A. Proof of Technical Lemmas

In this section, we provide complete proofs for the lemmas in Section 3 and Section 4.

A.1. Proof of Lemma 3.6

We provide a proof for an expanded version of Lemma 3.6.

Lemma A.1 *If f is ℓ-smooth and $\mathcal{Y}$ is bounded, we have*

1. *$\Phi_{1/2\ell}(\mathbf{x})$ and $\mathrm{prox}_{\Phi/2\ell}(\mathbf{x})$ are well-defined for $\forall \mathbf{x} \in \mathbb{R}^m$.*
2. *$\Phi(\mathrm{prox}_{\Phi/2\ell}(\mathbf{x})) \le \Phi(\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^m$.*
3. *$\Phi_{1/2\ell}$ is ℓ-smooth with $\nabla \Phi_{1/2\ell}(\mathbf{x}) = 2\ell(\mathbf{x} - \mathrm{prox}_{\Phi/2\ell}(\mathbf{x}))$.*
4. *$\Phi_{1/2\ell}(\mathbf{x}') - \Phi_{1/2\ell}(\mathbf{x}) - (\mathbf{x}' - \mathbf{x})^\top \nabla \Phi_{1/2\ell}(\mathbf{x}) \le (\ell/2)\|\mathbf{x}' - \mathbf{x}\|^2$ for any $\mathbf{x}', \mathbf{x} \in \mathbb{R}^m$.*

Proof. By the definition of Φ, we have

$$\Psi(\mathbf{x}) \doteq \Phi(\mathbf{x}) + \frac{\ell\|\mathbf{x}\|^2}{2} = \max_{\mathbf{y}\in\mathcal{Y}} \{f(\mathbf{x},\mathbf{y}) + \frac{\ell\|\mathbf{x}\|^2}{2}\}.$$

Since f is ℓ-smooth, $f(\mathbf{x},\mathbf{y}) + (\ell/2)\|\mathbf{x}\|^2$ is convex in $\mathbf{x}$ for any $\mathbf{y} \in \mathcal{Y}$. Since $\mathcal{Y}$ is bounded, Danskin's theorem (Rockafellar, 2015) implies that $\Psi(\mathbf{x})$ is convex. Putting these pieces together yields that $\Phi(\mathbf{w}) + \ell\|\mathbf{w} - \mathbf{x}\|^2$ is $(\ell/2)$-strongly convex. This implies that $\Phi_{1/2\ell}(\mathbf{x})$ and $\mathrm{prox}_{\Phi/2\ell}(\mathbf{x})$ are well-defined. Furthermore, by the definition of $\mathrm{prox}_{\Phi/2\ell}(\mathbf{x})$, we have

$$\Phi(\mathrm{prox}_{\Phi/2\ell}(\mathbf{x})) \le \Phi_{1/2\ell}(\mathrm{prox}_{\Phi/2\ell}(\mathbf{x})) \le \Phi(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^m.$$

Moreover, Davis & Drusvyatskiy (2019, Lemma 2.2) implies that $\Phi_{1/2\ell}$ is ℓ-smooth with

$$\nabla \Phi_{1/2\ell}(\mathbf{x}) = 2\ell(\mathbf{x} - \mathrm{prox}_{\Phi/2\ell}(\mathbf{x})).$$

Finally, it follows from Nesterov (2013, Theorem 2.1.5) that $\Phi_{1/2\ell}$ satisfies the last inequality. □

A.2. Proof of Lemma 3.8

Denote $\hat{\mathbf{x}} := \mathrm{prox}_{\Phi/2\ell}(\mathbf{x})$, we have $\nabla \Phi_{1/2\ell}(\mathbf{x}) = 2\ell(\mathbf{x} - \hat{\mathbf{x}})$ (cf. Lemma 3.6) and hence $\|\hat{\mathbf{x}} - \mathbf{x}\| = \|\nabla \Phi_{1/2\ell}(\mathbf{x})\|/2\ell$. Furthermore, the optimality condition for $\mathrm{prox}_{\Phi/2\ell}(\mathbf{x})$ implies that $2\ell(\mathbf{x} - \hat{\mathbf{x}}) \in \partial\Phi(\hat{\mathbf{x}})$. Putting these pieces together yields that $\min_{\xi \in \partial\Phi(\hat{\mathbf{x}})} \|\xi\| \le \|\nabla \Phi_{1/2\ell}(\mathbf{x})\|$.

A.3. Proof of Lemma 4.3

Since $f(\mathbf{x},\mathbf{y})$ is strongly concave in $\mathbf{y}$ for each $\mathbf{x} \in \mathbb{R}^m$, a function $\mathbf{y}^\star(\cdot)$ is unique and well-defined. Then we claim that $\mathbf{y}^\star(\cdot)$ is κ-Lipschitz. Indeed, let $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^m$, the optimality of $\mathbf{y}^\star(\mathbf{x}_1)$ and $\mathbf{y}^\star(\mathbf{x}_2)$ implies that

$$(\mathbf{y} - \mathbf{y}^\star(\mathbf{x}_1))^\top \nabla_{\mathbf{y}} f(\mathbf{x}_1, \mathbf{y}^\star(\mathbf{x}_1)) \le 0, \qquad \forall \mathbf{y} \in \mathcal{Y}, \tag{4}$$

$$(\mathbf{y} - \mathbf{y}^\star(\mathbf{x}_2))^\top \nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y}^\star(\mathbf{x}_2)) \le 0, \qquad \forall \mathbf{y} \in \mathcal{Y}. \tag{5}$$

Letting $\mathbf{y} = \mathbf{y}^\star(\mathbf{x}_2)$ in (4) and $\mathbf{y} = \mathbf{y}^\star(\mathbf{x}_1)$ in (5) and summing the resulting two inequalities yields

$$(\mathbf{y}^\star(\mathbf{x}_2) - \mathbf{y}^\star(\mathbf{x}_1))^\top (\nabla_{\mathbf{y}} f(\mathbf{x}_1, \mathbf{y}^\star(\mathbf{x}_1)) - \nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y}^\star(\mathbf{x}_2))) \le 0. \tag{6}$$

Recall that $f(\mathbf{x}_1, \cdot)$ is μ-strongly concave, we have

$$(\mathbf{y}^\star(\mathbf{x}_2) - \mathbf{y}^\star(\mathbf{x}_1))^\top (\nabla_{\mathbf{y}} f(\mathbf{x}_1, \mathbf{y}^\star(\mathbf{x}_2)) - \nabla_{\mathbf{y}} f(\mathbf{x}_1, \mathbf{y}^\star(\mathbf{x}_1))) + \mu\|\mathbf{y}^\star(\mathbf{x}_2) - \mathbf{y}^\star(\mathbf{x}_1)\|^2 \le 0. \tag{7}$$

Then we conclude the desired result by combining (6) and (7) with ℓ-smoothness of f, i.e.,

$$\begin{aligned}\mu \left\|\mathbf{y}^\star(\mathbf{x}_2) - \mathbf{y}^\star(\mathbf{x}_1)\right\|^2 &\leq \left(\mathbf{y}^\star(\mathbf{x}_2) - \mathbf{y}^\star(\mathbf{x}_1)\right)^\top \left(\nabla_{\mathbf{y}} f(\mathbf{x}_2, \mathbf{y}^\star(\mathbf{x}_2)) - \nabla_{\mathbf{y}} f(\mathbf{x}_1, \mathbf{y}^\star(\mathbf{x}_2))\right) \\ &\leq \ell \left\|\mathbf{y}^\star(\mathbf{x}_2) - \mathbf{y}^\star(\mathbf{x}_1)\right\| \left\|\mathbf{x}_2 - \mathbf{x}_1\right\|.\end{aligned}$$

Since $\mathbf{y}^\star(\mathbf{x})$ is unique and $\mathcal{Y}$ is convex and bounded, we conclude from Danskin's theorem (Rockafellar, 2015) that Φ is differentiable with $\nabla\Phi(\mathbf{x}) = \nabla_{\mathbf{x}} f\left(\mathbf{x}, \mathbf{y}^\star(\mathbf{x})\right)$. Since $\nabla\Phi(\mathbf{x}) = \nabla_{\mathbf{x}} f\left(\mathbf{x}, \mathbf{y}^\star(\mathbf{x})\right)$, we have

$$\left\|\nabla\Phi(\mathbf{x}) - \nabla\Phi(\mathbf{x}')\right\| \;=\; \left\|\nabla_{\mathbf{x}} f\left(\mathbf{x}, \mathbf{y}^\star(\mathbf{x})\right) - \nabla_{\mathbf{x}} f\left(\mathbf{x}', \mathbf{y}^\star(\mathbf{x}')\right)\right\| \;\leq\; \ell\left(\left\|\mathbf{x} - \mathbf{x}'\right\| + \left\|\mathbf{y}^\star(\mathbf{x}) - \mathbf{y}^\star(\mathbf{x}')\right\|\right).$$

Since $\mathbf{y}^\star(\cdot)$ is κ-Lipschitz, we conclude the desired result by plugging $\left\|\mathbf{y}^\star(\mathbf{x}) - \mathbf{y}^\star(\mathbf{x}')\right\| \leq \kappa$. Since $\kappa \geq 1$, Φ is $2\kappa\ell$-smooth. The last inequality follows from Nesterov (2013, Theorem 2.1.5).

A.4. Proof of Lemma 4.7

By the proof in Lemma A.1, Φ is ℓ-weakly convex and $\partial\Phi(\mathbf{x}) = \partial\Psi(\mathbf{x}) - \ell\mathbf{x}$ where $\Psi(\mathbf{x}) = \max_{\mathbf{y}\in\mathcal{Y}}\{f(\mathbf{x}, \mathbf{y}) + (\ell/2)\|\mathbf{x}\|^2\}$. Since $f(\mathbf{x}, \mathbf{y}) + (\ell/2)\|\mathbf{x}\|^2$ is convex in $\mathbf{x}$ for each $\mathbf{y} \in \mathcal{Y}$ and $\mathcal{Y}$ is bounded, Danskin's theorem implies that $\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}^\star(\mathbf{x})) + \ell\mathbf{x} \in \partial\Psi(\mathbf{x})$. Putting these pieces together yields that $\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}^\star(\mathbf{x})) \in \partial\Phi(\mathbf{x})$.

A.5. Proof of Lemma on Stochastic Gradient

The following lemma establishes some properties of the stochastic gradients sampled at each iteration.

Lemma A.2 *$\frac{1}{M}\sum_{i=1}^M G_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t, \xi_i)$ and $\frac{1}{M}\sum_{i=1}^M G_{\mathbf{y}}(\mathbf{x}_t, \mathbf{y}_t, \xi_i)$ are unbiased and have bounded variance,*

$$\begin{aligned}&\mathbb{E}\left[\tfrac{1}{M}\textstyle\sum_{i=1}^M G_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t, \xi_i)\right] = \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \quad \mathbb{E}\left[\left\|\tfrac{1}{M}\textstyle\sum_{i=1}^M G_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t, \xi_i)\right\|^2\right] \leq \left\|\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\right\|^2 + \tfrac{\sigma^2}{M}, \\ &\mathbb{E}\left[\tfrac{1}{M}\textstyle\sum_{i=1}^M G_{\mathbf{y}}(\mathbf{x}_t, \mathbf{y}_t, \xi_i)\right] = \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t), \quad \mathbb{E}\left[\left\|\tfrac{1}{M}\textstyle\sum_{i=1}^M G_{\mathbf{y}}(\mathbf{x}_t, \mathbf{y}_t, \xi_i)\right\|^2\right] \leq \left\|\nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t)\right\|^2 + \tfrac{\sigma^2}{M}.\end{aligned}$$

Proof. Since $G = (G_{\mathbf{x}}, G_{\mathbf{y}})$ is unbiased, we have

$$\mathbb{E}\left[\frac{1}{M}\sum_{i=1}^M G_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t, \xi_i)\right] \;=\; \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \quad \mathbb{E}\left[\frac{1}{M}\sum_{i=1}^M G_{\mathbf{y}}(\mathbf{x}_t, \mathbf{y}_t, \xi_i)\right] \;=\; \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t).$$

Furthermore, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^M G_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t, \xi_i) - \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\right\|^2\right] &= \frac{\sum_{i=1}^M \mathbb{E}\left[\left\|G_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t, \xi_i) - \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\right\|^2\right]}{M^2} \leq \frac{\sigma^2}{M}, \\ \mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^M G_{\mathbf{y}}(\mathbf{x}_t, \mathbf{y}_t, \xi_i) - \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t)\right\|^2\right] &= \frac{\sum_{i=1}^M \mathbb{E}\left[\left\|G_{\mathbf{y}}(\mathbf{x}_t, \mathbf{y}_t, \xi_i) - \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t)\right\|^2\right]}{M^2} \leq \frac{\sigma^2}{M}.\end{aligned}$$

Putting these pieces together yields the desired result. □

B. Proof for Propositions 4.11 and 4.12

In this section, we provide the detailed proof of Propositions 4.11 and 4.12.

Proof of Proposition 4.11: Assume that a point $\hat{\mathbf{x}}$ satisfies that $\|\nabla\Phi(\hat{\mathbf{x}})\| \leq \epsilon$, the optimization problem $\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y})$ is strongly concave (cf. Assumption 4.2) and $\mathbf{y}^\star(\hat{\mathbf{x}})$ is uniquely defined. We apply gradient descent for solving such problem and obtain a point $\mathbf{y}' \in \mathcal{Y}$ satisfying that

$$\left\|\mathcal{P}_{\mathcal{Y}}(\mathbf{y}' + (1/\ell)\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \mathbf{y}')) - \mathbf{y}'\right\| \leq \epsilon/\ell, \quad \left\|\mathbf{y}' - \mathbf{y}^\star(\hat{\mathbf{x}})\right\| \leq \epsilon.$$

If $\|\nabla\Phi(\hat{\mathbf{x}})\| \leq \epsilon$, we have

$$\|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \mathbf{y}')\| \;\leq\; \|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \mathbf{y}') - \nabla\Phi(\hat{\mathbf{x}})\| + \|\nabla\Phi(\hat{\mathbf{x}})\| \;=\; \|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \mathbf{y}') - \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \mathbf{y}^\star(\hat{\mathbf{x}}))\| + \epsilon.$$

Since $f(\cdot,\cdot)$ is ℓ-smooth, we have

$$\|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \mathbf{y}')\| \;\leq\; \ell\|\mathbf{y}' - \mathbf{y}^\star(\hat{\mathbf{x}})\|\| + \epsilon \;=\; O(\epsilon).$$

The required number of gradient evaluations is $\mathcal{O}(\kappa\log(1/\epsilon))$. This argument holds for applying stochastic gradient with proper stepsize and the required number of stochastic gradient evaluations is $\mathcal{O}(1/\epsilon^2)$.

