The Complexity of Nonconvex-Strongly-Concave Minimax Optimization (Supplementary Materials)

Siqi Zhang[1] **Junchi Yang**[1] **Cristóbal Guzmán**[2] **Negar Kiyavash**[3] **Niao He**[4]

[1]University of Illinois at Urbana-Champaign (UIUC)
[2]University of Twente & Pontificia Universidad Católica de Chile
[3]École polytechnique fédérale de Lausanne (EPFL)
[4]ETH Zürich

A NOTATIONS

For convenience, we summarize some of the notations used in the paper.

- SC / C / NC / WC: strongly convex, convex, nonconvex, weakly-convex.
- FS: finite-sum.
- L-S: L-Lipschitz smooth. L-IS / AS: L-Lipschitz individual / averaged smoothness.
- SOTA: state-of-the-art, LB / UB: lower / upper bound
- FO / IFO: first-order oracle, incremental first-order oracle, denoted by $\mathbb{O}_{\text{FO}}$ and $\mathbb{O}_{\text{IFO}}$.
- $\mathcal{A}$:linear-span first-order algorithm class.
- $\Phi(x)$, $\Psi(y)$: primal and dual functions of $f(x, y)$.
- $\nabla_x f$, $\nabla_y f$: gradients of a function F with respect to x and y. Also we set $\nabla f = (\nabla_x f, \nabla_y f)$.
- $\nabla^2_{xx} f$, $\nabla^2_{xy} f$, $\nabla^2_{yx} F$, $\nabla^2_{yy} f$: the Hessian of $F(x, y)$ with respect to different components.
- $\{\mathbf{U}^{(i)}\}_{i=1}^n \in \mathbf{Orth}(a, b, n)$: a matrix sequence where if for each $i, j \in [1, n]$ and $i \neq j$, $\mathbf{U}^{(i)}, \mathbf{U}^{(j)} \in \mathbb{R}^{a\times b}$ and $\mathbf{U}^{(i)}(\mathbf{U}^{(i)})^\top = \mathbf{I} \in \mathbb{R}^{a\times a}$ and $\mathbf{U}^{(i)}(\mathbf{U}^{(j)})^\top = \mathbf{0} \in \mathbb{R}^{a\times a}$. Sometimes we use $u^{(i)} \triangleq \mathbf{U}^{(i)}x$.
- e_i: unit vector with the i-th element as 1.
- 0: zero scalars or vectors.
- $\mathcal{X}_k = \text{Span}\{e_1, e_2, \cdots, e_k\}, \mathcal{Y}_k = \text{Span}\{e_{d+1}, e_d, \cdots, e_{d-k+2}\}, \mathcal{X}_0 = \mathcal{Y}_0 = \{0\}$.
- $a \vee b \triangleq \max\{a, b\}$, $a \wedge b \triangleq \min\{a, b\}$.
- $\|\cdot\|$: ℓ_2-norm.
- $\mathbb{N}^+$: all positive integers.
- $\mathbb{N}$: all nonnegative integers.
- $\mathbf{dom}\, f$: the domain of a function f.
- $d_1, d_2 \in \mathbb{N}^+$: dimension numbers of x and y.
- x_d: the d-th coordinate of x, x^t: the variable x in the t-th iteration (in Section 3 and Appendix C only)

B USEFUL LEMMAS AND PROOFS OF SECTION 2

Lemma B.1 (Lemma B.2[Lin et al., 2020]) *Assume $f(\cdot, y)$ is μ_x-strongly convex for $\forall y \in \mathbb{R}^{d_2}$ and $f(x, \cdot)$ is μ_y-strongly concave for $\forall x \in \mathbb{R}^{d_1}$ (we will later refer to this as (μ_x, μ_y)-SC-SC)) and f is L-Lipschitz smooth. Then we have*

a) $y^*(x) = \arg\max_{y\in\mathbb{R}^{d_2}} f(x, y)$ is $\frac{L}{\mu_y}$-Lipschitz;

b) $\Phi(x) = \max_{y\in\mathbb{R}^{d_2}} f(x,y)$ *is* $\frac{2L^2}{\mu_y}$*-Lipschitiz smooth and* μ_x*-strongly convex with* $\nabla\Phi(x) = \nabla_x f(x, y^*(x))$*;*

c) $x^*(y) = \arg\min_{x\in\mathbb{R}^{d_1}} f(x,y)$ *is* $\frac{L}{\mu_x}$*-Lipschitz;*

d) $\Psi(y) = \min_{x\in\mathbb{R}^{d_1}} f(x,y)$ *is* $\frac{2L^2}{\mu_x}$*-Lipschitiz smooth and* μ_y*-strongly concave with* $\nabla\Psi(y) = \nabla_y f(x^*(y), y)$*.*

