A. Proof of Theorem 1

The proof of Theorem 1 is inspired by Sinha et al. (2018). Before we prove this theorem, we need the following two technical lemmas.

Lemma 1. *Under Assumptions 1 and 2, we have $L_S(\boldsymbol{\theta})$ is L-smooth where $L = L_{\theta x}L_{x\theta}/\mu + L_{\theta\theta}$, i.e., for any $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, it holds*

$$L_S(\boldsymbol{\theta}_1) \le L_S(\boldsymbol{\theta}_2) + \langle \nabla L_S(\boldsymbol{\theta}_2), \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\rangle + \frac{L}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2,$$
$$\|\nabla L_S(\boldsymbol{\theta}_1) - \nabla L_S(\boldsymbol{\theta}_2)\|_2 \le L\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$$

Proof. By Assumption 2, we have for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$, and $\mathbf{x}_i^*(\boldsymbol{\theta}_1), \mathbf{x}_i^*(\boldsymbol{\theta}_2)$, we have

$$\begin{aligned} f(\boldsymbol{\theta}_2, \mathbf{x}_i^*(\boldsymbol{\theta}_1)) &\le f(\boldsymbol{\theta}_2, \mathbf{x}_i^*(\boldsymbol{\theta}_2)) + \langle \nabla_{\mathbf{x}} f(\boldsymbol{\theta}_2, \mathbf{x}_i^*(\boldsymbol{\theta}_2)), \mathbf{x}_i^*(\boldsymbol{\theta}_1) - \mathbf{x}_i^*(\boldsymbol{\theta}_2)\rangle - \frac{\mu}{2}\|\mathbf{x}_i^*(\boldsymbol{\theta}_1) - \mathbf{x}_i^*(\boldsymbol{\theta}_2)\|_2^2 \\ &\le f(\boldsymbol{\theta}_2, \mathbf{x}_i^*(\boldsymbol{\theta}_2)) - \frac{\mu}{2}\|\mathbf{x}_i^*(\boldsymbol{\theta}_1) - \mathbf{x}_i^*(\boldsymbol{\theta}_2)\|_2^2, \end{aligned} \tag{6}$$

where the inequality follows from $\langle \nabla_{\mathbf{x}} f(\boldsymbol{\theta}_2, \mathbf{x}_i^*(\boldsymbol{\theta}_2)), \mathbf{x}_i^*(\boldsymbol{\theta}_1) - \mathbf{x}_i^*(\boldsymbol{\theta}_2)\rangle \le 0$. In addition, we have

$$f(\boldsymbol{\theta}_2, \mathbf{x}_i^*(\boldsymbol{\theta}_2)) \le f(\boldsymbol{\theta}_2, \mathbf{x}_i^*(\boldsymbol{\theta}_1)) + \langle \nabla_{\mathbf{x}} f(\boldsymbol{\theta}_2, \mathbf{x}_i^*(\boldsymbol{\theta}_1)), \mathbf{x}_i^*(\boldsymbol{\theta}_2) - \mathbf{x}_i^*(\boldsymbol{\theta}_1)\rangle - \frac{\mu}{2}\|\mathbf{x}_i^*(\boldsymbol{\theta}_1) - \mathbf{x}_i^*(\boldsymbol{\theta}_2)\|_2^2 \tag{7}$$

Combining (6) and (7), we obtain

$$\begin{aligned} \mu\|\mathbf{x}_i^*(\boldsymbol{\theta}_1) - \mathbf{x}_i^*(\boldsymbol{\theta}_2)\|_2^2 &\le \langle \nabla_{\mathbf{x}} f(\boldsymbol{\theta}_2, \mathbf{x}_i^*(\boldsymbol{\theta}_1)), \mathbf{x}_i^*(\boldsymbol{\theta}_2) - \mathbf{x}_i^*(\boldsymbol{\theta}_1)\rangle \\ &\le \langle \nabla_{\mathbf{x}} f(\boldsymbol{\theta}_2, \mathbf{x}_i^*(\boldsymbol{\theta}_1)) - \nabla_{\mathbf{x}} f(\boldsymbol{\theta}_1, \mathbf{x}_i^*(\boldsymbol{\theta}_1)), \mathbf{x}_i^*(\boldsymbol{\theta}_2) - \mathbf{x}_i^*(\boldsymbol{\theta}_1)\rangle \\ &\le \|\nabla_{\mathbf{x}} f(\boldsymbol{\theta}_2, \mathbf{x}_i^*(\boldsymbol{\theta}_1)) - \nabla_{\mathbf{x}} f(\boldsymbol{\theta}_1, \mathbf{x}_i^*(\boldsymbol{\theta}_1))\|_2\|\mathbf{x}_i^*(\boldsymbol{\theta}_2) - \mathbf{x}_i^*(\boldsymbol{\theta}_1)\|_2 \\ &\le L_{x\theta}\|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|_2\|\mathbf{x}_i^*(\boldsymbol{\theta}_2) - \mathbf{x}_i^*(\boldsymbol{\theta}_1)\|_2 \end{aligned} \tag{8}$$