Conversely, if a point $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ satisfies $\|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})\| \leq \epsilon/\kappa$ and $\|\mathcal{P}_{\mathcal{Y}}(\hat{\mathbf{y}} + (1/\ell)\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})) - \hat{\mathbf{y}}\| \leq \epsilon/\kappa\ell$, then we have

$$\|\nabla\Phi(\hat{\mathbf{x}})\| \;\leq\; \|\nabla\Phi(\hat{\mathbf{x}}) - \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})\| + \|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})\| \;\leq\; \ell\|\hat{\mathbf{y}} - \mathbf{y}^\star(\hat{\mathbf{x}})\| + \epsilon/\kappa.$$

Since $f(\hat{\mathbf{x}}, \cdot)$ is μ-strongly-concave over $\mathcal{Y}$, the global error bound condition holds (Drusvyatskiy & Lewis, 2018) and

$$\mu\|\hat{\mathbf{y}} - \mathbf{y}^\star(\hat{\mathbf{x}})\| \;\leq\; \ell\|\mathcal{P}_{\mathcal{Y}}(\hat{\mathbf{y}} + (1/\ell)\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})) - \hat{\mathbf{y}}\| \;\leq\; \epsilon/\kappa.$$

Therefore, we conclude that

$$\|\nabla\Phi(\hat{\mathbf{x}})\| \;\leq\; \epsilon + \epsilon/\kappa \;=\; O(\epsilon).$$

This completes the proof.

B.1. Proof of Proposition 4.12

Assume that a point $\hat{\mathbf{x}}$ satisfies that $\|\nabla\Phi_{1/2\ell}(\hat{\mathbf{x}})\| \leq \epsilon$, the minimax optimization problem $\min_{\mathbf{x}\in\mathbb{R}^m}\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y}) + \ell\|\mathbf{x} - \hat{\mathbf{x}}\|^2$ is strongly-convex-concave (cf. Assumption 4.6) and $\mathbf{x}^\star(\hat{\mathbf{x}}) = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^m}\Phi(\mathbf{x}) + \ell\|\mathbf{x} - \hat{\mathbf{x}}\|^2$ is uniquely defined. We apply extragradient algorithm for solving such problem and obtain a point $(\mathbf{x}', \mathbf{y}')$ satisfying that

$$\|\nabla_{\mathbf{x}} f(\mathbf{x}', \mathbf{y}') + 2\ell(\mathbf{x}' - \hat{\mathbf{x}})\| \leq \epsilon, \quad \|\mathcal{P}_{\mathcal{Y}}(\mathbf{y}' + (1/\ell)\nabla_{\mathbf{y}} f(\mathbf{x}', \mathbf{y}')) - \mathbf{y}'\| \leq \epsilon/\ell, \quad \|\mathbf{x}' - \mathbf{x}^\star(\hat{\mathbf{x}})\| \leq \epsilon.$$

Since $\|\nabla\Phi_{1/2\ell}(\hat{\mathbf{x}})\| \leq \epsilon$, we have

$$\begin{aligned}\|\nabla_{\mathbf{x}} f(\mathbf{x}', \mathbf{y}')\| \;&\leq\; \|\nabla_{\mathbf{x}} f(\mathbf{x}', \mathbf{y}') + 2\ell(\mathbf{x}' - \hat{\mathbf{x}})\| + 2\ell\|\mathbf{x}' - \hat{\mathbf{x}}\| \;=\; \epsilon + 2\ell\|\mathbf{x}' - \mathbf{x}^\star(\hat{\mathbf{x}})\| + 2\ell\|\mathbf{x}^\star(\hat{\mathbf{x}}) - \hat{\mathbf{x}}\| \\ &\leq\; (2\ell+1)\epsilon + \|\nabla\Phi_{1/2\ell}(\hat{\mathbf{x}})\| \;=\; O(\epsilon).\end{aligned}$$

The required number of gradient evaluations is $O(\epsilon^{-2})$ (Mokhtari et al., 2019a). This argument holds for applying stochastic mirror-prox algorithm and the required number of stochastic gradient evaluations is $O(\epsilon^{-4})$ (Juditsky et al., 2011).

Conversely, we have $\|\nabla\Phi_{1/2\ell}(\hat{\mathbf{x}})\|^2 = 4\ell^2\|\hat{\mathbf{x}} - \mathbf{x}^\star(\hat{\mathbf{x}})\|^2$. Since $\Phi(\cdot) + \ell\|\cdot - \hat{\mathbf{x}}\|^2$ is $\ell/2$-strongly-convex, we have

$$\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}^\star(\hat{\mathbf{x}}), \mathbf{y}) - \ell\|\mathbf{x}^\star(\hat{\mathbf{x}}) - \hat{\mathbf{x}}\|^2 \;=\; \Phi(\hat{\mathbf{x}}) - \Phi(\mathbf{x}^\star(\hat{\mathbf{x}})) - \ell\|\mathbf{x}^\star(\hat{\mathbf{x}}) - \hat{\mathbf{x}}\|^2 \;\geq\; \frac{\ell\|\hat{\mathbf{x}} - \mathbf{x}^\star(\hat{\mathbf{x}})\|^2}{4}.$$

If a point $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ satisfies $\|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})\| \leq \epsilon$ and $\|\mathcal{P}_{\mathcal{Y}}(\hat{\mathbf{y}} + (1/\ell)\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})) - \hat{\mathbf{y}}\| \leq \epsilon^2/\ell$, we have

$$\begin{aligned}&\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}^\star(\hat{\mathbf{x}}), \mathbf{y}) - \ell\|\mathbf{x}^\star(\hat{\mathbf{x}}) - \hat{\mathbf{x}}\|^2 \\ \leq\;& \ell D\|\mathcal{P}_{\mathcal{Y}}(\hat{\mathbf{y}} + (1/\ell)\nabla_{\mathbf{y}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})) - \hat{\mathbf{y}}\| + \|\hat{\mathbf{x}} - \mathbf{x}^\star(\hat{\mathbf{x}})\|\|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})\| - \frac{\ell\|\hat{\mathbf{x}} - \mathbf{x}^\star(\hat{\mathbf{x}})\|^2}{4} \\ \leq\;& \epsilon^2 D + \frac{\|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}, \hat{\mathbf{y}})\|^2}{\ell} \;=\; O(\epsilon^2).\end{aligned}$$

Putting these pieces together yields that $\|\nabla\Phi_{1/2\ell}(\hat{\mathbf{x}})\| = O(\epsilon)$. This completes the proof.

C. Proof of Theorems in Section 4.1

In this section, we first specify the choice of parameters in Theorems 4.4 and 4.5. Then we present the proof of the main theorems in Section 4.1 with several technical lemmas. Note first that the case of $\ell D \lesssim \epsilon$ is trivial. Indeed, this means that the set $\mathcal{Y}$ is sufficiently small such that a single gradient ascent step is enough for approaching the ϵ-neighborhood of the optimal solution. In this case, the nonconvex-strongly-concave minimax problem reduces to a nonconvex smooth minimization problem, which has been studied extensively in the existing literature.

C.1. Choice of Parameters in Theorem 4.4 and 4.5

In this subsection, we present the full version of Theorems 4.4 and 4.5 with the detailed choice of $\eta_{\mathbf{x}}$, $\eta_{\mathbf{y}}$ and M which are important to subsequent analysis.

Theorem C.1 *Under Assumption 4.2 and letting the step sizes $\eta_{\mathbf{x}} > 0$ and $\eta_{\mathbf{y}} > 0$ be chosen as $\eta_{\mathbf{x}} = 1/[16(\kappa+1)^2\ell]$ and $\eta_{\mathbf{y}} = 1/\ell$, the iteration complexity of Algorithm 1 to return an ϵ-stationary point is bounded by*

$$O\left(\frac{\kappa^2\ell\Delta_\Phi + \kappa\ell^2 D^2}{\epsilon^2}\right),$$

which is also the total gradient complexity of the algorithm.

Theorem C.2 *Under Assumptions 4.1 and 4.2 and letting the step sizes $\eta_{\mathbf{x}} > 0$ and $\eta_{\mathbf{y}} > 0$ be the same in Theorem 4.4 with the batch size $M = \max\{1, 26\kappa\sigma^2\epsilon^{-2}\}$, the number of iterations required by Algorithm 2 to return an ϵ-stationary point is bounded by $O((\kappa^2\ell\Delta_\Phi + \kappa\ell^2 D^2)\epsilon^{-2})$ which gives the total gradient complexity of the algorithm:*

$$O\left(\frac{\kappa^2\ell\Delta_\Phi + \kappa\ell^2 D^2}{\epsilon^2}\max\left\{1, \frac{\kappa\sigma^2}{\epsilon^2}\right\}\right).$$

C.2. Proof of Technical Lemmas

In this subsection, we present three key lemmas which are important for the subsequent analysis.

Lemma C.3 *For two-time-scale GDA, the iterates $\{\mathbf{x}_t\}_{t\geq 1}$ satisfies the following inequality,*

$$\Phi(\mathbf{x}_t) \leq \Phi(\mathbf{x}_{t-1}) - \left(\frac{\eta_{\mathbf{x}}}{2} - 2\eta_{\mathbf{x}}^2\kappa\ell\right)\|\nabla\Phi(\mathbf{x}_{t-1})\|^2 + \left(\frac{\eta_{\mathbf{x}}}{2} + 2\eta_{\mathbf{x}}^2\kappa\ell\right)\|\nabla\Phi(\mathbf{x}_{t-1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2.$$

For two-time-scale SGDA, the iterates $\{\mathbf{x}_t\}_{t\geq 1}$ satisfy the following inequality:

$$\begin{aligned}\mathbb{E}\left[\Phi(\mathbf{x}_t)\right] \leq\ & \mathbb{E}\left[\Phi(\mathbf{x}_{t-1})\right] - \left(\frac{\eta_{\mathbf{x}}}{2} - 2\eta_{\mathbf{x}}^2\kappa\ell\right)\mathbb{E}\left[\|\nabla\Phi(\mathbf{x}_{t-1})\|^2\right] \\ & + \left(\frac{\eta_{\mathbf{x}}}{2} + 2\eta_{\mathbf{x}}^2\kappa\ell\right)\mathbb{E}\left[\|\nabla\Phi(\mathbf{x}_{t-1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2\right] + \frac{\eta_{\mathbf{x}}^2\kappa\ell\sigma^2}{M}.\end{aligned}$$

Proof. We first consider the deterministic setting. Since Φ is $(\ell + \kappa\ell)$-smooth, we have

$$\Phi(\mathbf{x}_t) - \Phi(\mathbf{x}_{t-1}) - (\mathbf{x}_t - \mathbf{x}_{t-1})^\top\nabla\Phi(\mathbf{x}_{t-1}) \leq \kappa\ell\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2. \tag{8}$$

Plugging $\mathbf{x}_t - \mathbf{x}_{t-1} = -\eta_{\mathbf{x}}\nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})$ into (8) yields that

$$\begin{aligned}\Phi(\mathbf{x}_t) \leq\ & \Phi(\mathbf{x}_{t-1}) - \eta_{\mathbf{x}}\|\nabla\Phi(\mathbf{x}_{t-1})\|^2 + \eta_{\mathbf{x}}^2\kappa\ell\|\nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2 \\ & + \eta_{\mathbf{x}}\left(\nabla\Phi(\mathbf{x}_{t-1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\right)^\top\nabla\Phi(\mathbf{x}_{t-1}).\end{aligned} \tag{9}$$

By Young's inequality, we have

$$\begin{aligned}& \left(\nabla\Phi(\mathbf{x}_{t-1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\right)^\top\nabla\Phi(\mathbf{x}_{t-1}) \\ \leq\ & \frac{\|\nabla\Phi(\mathbf{x}_{t-1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2 + \|\nabla\Phi(\mathbf{x}_{t-1})\|^2}{2}.\end{aligned} \tag{10}$$

By the Cauchy-Schwartz inequality, we have

$$\|\nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2 \leq 2\left(\|\nabla\Phi(\mathbf{x}_{t-1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2 + \|\nabla\Phi(\mathbf{x}_{t-1})\|^2\right). \tag{11}$$

Plugging (10) and (11) into (9) yields the first desired inequality.