Lemma B.2 *Under the same assumptions as Lemma B.1, we have*

a) $\mathrm{gap}_f(x,y) \le \frac{L^2}{\mu_y}\|x-x^*\|^2 + \frac{L^2}{\mu_x}\|y-y^*\|^2$*, where* (x^*,y^*) *is the optimal solution to* $\min_{x\in\mathbb{R}^{d_1}}\max_{y\in\mathbb{R}^{d_2}} f(x,y)$*.*

b) $\mathrm{gap}_f(x,y) \le \frac{1}{2\mu_x}\|\nabla_x f(x,y)\|^2 + \frac{1}{2\mu_y}\|\nabla_y f(x,y)\|^2$*.*

c) $\frac{\mu_x}{2}\|x-x^*\|^2 + \frac{\mu_y}{2}\|y-y^*\|^2 \le \mathrm{gap}_f(x,y)$*.*

d) $\|\nabla_x f(x,y)\|^2 + \|\nabla_y f(x,y)\|^2 \le 4L^2(\|x-x^*\|^2 + \|y-y^*\|^2)$*.*

Proof

a) Because $\Phi(x)$ is $\frac{2L^2}{\mu_y}$-smooth by Lemma B.1 and $\nabla\Phi(x^*) = 0$, we have $\Phi(x) - \Phi(x^*) \le \frac{L^2}{\mu_y}\|x-x^*\|^2$. Similarly, because $\Psi(y)$ is $\frac{2L^2}{\mu_x}$-smooth and $\Psi(y^*) = 0$, we have $\Psi(y^*) - \Psi(y) \le \frac{L^2}{\mu_x}\|y-y^*\|^2$. We reach the conclusion by noting that $\mathrm{gap}_f(x,y) = \Phi(x) - \Psi(y)$ and $\Phi(x^*) = \Psi(y^*)$.

b) Because $f(\cdot,y)$ is μ_x-strongly-convex and $\nabla_x f(x^*(y),y) = 0$, we have $f(x,y) - \min_x f(x,y) \le \langle \nabla_x f(x^*(y),y), x - x^*(y)\rangle + \frac{1}{2\mu_x}\|\nabla_x f(x,y) - \nabla_x f(x^*(y),y)\|^2 \le \frac{1}{2\mu_x}\|\nabla_x f(x,y)\|^2$. Similarly, we have $\max_y f(x,y) - f(x,y) \le \frac{1}{2\mu_y}\|\nabla_y f(x,y)\|^2$. Then we note that $\mathrm{gap}_f(x,y) = \max_y f(x,y) - f(x,y) + f(x,y) - \min_x f(x,y)$.

c) Because $\Phi(x)$ is μ_x strongly-convex and $\nabla\Phi(x^*) = 0$, we have $\Phi(x) \ge \Phi(x^*) + \frac{\mu_x}{2}\|x-x^*\|^2$. Similarly, because $\Psi(y)$ is μ_y strongly-concave and $\nabla\Psi(y^*) = 0$, we have $\Psi(y^*) - \Psi(y) \ge \frac{\mu_y}{2}\|y-y^*\|^2$.

d) By definition of Lipschitz smoothness, $\|\nabla_x f(x,y)\|^2 = \|\nabla_x f(x,y) - \nabla_x f(x^*,y^*)\|^2 \le L^2(\|x-x^*\| + \|y-y^*\|)^2 \le 2L^2(\|x-x^*\|^2 + \|y-y^*\|^2)$ and $\|\nabla_y f(x,y)\|^2 = \|\nabla_y f(x,y) - \nabla_y f(x^*,y^*)\|^2 \le L^2(\|x-x^*\| + \|y-y^*\|)^2 \le 2L^2(\|x-x^*\|^2 + \|y-y^*\|^2)$.

■

Proof of Proposition 2.1

Proof (a) and (b) directly follow from the definition of averaged smoothness and individual smoothness.

(c) Denote

$$\bar{f}(x,y) = f(x,y) + \frac{\tau_x}{2}\|x-\tilde{x}\|^2 - \frac{\tau_y}{2}\|y-\tilde{y}\|^2 = \frac{1}{n}\sum_{i=1}^n \left[f_i(x,y) + \frac{\tau_x}{2}\|x-\tilde{x}\|^2 - \frac{\tau_y}{2}\|y-\tilde{y}\|^2\right] \triangleq \frac{1}{n}\sum_{i=1}^n \bar{f}_i(x,y),$$

where $\bar{f}_i(x,y) = f_i(x,y) + \frac{\tau_x}{2}\|x-\tilde{x}\|^2 - \frac{\tau_y}{2}\|y-\tilde{y}\|^2$. Note that for any (x_1,y_1) and (x_2,y_2),

$$\|\nabla_x \bar{f}_i(x_1,y_1) - \nabla_x \bar{f}_i(x_2,y_2)\|^2 \le 2\|\nabla_x f_i(x_1,y_1) - \nabla_x f_i(x_2,y_2)\|^2 + 2\tau_x^2\|x_1-x_2\|^2,$$
$$\|\nabla_y \bar{f}_i(x_1,y_1) - \nabla_y \bar{f}_i(x_2,y_2)\|^2 \le 2\|\nabla_y f_i(x_1,y_1) - \nabla_y f_i(x_2,y_2)\|^2 + 2\tau_y^2\|y_1-y_2\|^2.$$

