^{S}\eta_j(1-c_1)^{S+1-j} \\
&= (1-c_1)^S E[\Delta_0] + \frac{12LB^2}{5}\sum_{j=1}^{S}\eta_j^2 {I'}_j^2 (1-c_1)^{S+1-j} + \frac{6\sigma^2}{5K}\sum_{j=1}^{S}\eta_j(1-c_1)^{S+1-j}.
\end{aligned} \tag{51}$$

We then have

$$\begin{aligned}
E[\Delta_S] &\leq (1-c_1)^S E[\Delta_0] + \left(\frac{12LB^2}{5K} + \frac{6\sigma^2}{5K}\right)\sum_{j=1}^{S}\eta_j(1-c_1)^{S+1-j} \\
&\leq \exp(-c_1 S)\Delta_0 + \left(\frac{12LB^2}{5K} + \frac{6\sigma^2}{5K}\right)\sum_{j=1}^{S}\eta_j \exp(-c_1(S+1-j)) \\
&= \exp(-c_1 S)\Delta_0 + \left(\frac{12LB^2}{5} + \frac{6\sigma^2}{5}\right)\eta_0 S\exp(-c_1 S).
\end{aligned} \tag{52}$$

To achieve $E[\Delta_S] \leq \epsilon$, it suffices to make

$$\exp(-c_1 S)\Delta_0 \leq \epsilon/2 \tag{53}$$

and

$$\left(\frac{12LB^2}{5} + \frac{6\sigma^2}{5}\right)\eta_0 S\exp(-c_1 S) \leq \epsilon/2. \tag{54}$$

So, it suffices to make

$$S \geq c_1^{-1}\max\left\{\log\left(\frac{2\Delta_0}{\epsilon}\right), \log S + \log\left[\frac{2\eta_0}{\epsilon}\frac{12LB^2+6\sigma^2}{5}\right]\right\}. \tag{55}$$

Taking summation of iteration over $s = 1, ..., S$, we have the total iteration complexity as

$$\begin{aligned}
T &= \sum_{s=1}^{S} T_s \leq \frac{2}{L\eta_0 K}\frac{\exp(c_1 S)-1}{\exp(c_1)-1} \leq \frac{2}{L\eta_0 K}\frac{5L+\mu}{\mu}\exp(c_1 S) \\
&= \tilde{O}\left(\max\left(\frac{\Delta_0}{\mu\epsilon\eta_0 K}, \frac{S(LB^2+\sigma^2)}{\mu\epsilon K}\right)\right) = \tilde{O}\left(\max\left(\frac{1}{\mu\epsilon\eta_0 K}, \frac{1}{\mu^2 K\epsilon}\right)\right).
\end{aligned} \tag{56}$$

To analyze the total communication complexity, we will analyze two cases: (1) $\frac{1}{K\sqrt{\eta_0}} > 1$; (2) $\frac{1}{K\sqrt{\eta_0}} \leq 1$.

(1) If $\frac{1}{K\sqrt{\eta_0}} > 1$, thus $I_s = \max(1, \frac{1}{K\sqrt{\eta_0}} \exp(\frac{c_1(s-1)}{2})) = \frac{1}{K\sqrt{\eta_0}} \exp(\frac{c_1(s-1)}{2})$ for any $s \geq 1$.

Total number of communications:

$$\begin{aligned}
&\sum_{s=1}^{S} \frac{T_s}{I_s} = \sum_{s=1}^{S} \frac{2}{L\eta_0^{1/2}} \exp\left(\frac{c_1(s-1)}{2}\right) = \frac{2}{L\eta_0^{1/2}} \frac{\exp(c_1 S/2) - 1}{\exp(c_1/2) - 1} \\
&= \tilde{O}\bigg(\max\left(\frac{(2\Delta_0/\epsilon)^{1/2}}{\mu\eta_0^{1/2}}, \frac{(S(6LB^2 + 6\sigma^2)^{1/2}}{\mu\epsilon^{1/2}} \right) \bigg) = \tilde{O}\left(\frac{\Delta_0^{1/2}}{\mu(\eta_0\epsilon)^{1/2}}, \frac{L^{1/2}}{\mu^{3/2}\epsilon^{1/2}} \right).
\end{aligned} \tag{57}$$

(2) If $\frac{1}{K\sqrt{\eta_0}} \leq 1$, thus $I_s = 1$ for $s \leq \left\lceil 2d^{-1}\log(K\sqrt{\eta_0}) + 1 \right\rceil := S_1$ and $I_s = \frac{1}{K\sqrt{\eta_0}} \exp\left(\frac{s-1}{2}\right)$ for $s > \frac{2(5+\mu/L)}{\mu/L} \log(K\sqrt{\eta_0}) + 1$.

Obviously, $S_1 \leq \frac{2(5+\mu/L)}{\mu/L} \log(K\sqrt{\eta_0}) + 2$. The number of iterations from $s = 1$ to S_1 is

$$\begin{aligned}
\sum_{s=1}^{S_1} T_s &= \sum_{s=1}^{S_1} \frac{2}{\eta_0 LK} \exp(c_1(s-1)) \\
&= \frac{2}{\eta_0 LK} \frac{\exp(c_1 S_1) - 1}{\exp(c_1) - 1} \\
&\leq c_1^{-1} \frac{2}{\eta_0 LK} \exp\left(2\log(K\sqrt{\eta_0}) + 2c\right) \\
&= c_1^{-1} \frac{2}{\eta_0 LK} K^2 \eta_0 \exp\left(\frac{2\mu/L}{5 + \mu/L}\right) \\
&\leq c_1^{-1} 2K \exp(2).
\end{aligned} \tag{58}$$

Thus, the total number of communications is

$$\begin{aligned}
&\sum_{s=1}^{S_1} T_s + \sum_{s=S_1+1}^{S} \frac{T_s}{I_s} \\
&= c_1^{-1} 2K \exp(2) + \sum_{s=S_1+1}^{S} \frac{2}{L\eta_0^{1/2}} \exp\left(\frac{s-1}{2} \frac{\mu/L}{5 + \mu/L}\right) \\
&\leq c_1^{-1} 2K \exp(2) + \sum_{s=1}^{S} \frac{2}{L\eta_0^{1/2}} \exp\left(\frac{s-1}{2} \frac{\mu/L}{5 + \mu/L}\right) \\
&\leq c_1^{-1} 2K \exp(2) + \frac{2}{L\eta_0^{1/2}} \frac{\exp(\frac{S}{2} \frac{\mu/L}{5+\mu/L}) - 1}{\exp(\frac{\mu/L}{2(5+\mu/L)}) - 1} \\
&\in O\left(\max\left(\frac{K}{\mu} + \frac{1}{\mu\eta_0^{1/2}\epsilon^{1/2}}, \frac{K}{\mu} + \frac{1}{\mu^{3/2}\epsilon^{1/2}} \right) \right).
\end{aligned} \tag{59}$$