We proceed to the stochastic setting. Plugging $\mathbf{x}_t - \mathbf{x}_{t-1} = -\eta_{\mathbf{x}} \left(\frac{1}{M} \sum_{i=1}^{M} G_{\mathbf{x}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi_i) \right)$ into (8) yields that

$$\begin{aligned} \Phi(\mathbf{x}_t) \;\leq\; & \Phi(\mathbf{x}_{t-1}) - \eta_{\mathbf{x}} \left\| \nabla \Phi(\mathbf{x}_{t-1}) \right\|^2 + \eta_{\mathbf{x}}^2 \kappa \ell \left\| \frac{1}{M} \sum_{i=1}^{M} G_{\mathbf{x}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi_i) \right\|^2 \\ & + \eta_{\mathbf{x}} \left(\nabla \Phi(\mathbf{x}_{t-1}) - \left(\frac{1}{M} \sum_{i=1}^{M} G_{\mathbf{x}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi_i) \right) \right)^{\top} \nabla \Phi(\mathbf{x}_t). \end{aligned}$$

Taking an expectation on both sides, conditioned on $(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})$, yields that

$$\begin{aligned} \mathbb{E}\left[\Phi(\mathbf{x}_t) \mid \mathbf{x}_{t-1}, \mathbf{y}_{t-1} \right] \;\leq\; & \Phi(\mathbf{x}_{t-1}) - \eta_{\mathbf{x}} \left\| \nabla \Phi(\mathbf{x}_{t-1}) \right\|^2 + \eta_{\mathbf{x}}^2 \kappa \ell \left\| \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \right\|^2 \\ & + \eta_{\mathbf{x}} \left(\nabla \Phi(\mathbf{x}_{t-1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \right)^{\top} \nabla \Phi(\mathbf{x}_{t-1}) + \eta_{\mathbf{x}}^2 \kappa \ell \left\| \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \right\|^2 \\ & + \eta_{\mathbf{x}}^2 \kappa \ell \mathbb{E} \left[\left\| \frac{1}{M} \sum_{i=1}^{M} G_{\mathbf{x}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi_i) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \right\|^2 \mid \mathbf{x}_{t-1}, \mathbf{y}_{t-1} \right]. \end{aligned} \tag{12}$$

Plugging (10) and (11) into (12) and taking the expectation of both sides yields the second desired inequality. □

Lemma C.4 *For two-time-scale GDA, let $\delta_t = \|\mathbf{y}^\star(\mathbf{x}_t) - \mathbf{y}_t\|^2$, the following statement holds true,*

$$\delta_t \;\leq\; \left(1 - \frac{1}{2\kappa} + 4\kappa^3 \ell^2 \eta_{\mathbf{x}}^2 \right) \delta_{t-1} + 4\kappa^3 \eta_{\mathbf{x}}^2 \left\| \nabla \Phi(\mathbf{x}_{t-1}) \right\|^2.$$

For two-time-scale SGDA, let $\delta_t = \mathbb{E}[\|\mathbf{y}^\star(\mathbf{x}_t) - \mathbf{y}_t\|^2]$, the following statement holds true,

$$\delta_t \;\leq\; \left(1 - \frac{1}{2\kappa} + 4\kappa^3 \ell^2 \eta_{\mathbf{x}}^2 \right) \delta_{t-1} + 4\kappa^3 \eta_{\mathbf{x}}^2 \mathbb{E}\left[\left\| \nabla \Phi(\mathbf{x}_{t-1}) \right\|^2 \right] + \frac{2\sigma^2 \kappa^3 \eta_{\mathbf{x}}^2}{M} + \frac{\sigma^2}{\ell^2 M}.$$

Proof. We first prove the deterministic setting. Since $f(\mathbf{x}_t, \cdot)$ is μ-strongly concave and $\eta_{\mathbf{y}} = 1/\ell$, we have

$$\left\| \mathbf{y}^\star(\mathbf{x}_{t-1}) - \mathbf{y}_t \right\|^2 \;\leq\; \left(1 - \frac{1}{\kappa} \right) \delta_{t-1}. \tag{13}$$

By Young's inequality, we have

$$\begin{aligned} \delta_t \;\leq\; & \left(1 + \frac{1}{2(\kappa - 1)} \right) \left\| \mathbf{y}^\star(\mathbf{x}_{t-1}) - \mathbf{y}_t \right\|^2 + (1 + 2(\kappa - 1)) \left\| \mathbf{y}^\star(\mathbf{x}_t) - \mathbf{y}^\star(\mathbf{x}_{t-1}) \right\|^2 \\ \leq\; & \left(\frac{2\kappa - 1}{2\kappa - 2} \right) \left\| \mathbf{y}^\star(\mathbf{x}_{t-1}) - \mathbf{y}_t \right\|^2 + 2\kappa \left\| \mathbf{y}^\star(\mathbf{x}_t) - \mathbf{y}^\star(\mathbf{x}_{t-1}) \right\|^2 \\ \overset{(13)}{\leq}\; & \left(1 - \frac{1}{2\kappa} \right) \delta_{t-1} + 2\kappa \left\| \mathbf{y}^\star(\mathbf{x}_t) - \mathbf{y}^\star(\mathbf{x}_{t-1}) \right\|^2. \end{aligned}$$

Since $\mathbf{y}^\star(\cdot)$ is κ-Lipschitz, $\|\mathbf{y}^\star(\mathbf{x}_t) - \mathbf{y}^\star(\mathbf{x}_{t-1})\| \leq \kappa \|\mathbf{x}_t - \mathbf{x}_{t-1}\|$. Furthermore, we have

$$\left\| \mathbf{x}_t - \mathbf{x}_{t-1} \right\|^2 \;=\; \eta_{\mathbf{x}}^2 \|\nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2 \;\leq\; 2\eta_{\mathbf{x}}^2 \ell^2 \delta_{t-1} + 2\eta_{\mathbf{x}}^2 \|\nabla \Phi(\mathbf{x}_{t-1})\|^2.$$

Putting these pieces together yields the first desired inequality.

We proceed to the stochastic setting. Since $f(\mathbf{x}_t, \cdot)$ is μ-strongly concave and $\eta_{\mathbf{y}} = 1/\ell$, we have

$$\mathbb{E}\left[\left\| \mathbf{y}^\star(\mathbf{x}_{t-1}) - \mathbf{y}_t \right\|^2 \right] \;\leq\; \left(1 - \frac{1}{\kappa} \right) \delta_{t-1} + \frac{\sigma^2}{\ell^2 M}. \tag{14}$$

By Young's inequality, we have

$$\begin{aligned}\delta_t &\leq \left(1+\frac{1}{2(\kappa-1)}\right)\mathbb{E}[\|\mathbf{y}^\star(\mathbf{x}_{t-1})-\mathbf{y}_t\|^2]+(1+2(\kappa-1))\,\mathbb{E}[\|\mathbf{y}^\star(\mathbf{x}_t)-\mathbf{y}^\star(\mathbf{x}_{t-1})\|^2]\\ &\leq \left(\frac{2\kappa-1}{2\kappa-2}\right)\mathbb{E}[\|\mathbf{y}^\star(\mathbf{x}_{t-1})-\mathbf{y}_t\|^2]+2\kappa\mathbb{E}[\|\mathbf{y}^\star(\mathbf{x}_t)-\mathbf{y}^\star(\mathbf{x}_{t-1})\|^2]\\ &\overset{(14)}{\leq} \left(1-\frac{1}{2\kappa}\right)\delta_{t-1}+2\kappa\mathbb{E}[\|\mathbf{y}^\star(\mathbf{x}_t)-\mathbf{y}^\star(\mathbf{x}_{t-1})\|^2]+\frac{\sigma^2}{\ell^2 M}.\end{aligned}$$

Since $\mathbf{y}^\star(\cdot)$ is κ-Lipschitz, $\|\mathbf{y}^\star(\mathbf{x}_t)-\mathbf{y}^\star(\mathbf{x}_{t-1})\|\leq\kappa\|\mathbf{x}_t-\mathbf{x}_{t-1}\|$. Furthermore, we have

$$\mathbb{E}[\|\mathbf{x}_t-\mathbf{x}_{t-1}\|^2] \;=\; \eta_{\mathbf{x}}^2\mathbb{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}G_{\mathbf{x}}(\mathbf{x}_{t-1},\mathbf{y}_{t-1},\xi_i)\right\|^2\right] \;\leq\; 2\eta_{\mathbf{x}}^2\ell^2\delta_{t-1}+2\eta_{\mathbf{x}}^2\mathbb{E}[\|\nabla\Phi(\mathbf{x}_{t-1})\|^2]+\frac{\eta_{\mathbf{x}}^2\sigma^2}{M}.$$

Putting these pieces together yields the second desired inequality. □

Lemma C.5 *For two-time-scale GDA, let $\delta_t=\|\mathbf{y}^\star(\mathbf{x}_t)-\mathbf{y}_t\|^2$, the following statement holds true,*

$$\Phi(\mathbf{x}_t) \;\leq\; \Phi(\mathbf{x}_{t-1})-\frac{7\eta_{\mathbf{x}}}{16}\|\nabla\Phi(\mathbf{x}_{t-1})\|^2+\frac{9\eta_{\mathbf{x}}\ell^2\delta_{t-1}}{16}.$$

For two-time-scale SGDA, let $\delta_t=\mathbb{E}[\|\mathbf{y}^\star(\mathbf{x}_t)-\mathbf{y}_t\|^2]$, the following statement holds true,

$$\mathbb{E}\left[\Phi(\mathbf{x}_t)\right] \;\leq\; \mathbb{E}\left[\Phi(\mathbf{x}_{t-1})\right]-\frac{7\eta_{\mathbf{x}}}{16}\mathbb{E}\left[\|\nabla\Phi(\mathbf{x}_{t-1})\|^2\right]+\frac{9\eta_{\mathbf{x}}\ell^2\delta_{t-1}}{16}+\frac{\eta_{\mathbf{x}}^2\kappa\ell\sigma^2}{M}.$$

Proof. For two-time-scale GDA and SGDA, $\eta_{\mathbf{x}}=1/16(\kappa+1)\ell$ and hence

$$\frac{7\eta_{\mathbf{x}}}{16} \;\leq\; \frac{\eta_{\mathbf{x}}}{2}-2\eta_{\mathbf{x}}^2\kappa\ell \;\leq\; \frac{\eta_{\mathbf{x}}}{2}+2\eta_{\mathbf{x}}^2\kappa\ell \;\leq\; \frac{9\eta_{\mathbf{x}}}{16}. \tag{15}$$

Combining (15) with the first inequality in Lemma C.3 yields that

$$\Phi(\mathbf{x}_t) \;\leq\; \Phi(\mathbf{x}_{t-1})-\frac{7\eta_{\mathbf{x}}}{16}\|\nabla\Phi(\mathbf{x}_{t-1})\|^2+\frac{9\eta_{\mathbf{x}}}{16}\|\nabla\Phi(\mathbf{x}_{t-1})-\nabla_{\mathbf{x}}f(\mathbf{x}_{t-1},\mathbf{y}_{t-1})\|^2.$$

Since $\nabla\Phi(\mathbf{x}_{t-1})=\nabla_{\mathbf{x}}f\left(\mathbf{x}_{t-1},\mathbf{y}^\star(\mathbf{x}_{t-1})\right)$, we have

$$\|\nabla\Phi(\mathbf{x}_{t-1})-\nabla_{\mathbf{x}}f(\mathbf{x}_{t-1},\mathbf{y}_{t-1})\|^2 \;\leq\; \ell^2\|\mathbf{y}^\star(\mathbf{x}_{t-1})-\mathbf{y}_{t-1}\|^2 \;=\; \ell^2\delta_{t-1}.$$

Putting these pieces together yields the first desired inequality.

We proceed to the stochastic setting, combining (15) with the second inequality in Lemma C.3 yields that

$$\mathbb{E}\left[\Phi(\mathbf{x}_t)\right] \;\leq\; \mathbb{E}\left[\Phi(\mathbf{x}_{t-1})\right]-\frac{7\eta_{\mathbf{x}}}{16}\mathbb{E}\left[\|\nabla\Phi(\mathbf{x}_{t-1})\|^2\right]+\frac{9\eta_{\mathbf{x}}}{16}\mathbb{E}\left[\|\nabla\Phi(\mathbf{x}_{t-1})-\nabla_{\mathbf{x}}f(\mathbf{x}_{t-1},\mathbf{y}_{t-1})\|^2\right]+\frac{\eta_{\mathbf{x}}^2\kappa\ell\sigma^2}{M}.$$

Since $\nabla\Phi(\mathbf{x}_{t-1})=\nabla_{\mathbf{x}}f\left(\mathbf{x}_{t-1},\mathbf{y}^\star(\mathbf{x}_{t-1})\right)$, we have

$$\mathbb{E}\left[\|\nabla\Phi(\mathbf{x}_{t-1})-\nabla_{\mathbf{x}}f(\mathbf{x}_{t-1},\mathbf{y}_{t-1})\|^2\right] \;\leq\; \ell^2\mathbb{E}\left[\|\mathbf{y}^\star(\mathbf{x}_{t-1})-\mathbf{y}_{t-1}\|^2\right] \;=\; \ell^2\delta_{t-1}.$$

Putting these pieces together yields the second desired inequality. □

C.3. Proof of Theorem C.1

Throughout this subsection, we define $\gamma = 1 - 1/2\kappa + 4\kappa^3\ell^2\eta_{\mathbf{x}}^2$. Performing the first inequality in Lemma C.4 recursively yields that

$$\begin{aligned}\delta_t &\leq \gamma^t\delta_0 + 4\kappa^3\eta_{\mathbf{x}}^2\left(\sum_{j=0}^{t-1}\gamma^{t-1-j}\left\|\nabla\Phi(\mathbf{x}_j)\right\|^2\right) \\ &\leq \gamma^t D^2 + 4\kappa^3\eta_{\mathbf{x}}^2\left(\sum_{j=0}^{t-1}\gamma^{t-1-j}\left\|\nabla\Phi(\mathbf{x}_j)\right\|^2\right).\end{aligned} \tag{16}$$

Combining (16) with the first inequality in Lemma C.5 yields that,

$$\Phi(\mathbf{x}_t) \leq \Phi(\mathbf{x}_{t-1}) - \frac{7\eta_{\mathbf{x}}}{16}\left\|\nabla\Phi(\mathbf{x}_{t-1})\right\|^2 + \frac{9\eta_{\mathbf{x}}\ell^2\gamma^{t-1}D^2}{16} + \frac{9\eta_{\mathbf{x}}^3\ell^2\kappa^3}{4}\left(\sum_{j=0}^{t-2}\gamma^{t-2-j}\left\|\nabla\Phi(\mathbf{x}_j)\right\|^2\right). \tag{17}$$

Summing up (17) over $t = 1, 2, \ldots, T+1$ and rearranging the terms yields that

$$\Phi(\mathbf{x}_{T+1}) \leq \Phi(\mathbf{x}_0) - \frac{7\eta_{\mathbf{x}}}{16}\sum_{t=0}^{T}\left\|\nabla\Phi(\mathbf{x}_t)\right\|^2 + \frac{9\eta_{\mathbf{x}}\ell^2D^2}{16}\left(\sum_{t=0}^{T}\gamma^t\right) + \frac{9\eta_{\mathbf{x}}^3\ell^2\kappa^3}{4}\left(\sum_{t=1}^{T+1}\sum_{j=0}^{t-2}\gamma^{t-2-j}\left\|\nabla\Phi(\mathbf{x}_j)\right\|^2\right).$$

Since $\eta_{\mathbf{x}} = 1/16(\kappa+1)^2\ell$, we have $\gamma \leq 1 - \frac{1}{4\kappa}$ and $\frac{9\eta_{\mathbf{x}}^3\ell^2\kappa^3}{4} \leq \frac{9\eta_{\mathbf{x}}}{1024\kappa}$. This implies that $\sum_{t=0}^{T}\gamma^t \leq 4\kappa$ and

$$\sum_{t=1}^{T+1}\sum_{j=0}^{t-2}\gamma^{t-2-j}\left\|\nabla\Phi(\mathbf{x}_j)\right\|^2 \leq 4\kappa\left(\sum_{t=0}^{T}\left\|\nabla\Phi(\mathbf{x}_t)\right\|^2\right)$$

Putting these pieces together yields that

$$\Phi(\mathbf{x}_{T+1}) \leq \Phi(\mathbf{x}_0) - \frac{103\eta_{\mathbf{x}}}{256}\left(\sum_{t=0}^{T}\left\|\nabla\Phi(\mathbf{x}_t)\right\|^2\right) + \frac{9\eta_{\mathbf{x}}\kappa\ell^2D^2}{4}.$$

By the definition of Δ_Φ, we have

$$\frac{1}{T+1}\left(\sum_{t=0}^{T}\left\|\nabla\Phi(\mathbf{x}_t)\right\|^2\right) \leq \frac{256(\Phi(\mathbf{x}_0) - \Phi(\mathbf{x}_{T+1}))}{103\eta_{\mathbf{x}}(T+1)} + \frac{576\kappa\ell^2D^2}{103(T+1)} \leq \frac{128\kappa^2\ell\Delta_\Phi + 5\kappa\ell^2D^2}{T+1}.$$

This implies that the number of iterations required by Algorithm 1 to return an ϵ-stationary point is bounded by

$$O\left(\frac{\kappa^2\ell\Delta_\Phi + \kappa\ell^2D^2}{\epsilon^2}\right),$$

which gives the same total gradient complexity.