where the second inequality holds because $\langle \nabla_{\mathbf{x}} f(\boldsymbol{\theta}_1, \mathbf{x}_i^*(\boldsymbol{\theta}_1)), \mathbf{x}_i^*(\boldsymbol{\theta}_2) - \mathbf{x}_i^*(\boldsymbol{\theta}_1)\rangle \le 0$, the third inequality follows from CauchySchwarz inequality, and the last inequality holds due to Assumption 1. (8) immediately yields

$$\|\mathbf{x}_i^*(\boldsymbol{\theta}_1) - \mathbf{x}_i^*(\boldsymbol{\theta}_2)\|_2 \le \frac{L_{x\theta}}{\mu}\|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|_2. \tag{9}$$

Then we have for $i \in [n]$,

$$\begin{aligned} \|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_1, \mathbf{x}_i^*(\boldsymbol{\theta}_1)) - \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_2, \mathbf{x}_i^*(\boldsymbol{\theta}_2))\|_2 &\le \|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_1, \mathbf{x}_i^*(\boldsymbol{\theta}_1)) - \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_1, \mathbf{x}_i^*(\boldsymbol{\theta}_2))\|_2 \\ &\quad + \|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_1, \mathbf{x}_i^*(\boldsymbol{\theta}_2)) - \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_2, \mathbf{x}_i^*(\boldsymbol{\theta}_2))\|_2 \\ &\le L_{\theta x}\|\mathbf{x}_i^*(\boldsymbol{\theta}_1) - \mathbf{x}_i^*(\boldsymbol{\theta}_2)\|_2 + L_{\theta\theta}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 \\ &= \left(\frac{L_{\theta x}L_{x\theta}}{\mu} + L_{\theta\theta}\right)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2 \end{aligned} \tag{10}$$

where the first inequality follows from triangle inequality, the second inequality holds due to Assumption 1, and the last inequality is due to (10). Finally, by the definition of $L_S(\boldsymbol{\theta})$, we have

$$\begin{aligned} \|\nabla L_S(\boldsymbol{\theta}_1) - \nabla L_S(\boldsymbol{\theta}_2)\|_2 &\le \left\|\frac{1}{n}\sum_{i=1}^n \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_1, \mathbf{x}_i^*(\boldsymbol{\theta}_1)) - \frac{1}{n}\sum_{i=1}^n \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_2, \mathbf{x}_i^*(\boldsymbol{\theta}_2))\right\|_2 \\ &\le \frac{1}{n}\sum_{i=1}^n \|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_1, \mathbf{x}_i^*(\boldsymbol{\theta}_1)) - \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_2, \mathbf{x}_i^*(\boldsymbol{\theta}_2))\|_2 \\ &\le \left(\frac{L_{\theta x}L_{x\theta}}{\mu} + L_{\theta\theta}\right)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2, \end{aligned}$$

where the last inequality follows from (10). This completes the proof. □

Lemma 2. *Under Assumptions 1 and 2, the approximate stochastic gradient* $\hat{\mathbf{g}}(\theta)$ *satisfies*

$$\|\hat{\mathbf{g}}(\boldsymbol{\theta}) - \mathbf{g}(\boldsymbol{\theta})\|_2 \le L_{\theta x}\sqrt{\frac{\delta}{\mu}}. \tag{11}$$

Proof. We have

$$\begin{aligned}\|\hat{\mathbf{g}}(\boldsymbol{\theta}) - \mathbf{g}(\boldsymbol{\theta})\|_2 &= \left\| \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}(\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, \hat{\mathbf{x}}_i(\boldsymbol{\theta})) - \nabla \bar{f}_i(\boldsymbol{\theta}))\right\|_2 \\ &\le \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\left\|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, \hat{\mathbf{x}}_i(\boldsymbol{\theta})) - \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, \mathbf{x}_i^*(\boldsymbol{\theta}))\right\|_2 \\ &\le \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} L_{\theta x}\|\hat{\mathbf{x}}_i(\boldsymbol{\theta}) - \mathbf{x}_i^*(\boldsymbol{\theta})\|_2, \end{aligned} \tag{12}$$

