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 $L_S(\theta)$ is $L$-smooth where $L = L_{\theta x} / \mu + L_{\theta \theta}$, i.e., for any $\theta_1$ and $\theta_2$, it holds

$$L_S(\theta_1) \leq L_S(\theta_2) + \langle \nabla L_S(\theta_2), \theta_1 - \theta_2 \rangle + \frac{L}{2} \|	heta_1 - \theta_2\|^2,$$

$$\|
abla L_S(\theta_1) - \nabla L_S(\theta_2)\|_2 \leq L\|	heta_1 - \theta_2\|_2$$

**Proof.** By Assumption 2, we have for any $\theta_1, \theta_2$, and $x^*_i(\theta_1), x^*_i(\theta_2)$, we have

$$f(\theta_2, x^*_i(\theta_1)) \leq f(\theta_2, x^*_i(\theta_2)) + \langle \nabla_x f(\theta_2, x^*_i(\theta_2)), x^*_i(\theta_1) - x^*_i(\theta_2) \rangle - \frac{\mu}{2} \|x^*_i(\theta_1) - x^*_i(\theta_2)\|^2,$$

where the inequality follows from $\langle \nabla_x f(\theta_2, x^*_i(\theta_2)), x^*_i(\theta_1) - x^*_i(\theta_2) \rangle \leq 0$. In addition, we have

$$f(\theta_2, x^*_i(\theta_2)) \leq f(\theta_2, x^*_i(\theta_1)) + \langle \nabla_x f(\theta_2, x^*_i(\theta_1)), x^*_i(\theta_2) - x^*_i(\theta_1) \rangle - \frac{\mu}{2} \|x^*_i(\theta_1) - x^*_i(\theta_2)\|^2$$

Combining (6) and (7), we obtain

$$\mu \|x^*_i(\theta_1) - x^*_i(\theta_2)\|^2 \leq \|\nabla_x f(\theta_2, x^*_i(\theta_1)), x^*_i(\theta_2) - x^*_i(\theta_1)\|^2 - \langle \nabla_x f(\theta_2, x^*_i(\theta_2)), x^*_i(\theta_1) - x^*_i(\theta_2) \rangle\| \|x^*_i(\theta_2) - x^*_i(\theta_1)\|_2$$

where the second inequality holds because $\langle \nabla_x f(\theta_2, x^*_i(\theta_1)), x^*_i(\theta_2) - x^*_i(\theta_1) \rangle \leq 0$. The third inequality follows from CauchySchwarz inequality, and the last inequality holds due to Assumption 1. (8) immediately yields

$$\|x^*_i(\theta_1) - x^*_i(\theta_2)\|_2 \leq \frac{L_{\theta x}}{\mu} \|\theta_2 - \theta_1\|_2. \quad (9)$$

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

$$\|\nabla_{\theta} f(\theta_1, x^*_i(\theta_1)) - \nabla_{\theta} f(\theta_2, x^*_i(\theta_2))\|_2 \leq \|\nabla_{\theta} f(\theta_1, x^*_i(\theta_1)) - \nabla_{\theta} f(\theta_1, x^*_i(\theta_2))\|_2 + \|\nabla_{\theta} f(\theta_1, x^*_i(\theta_2)) - \nabla_{\theta} f(\theta_2, x^*_i(\theta_2))\|_2$$

$$\leq L_{\theta\theta} \|x^*_i(\theta_1) - x^*_i(\theta_2)\|_2 + L_{\theta \theta} \|\theta_1 - \theta_2\|_2$$

$$= \left( \frac{L_{\theta x} \mu + L_{\theta \theta}}{\mu} \right) \|\theta_1 - \theta_2\|_2 \quad (10)$$

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 $L_S(\theta)$, we have

$$\|\nabla L_S(\theta_1) - \nabla L_S(\theta_2)\|_2 \leq \left| \frac{1}{n} \sum_{i=1}^{n} \nabla_{\theta} f(\theta_1, x^*_i(\theta_1)) - \frac{1}{n} \sum_{i=1}^{n} \nabla_{\theta} f(\theta_2, x^*_i(\theta_2)) \right|_2$$

$$\leq \frac{1}{\mu} \sum_{i=1}^{n} \|\nabla_{\theta} f(\theta_1, x^*_i(\theta_1)) - \nabla_{\theta} f(\theta_2, x^*_i(\theta_2))\|_2$$

$$\leq \left( \frac{L_{\theta x} \mu + L_{\theta \theta}}{\mu} \right) \|\theta_1 - \theta_2\|_2,$$

where the last inequality follows from (10). This completes the proof. □
Lemma 2. Under Assumptions 1 and 2, the approximate stochastic gradient $\hat{g}(\theta)$ satisfies

$$\|\hat{g}(\theta) - g(\theta)\|_2 \leq L_{\theta x} \sqrt{\frac{\delta}{\mu}}.$$  \hspace{1cm} (11)