C.4. Proof of Theorem C.2

Throughout this subsection, we define $\gamma = 1 - 1/2\kappa + 4\kappa^3\ell^2\eta_{\mathbf{x}}^2$. Performing the second inequality in Lemma C.4 recursively together with $\delta_0 \leq D^2$ yields that

$$\delta_t \leq \gamma^t D^2 + 4\kappa^3\eta_{\mathbf{x}}^2\left(\sum_{j=0}^{t-1}\gamma^{t-1-j}\mathbb{E}[\left\|\nabla\Phi(\mathbf{x}_j)\right\|^2]\right) + \left(\frac{2\sigma^2\kappa^3\eta_{\mathbf{x}}^2}{M} + \frac{\sigma^2}{\ell^2 M}\right)\left(\sum_{j=0}^{t-1}\gamma^{t-1-j}\right). \tag{18}$$

Combining (18) with the second inequality in Lemma C.5 yields that,

$$\begin{aligned}\mathbb{E}\left[\Phi(\mathbf{x}_t)\right]] \quad &\leq \quad \mathbb{E}\left[\Phi(\mathbf{x}_{t-1})\right] - \frac{7\eta_{\mathbf{x}}}{16}\mathbb{E}\left[\left\|\nabla\Phi(\mathbf{x}_{t-1})\right\|^2\right] + \frac{9\eta_{\mathbf{x}}\ell^2\gamma^{t-1}D^2}{16} + \frac{\eta_{\mathbf{x}}^2\kappa\ell\sigma^2}{M} \\ &\quad + \frac{9\eta_{\mathbf{x}}^3\ell^2\kappa^3}{4}\left(\sum_{j=0}^{t-2}\gamma^{t-2-j}\mathbb{E}\left[\left\|\nabla\Phi(\mathbf{x}_j)\right\|^2\right]\right) + \frac{9\eta_{\mathbf{x}}\ell^2}{16}\left(\frac{2\sigma^2\kappa^3\eta_{\mathbf{x}}^2}{M} + \frac{\sigma^2}{\ell^2 M}\right)\left(\sum_{j=0}^{t-2}\gamma^{t-2-j}\right).\end{aligned} \tag{19}$$

Summing up (19) over $t = 1, 2, \ldots, T+1$ and rearranging the terms yields that

$$\begin{aligned}\mathbb{E}\left[\Phi(\mathbf{x}_{T+1})\right] \quad &\leq \quad \Phi(\mathbf{x}_0) - \frac{7\eta_{\mathbf{x}}}{16}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla\Phi(\mathbf{x}_t)\right\|^2\right] + \frac{9\eta_{\mathbf{x}}\ell^2 D^2}{16}\left(\sum_{t=0}^{T}\gamma^t\right) \\ &\quad + \frac{\eta_{\mathbf{x}}^2\kappa\ell\sigma^2(T+1)}{M} + \frac{9\eta_{\mathbf{x}}^3\ell^2\kappa^3}{4}\left(\sum_{t=1}^{T+1}\sum_{j=0}^{t-2}\gamma^{t-2-j}\mathbb{E}\left[\left\|\nabla\Phi(\mathbf{x}_j)\right\|^2\right]\right) \\ &\quad + \frac{9\eta_{\mathbf{x}}\ell^2}{16}\left(\frac{2\sigma^2\kappa^3\eta_{\mathbf{x}}^2}{M} + \frac{\sigma^2}{\ell^2 M}\right)\left(\sum_{t=1}^{T+1}\sum_{j=0}^{t-2}\gamma^{t-2-j}\right).\end{aligned}$$

Since $\eta_{\mathbf{x}} = 1/16(\kappa+1)^2\ell$, we have $\gamma \leq 1 - \frac{1}{4\kappa}$ and $\frac{9\eta_{\mathbf{x}}^3\ell^2\kappa^3}{4} \leq \frac{9\eta_{\mathbf{x}}}{1024\kappa}$ and $\frac{2\sigma^2\kappa^3\eta_{\mathbf{x}}^2}{M} \leq \frac{\sigma^2}{\ell^2 M}$. This implies that $\sum_{t=0}^{T}\gamma^t \leq 4\kappa$ and

$$\begin{aligned}\sum_{t=1}^{T+1}\sum_{j=0}^{t-2}\gamma^{t-2-j}\mathbb{E}\left[\left\|\nabla\Phi(\mathbf{x}_j)\right\|^2\right] \quad &\leq \quad 4\kappa\left(\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla\Phi(\mathbf{x}_t)\right\|^2\right]\right), \\ \left(\sum_{t=1}^{T+1}\sum_{j=0}^{t-2}\gamma^{t-1-j}\right) \quad &\leq \quad 4\kappa(T+1).\end{aligned}$$

Putting these pieces together yields that

$$\mathbb{E}\left[\Phi(\mathbf{x}_{T+1})\right] \leq \Phi(\mathbf{x}_0) - \frac{103\eta_{\mathbf{x}}}{256}\left(\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla\Phi(\mathbf{x}_t)\right\|^2\right]\right) + \frac{9\eta_{\mathbf{x}}\kappa\ell^2 D^2}{4} + \frac{\eta_{\mathbf{x}}\sigma^2(T+1)}{16\kappa M} + \frac{9\eta_{\mathbf{x}}\kappa\sigma^2(T+1)}{2M}.$$

By the definition of Δ_Φ, we have

$$\begin{aligned}\frac{1}{T+1}\left(\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla\Phi(\mathbf{x}_t)\right\|^2\right]\right) \quad &\leq \quad \frac{256(\Phi(\mathbf{x}_0) - \mathbb{E}\left[\Phi(\mathbf{x}_{T+1})\right])}{103\eta_{\mathbf{x}}(T+1)} + \frac{576\kappa\ell^2 D^2}{103(T+1)} + \frac{16\sigma^2}{103\kappa M} + \frac{1152\kappa\sigma^2}{103M} \\ &\leq \quad \frac{2\Delta_\Phi}{\eta_{\mathbf{x}}(T+1)} + \frac{5\kappa\ell^2 D^2}{T+1} + \frac{13\kappa\sigma^2}{M} \\ &\leq \quad \frac{128\kappa^2\ell\Delta_\Phi + 5\kappa\ell^2 D^2}{T+1} + \frac{13\sigma^2\kappa}{M}.\end{aligned}$$

This implies that the number of iterations required by Algorithm 2 to return an ϵ-stationary point is bounded by

$$O\left(\frac{\kappa^2\ell\Delta_\Phi + \kappa\ell^2 D^2}{\epsilon^2}\right).$$

iterations, which gives the total gradient complexity of the algorithm:

$$O\left(\frac{\kappa^2\ell\Delta_\Phi + \kappa\ell^2 D^2}{\epsilon^2}\max\left\{1,\ \frac{\kappa\sigma^2}{\epsilon^2}\right\}\right).$$

This completes the proof.

D. Proof of Theorems in Section 4.2

In this section, we first specify the choice of parameters in Theorems 4.8 and 4.9. Then we present the proof of main theorems in Section 4.2 with several technical lemmas. Differently from the previous section, we include the case of $\ell D \lesssim \varepsilon$ in the analysis for nonconvex-concave minimax problems.

D.1. Choice of Parameters in Theorem 4.8 and 4.9

In this subsection, we present the full version of Theorems 4.8 and 4.9 with the detailed choice of $\eta_{\mathbf{x}}$, $\eta_{\mathbf{y}}$ and M which are important to subsequent analysis.

Theorem D.1 *Under Assumption 4.6 and letting the step sizes $\eta_{\mathbf{x}} > 0$ and $\eta_{\mathbf{y}} > 0$ be chosen as $\eta_{\mathbf{x}} = \min\{\epsilon^2/[16\ell L^2], \epsilon^4/[4096\ell^3 L^2 D^2]\}$ and $\eta_{\mathbf{y}} = 1/\ell$, the iterations complexity of Algorithm 1 to return an ϵ-stationary point is bounded by*

$$O\left(\frac{\ell^3 L^2 D^2 \widehat{\Delta}_\Phi}{\epsilon^6} + \frac{\ell^3 D^2 \widehat{\Delta}_0}{\epsilon^4}\right).$$

which is also the total gradient complexity of the algorithm.

Theorem D.2 *Under Assumptions 4.1 and 4.6 and letting the step sizes $\eta_{\mathbf{x}} > 0$ and $\eta_{\mathbf{y}} > 0$ be chosen as $\eta_{\mathbf{y}} = \min\{1/2\ell, \epsilon^2/[16\ell\sigma^2]\}$ and $\eta_{\mathbf{x}} = \min\{\epsilon^2/[16\ell(L^2+\sigma^2)], \epsilon^4/[8192\ell^3 D^2 L\sqrt{L^2+\sigma^2}],\ \epsilon^6/[65536\ell^3 D^2\sigma^2 L\sqrt{L^2+\sigma^2}]\}$ with a batch size $M = 1$, the iteration complexity of Algorithm 2 to return an ϵ-stationary point is bounded by*

$$O\left(\left(\frac{\ell^3\left(L^2+\sigma^2\right) D^2 \widehat{\Delta}_\Phi}{\epsilon^6} + \frac{\ell^3 D^2 \widehat{\Delta}_0}{\epsilon^4}\right)\max\left\{1,\ \frac{\sigma^2}{\epsilon^2}\right\}\right),$$

which is also the total gradient complexity of the algorithm.

D.2. Proof of Technical Lemmas

In this subsection, we present three key lemmas which are important for the subsequent analysis.

Lemma D.3 *For two-time-scale GDA, let $\Delta_t = \Phi(\mathbf{x}_t) - f(\mathbf{x}_t, \mathbf{y}_t)$, the following statement holds true,*

$$\Phi_{1/2\ell}(\mathbf{x}_t) \ \le\ \Phi_{1/2\ell}(\mathbf{x}_{t-1}) + 2\eta_{\mathbf{x}}\ell\Delta_{t-1} - \frac{\eta_{\mathbf{x}}}{4}\left\|\nabla\Phi_{1/2\ell}(\mathbf{x}_{t-1})\right\|^2 + \eta_{\mathbf{x}}^2\ell L^2.$$

For two-time-scale SGDA, let $\Delta_t = \mathbb{E}\left[\Phi(\mathbf{x}_t) - f(\mathbf{x}_t, \mathbf{y}_t)\right]$, the following statement holds true,

$$\mathbb{E}\left[\Phi_{1/2\ell}(\mathbf{x}_t)\right] \ \le\ \mathbb{E}\left[\Phi_{1/2\ell}(\mathbf{x}_{t-1})\right] + 2\eta_{\mathbf{x}}\ell\Delta_{t-1} - \frac{\eta_{\mathbf{x}}}{4}\mathbb{E}\left[\left\|\nabla\Phi_{1/2\ell}(\mathbf{x}_{t-1})\right\|^2\right] + \eta_{\mathbf{x}}^2\ell\left(L^2+\sigma^2\right).$$

Proof. We first consider the deterministic setting. Let $\hat{\mathbf{x}}_{t-1} = \mathrm{prox}_{\Phi/2\ell}(\mathbf{x}_{t-1})$, we have

$$\Phi_{1/2\ell}(\mathbf{x}_t) \ \le\ \Phi(\hat{\mathbf{x}}_{t-1}) + \ell\left\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_t\right\|^2 \tag{20}$$

Since $f(\cdot, \mathbf{y})$ is L-Lipschitz for any $\mathbf{y} \in \mathcal{Y}$, we have

$$\begin{aligned}\left\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_t\right\|^2 &= \left\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1} + \eta_{\mathbf{x}}\nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\right\|^2 \\ &\le \left\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\right\|^2 + 2\eta_{\mathbf{x}}\left\langle\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}, \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\right\rangle + \eta_{\mathbf{x}}^2 L^2.\end{aligned} \tag{21}$$

Plugging (21) into (20) yields that

$$\Phi_{1/2\ell}(\mathbf{x}_t) \ \le\ \Phi_{1/2\ell}(\mathbf{x}_{t-1}) + 2\eta_{\mathbf{x}}\ell\left\langle\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}, \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\right\rangle + \eta_{\mathbf{x}}^2\ell L^2. \tag{22}$$

Since f is ℓ-smooth, we have

$$\left\langle\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}, \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\right\rangle \ \le\ f(\hat{\mathbf{x}}_{t-1}, \mathbf{y}_{t-1}) - f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) + \frac{\ell}{2}\left\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\right\|^2. \tag{23}$$

Furthermore, $\Phi(\hat{\mathbf{x}}_{t-1}) \geq f(\hat{\mathbf{x}}_{t-1}, \mathbf{y}_{t-1})$. By the definition of Δ_t, we have

$$f(\hat{\mathbf{x}}_{t-1}, \mathbf{y}_{t-1}) - f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \;\leq\; \Phi(\hat{\mathbf{x}}_{t-1}) - f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \;\leq\; \Delta_{t-1} - \frac{\ell}{2}\left\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\right\|^2. \tag{24}$$

Plugging (23) and (24) into (22) together with $\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\| = \|\nabla\Phi_{1/2\ell}(\mathbf{x}_{t-1})\|/2\ell$ yields the first desired inequality.