where the first inequality follows from triangle inequality, and the second inequality holds due to Assumption 1. By Assumption 2, we have for any $\boldsymbol{\theta}$, and $\mathbf{x}_i^*(\boldsymbol{\theta}), \hat{\mathbf{x}}_i(\boldsymbol{\theta})$, we have

$$\mu\|\mathbf{x}_i^*(\boldsymbol{\theta}) - \hat{\mathbf{x}}_i(\boldsymbol{\theta})\|_2^2 \le \langle \nabla_{\mathbf{x}} f(\boldsymbol{\theta}, \mathbf{x}_i^*(\boldsymbol{\theta})) - \nabla_{\mathbf{x}} f(\boldsymbol{\theta}, \hat{\mathbf{x}}_i(\boldsymbol{\theta})), \hat{\mathbf{x}}_i(\boldsymbol{\theta}) - \mathbf{x}_i^*(\boldsymbol{\theta})\rangle. \tag{13}$$

Since $\hat{\mathbf{x}}_i(\boldsymbol{\theta})$ is a δ-approximate maximizer of $f(\boldsymbol{\theta}, \hat{\mathbf{x}}_i(\boldsymbol{\theta}))$, we have

$$\langle \mathbf{x}_i^*(\boldsymbol{\theta}) - \hat{\mathbf{x}}_i(\boldsymbol{\theta}), \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, \hat{\mathbf{x}}_i(\boldsymbol{\theta}))\rangle \le \delta. \tag{14}$$

In addition, we have

$$\langle \hat{\mathbf{x}}_i(\boldsymbol{\theta}) - \mathbf{x}_i^*(\boldsymbol{\theta}), \nabla_{\mathbf{x}} f(\boldsymbol{\theta}, \mathbf{x}_i^*(\boldsymbol{\theta}))\rangle \le 0. \tag{15}$$

Combining (14) and (15) gives rise to

$$\langle \hat{\mathbf{x}}_i(\boldsymbol{\theta}) - \mathbf{x}_i^*(\boldsymbol{\theta}), \nabla_{\mathbf{x}} f(\boldsymbol{\theta}, \mathbf{x}_i^*(\boldsymbol{\theta})) - \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, \hat{\mathbf{x}}_i(\boldsymbol{\theta}))\rangle \le \delta. \tag{16}$$

Substitute (16) into (13), we obtain

$$\mu\|\mathbf{x}_i^*(\boldsymbol{\theta}) - \hat{\mathbf{x}}_i(\boldsymbol{\theta})\|_2^2 \le \delta,$$

which immediately yields

$$\|\mathbf{x}_i^*(\boldsymbol{\theta}) - \hat{\mathbf{x}}_i(\boldsymbol{\theta})\|_2 \le \sqrt{\frac{\delta}{\mu}}. \tag{17}$$

Substitute (17) into (12), we obtain

$$\|\hat{\mathbf{g}}(\boldsymbol{\theta}) - \mathbf{g}(\boldsymbol{\theta})\|_2 \le L_{\theta x}\sqrt{\frac{\delta}{\mu}},$$

which completes the proof. □

Now we are ready to prove Theorem 1.

Proof of Theorem 1. Let $\bar{f}(\boldsymbol{\theta}) = 1/n \sum_{i=1}^n \min_{\mathbf{x}_i} f(\boldsymbol{\theta}, \mathbf{x}_i) = 1/n \sum_{i=1}^n f(\boldsymbol{\theta}, \mathbf{x}_i^*)$. By Lemma 1, we have