Therefore,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \left\|\nabla\bar{f}_i(x_1,y_1) - \nabla\bar{f}_i(x_2,y_2)\right\|^2 &\le \frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x_1,y_1) - \nabla f_i(x_2,y_2)\|^2 + 2[\tau_x^2\|x_1-x_2\|^2 + \tau_y^2\|y_1-y_2\|^2] \\ &\le \left(2L^2 + 2\max\{\tau_x^2,\tau_y^2\}\right)\left(\|x_1-x_2\|^2 + \|y_1-y_2\|^2\right).\end{aligned}$$

■

An important trick to transform the basic hard instance into the final hard instance is scaling, which will preserve the smoothness of the original function while extend the domain of the function to a high dimension, i.e., enlarging d, which helps to increase the lower bound. The properties of scaling is summarized in the following lemma.

Lemma B.3 (Scaling and Smoothness) *For a function $\bar{g}(x,y)$ defined on $\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, if $\bar{g}$ is L-smooth, then for the following scaled function:*

$$g(x,y) = \eta^2 \bar{g}\left(\frac{x}{\eta}, \frac{y}{\eta}\right), \tag{1}$$

then g is also L-smooth. Furthermore if the function $\bar{g}$ has a finite-sum form: $\bar{g}(x,y) = \frac{1}{n}\sum_{i=1}^n \bar{g}_i(x,y)$, if $\{\bar{g}_i\}_{i=1}^n$ is L-averaged smooth, then for the following functions:

$$g_i(x,y) = \eta^2 \bar{g}_i\left(\frac{x}{\eta}, \frac{y}{\eta}\right), \quad \textit{and} \quad g(x,y) = \frac{1}{n}\sum_{i=1}^n g_i(x,y) = \frac{1}{n}\sum_{i=1}^n \eta^2 \bar{g}_i\left(\frac{x}{\eta}, \frac{y}{\eta}\right), \tag{2}$$

$\{g_i\}_{i=1}^n$ is also L-averaged smooth. If we further assume $\{\bar{g}_i\}_{i=1}^n$ is L-individually smooth, then $\{g_i\}_{i=1}^n$ is also L-individually smooth.

Proof For the first statement, note that $\nabla g(x,y) = \eta \nabla \bar{g}\left(\frac{x}{\eta}, \frac{y}{\eta}\right)$, so for any $(x_1,y_1),(x_2,y_2) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$,

$$\begin{aligned} \|\nabla_x g(x_1,y_1) - \nabla_x g(x_2,y_2)\| &= \left\|\eta \nabla_x g\left(\frac{x_1}{\eta}, \frac{y_1}{\eta}\right) - \eta \nabla_x g\left(\frac{x_2}{\eta}, \frac{y_2}{\eta}\right)\right\| \\ &\leq \eta L\left(\left\|\frac{x_1}{\eta} - \frac{x_2}{\eta}\right\| + \left\|\frac{y_1}{\eta} - \frac{y_2}{\eta}\right\|\right) = L(\|x_1 - x_2\| + \|y_1 - y_2\|), \end{aligned} \tag{3}$$

similar conclusion also holds for $\nabla_y g$, which verifies the first conclusion.

For the averaged smooth finite-sum statement, note that $\nabla g_i(x,y) = \eta \nabla \bar{g}_i\left(\frac{x}{\eta}, \frac{y}{\eta}\right)$, so for any $(x_1,y_1),(x_2,y_2) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$,

$$\begin{aligned} &\mathbb{E}\left[\|\nabla g_i(x_1,y_1) - \nabla g_i(x_2,y_2)\|^2\right] \\ &= \mathbb{E}\left[\left\|\eta \nabla \bar{g}_i\left(\frac{x_1}{\eta}, \frac{y_1}{\eta}\right) - \eta \nabla \bar{g}_i\left(\frac{x_2}{\eta}, \frac{y_2}{\eta}\right)\right\|^2\right] \\ &= \eta^2 \mathbb{E}\left[\left\|\nabla \bar{g}_i\left(\frac{x_1}{\eta}, \frac{y_1}{\eta}\right) - \nabla \bar{g}_i\left(\frac{x_2}{\eta}, \frac{y_2}{\eta}\right)\right\|^2\right] \\ &\leq \eta^2 L^2\left(\left\|\frac{x_1}{\eta} - \frac{x_2}{\eta}\right\|^2 + \left\|\frac{y_1}{\eta} - \frac{y_2}{\eta}\right\|^2\right) = L^2\left(\|x_1 - x_2\|^2 + \|y_1 - y_2\|^2\right), \end{aligned} \tag{4}$$

so $\{g_i\}_{i=1}^n$ is L-averaged smooth.