We proceed to the stochastic setting. Indeed, we have

$$\left\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_t\right\|^2 \;\leq\; \left\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\right\|^2 + \eta_{\mathbf{x}}^2\left\|\frac{1}{M}\sum_{i=1}^{M} G_{\mathbf{x}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi_i)\right\| + 2\eta_{\mathbf{x}}\left\langle \hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}, \frac{1}{M}\sum_{i=1}^{M} G_{\mathbf{x}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi_i)\right\rangle.$$

Taking an expectation of both sides of the above inequality, conditioned on $(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})$, together with Lemma A.2 and the Lipschitz property of $f(\cdot, \mathbf{y}_{t-1})$ yields that

$$\begin{aligned}\mathbb{E}\left[\left\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_t\right\|^2 \mid \mathbf{x}_{t-1}, \mathbf{y}_{t-1}\right] \;\leq\;& \left\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\right\|^2 + 2\eta_{\mathbf{x}}\left\langle \hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}, \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\right\rangle + \eta_{\mathbf{x}}^2 L^2 \\ &+\eta_{\mathbf{x}}^2\mathbb{E}\left[\left\|\nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) - \frac{1}{M}\sum_{i=1}^{M} G_{\mathbf{x}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi_i)\right\|^2 \mid \mathbf{x}_{t-1}, \mathbf{y}_{t-1}\right].\end{aligned}$$

Taking the expectation of both sides together with Lemma A.2 yields that

$$\mathbb{E}\left[\left\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_t\right\|^2\right] \;\leq\; \mathbb{E}\left[\left\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\right\|^2\right] + 2\eta_{\mathbf{x}}\mathbb{E}\left[\left\langle \hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}, \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\right\rangle\right] + \eta_{\mathbf{x}}^2\left(L^2 + \sigma^2\right).$$

Combining with (23) and (24) yields that

$$\begin{aligned}\mathbb{E}\left[\Phi_{1/2\ell}(\mathbf{x}_t)\right] \;&\leq\; \mathbb{E}\left[\Phi_{1/2\ell}(\mathbf{x}_{t-1})\right] + 2\eta_{\mathbf{x}}\mathbb{E}\left[\left\langle \hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}, \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\right\rangle\right] + \eta_{\mathbf{x}}^2\ell\left(L^2 + \sigma^2\right) \\ &\leq\; \mathbb{E}\left[\Phi_{1/2\ell}(\mathbf{x}_{t-1})\right] + 2\eta_{\mathbf{x}}\ell\Delta_{t-1} - \eta_{\mathbf{x}}\ell^2\mathbb{E}\left[\left\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\right\|^2\right] + \eta_{\mathbf{x}}^2\ell\left(L^2 + \sigma^2\right).\end{aligned}$$

This together with $\|\hat{\mathbf{x}}_{t-1} - \mathbf{x}_{t-1}\| = \|\nabla\Phi_{1/2\ell}(\mathbf{x}_{t-1})\|/2\ell$ yields the second desired inequality. □

Lemma D.4 *For two-time-scale GDA, let $\Delta_t = \Phi(\mathbf{x}_t) - f(\mathbf{x}_t, \mathbf{y}_t)$, the following statement holds true for $\forall s \leq t-1$,*

$$\Delta_{t-1} \;\leq\; \eta_{\mathbf{x}}L^2(2t - 2s - 1) + \frac{\ell}{2}\left(\left\|\mathbf{y}_{t-1} - \mathbf{y}^\star(\mathbf{x}_s)\right\|^2 - \left\|\mathbf{y}_t - \mathbf{y}^\star(\mathbf{x}_s)\right\|^2\right) + \left(f(\mathbf{x}_t, \mathbf{y}_t) - f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\right).$$

For two-time-scale SGDA, let $\Delta_t = \mathbb{E}\left[\Phi(\mathbf{x}_t) - f(\mathbf{x}_t, \mathbf{y}_t)\right]$, the following statement holds true for $\forall s \leq t-1$,

$$\begin{aligned}\Delta_{t-1} \;\leq\;& \eta_{\mathbf{x}}L\sqrt{L^2 + \sigma^2}(2t - 2s - 1) + \frac{1}{2\eta_{\mathbf{y}}}\left(\mathbb{E}\left[\left\|\mathbf{y}_{t-1} - \mathbf{y}^\star(\mathbf{x}_s)\right\|^2\right] - \mathbb{E}\left[\left\|\mathbf{y}_t - \mathbf{y}^\star(\mathbf{x}_s)\right\|^2\right]\right) \\ &+\mathbb{E}\left[f(\mathbf{x}_t, \mathbf{y}_t) - f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\right] + \frac{\eta_{\mathbf{y}}\sigma^2}{2}.\end{aligned}$$

Proof. We first consider the deterministic setting. For any $\mathbf{y} \in \mathcal{Y}$, the convexity of $\mathcal{Y}$ and the update formula of $\mathbf{y}_t$ imply that

$$\left(\mathbf{y} - \mathbf{y}_t\right)^\top\left(\mathbf{y}_t - \mathbf{y}_{t-1} - \eta_{\mathbf{y}}\nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\right) \;\geq\; 0.$$

Rearranging the inequality yields that

$$\left\|\mathbf{y} - \mathbf{y}_t\right\|^2 \;\leq\; 2\eta_{\mathbf{y}}(\mathbf{y}_{t-1} - \mathbf{y})^\top\nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) + 2\eta_{\mathbf{y}}(\mathbf{y}_t - \mathbf{y}_{t-1})^\top\nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) + \left\|\mathbf{y} - \mathbf{y}_{t-1}\right\|^2 - \left\|\mathbf{y}_t - \mathbf{y}_{t-1}\right\|^2.$$

Since $f(\mathbf{x}_{t-1}, \cdot)$ is concave and ℓ-smooth and $\eta_{\mathbf{y}} = 1/\ell$, we have

$$f(\mathbf{x}_{t-1}, \mathbf{y}) - f(\mathbf{x}_{t-1}, \mathbf{y}_t) \;\leq\; \frac{\ell}{2}\left(\left\|\mathbf{y} - \mathbf{y}_{t-1}\right\|^2 - \left\|\mathbf{y} - \mathbf{y}_t\right\|^2\right).$$

Plugging $\mathbf{y} = \mathbf{y}^\star(\mathbf{x}_s)$ ($s \le t-1$) in the above inequality yields that

$$f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_s)) - f(\mathbf{x}_{t-1}, \mathbf{y}_t) \;\le\; \frac{\ell}{2}\left(\|\mathbf{y}_{t-1} - \mathbf{y}^\star(\mathbf{x}_s)\|^2 - \|\mathbf{y}_t - \mathbf{y}^\star(\mathbf{x}_s)\|^2\right).$$

By the definition of Δ_{t-1}, we have

$$\begin{aligned}\Delta_{t-1} \;\le\;& (f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_{t-1})) - f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_s))) + (f(\mathbf{x}_t, \mathbf{y}_t) - f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})) + (f(\mathbf{x}_{t-1}, \mathbf{y}_t) - f(\mathbf{x}_t, \mathbf{y}_t))\\ &+\frac{\ell}{2}\left(\|\mathbf{y}_{t-1} - \mathbf{y}^\star(\mathbf{x}_s)\|^2 - \|\mathbf{y}_t - \mathbf{y}^\star(\mathbf{x}_s)\|^2\right).\end{aligned}$$

Since $f(\mathbf{x}_s, \mathbf{y}^\star(\mathbf{x}_s)) \ge f(\mathbf{x}_s, \mathbf{y})$ for $\forall \mathbf{y} \in \mathcal{Y}$, we have

$$\begin{aligned}&f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_{t-1})) - f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_s))\\ &\le\; f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_{t-1})) - f(\mathbf{x}_s, \mathbf{y}^\star(\mathbf{x}_{t-1})) + f(\mathbf{x}_s, \mathbf{y}^\star(\mathbf{x}_{t-1})) - f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_s))\\ &\le\; f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_{t-1})) - f(\mathbf{x}_s, \mathbf{y}^\star(\mathbf{x}_{t-1})) + f(\mathbf{x}_s, \mathbf{y}^\star(\mathbf{x}_s)) - f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_s)).\end{aligned} \tag{25}$$

Since $f(\cdot, \mathbf{y})$ is L-Lipschitz for any $\mathbf{y} \in \mathcal{Y}$, we have

$$\begin{aligned}f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_{t-1})) - f(\mathbf{x}_s, \mathbf{y}^\star(\mathbf{x}_{t-1})) &\le L\|\mathbf{x}_{t-1} - \mathbf{x}_s\| \le \eta_{\mathbf{x}} L^2 (t-1-s),\\ f(\mathbf{x}_s, \mathbf{y}^\star(\mathbf{x}_s)) - f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_s)) &\le L\|\mathbf{x}_{t-1} - \mathbf{x}_s\| \le \eta_{\mathbf{x}} L^2 (t-1-s)\\ f(\mathbf{x}_{t-1}, \mathbf{y}_t) - f(\mathbf{x}_t, \mathbf{y}_t) &\le L\|\mathbf{x}_{t-1} - \mathbf{x}_t\| \le \eta_{\mathbf{x}} L^2.\end{aligned}$$

Putting these pieces together yields the first desired inequality.

We proceed to the stochastic setting. For $\forall \mathbf{y} \in \mathcal{Y}$, we use the similar argument and obtain that

$$\begin{aligned}\|\mathbf{y} - \mathbf{y}_t\|^2 \;\le\;& 2\eta_{\mathbf{y}}(\mathbf{y}_{t-1} - \mathbf{y})^\top G_{\mathbf{y}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi) + 2\eta_{\mathbf{y}}(\mathbf{y}_t - \mathbf{y}_{t-1})^\top \nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\\ &+2\eta_{\mathbf{y}}(\mathbf{y}_t - \mathbf{y}_{t-1})^\top (G_{\mathbf{y}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi) - \nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})) + \|\mathbf{y} - \mathbf{y}_{t-1}\|^2 - \|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2.\end{aligned}$$

Using the Young's inequality, we have

$$\eta_{\mathbf{y}}(\mathbf{y}_t - \mathbf{y}_{t-1})^\top (G_{\mathbf{y}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi) - \nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})) \;\le\; \frac{\|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2}{4} + \eta_{\mathbf{y}}^2 \|G_{\mathbf{y}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi) - \nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2.$$

Taking an expectation of both sides of the above equality, conditioned on $(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})$, together with Lemma A.2 yields that

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{y} - \mathbf{y}_t\|^2 \mid \mathbf{x}_{t-1}, \mathbf{y}_{t-1}\right]\\ \le\;& 2\eta_{\mathbf{y}} (\mathbf{y}_{t-1} - \mathbf{y})^\top \nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) + 2\eta_{\mathbf{y}} \mathbb{E}\left[(\mathbf{y}_t - \mathbf{y}_{t-1})^\top \nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \mid \mathbf{x}_{t-1}, \mathbf{y}_{t-1}\right]\\ &+2\eta_{\mathbf{y}}^2 \mathbb{E}\left[\|\nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) - G_{\mathbf{y}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi)\|^2 \mid \mathbf{x}_{t-1}, \mathbf{y}_{t-1}\right] + \|\mathbf{y} - \mathbf{y}_{t-1}\|^2 - \frac{\mathbb{E}\left[\|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2 \mid \mathbf{x}_{t-1}, \mathbf{y}_{t-1}\right]}{2}.\end{aligned}$$

Taking the expectation of both sides together with Lemma A.2 yields that

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{y} - \mathbf{y}_t\|^2\right] \;\le\;& 2\eta_{\mathbf{y}} \mathbb{E}\left[(\mathbf{y}_{t-1} - \mathbf{y})^\top \nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) + (\mathbf{y}_t - \mathbf{y}_{t-1})^\top \nabla_{\mathbf{y}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\right]\\ &+\mathbb{E}\left[\|\mathbf{y} - \mathbf{y}_{t-1}\|^2\right] - \frac{\mathbb{E}\left[\|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2\right]}{2} + \eta_{\mathbf{y}}^2 \sigma^2.\end{aligned}$$

Since $f(\mathbf{x}_{t-1}, \cdot)$ is concave and ℓ-smooth, $\mathcal{Y}$ is convex and $\eta_{\mathbf{y}} \le 1/2\ell$, we have

$$\mathbb{E}\left[\|\mathbf{y} - \mathbf{y}_t\|^2\right] \;\le\; \mathbb{E}\left[\|\mathbf{y} - \mathbf{y}_{t-1}\|^2\right] + 2\eta_{\mathbf{y}} (f(\mathbf{x}_{t-1}, \mathbf{y}_t) - f(\mathbf{x}_{t-1}, \mathbf{y})) + \eta_{\mathbf{y}}^2 \sigma^2.$$

Plugging $\mathbf{y} = \mathbf{y}^\star(\mathbf{x}_s)$ ($s \le t-1$) in the above inequality yields that

$$\mathbb{E}\left[f(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_s)) - f(\mathbf{x}_{t-1}, \mathbf{y}_t)\right] \;\le\; \frac{1}{2\eta_{\mathbf{y}}}\left(\mathbb{E}\left[\|\mathbf{y}_{t-1} - \mathbf{y}^\star(\mathbf{x}_s)\|^2\right] - \mathbb{E}\left[\|\mathbf{y}_t - \mathbf{y}^\star(\mathbf{x}_s)\|^2\right]\right) + \frac{\eta_{\mathbf{y}} \sigma^2}{2}.$$