Proof. We have

$$\|\hat{g}(\theta) - g(\theta)\|_2 = \left\| \frac{1}{|B|} \sum_{i \in B} (\nabla_{\theta} f(\theta, \hat{x}_i(\theta)) - \nabla_{\theta} \hat{f}_i(\theta)) \right\|_2$$

$$\leq \frac{1}{|B|} \sum_{i \in B} \| \nabla_{\theta} f(\theta, \hat{x}_i(\theta)) - \nabla_{\theta} f(\theta, x_i^*(\theta)) \|_2$$

$$\leq \frac{1}{|B|} \sum_{i \in B} L_{\theta x} \| \hat{x}_i(\theta) - x_i^*(\theta) \|_2,$$  \hspace{1cm} (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 $\theta$, and $x_i^*(\theta), \hat{x}_i(\theta)$, we have

$$\mu \| x_i^*(\theta) - \hat{x}_i(\theta) \|_2^2 \leq \langle \nabla_x f(\theta, x_i^*(\theta)) - \nabla_x f(\theta, \hat{x}_i(\theta)), \hat{x}_i(\theta) - x_i^*(\theta) \rangle.$$  \hspace{1cm} (13)

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

$$\langle x_i^*(\theta) - \hat{x}_i(\theta), \nabla_{\theta} f(\theta, \hat{x}_i(\theta)) \rangle \leq \delta.$$  \hspace{1cm} (14)

In addition, we have

$$\langle \hat{x}_i(\theta) - x_i^*(\theta), \nabla_x f(\theta, x_i^*(\theta)) \rangle \leq 0.$$  \hspace{1cm} (15)

Combining (14) and (15) gives rise to

$$\langle \hat{x}_i(\theta) - x_i^*(\theta), \nabla_x f(\theta, x_i^*(\theta)) - \nabla_{\theta} f(\theta, \hat{x}_i(\theta)) \rangle \leq \delta.$$  \hspace{1cm} (16)

Substitute (16) into (13), we obtain

$$\mu \| x_i^*(\theta) - \hat{x}_i(\theta) \|_2^2 \leq \delta,$$

which immediately yields

$$\| x_i^*(\theta) - \hat{x}_i(\theta) \|_2 \leq \sqrt{\frac{\delta}{\mu}}.$$  \hspace{1cm} (17)

Substitute (17) into (12), we obtain

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

which completes the proof. \hfill \Box

Now we are ready to prove Theorem 1.
Proof of Theorem 1. Let $\bar{f}(\theta) = 1/n \sum_{i=1}^{n} \min_{x_i} f(\theta, x_i)$, and $f(\theta, x_i) = 1/n \sum_{i=1}^{n} f(\theta, x_i)$. By Lemma 1, we have

$$L_S(\theta^{t+1}) \leq L_S(\theta^t) + \langle \nabla L_S(\theta^t), \theta^{t+1} - \theta^t \rangle + \frac{L}{2} \| \theta^{t+1} - \theta^t \|^2$$

$$= L_S(\theta^t) - \eta_t \| \nabla L_S(\theta^t) \|^2_2 + \frac{L \eta_t^2}{2} \| \tilde{g}(\theta^t) \|^2_2 + \eta_t \langle \nabla L_S(\theta^{t+1}), \nabla L_S(\theta^{t+1}) - \tilde{g}(\theta^t) \rangle$$

$$= L_S(\theta^t) - \eta_t \left(1 - \frac{L \eta_t}{2}\right) \| \nabla L_S(\theta^t) \|^2_2 + \eta_t \left(1 - \frac{L \eta_t}{2}\right) \langle \nabla L_S(\theta^t), \nabla L_S(\theta^t) - \tilde{g}(\theta^t) \rangle$$

$$+ \frac{L \eta_t^2}{2} \| \tilde{g}(\theta^t) \|^2_2$$

$$\leq L_S(\theta^t) - \eta_t \left(1 - \frac{L \eta_t}{2}\right) \| \nabla L_S(\theta^t) \|^2_2 + \eta_t \left(1 - \frac{L \eta_t}{2}\right) \| \tilde{g}(\theta^t) \|^2_2 + \frac{L \eta_t^2}{2} \| \tilde{g}(\theta^t) - g(\theta^t) \|^2_2 + \| g(\theta^t) - \nabla L_S(\theta^t) \|^2_2$$

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

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

(18)

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

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

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

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

This completes the proof. \qed