For the individually smooth case statement, note that each g_i is a scaled version of $\bar{g}_i$, which is L-smooth, by the conclusion for the first statement, it implies that g_i is also L-smooth, which concludes the proof. ■

C PROOF OF NC-SC LOWER BOUND

Similar to Section 3 in the main text, here in this section only, we denote x_d as the d-th coordinate of x and x^t as the variable x in the t-th iteration.

C.1 DETERMINISTIC NC-SC LOWER BOUND

We start from the proof several important lemmas, then proceed to the analysis of Theorem 3.1.

C.1.1 Proof of Lemma 3.1

Proof Recall the definition of F_d in (7), define $\Gamma_d(x) \triangleq \sum_{i=1}^d \Gamma(x_i)$, note that $x_i^2 = x^\top e_i e_i^\top x$, and

$$\begin{aligned} &\nabla_x F_d(x,y;\lambda,\alpha) = \lambda_1 B_d \top y - \frac{\lambda_1^2\sqrt{\alpha}}{2\lambda_2} e_1 + \frac{\lambda_1^2\alpha}{2\lambda_2}\nabla\Gamma_d(x) - \frac{\lambda_1^2\alpha}{2\lambda_2} e_{d+1}e_{d+1}^\top x \\ &\nabla_y F_d(x,y;\lambda,\alpha) = \lambda_1 B_d x - 2\lambda_2 y, \end{aligned} \tag{5}$$

where $\nabla\Gamma_d(x) = (\nabla\Gamma(x_1), \nabla\Gamma(x_2), \cdots, \nabla\Gamma(x_d))^\top$. Then for the matrix norm of B_d, note that $\alpha \in [0,1]$ and

$$\begin{aligned} \|B_d x\| &= \sqrt{x_{d+1}^2 + (x_d - x_{d+1})^2 + \cdots + (x_1 - x_2)^2 + (\sqrt[4]{\alpha}x_1)^2} \\ &\le \sqrt{x_{d+1}^2 + 2\big(x_d^2 + x_{d+1}^2 + x_{d-1}^2 + x_d^2 + \cdots + x_2^2 + x_3^2 + x_1^2 + x_2^2\big) + x_1^2} \\ &\le \sqrt{4\big(x_{d+1}^2 + x_d^2 + x_{d-1}^2 + \cdots + x_2^2 + x_1^2\big)} = 2\|x\|, \end{aligned} \tag{6}$$

similarly we have $\left\|B_d^T y\right\| \le 2\|y\|$. Denote $C_\gamma \triangleq 360$,[1] so because $0 \le \alpha \le 1$ and $\|B_d\| \le 2$, we have ($\|\cdot\|$ here denotes the spectral norm of a matrix)

$$\|\nabla^2_{xx}F_d\| \le \frac{\lambda_1^2}{2\lambda_2}(C_\gamma\alpha + \alpha) \le \frac{400\lambda_1^2\alpha}{2\lambda_2} = \frac{200\lambda_1^2\alpha}{\lambda_2}, \quad \|\nabla^2_{xy}F_d\| \le 2\lambda_1, \quad \|\nabla^2_{yx}F_d\| \le 2\lambda_1, \quad \|\nabla^2_{yy}F_d\| = 2\lambda_2, \tag{7}$$

which proves the first two statements (i) and (ii).

For (iii), due to the structure of B_d and concerning the activation status defined in $\mathcal{X}_k$ and $\mathcal{Y}_k$, it is easy to verify that if $x \in \mathcal{X}_{k_1}, y \in \mathcal{Y}_{k_2}$ for $k_1, k_2 \in \mathbb{N}$ and $k_1, k_2 \le d$, we have

$$B_d x \in \mathcal{Y}_{k_1}, \quad B_d^\top y \in \mathcal{X}_{k_2+1}.$$

Since the remaining components in the gradient do not affect the activation with the initial point $(0,0) \in \mathbb{R}^{d+1} \times \mathbb{R}^{d+2}$, this proves (iii).