By the definition of Δ_{t-1}, we have

$$\begin{aligned}\Delta_{t-1} &\leq \mathbb{E}\left[f(\mathbf{x}_{t-1},\mathbf{y}^\star(\mathbf{x}_{t-1})) - f(\mathbf{x}_{t-1},\mathbf{y}^\star(\mathbf{x}_s)) + (f(\mathbf{x}_t,\mathbf{y}_t) - f(\mathbf{x}_{t-1},\mathbf{y}_{t-1})) + (f(\mathbf{x}_{t-1},\mathbf{y}_t) - f(\mathbf{x}_t,\mathbf{y}_t))\right] \\ &\quad + \frac{\eta_{\mathbf{y}}\sigma^2}{2} + \frac{1}{2\eta_{\mathbf{y}}}\left(\mathbb{E}\left[\|\mathbf{y}_{t-1} - \mathbf{y}^\star(\mathbf{x}_s)\|^2\right] - \mathbb{E}\left[\|\mathbf{y}_t - \mathbf{y}^\star(\mathbf{x}_s)\|^2\right]\right).\end{aligned}$$

By the fact that $f(\cdot,\mathbf{y})$ is L-Lipschitz for $\forall \mathbf{y}\in\mathcal{Y}$ and Lemma A.2, we have

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_{t-1},\mathbf{y}^\star(\mathbf{x}_{t-1})) - f(\mathbf{x}_s,\mathbf{y}^\star(\mathbf{x}_{t-1}))\right] &\leq \eta_{\mathbf{x}} L\sqrt{L^2+\sigma^2}(t-1-s), \\ \mathbb{E}\left[f(\mathbf{x}_s,\mathbf{y}^\star(\mathbf{x}_s)) - f(\mathbf{x}_{t-1},\mathbf{y}^\star(\mathbf{x}_s))\right] &\leq \eta_{\mathbf{x}} L\sqrt{L^2+\sigma^2}(t-1-s), \\ \mathbb{E}\left[f(\mathbf{x}_{t-1},\mathbf{y}_t) - f(\mathbf{x}_t,\mathbf{y}_t)\right] &\leq \eta_{\mathbf{x}} L\sqrt{L^2+\sigma^2}.\end{aligned}$$

Putting these pieces together with (25) yields the second desired inequality. □

Without loss of generality, we assume that $B \leq T+1$ such that $(T+1)/B$ is an integer. The following lemma provides an upper bound for $\frac{1}{T+1}(\sum_{t=0}^T \Delta_t)$ for two-time-scale GDA and SGDA using a localization technique.

Lemma D.5 *For two-time-scale GDA, let $\Delta_t = \Phi(\mathbf{x}_t) - f(\mathbf{x}_t,\mathbf{y}_t)$, the following statement holds true,*

$$\frac{1}{T+1}\left(\sum_{t=0}^T \Delta_t\right) \leq \eta_{\mathbf{x}} L^2(B+1) + \frac{\ell D^2}{2B} + \frac{\widehat{\Delta}_0}{T+1}.$$

For two-time-scale SGDA, let $\Delta_t = \mathbb{E}\left[\Phi(\mathbf{x}_t) - f(\mathbf{x}_t,\mathbf{y}_t)\right]$, the following statement holds true,

$$\frac{1}{T+1}\left(\sum_{t=0}^T \Delta_t\right) \leq \eta_{\mathbf{x}} L\sqrt{L^2+\sigma^2}(B+1) + \frac{D^2}{2B\eta_{\mathbf{y}}} + \frac{\eta_{\mathbf{y}}\sigma^2}{2} + \frac{\widehat{\Delta}_0}{T+1}.$$

Proof. We first consider the deterministic setting. In particular, we divide $\{\Delta_t\}_{t=0}^T$ into several blocks in which each block contains at most B terms, given by

$$\{\Delta_t\}_{t=0}^{B-1},\ \{\Delta_t\}_{t=B}^{2B-1},\ldots,\{\Delta_t\}_{T-B+1}^{T}.$$

Then we have

$$\frac{1}{T+1}\left(\sum_{t=0}^T \Delta_t\right) \leq \frac{B}{T+1}\left[\sum_{j=0}^{(T+1)/B-1}\left(\frac{1}{B}\sum_{t=jB}^{(j+1)B-1}\Delta_t\right)\right]. \tag{26}$$

Furthermore, letting $s=0$ in the first inequality in Lemma (D.4) yields that

$$\begin{aligned}\sum_{t=0}^{B-1}\Delta_t &\leq \eta_{\mathbf{x}} L^2B^2 + \frac{\ell}{2}\|\mathbf{y}_0 - \mathbf{y}^\star(\mathbf{x}_0)\|^2 + (f(\mathbf{x}_B,\mathbf{y}_B) - f(\mathbf{x}_0,\mathbf{y}_0)) \\ &\leq \eta_{\mathbf{x}} L^2B^2 + \frac{\ell D^2}{2} + (f(\mathbf{x}_B,\mathbf{y}_B) - f(\mathbf{x}_0,\mathbf{y}_0)).\end{aligned} \tag{27}$$

Similarly, letting $s = jB$ yields that, for $1 \leq j \leq \frac{T+1}{B} - 1$,

$$\sum_{t=jB}^{(j+1)B-1}\Delta_t \leq \eta_{\mathbf{x}} L^2B^2 + \frac{\ell D^2}{2} + (f(\mathbf{x}_{jB+B},\mathbf{y}_{jB+B}) - f(\mathbf{x}_{jB},\mathbf{y}_{jB})). \tag{28}$$

Plugging (27) and (28) into (26) yields

$$\frac{1}{T+1}\left(\sum_{t=0}^T \Delta_t\right) \leq \eta_{\mathbf{x}} L^2B + \frac{\ell D^2}{2B} + \frac{f(\mathbf{x}_{T+1},\mathbf{y}_{T+1}) - f(\mathbf{x}_0,\mathbf{y}_0)}{T+1}. \tag{29}$$

Since $f(\cdot, \mathbf{y})$ is L-Lipschitz for any $\mathbf{y} \in \mathcal{Y}$, we have

$$\begin{aligned} f(\mathbf{x}_{T+1}, \mathbf{y}_{T+1}) - f(\mathbf{x}_0, \mathbf{y}_0) &= f(\mathbf{x}_{T+1}, \mathbf{y}_{T+1}) - f(\mathbf{x}_0, \mathbf{y}_{T+1}) + f(\mathbf{x}_0, \mathbf{y}_{T+1}) - f(\mathbf{x}_0, \mathbf{y}_0) \\ &\leq \eta_{\mathbf{x}} L^2 (T+1) + \widehat{\Delta}_0. \end{aligned} \tag{30}$$

Plugging (30) into (29) yields the desired inequality. As for the stochastic case, letting $s = jB$ in the second inequality in Lemma D.4 yields that

$$\sum_{t=jB}^{(j+1)B-1} \Delta_t \leq \eta_{\mathbf{x}} L \sqrt{L^2 + \sigma^2} B^2 + \frac{D^2}{2\eta_{\mathbf{y}}} + \frac{\eta_{\mathbf{y}} \sigma^2}{2}, \qquad 0 \leq j \leq \frac{T+1}{B} - 1. \tag{31}$$

Using the similar argument with (31) and (26) yields the second desired inequality. □

D.3. Proof of Theorem D.1

Summing up the first inequality in Lemma D.3 over $t = 1, 2, \ldots, T+1$ yields that

$$\Phi_{1/2\ell}(\mathbf{x}_{T+1}) \leq \Phi_{1/2\ell}(\mathbf{x}_0) + 2\eta_{\mathbf{x}} \ell \left(\sum_{t=0}^{T} \Delta_t \right) - \frac{\eta_{\mathbf{x}}}{4} \left(\sum_{t=0}^{T} \left\| \nabla \Phi_{1/2\ell}(\mathbf{x}_t) \right\|^2 \right) + \eta_{\mathbf{x}}^2 \ell L^2 (T+1).$$

Combining the above inequality with the first inequality in Lemma D.5 yields that

$$\begin{aligned} \Phi_{1/2\ell}(\mathbf{x}_{T+1}) \leq\ & \Phi_{1/2\ell}(\mathbf{x}_0) + 2\eta_{\mathbf{x}} \ell (T+1) \left(\eta_{\mathbf{x}} L^2 (B+1) + \frac{\ell D^2}{2B} \right) + 2\eta_{\mathbf{x}} \ell \widehat{\Delta}_0 \\ & - \frac{\eta_{\mathbf{x}}}{4} \left(\sum_{t=0}^{T} \left\| \nabla \Phi_{1/2\ell}(\mathbf{x}_t) \right\|^2 \right) + \eta_{\mathbf{x}}^2 \ell L^2 (T+1). \end{aligned}$$

By the definition of $\widehat{\Delta}_\Phi$, we have

$$\frac{1}{T+1} \left(\sum_{t=0}^{T} \left\| \nabla \Phi_{1/2\ell}(\mathbf{x}_t) \right\|^2 \right) \leq \frac{4\widehat{\Delta}_\Phi}{\eta_{\mathbf{x}}(T+1)} + 8\ell \left(\eta_{\mathbf{x}} (B+1) L^2 + \frac{\ell D^2}{2B} \right) + \frac{8\ell \widehat{\Delta}_0}{T+1} + 4\eta_{\mathbf{x}} \ell L^2.$$

Letting $B = 1$ for $D = 0$ and $B = \frac{D}{2L} \sqrt{\frac{\ell}{\eta_{\mathbf{x}}}}$ for $D > 0$, we have

$$\frac{1}{T+1} \left(\sum_{t=0}^{T} \left\| \nabla \Phi_{1/2\ell}(\mathbf{x}_t) \right\|^2 \right) \leq \frac{4\widehat{\Delta}_\Phi}{\eta_{\mathbf{x}}(T+1)} + \frac{8\ell \widehat{\Delta}_0}{T+1} + 16\ell L D \sqrt{\ell \eta_{\mathbf{x}}} + 4\eta_{\mathbf{x}} \ell L^2.$$

Since $\eta_{\mathbf{x}} = \min \left\{ \frac{\epsilon^2}{16\ell L^2}, \frac{\epsilon^4}{4096 \ell^3 L^2 D^2} \right\}$, we have

$$\frac{1}{T+1} \left(\sum_{t=0}^{T} \left\| \nabla \Phi_{1/2\ell}(\mathbf{x}_t) \right\|^2 \right) \leq \frac{4\widehat{\Delta}_\Phi}{\eta_{\mathbf{x}}(T+1)} + \frac{8\ell \widehat{\Delta}_0}{T+1} + \frac{\epsilon^2}{2}.$$

This implies that the number of iterations required by Algorithm 1 to return an ϵ-stationary point is bounded by

$$O\left(\left(\frac{\ell L^2 \widehat{\Delta}_\Phi}{\epsilon^4} + \frac{\ell \widehat{\Delta}_0}{\epsilon^2} \right) \max \left\{ 1, \frac{\ell^2 D^2}{\epsilon^2} \right\} \right),$$

which gives the same total gradient complexity.

D.4. Proof of Theorem D.2

Summing up the second inequality in Lemma D.3 over $t = 1, 2, \ldots, T+1$ yields that

$$\mathbb{E}\left[\Phi_{1/2\ell}(\mathbf{x}_{T+1})\right] \leq \Phi_{1/2\ell}(\mathbf{x}_0) + 2\eta_{\mathbf{x}}\ell\sum_{t=0}^{T}\Delta_t - \frac{\eta_{\mathbf{x}}}{4}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla\Phi_{1/2\ell}(\mathbf{x}_t)\right\|^2\right] + \eta_{\mathbf{x}}^2\ell\left(L^2+\sigma^2\right)(T+1).$$

Combining the above inequality with the second inequality in Lemma D.5 yields that

$$\begin{aligned}\mathbb{E}\left[\Phi_{1/2\ell}(\mathbf{x}_{T+1})\right] \leq\ & \Phi_{1/2\ell}(\mathbf{x}_0) + 2\eta_{\mathbf{x}}\ell(T+1)\left(\eta_{\mathbf{x}}L\sqrt{L^2+\sigma^2}(B+1) + \frac{D^2}{2B\eta_{\mathbf{y}}} + \frac{\eta_{\mathbf{y}}\sigma^2}{2}\right) + 2\eta_{\mathbf{x}}\ell\widehat{\Delta}_0 \\ & -\frac{\eta_{\mathbf{x}}}{4}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla\Phi_{1/2\ell}(\mathbf{x}_t)\right\|^2\right] + \eta_{\mathbf{x}}^2\ell\left(L^2+\sigma^2\right)(T+1).\end{aligned}$$

By the definition of $\widehat{\Delta}_\Phi$, we have

$$\begin{aligned}\frac{1}{T+1}\left(\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla\Phi_{1/2\ell}(\mathbf{x}_t)\right\|^2\right]\right) \leq\ & \frac{4\widehat{\Delta}_\Phi}{\eta_{\mathbf{x}}(T+1)} + 8\ell\left(\eta_{\mathbf{x}}L\sqrt{L^2+\sigma^2}(B+1) + \frac{D^2}{2B\eta_{\mathbf{y}}} + \frac{\eta_{\mathbf{y}}\sigma^2}{2}\right) \\ & + \frac{8\ell\widehat{\Delta}_0}{T+1} + 4\eta_{\mathbf{x}}\ell\left(L^2+\sigma^2\right).\end{aligned}$$

Letting $B = 1$ for $D = 0$ and $B = \frac{D}{2}\sqrt{\frac{1}{\eta_{\mathbf{x}}\eta_{\mathbf{y}}L\sqrt{L^2+\sigma^2}}}$ for $D > 0$, we have

$$\frac{1}{T+1}\left(\sum_{t=0}^{T}\left\|\nabla\Phi_{1/2\ell}(\mathbf{x}_t)\right\|^2\right) \leq \frac{4\widehat{\Delta}_\Phi}{\eta_{\mathbf{x}}(T+1)} + \frac{8\ell\widehat{\Delta}_0}{T+1} + 16\ell D\sqrt{\frac{\eta_{\mathbf{x}}L\sqrt{L^2+\sigma^2}}{\eta_{\mathbf{y}}}} + 4\eta_{\mathbf{y}}\ell\sigma^2 + 4\eta_{\mathbf{x}}\ell\left(L^2+\sigma^2\right).$$

Since $\eta_{\mathbf{x}} = \min\left\{\frac{\epsilon^2}{16\ell(L^2+\sigma^2)}, \frac{\epsilon^4}{8192\ell^3D^2L\sqrt{L^2+\sigma^2}}, \frac{\epsilon^6}{65536\ell^3D^2\sigma^2L\sqrt{L^2+\sigma^2}}\right\}$ and $\eta_{\mathbf{y}} = \min\left\{\frac{1}{2\ell}, \frac{\epsilon^2}{16\ell\sigma^2}\right\}$, we have

$$\frac{1}{T+1}\left(\sum_{t=0}^{T}\left\|\nabla\Phi_{1/2\ell}(\mathbf{x}_t)\right\|^2\right) \leq \frac{4\widehat{\Delta}_\Phi}{\eta_{\mathbf{x}}(T+1)} + \frac{8\ell\widehat{\Delta}_0}{T+1} + \frac{3\epsilon^2}{4}.$$

This implies that the number of iterations required by Algorithm 2 to return an ϵ-stationary point is bounded by

$$O\left(\left(\frac{\ell\left(L^2+\sigma^2\right)\widehat{\Delta}_\Phi}{\epsilon^4} + \frac{\ell\widehat{\Delta}_0}{\epsilon^2}\right)\max\left\{1, \frac{\ell^2D^2}{\epsilon^2}, \frac{\ell^2D^2\sigma^2}{\epsilon^4}\right\}\right),$$

which gives the same total gradient complexity.