$$
\begin{aligned}
L_S(\boldsymbol{\theta}^{t+1}) &\leq L_S(\boldsymbol{\theta}^t) + \langle \nabla L_S(\boldsymbol{\theta}^t), \boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t \rangle + \frac{L}{2}\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\|_2^2 \\
&= L_S(\boldsymbol{\theta}^t) - \eta_t \|\nabla L_S(\boldsymbol{\theta}^t)\|_2^2 + \frac{L\eta_t^2}{2}\|\hat{\mathbf{g}}(\boldsymbol{\theta}^t)\|_2^2 + \eta_t \langle \nabla L_S(\boldsymbol{\theta}^{t+1}), \nabla L_S(\boldsymbol{\theta}^{t+1}) - \hat{\mathbf{g}}(\boldsymbol{\theta}^t)\rangle \\
&= L_S(\boldsymbol{\theta}^t) - \eta_t\left(1 - \frac{L\eta_t}{2}\right)\|\nabla L_S(\boldsymbol{\theta}^t)\|_2^2 + \eta_t\left(1 - \frac{L\eta_t}{2}\right)\langle \nabla L_S(\boldsymbol{\theta}^t), \nabla L_S(\boldsymbol{\theta}^t) - \hat{\mathbf{g}}(\boldsymbol{\theta}^t)\rangle \\
&\quad + \frac{L\eta_t^2}{2}\|\hat{\mathbf{g}}(\boldsymbol{\theta}^t) - \nabla L_S(\boldsymbol{\theta}^t)\|_2^2 \\
&= L_S(\boldsymbol{\theta}^t) - \eta_t\left(1 - \frac{L\eta_t}{2}\right)\|\nabla L_S(\boldsymbol{\theta}^t)\|_2^2 + \eta_t\left(1 - \frac{L\eta_t}{2}\right)\langle \nabla L_S(\boldsymbol{\theta}^t), \mathbf{g}(\boldsymbol{\theta}^t) - \hat{\mathbf{g}}(\boldsymbol{\theta}^t)\rangle \\
&\quad + \eta_t\left(1 - \frac{L\eta_t}{2}\right)\langle \nabla L_S(\boldsymbol{\theta}^t), \nabla L_S(\boldsymbol{\theta}^t) - \mathbf{g}(\boldsymbol{\theta}^t)\rangle + \frac{L\eta_t^2}{2}\|\hat{\mathbf{g}}(\boldsymbol{\theta}^t) - \mathbf{g}(\boldsymbol{\theta}^t) + \mathbf{g}(\boldsymbol{\theta}^t) - \nabla L_S(\boldsymbol{\theta}^t)\|_2^2 \\
&\leq L_S(\boldsymbol{\theta}^t) - \frac{\eta_t}{2}\left(1 - \frac{L\eta_t}{2}\right)\|\nabla L_S(\boldsymbol{\theta}^t)\|_2^2 + \frac{\eta_t}{2}\left(1 - \frac{L\eta_t}{2}\right)\|\hat{\mathbf{g}}(\boldsymbol{\theta}^t) - \mathbf{g}(\boldsymbol{\theta}^t)\|_2^2 \\
&\quad + \eta_t\left(1 + \frac{L\eta_t}{2}\right)\langle \nabla L_S(\boldsymbol{\theta}^t), \nabla L_S(\boldsymbol{\theta}^t) - \mathbf{g}(\boldsymbol{\theta}^t)\rangle + L\eta_t^2\big(\|\hat{\mathbf{g}}(\boldsymbol{\theta}^t) - \mathbf{g}(\boldsymbol{\theta}^t)\|_2^2 + \|\mathbf{g}(\boldsymbol{\theta}^t) - \nabla L_S(\boldsymbol{\theta}^t)\|_2^2\big)
\end{aligned}
$$

Taking expectation on both sides of the above inequality conditioned on $\boldsymbol{\theta}^t$, we have

$$
\mathbb{E}[L_S(\boldsymbol{\theta}^{t+1}) - L_S(\boldsymbol{\theta}^t)|\boldsymbol{\theta}^t] \leq -\frac{\eta_t}{2}\left(1 - \frac{L\eta_t}{2}\right)\|\nabla L_S(\boldsymbol{\theta}^t)\|_2^2 + \frac{\eta_t}{2}\left(1 + \frac{3L\eta_t}{2}\right)\frac{L_{\theta x}^2\delta}{\mu} + L\eta_t^2\sigma^2 \tag{18}
$$

where we used the fact that $\mathbb{E}[\mathbf{g}(\boldsymbol{\theta}^t)] = \nabla L_S(\boldsymbol{\theta}^t)$, Assumption 3, and Lemma 2. Taking telescope sum of (18) over $t = 0, \ldots, T-1$, we obtain that

$$
\sum_{t=0}^{T-1}\frac{\eta_t}{2}\left(1 - \frac{L\eta_t}{2}\right)\mathbb{E}\big[\|\nabla L_S(\boldsymbol{\theta}^t)\|_2^2\big] \leq \mathbb{E}[L_S(\boldsymbol{\theta}^0) - L_S(\boldsymbol{\theta}^T)] + \sum_{t=0}^{T-1}\frac{\eta_t}{2}\left(1 + \frac{3L\eta_t}{2}\right)\frac{L_{\theta x}^2\delta}{\mu} + L\sum_{t=0}^{T-1}\eta_t^2\sigma^2
$$

Choose $\eta_t = \eta = \min(1/L, \sqrt{\Delta/TL\sigma^2})$ where $L = L_{\theta x}L_{x\theta}/\mu + L_{\theta\theta}$, we can show that

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla L_S(\boldsymbol{\theta}^t)\|_2^2\big] \leq 4\sigma\sqrt{\frac{L\Delta}{T}} + \frac{5L_{\theta x}^2\delta}{\mu}.
$$

This completes the proof. □