For (iv), by substituting the parameter settings, we have $\frac{200\lambda_1^2\alpha}{\lambda_2} = L$, $2\lambda_1 = L$ and $2\lambda_2 = \mu$, so the function F_d is μ-strongly concave in y and L-Lipschitz smooth, which concludes the proof. ■

C.1.2 Proof of Lemma 3.2

Proof Recall the primal function Φ_d of F_d (7):

$$\Phi_d(x;\lambda,\alpha) = \underbrace{\frac{\lambda_1^2}{2\lambda_2}\left(\frac{1}{2}x^\top A_d x - \sqrt{\alpha}x_1 + \frac{\sqrt{\alpha}}{2} + \alpha\sum_{i=1}^d \Gamma(x_i)\right)}_{\triangleq\Phi_{d1}(x)} + \underbrace{\frac{(1-\alpha)\lambda_1^2}{4\lambda_2}x_{d+1}^2}_{\triangleq\Phi_{d2}(x)}. \tag{8}$$

For the first statement, because $x_d = x_{d+1} = 0$, we have

$$\nabla\Phi_d(x;\lambda,\alpha) = \nabla\Phi_{d1}(x;\lambda,\alpha) + \nabla\Phi_{d2}(x;\lambda,\alpha) = \nabla\Phi_{d1}(x;\lambda,\alpha), \tag{9}$$

which corresponds to the hard instance in [Carmon et al., 2019, Equation 9] with an extra coefficient $\frac{\lambda_1^2}{2\lambda_2}$, then we apply [Carmon et al., 2019, Lemma 3] therein to attain the desired large gradient norm result, i.e.

$$\|\nabla\Phi_d(x;\lambda,\alpha)\| \ge \frac{\lambda_1^2}{2\lambda_2} \times \frac{\alpha^{\frac{3}{4}}}{4} = \frac{\lambda_1^2}{8\lambda_2}\alpha^{\frac{3}{4}}. \tag{10}$$

[1] The choice of C_γ follows the setting in [Zhou and Gu, 2019, Proposition 3.11], which is an upper bound of the Lipschitz smoothness parameter of $\Gamma_d(x)$ in [Carmon et al., 2019, Lemma 2].

For the second statement, we have

$$\begin{aligned}
&\Phi_d(0;\lambda,\alpha) - \inf_{x\in\mathbb{R}^{d+1}} \Phi_d(x;\lambda,\alpha)\\
&= \Phi_{d1}(0;\lambda,\alpha) - \inf_{x\in\mathbb{R}^{d+1}} \left[\Phi_{d1}(x;\lambda,\alpha) + \Phi_{d2}(x;\lambda,\alpha)\right]\\
&\leq \Phi_{d1}(0;\lambda,\alpha) - \inf_{x\in\mathbb{R}^{d+1}} \Phi_{d1}(x;\lambda,\alpha)\\
&\leq \frac{\lambda_1^2}{2\lambda_2}\left(\frac{\sqrt{\alpha}}{2} + 10\alpha d\right),
\end{aligned} \tag{11}$$

where the first inequality uses that $\Phi_{d2}(x;\lambda,\alpha) \geq 0$ because $\alpha \in [0,1]$, and the last inequality applies [Carmon et al., 2019, Lemma 4], which proves the second statement. ■

C.1.3 Proof of Theorem 3.1

The complexity for deterministic nonconvex-strongly-concave problems is defined as

$$\begin{aligned}
\mathrm{Compl}_\epsilon\left(\mathcal{F}_{\mathrm{NCSC}}^{L,\mu,\Delta}, \mathcal{A}, \mathbb{O}_{\mathrm{FO}}\right) &\triangleq \sup_{f\in\mathcal{F}_{\mathrm{NCSC}}^{L,\mu,\Delta}} \inf_{\mathtt{A}\in\mathcal{A}(\mathbb{O}_{\mathrm{FO}})} T_\epsilon(f,\mathtt{A})\\
&= \sup_{f\in\mathcal{F}_{\mathrm{NCSC}}^{L,\mu,\Delta}} \inf_{\mathtt{A}\in\mathcal{A}(\mathbb{O}_{\mathrm{FO}})} \inf\left\{T\in\mathbb{N} \,\middle|\, \left\|\nabla\Phi(x^T)\right\| \leq \epsilon\right\}.
\end{aligned} \tag{12}$$

As a helper lemma, we first discuss the primal function of the scaled hard instance.

Lemma C.1 (Primal of the Scaled Hard Instance) *With the function F_d defined in* (7)*, Φ_d defined in* (9) *and any $\eta \in \mathbb{R}$, for the following function:*

$$f(x,y) = \eta^2 F_d\left(\frac{x}{\eta}, \frac{y}{\eta}; \lambda, \alpha\right), \tag{13}$$

then for its primal function $\Phi(x) \triangleq \max_{y\in\mathbb{R}^{d+2}} f(x,y)$, we have