E. Results for GDmax and SGDmax

For the sake of completeness, we present GDmax and SGDmax in Algorithm 3 and 4. For any given $\mathbf{x}_t \in \mathbb{R}^m$, the max-oracle approximately solves $\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}_t, \mathbf{y})$ at each iteration. Although GDmax and SGDmax are easier to understand, they have two disadvantages over two-time-scale GDA and SGDA: 1) Both GDmax and SGDmax are nested-loop algorithms. Since it is difficult to pre-determine the number iterations for the inner loop, these algorithms are not favorable in practice; 2) In the general setting where $f(\mathbf{x}, \cdot)$ is nonconcave, GDmax and SGDmax are inapplicable as we can not efficiently solve the maximization problem to a global optimum. Nonetheless, we present the complexity bound for GDmax and SGDmax for the sake of completeness. Note that a portion of results have been derived before (Jin et al., 2019; Nouiehed et al., 2019) and our proof depends on the same techniques.

For nonconvex-strongly-convex problems, the target is to find an ϵ-stationary point (cf. Definition 3.3) given only gradient (or stochastic gradient) access to f. Denote $\Delta_\Phi = \Phi(\mathbf{x}_0) - \min_{\mathbf{x}\in\mathbb{R}^m}\Phi(\mathbf{x})$, we present the gradient complexity for GDmax in the following theorem.

Theorem E.1 *Under Assumption 4.2 and letting the step size $\eta_{\mathbf{x}} > 0$ and the tolerance for the max-oracle $\zeta > 0$ be $\eta_{\mathbf{x}} = 1/[8\kappa\ell]$ and $\zeta = \epsilon^2/[6\ell]$, the number of iterations required by Algorithm 3 to return an ϵ-stationary point is bounded by $O(\kappa\ell\Delta_\Phi\epsilon^{-2})$. Furthermore, the ζ-accurate max-oracle can be realized by gradient ascent (GA) with the stepsize $\eta_{\mathbf{y}} = 1/\ell$ for $O(\kappa\log(\ell D^2/\zeta))$ iterations, which gives the total gradient complexity of the algorithm:*

$$O\left(\frac{\kappa^2\ell\Delta_\Phi}{\epsilon^2}\log\left(\frac{\ell D}{\epsilon}\right)\right).$$

Theorem E.1 shows that, if we alternate between one-step gradient descent over $\mathbf{x}$ and $O(\kappa\log(\ell D/\epsilon))$ gradient ascent steps over $\mathbf{y}$ with a pair of proper learning rates $(\eta_{\mathbf{x}}, \eta_{\mathbf{y}})$, we find at least one stationary point of Φ within $O(\kappa^2\epsilon^{-2}\log(\ell/\epsilon))$ gradient evaluations. Then we present similar guarantees when only stochastic gradients are available in the following theorem.

Theorem E.2 *Under Assumption 4.1 and 4.2 and letting the step size $\eta_{\mathbf{x}} > 0$ and the tolerance for the max-oracle $\zeta > 0$ be the same in Theorem E.1 with the batch size $M = \max\{1, 12\kappa\sigma^2\epsilon^{-2}\}$, the number of iterations required by Algorithm 4 to return an ϵ-stationary point is bounded by $O(\kappa\ell\Delta_\Phi\epsilon^{-2})$. Furthermore, the ζ-accurate max-oracle can be realized by mini-batch stochastic gradient ascent (SGA) with the step size $\eta_{\mathbf{y}} = 1/\ell$ and the mini-batch size $M = \max\{1, 2\sigma^2\kappa\ell^{-1}\zeta^{-1}\}$ for $O(\kappa\log(\ell D^2/\zeta)\max\{1, 2\sigma^2\kappa\ell^{-1}\zeta^{-1}\})$ gradient evaluations, which gives the total gradient complexity of the algorithm:*

$$O\left(\frac{\kappa^2\ell\Delta_\Phi}{\epsilon^2}\log\left(\frac{\ell D}{\epsilon}\right)\max\left\{1,\ \frac{\kappa\sigma^2}{\epsilon^2}\right\}\right).$$

The sample size $M = O(\kappa\sigma^2\epsilon^{-2})$ guarantees that the variance is less than ϵ^2/κ so that the average stochastic gradients over the batch are sufficiently close to the true gradients $\nabla_{\mathbf{x}} f$ and $\nabla_{\mathbf{y}} f$.

We now proceed to the theoretical guarantee for GDmax and SGDmax algorithms for nonconvex-concave problems. The target is to find an ϵ-stationary point of a weakly convex function (Definition 3.7) given only gradient (or stochastic gradient) access to f. Denote $\widehat{\Delta}_\Phi = \Phi_{1/2\ell}(\mathbf{x}_0) - \min_{\mathbf{x}\in\mathbb{R}^m}\Phi_{1/2\ell}(\mathbf{x})$, we present the gradient complexity for GDmax and SGDmax in the following two theorems.

Theorem E.3 *Under Assumption 4.6 and letting the step size $\eta_{\mathbf{x}} > 0$ and the tolerance for the max-oracle $\zeta > 0$ be $\eta_{\mathbf{x}} = \epsilon^2/[\ell L^2]$ and $\zeta = \epsilon^2/[24\ell]$, the number of iterations required by Algorithm 3 to return an ϵ-stationary point is bounded by $O(\ell L^2\widehat{\Delta}_\Phi\epsilon^{-4})$. Furthermore, the ζ-accurate max-oracle is realized by GA with the step size $\eta_{\mathbf{y}} = 1/2\ell$ for $O(\ell D^2/\zeta)$ iterations, which gives the total gradient complexity of the algorithm:*

$$O\left(\frac{\ell^3 L^2 D^2\widehat{\Delta}_\Phi}{\epsilon^6}\right).$$

Theorem E.4 *Under Assumptions 4.1 and 4.6 and letting the tolerance for the max-oracle $\zeta > 0$ be chosen as the same as in Theorem E.3 with a step size $\eta_{\mathbf{x}} > 0$ and a batch size $M > 0$ given by $\eta_{\mathbf{x}} = \epsilon^2/[\ell(L^2+\sigma^2)]$ and $M = 1$, the number of iterations required by Algorithm 4 to return an ϵ-stationary point is bounded by $O(\ell(L^2+\sigma^2)\widehat{\Delta}_\Phi\epsilon^{-4})$. Furthermore, the ζ-accurate max-oracle is realized by SGA with the step size $\eta_{\mathbf{y}} = \min\{1/2\ell, \epsilon^2/[\ell\sigma^2]\}$ and a batch size $M = 1$ for $O(\ell D^2\zeta^{-1}\max\{1, \sigma^2\ell^{-1}\zeta^{-1}\})$ iterations, which gives the following total gradient complexity of the algorithm:*

$$O\left(\frac{\ell^3(L^2+\sigma^2)D^2\widehat{\Delta}_\Phi}{\epsilon^6}\max\left\{1,\ \frac{\sigma^2}{\epsilon^2}\right\}\right).$$

When $\sigma^2 \lesssim \varepsilon^2$, the stochastic gradients are sufficiently close to the true gradients $\nabla_{\mathbf{x}} f$ and $\nabla_{\mathbf{y}} f$ and the gradient complexity of SGDmax matches that of GDmax.

E.1. Proof of Theorem E.1

We present the gradient complexity bound of the gradient-ascent-based ζ-accurate max-oracle in the following lemma.

Algorithm 3 Gradient Descent with Max-oracle (GDmax)

Input: initial point $\mathbf{x}_0$, learning rate $\eta_{\mathbf{x}}$ and max-oracle accuracy ζ.
for $t = 1, 2, \ldots$ **do**
find $\mathbf{y}_{t-1} \in \mathcal{Y}$ so that $f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \geq \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}_{t-1}, \mathbf{y}) - \zeta$.
$\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta_{\mathbf{x}} \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})$.

Algorithm 4 Stochastic Gradient Descent with Max-oracle (SGDmax)

Input: initial point $\mathbf{x}_0$, learning rate $\eta_{\mathbf{x}}$ and max-oracle accuracy ζ.
for $t = 1, 2, \ldots$ **do**
Draw a collection of i.i.d. data samples $\{\xi_i\}_{i=1}^M$.
find $\mathbf{y}_{t-1} \in \mathcal{Y}$ so that $\mathbb{E}[f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \mid \mathbf{x}_{t-1}] \geq \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}_{t-1}, \mathbf{y}) - \zeta$.
$\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta_{\mathbf{x}} \left(\frac{1}{M} \sum_{i=1}^M G_{\mathbf{x}} \left(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}, \xi_i \right) \right)$.

Lemma E.5 *Let $\zeta > 0$ be given, the ζ-accurate max-oracle can be realized by running gradient ascent with a step size $\eta_{\mathbf{y}} = 1/\ell$ for*

$$O\left(\kappa \log\left(\frac{\ell D^2}{\zeta}\right)\right)$$

gradient evaluations. In addition, the output $\mathbf{y}$ satisfies $\|\mathbf{y}^\star - \mathbf{y}\|^2 \leq \zeta/\ell$, where $\mathbf{y}^\star$ is the exact maximizer.

Proof. Since $f(\mathbf{x}_t, \cdot)$ is μ-strongly concave, we have

$$\begin{aligned} f(\mathbf{x}_t, \mathbf{y}^\star(\mathbf{x}_t)) - f(\mathbf{x}_t, \mathbf{y}_t) &\leq \left(1 - \frac{1}{\kappa}\right)^{N_t} \frac{\ell D^2}{2}, \\ \|\mathbf{y}^\star(\mathbf{x}_t) - \mathbf{y}_t\|^2 &\leq \left(1 - \frac{1}{\kappa}\right)^{N_t} D^2. \end{aligned}$$

The first inequality implies that the number of iterations required is $O(\kappa \log(\ell D^2/\zeta))$ which is also the number of gradient evaluations. This, together with the second inequality, yields the other results. □

Proof of Theorem E.1: It is easy to find that the first descent inequality in Lemma C.3 is applicable to GDmax:

$$\Phi(\mathbf{x}_t) \;\leq\; \Phi(\mathbf{x}_{t-1}) - \left(\frac{\eta_{\mathbf{x}}}{2} - 2\eta_{\mathbf{x}}^2\kappa\ell\right) \|\nabla\Phi(\mathbf{x}_{t-1})\|^2 + \left(\frac{\eta_{\mathbf{x}}}{2} + 2\eta_{\mathbf{x}}^2\kappa\ell\right) \|\nabla\Phi(\mathbf{x}_{t-1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2. \tag{32}$$

Since $\nabla\Phi(\mathbf{x}_{t-1}) = \nabla_{\mathbf{x}} f\left(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_{t-1})\right)$, we have

$$\|\nabla\Phi(\mathbf{x}_{t-1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2 \;\leq\; \ell^2 \|\mathbf{y}^\star(\mathbf{x}_{t-1}) - \mathbf{y}_{t-1}\|^2 \;\leq\; \ell\zeta. \tag{33}$$

Since $\eta_{\mathbf{x}} = 1/8\kappa\ell$, we have

$$\frac{\eta_{\mathbf{x}}}{4} \;\leq\; \frac{\eta_{\mathbf{x}}}{2} - 2\eta_{\mathbf{x}}^2\kappa\ell \;\leq\; \frac{\eta_{\mathbf{x}}}{2} + 2\eta_{\mathbf{x}}^2\kappa\ell \;\leq\; \frac{3\eta_{\mathbf{x}}}{4}. \tag{34}$$

Plugging (33) and (34) into (32) yields that

$$\Phi(\mathbf{x}_t) \;\leq\; \Phi(\mathbf{x}_{t-1}) - \frac{\eta_{\mathbf{x}}}{4} \|\nabla\Phi(\mathbf{x}_{t-1})\|^2 + \frac{3\eta_{\mathbf{x}}\ell\zeta}{4}. \tag{35}$$

Summing up (35) over $t = 1, 2, \ldots, T+1$ and rearranging the terms yields that

$$\frac{1}{T+1} \sum_{t=0}^{T} \|\nabla\Phi(\mathbf{x}_t)\|^2 \;\leq\; \frac{4\left(\Phi(\mathbf{x}_0) - \Phi(\mathbf{x}_{T+1})\right)}{\eta_{\mathbf{x}}(T+1)} + 3\ell\zeta.$$

By the definition of $\eta_{\mathbf{x}}$ and Δ_Φ, we conclude that

$$\frac{1}{T+1} \sum_{t=0}^{T} \|\nabla\Phi(\mathbf{x}_t)\|^2 \;\leq\; \frac{32\kappa\ell\Delta_\Phi}{T+1} + 3\ell\zeta.$$

This implies that the number of iterations required by Algorithm 3 to return an ϵ-stationary point is bounded by

$$O\left(\frac{\kappa\ell\Delta_\Phi}{\epsilon^2}\right).$$

Combining Lemma E.5 gives the total gradient complexity of Algorithm 3:

$$O\left(\frac{\kappa^2\ell\Delta_\Phi}{\epsilon^2}\log\left(\frac{\ell D}{\epsilon}\right)\right).$$

This completes the proof.