$$\Phi(x) = \eta^2 \Phi_d\left(\frac{x}{\eta};\lambda,\alpha\right). \tag{14}$$

Proof Check the scaled function,

$$\begin{aligned}
&f(x,y)\\
&= \eta^2\left(\lambda_1\left\langle B_d\frac{x}{\eta}, \frac{y}{\eta}\right\rangle - \lambda_2\left\|\frac{y}{\eta}\right\|^2 - \frac{\lambda_1^2\sqrt{\alpha}}{2\lambda_2}\left\langle e_1, \frac{x}{\eta}\right\rangle + \frac{\lambda_1^2\alpha}{2\lambda_2}\sum_{i=1}^d \Gamma\left(\frac{x_i}{\eta}\right) - \frac{\lambda_1^2\alpha}{4\lambda_2}\left(\frac{x_{d+1}}{\eta}\right)^2 + \frac{\lambda_1^2\sqrt{\alpha}}{4\lambda_2}\right)\\
&= \lambda_1\langle B_d x, y\rangle - \lambda_2\|y\|^2 + \eta^2\left(-\frac{\lambda_1^2\sqrt{\alpha}}{2\lambda_2}\left\langle e_1, \frac{x}{\eta}\right\rangle + \frac{\lambda_1^2\alpha}{2\lambda_2}\sum_{i=1}^d \Gamma\left(\frac{x_i}{\eta}\right) - \frac{\lambda_1^2\alpha}{4\lambda_2}\left(\frac{x_{d+1}}{\eta}\right)^2 + \frac{\lambda_1^2\sqrt{\alpha}}{4\lambda_2}\right),
\end{aligned} \tag{15}$$

check the gradient over y and set it to be 0 to solve for $y^*(x)$, we have

$$\nabla_y f(x, y^*(x)) = \lambda_1 B_d x - 2\lambda_2 y^*(x) = 0 \quad\Longrightarrow\quad y^*(x) = \frac{\lambda_1}{2\lambda_2}B_d x, \tag{16}$$

so the primal function is

$$
\begin{aligned}
&\Phi(x) = f(x, y^*(x)) \\
&= \lambda_1 \langle B_d x, y^*(x) \rangle - \lambda_2 \|y^*(x)\|^2 + \eta^2 \left(-\frac{\lambda_1^2 \sqrt{\alpha}}{2\lambda_2} \left\langle e_1, \frac{x}{\eta} \right\rangle + \frac{\lambda_1^2 \alpha}{2\lambda_2} \sum_{i=1}^{d} \Gamma\left(\frac{x_i}{\eta}\right) - \frac{\lambda_1^2 \alpha}{4\lambda_2} \left(\frac{x_{d+1}}{\eta}\right)^2 + \frac{\lambda_1^2 \sqrt{\alpha}}{4\lambda_2} \right) \\
&= \frac{\lambda_1^2}{4\lambda_2} \|B_d x\|^2 + \eta^2 \left(-\frac{\lambda_1^2 \sqrt{\alpha}}{2\lambda_2} \left\langle e_1, \frac{x}{\eta} \right\rangle + \frac{\lambda_1^2 \alpha}{2\lambda_2} \sum_{i=1}^{d} \Gamma\left(\frac{x_i}{\eta}\right) - \frac{\lambda_1^2 \alpha}{4\lambda_2} \left(\frac{x_{d+1}}{\eta}\right)^2 + \frac{\lambda_1^2 \sqrt{\alpha}}{4\lambda_2} \right) \\
&= \eta^2 \left(\frac{\lambda_1^2}{4\lambda_2} \left\| B_d \frac{x}{\eta} \right\|^2 - \frac{\lambda_1^2 \sqrt{\alpha}}{2\lambda_2} \left\langle e_1, \frac{x}{\eta} \right\rangle + \frac{\lambda_1^2 \alpha}{2\lambda_2} \sum_{i=1}^{d} \Gamma\left(\frac{x_i}{\eta}\right) - \frac{\lambda_1^2 \alpha}{4\lambda_2} \left(\frac{x_{d+1}}{\eta}\right)^2 + \frac{\lambda_1^2 \sqrt{\alpha}}{4\lambda_2} \right) \\
&= \eta^2 \Phi_d\left(\frac{x}{\eta}; \lambda, \alpha\right),
\end{aligned} \tag{17}
$$

which concludes the proof. ■

Now we come to the formal statement and proof of the main theorem.

Theorem C.1 (Lower Bound for General NC-SC, Restate Theorem 3.1) *For any linear-span first-order algorithm* $\mathtt{A} \in \mathcal{A}$ *and parameters* $L, \mu, \Delta > 0$, *with a desired accuracy* $\epsilon > 0$, *for the following function* $f : \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} \to \mathbb{R}$*:*

$$f(x, y) \triangleq \eta^2 F_d\left(\frac{x}{\eta}, \frac{y}{\eta}; \lambda^*, \alpha\right), \tag{18}$$

where F_d *is defined in* (7), *with a primal function* $\Phi(x) \triangleq \max_{y \in \mathbb{R}^{d+1}} f(x, y)$, *for a small enough* $\epsilon > 0$ *satisfying*

$$\epsilon^2 \leq \min\left(\frac{\Delta L}{64000}, \frac{\Delta L \sqrt{\kappa}}{38400}\right),$$

if we set

$$\lambda^* = \left(\frac{L}{2}, \frac{\mu}{2}\right), \quad \eta = \frac{16\mu}{L^2} \alpha^{-3/4} \epsilon, \quad \alpha = \frac{\mu}{100L} \in [0, 1], \quad d = \left\lfloor \frac{\Delta L \sqrt{\kappa}}{12800} \epsilon^{-2} \right\rfloor \geq 3, \tag{19}$$

we have

- *The proposed function* $f \in \mathcal{F}_{\mathrm{NCSC}}^{L,\mu,\Delta}$.
- *To obtain a point* $\hat{x} \in \mathbb{R}^{d+1}$ *such that* $\|\nabla \Phi(\hat{x})\| \leq \epsilon$, *the number of FO queries required by the algorithm* $\mathtt{A} \in \mathcal{A}$ *is at least* $2d - 1 = \Omega\left(\sqrt{\kappa} \Delta L \epsilon^{-2}\right)$, *namely,*

$$\mathrm{Compl}_\epsilon\left(\mathcal{F}_{\mathrm{NCSC}}^{L,\mu,\Delta}, \mathcal{A}, \mathbb{O}_{\mathrm{FO}}\right) = \Omega\left(\sqrt{\kappa} \Delta L \epsilon^{-2}\right). \tag{20}$$

Proof First, we verify the smoothness and strong concavity of the function f. According to Lemma 3.1, $\alpha \leq \frac{\mu}{100L}$ implies that $F_d(x, y; \lambda^*, \alpha)$ is L-smooth and μ-strongly concave in y. Given that f is a scaled version of F_d, by Lemma B.3, it is easy to verify that f is also L-smooth and μ-strongly concave in y.

Then by Lemma C.1, we have

$$\Phi(x) = \eta^2 \Phi_d\left(\frac{x}{\eta}; \lambda^*, \alpha\right), \tag{21}$$

where Φ_d is defined in (9). Next we check the initial primal function gap, by Lemma 3.2 and parameter substitution,

$$\Phi(0) - \inf_x \Phi(x) = \eta^2 \left(\Phi_d(0) - \inf_x \Phi_d(x)\right) \leq \frac{\eta^2 L^2}{4\mu} \left(\frac{\sqrt{\alpha}}{2} + 10\alpha d\right) = \frac{64\mu}{L^2} \left(\frac{1}{2\alpha} + \frac{10d}{\sqrt{\alpha}}\right) \epsilon^2, \tag{22}$$

by substituting α and d into the RHS above, we have

$$
\begin{aligned}
\frac{64\mu}{L^2} \left(\frac{1}{2\alpha} + \frac{10d}{\sqrt{\alpha}}\right) \epsilon^2 &\leq \frac{64\mu}{L^2} \left(\frac{50L}{\mu} + 100\sqrt{\frac{L}{\mu}} \cdot \frac{\Delta L \sqrt{\kappa}}{12800} \epsilon^{-2}\right) \epsilon^2 \\
&\leq \frac{64}{L} \left(50 + \frac{\Delta L}{128} \epsilon^{-2}\right) \epsilon^2 \leq \frac{64}{L} \left(\frac{\Delta L}{64} \epsilon^{-2}\right) \epsilon^2 = \Delta.
\end{aligned} \tag{23}
$$

The second inequality holds because ϵ above is set to be small enough than $\frac{\Delta L}{6400}$. We conclude that $f \in \mathcal{F}_{\text{NCSC}}^{L,\mu,\Delta}$.

We now discuss the lower bound argument. Based on Lemma 3.2 and the setting of η, we have when $x_d = x_{d+1} = 0$,

$$\|\nabla \Phi(x)\| = \eta \left\| \nabla \Phi_d \left(\frac{x}{\eta}; \lambda^*, \alpha \right) \right\| \geq \frac{\eta L^2}{16\mu} \alpha^{3/4} = \epsilon. \tag{24}$$

So starting from $(x, y) = (0, 0) \in \mathbb{R}^{d+1} \times \mathbb{R}^{d+2}$, we cannot get the primal stationarity convergence at least until $x_d \neq 0$. By the "alternating zero-chain" mechanism[2] in Lemma 3.1, each update with the linear-span algorithm interacting with the FO oracle call will activate exactly one coordinate alternatively between x and y. Therefore the algorithm $\mathtt{A}$ requires at least $2d - 1$ queries to FO to activate the d-th element of x, i.e., x_d, which implies the lower bound is (note that ϵ is small enough such that $d \geq 3$)

$$2d - 1 = \Omega\big(\sqrt{\kappa} \Delta L \epsilon^{-2}\big), \tag{25}$$

which concludes the proof. Notice that this argument works even for randomized algorithms, as long as they satisfy the linear-span assumption. ■

C.2 AVERAGED SMOOTH FINITE-SUM NC-SC LOWER BOUND

Similar to the deterministic NC-SC case, here we still start from several important lemmas and proceed to the proof of Theorem 3.2.