E.2. Proof of Theorem E.2

We present the gradient complexity bound of the stochastic-gradient-ascent-based ζ-accurate max-oracle in terms of stochastic gradient in the following lemma.

Lemma E.6 *Let $\zeta > 0$ be given, the ζ-accurate max-oracle can be realized by running stochastic gradient ascent with a step size $\eta_{\mathbf{y}} = 1/\ell$ and a batch size $M = \max\{1, 2\sigma^2\kappa/\ell\zeta\}$ for*

$$O\left(\kappa\log\left(\frac{\ell D^2}{\zeta}\right)\max\left\{1,\ \frac{2\sigma^2\kappa}{\ell\zeta}\right\}\right)$$

stochastic gradient evaluations. In addition, the output $\mathbf{y}$ satisfies $\|\mathbf{y}^\star - \mathbf{y}\|^2 \le \zeta/\ell$ where $\mathbf{y}^\star$ is the exact maximizer.

Proof. Since $f(\mathbf{x}_t, \cdot)$ is μ-strongly concave, we have

$$\mathbb{E}\left[f(\mathbf{x}_t, \mathbf{y}^\star(\mathbf{x}_t)) - f(\mathbf{x}_t, \mathbf{y}_t)\right] \ \le\ \left(1-\frac{1}{\kappa}\right)^{N_t}\frac{\ell D^2}{2} + \frac{\eta_{\mathbf{y}}^2\ell\sigma^2}{M}\left(\sum_{j=0}^{N_{t-1}}\left(1-\mu\eta_{\mathbf{y}}\right)^{N_{t-1}-1-j}\right) \ \le\ \left(1-\frac{1}{\kappa}\right)^{N_t}\frac{\ell D^2}{2} + \frac{\sigma^2\kappa}{\ell M},$$

and

$$\mathbb{E}\left[\|\mathbf{y}^\star(\mathbf{x}_t)) - \mathbf{y}_t\|^2\right] \ \le\ \left(1-\frac{1}{\kappa}\right)^{N_t} D^2 + \frac{\eta_{\mathbf{y}}^2\sigma^2}{M}\left(\sum_{j=0}^{N_{t-1}}\left(1-\mu\eta_{\mathbf{y}}\right)^{N_{t-1}-1-j}\right) \ \le\ \left(1-\frac{1}{\kappa}\right)^{N_t}\frac{\ell D^2}{2} + \frac{\sigma^2\kappa}{\ell^2 M}.$$

The first inequality implies that the number of iterations is $O(\kappa\log(\ell D^2/\zeta))$ and the number of stochastic gradient evaluation is $O(\kappa\log(\ell D^2/\zeta)\max\{1, 2\sigma^2\kappa/\ell\zeta\})$. This together with the second inequality yields the other results. □

Proof of Theorem E.2: It is easy to find that the second descent inequality in Lemma C.3 is applicable to SGDmax:

$$\begin{aligned}\mathbb{E}\left[\Phi(\mathbf{x}_t)\right] \ \le\ & \mathbb{E}\left[\Phi(\mathbf{x}_{t-1})\right] - \left(\frac{\eta_{\mathbf{x}}}{2} - 2\eta_{\mathbf{x}}^2\kappa\ell\right)\mathbb{E}\left[\left\|\nabla\Phi(\mathbf{x}_{t-1})\right\|^2\right] \\ & + \left(\frac{\eta_{\mathbf{x}}}{2} + 2\eta_{\mathbf{x}}^2\kappa\ell\right)\mathbb{E}\left[\left\|\nabla\Phi(\mathbf{x}_{t-1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\right\|^2\right] + \frac{\eta_{\mathbf{x}}^2\kappa\ell\sigma^2}{M}.\end{aligned} \tag{36}$$

Since $\nabla\Phi(\mathbf{x}_{t-1}) = \nabla_{\mathbf{x}} f\left(\mathbf{x}_{t-1}, \mathbf{y}^\star(\mathbf{x}_{t-1})\right)$, we have

$$\mathbb{E}\left[\left\|\nabla\Phi(\mathbf{x}_t) - \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\right\|^2\right] \ \le\ \ell^2\mathbb{E}\left[\left\|\mathbf{y}^\star(\mathbf{x}_t) - \mathbf{y}_t\right\|^2\right] \ \le\ \ell\zeta. \tag{37}$$

Since $\eta_{\mathbf{x}} = 1/8\kappa\ell$, we have (34). Plugging (34) and (37) into (36) yields that

$$\mathbb{E}\left[\Phi(\mathbf{x}_t)\right] \ \le\ \mathbb{E}\left[\Phi(\mathbf{x}_{t-1})\right] - \frac{\eta_{\mathbf{x}}}{4}\mathbb{E}\left[\left\|\nabla\Phi(\mathbf{x}_{t-1})\right\|^2\right] + \frac{3\eta_{\mathbf{x}}\ell\zeta}{4} + \frac{\eta_{\mathbf{x}}^2\kappa\ell\sigma^2}{M}. \tag{38}$$

Summing up (38) over $t = 1, 2, \ldots, T+1$ and rearranging the terms yields that

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla\Phi(\mathbf{x}_t)\right\|^2\right] \ \le\ \frac{4\left(\Phi(\mathbf{x}_0) - \mathbb{E}\left[\Phi(\mathbf{x}_{T+1})\right]\right)}{\eta_{\mathbf{x}}(T+1)} + 3\ell\zeta + \frac{4\eta_{\mathbf{x}}\kappa\ell\sigma^2}{M}.$$

By the definition of $\eta_{\mathbf{x}}$ and Δ_Φ, we conclude that

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left[\|\nabla\Phi(\mathbf{x}_t)\|^2\right] \;\leq\; \frac{32\kappa\ell\Delta_\Phi}{T+1}+3\ell\zeta+\frac{\sigma^2}{2M}.$$

This implies that the number of iterations required by Algorithm 4 to return an ϵ-stationary point is bounded by

$$O\left(\frac{\kappa\ell\Delta_\Phi}{\epsilon^2}\right).$$

Note that the same batch set can be reused to construct the unbiased stochastic gradients for both $\nabla_{\mathbf{x}} f(\mathbf{x}_{t-1},\mathbf{y}_{t-1})$ and $\nabla_{\mathbf{y}} f(\mathbf{x}_{t-1},\mathbf{y}_{t-1})$ at each iteration. Combining Lemma E.6 gives the total gradient complexity of Algorithm 4:

$$O\left(\frac{\kappa^2\ell\Delta_\Phi}{\epsilon^2}\log\left(\frac{\sqrt{\kappa}\ell D}{\epsilon}\right)\max\left\{1,\;\frac{\sigma^2\kappa^2}{\epsilon^2}\right\}\right).$$

This completes the proof.

E.3. Proof of Theorem E.3

We present the gradient complexity bound of the gradient-ascent-based ζ-accurate max-oracle in the following lemma.

Lemma E.7 *Let $\zeta>0$ be given, the ζ-accurate max-oracle can be realized by running gradient ascent with a step size $\eta_{\mathbf{y}}=1/2\ell$ for*

$$O\left(\max\left\{1,\;\frac{2\ell D^2}{\zeta}\right\}\right)$$

gradient evaluations.

Proof. Since $f(\mathbf{x}_t,\cdot)$ is concave, we have

$$f(\mathbf{x}_t,\mathbf{y}^\star(\mathbf{x}_t))-f(\mathbf{x}_t,\mathbf{y}_t) \;\leq\; \frac{2\ell D^2}{N_t},$$

which implies that the number of iterations required is $\mathcal{O}\left(\max\left\{1,\;\frac{2\ell D^2}{\zeta}\right\}\right)$ which is the number of gradient evaluation. □

Proof of Theorem E.3: It is easy to find that the first descent inequality in Lemma D.3 is applicable to GDmax:

$$\Phi_{1/2\ell}(\mathbf{x}_t) \;\leq\; \Phi_{1/2\ell}(\mathbf{x}_{t-1})+2\eta_{\mathbf{x}}\ell\Delta_{t-1}-\frac{\eta_{\mathbf{x}}}{4}\left\|\nabla\Phi_{1/2\ell}(\mathbf{x}_{t-1})\right\|^2+\eta_{\mathbf{x}}^2\ell L^2. \tag{39}$$

Summing up (39) over $T=1,2,\ldots,T+1$ together with $\Delta_{t-1}\leq\zeta$ and rearranging the terms yields that

$$\frac{1}{T+1}\sum_{t=0}^{T}\left\|\nabla\Phi_{1/2\ell}(\mathbf{x}_t)\right\|^2 \;\leq\; \frac{4\left(\Phi_{1/2\ell}(\mathbf{x}_0)-\Phi_{1/2\ell}(\mathbf{x}_{T+1})\right)}{\eta_{\mathbf{x}}(T+1)}+8\ell\zeta+4\eta_{\mathbf{x}}\ell L^2.$$

By the definition of $\eta_{\mathbf{x}}$ and $\widehat{\Delta}_\Phi$, we have

$$\frac{1}{T+1}\sum_{t=0}^{T}\left\|\nabla\Phi_{1/2\ell}(\mathbf{x}_t)\right\|^2 \;\leq\; \frac{48\ell L^2\widehat{\Delta}_\Phi}{\epsilon^2(T+1)}+8\ell\zeta+\frac{\epsilon^2}{3}.$$

This implies that the number of iterations required by Algorithm 3 to return an ϵ-stationary point is bounded by

$$O\left(\frac{\ell L^2\widehat{\Delta}_\Phi}{\epsilon^4}\right).$$

Combining Lemma E.7 gives the total gradient complexity of Algorithm 3:

$$O\left(\frac{\ell L^2\widehat{\Delta}_\Phi}{\epsilon^4}\max\left\{1,\;\frac{\ell^2 D^2}{\epsilon^2}\right\}\right).$$

This completes the proof.

E.4. Proof of Theorem E.4

We present the gradient complexity bound of the stochastic-ascent-based ζ-accurate max-oracle in the following lemma.

Lemma E.8 *Let $\zeta > 0$ be given, the ζ-accurate max-oracle can be realized by running stochastic gradient ascent with a step size $\eta_{\mathbf{y}} = \min\{1/2\ell, \zeta/2\sigma^2\}$ and a batch size $M = 1$ for*

$$O\left(\max\left\{1,\ \frac{4\ell D^2}{\zeta},\ \frac{4\sigma^2 D^2}{\zeta^2}\right\}\right) \tag{40}$$

stochastic gradient evaluations.

Proof. Since $f(\mathbf{x}_t, \cdot)$ is concave and $\eta_{\mathbf{y}} = \min\{\frac{1}{2\ell}, \frac{\zeta}{2\sigma^2}\}$, we have

$$\mathbb{E}\left[f(\mathbf{x}_t, \mathbf{y}^\star(\mathbf{x}_t))\right] - \mathbb{E}\left[f(\mathbf{x}_t, \mathbf{y}_t)\right] \ \leq\ \frac{D^2}{\eta_{\mathbf{y}} N_t} + \eta_{\mathbf{y}}\sigma^2.$$

which implies that the number of iterations required is $O(\max\{1, 4\ell D^2\zeta^{-1}, 4\sigma^2 D^2\zeta^{-2}\})$ which is also the number of stochastic gradient evaluations since $M = 1$. □

Proof of Theorem E.4: It is easy to find that the second descent inequality in Lemma D.3 is applicable to SGDmax:

$$\mathbb{E}\left[\Phi_{1/2\ell}(\mathbf{x}_t)\right] \ \leq\ \mathbb{E}\left[\Phi_{1/2\ell}(\mathbf{x}_{t-1})\right] + 2\eta_{\mathbf{x}}\ell\Delta_{t-1} - \frac{\eta_{\mathbf{x}}}{4}\mathbb{E}\left[\left\|\nabla\Phi_{1/2\ell}(\mathbf{x}_{t-1})\right\|^2\right] + \eta_{\mathbf{x}}^2\ell\left(L^2 + \sigma^2\right). \tag{41}$$

Summing up (41) over $T = 1, 2, \ldots, T+1$ together with $\Delta_{t-1} \leq \zeta$ and rearranging the terms yields that

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla\Phi_{1/2\ell}(\mathbf{x}_t)\right\|^2\right] \ \leq\ \frac{4\left(\Phi_{1/2\ell}(\mathbf{x}_0) - \mathbb{E}\left[\Phi_{1/2\ell}(\mathbf{x}_{T+1})\right]\right)}{\eta_{\mathbf{x}}(T+1)} + 8\ell\zeta + 4\eta_{\mathbf{x}}\ell\left(L^2 + \sigma^2\right).$$

By the definition of $\eta_{\mathbf{x}}$ and $\widehat{\Delta}_\Phi$, we have

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left[\left\|\nabla\Phi_{1/2\ell}(\mathbf{x}_t)\right\|^2\right] \ \leq\ \frac{48\ell(L^2+\sigma^2)\widehat{\Delta}_\Phi}{\epsilon^2(T+1)} + 8\ell\zeta + \frac{\epsilon^2}{3}.$$

This implies that the number of iterations required by Algorithm 4 to return an ϵ-stationary point is bounded by

$$O\left(\frac{\ell(L^2+\sigma^2)\widehat{\Delta}_\Phi}{\epsilon^4}\right).$$

Combining Lemma E.8 gives the total gradient complexity of Algorithm 3:

$$O\left(\frac{\ell(L^2+\sigma^2)\widehat{\Delta}_\Phi}{\epsilon^4}\max\left\{1,\ \frac{\ell^2 D^2}{\epsilon^2},\ \frac{\ell^2 D^2\sigma^2}{\epsilon^4}\right\}\right).$$

This completes the proof.