C.2.1 Hard Instance Construction

Recall the (unscaled) hard instance in averaged smooth finite-sum case in (15): $H_d : \mathbb{R}^{d+2} \times \mathbb{R}^{d+1} \to \mathbb{R}$, $\Gamma_d^n : \mathbb{R}^{n(d+1)} \to \mathbb{R}$ and

$$\begin{aligned} H_d(x, y; \lambda, \alpha) &\triangleq \lambda_1 \langle B_d x, y \rangle - \lambda_2 \|y\|^2 - \frac{\lambda_1^2 \sqrt{\alpha}}{2\lambda_2} \langle e_1, x \rangle - \frac{\lambda_1^2 \alpha}{4\lambda_2} x_{d+1}^2 + \frac{\lambda_1^2 \sqrt{\alpha}}{4\lambda_2}, \\ \Gamma_d^n(x) &\triangleq \sum_{i=1}^{n} \sum_{j=i(d+1)-d}^{i(d+1)-1} \Gamma(x_j), \end{aligned} \tag{26}$$

then $\bar{f}_i, \bar{f} : \mathbb{R}^{n(d+1)} \times \mathbb{R}^{n(d+2)} \to \mathbb{R}$, $\{\mathbf{U}^{(i)}\}_{i=1}^n \in \mathbf{Orth}(d+1, n(d+1), n)$, $\{\mathbf{V}^{(i)}\}_{i=1}^n \in \mathbf{Orth}(d+2, n(d+2), n)$ and

$$\begin{aligned} \bar{f}_i(x, y) &\triangleq H_d\Big(\mathbf{U}^{(i)} x, \mathbf{V}^{(i)} y; \lambda, \alpha\Big) + \frac{\lambda_1^2 \alpha}{2n\lambda_2} \Gamma_d^n(x), \\ \bar{f}(x, y) &\triangleq \frac{1}{n} \sum_{i=1}^{n} \bar{f}_i(x, y) = \frac{1}{n} \sum_{i=1}^{n} \left[H_d\Big(\mathbf{U}^{(i)} x, \mathbf{V}^{(i)} y; \lambda, \alpha\Big) + \frac{\lambda_1^2 \alpha}{2n\lambda_2} \Gamma_d^n(x) \right]. \end{aligned} \tag{27}$$

i.e., by denoting $u^{(i)} \triangleq \mathbf{U}^{(i)} x$ and note that $\|y\|^2 = \sum_{i=1}^n \big\|\mathbf{V}^{(i)} y\big\|^2$,

$$\begin{aligned} &\bar{f}(x, y) \\ = &\frac{1}{n} \sum_{i=1}^{n} \left[\lambda_1 \Big\langle B_d \mathbf{U}^{(i)} x, \mathbf{V}^{(i)} y \Big\rangle - \lambda_2 \|\mathbf{V}^{(i)} y\|^2 - \frac{\lambda_1^2 \sqrt{\alpha}}{2\lambda_2} \Big\langle e_1, \mathbf{U}^{(i)} x \Big\rangle + \frac{\lambda_1^2 \alpha}{2n\lambda_2} \Gamma_d^n(x) - \frac{\lambda_1^2 \alpha}{4\lambda_2} \Big(u_{d+1}^{(i)}\Big)^2 + \frac{\lambda_1^2 \sqrt{\alpha}}{4\lambda_2} \right] \\ = &-\frac{\lambda_2}{n} \|y\|^2 + \frac{1}{n} \sum_{i=1}^{n} \left[\lambda_1 \Big\langle B_d \mathbf{U}^{(i)} x, \mathbf{V}^{(i)} y \Big\rangle - \frac{\lambda_1^2 \sqrt{\alpha}}{2\lambda_2} \Big\langle e_1, \mathbf{U}^{(i)} x \Big\rangle + \frac{\lambda_1^2 \alpha}{2n\lambda_2} \Gamma_d^n(x) - \frac{\lambda_1^2 \alpha}{4\lambda_2} \Big(u_{d+1}^{(i)}\Big)^2 + \frac{\lambda_1^2 \sqrt{\alpha}}{4\lambda_2} \right], \end{aligned} \tag{28}$$

so $\bar{f}$ is $\frac{2\lambda_2}{n}$-strongly concave in y. Recall the gradient of f_i:

$$\begin{aligned} \nabla_x f_i(x, y) &= \lambda_1 \Big(\mathbf{U}^{(i)}\Big)^\top B_d^\top \mathbf{V}^{(i)} y - \frac{\lambda_1^2 \sqrt{\alpha}}{2\lambda_2} \Big(\mathbf{U}^{(i)}\Big)^\top e_1 + \frac{\lambda_1^2 \alpha}{2n\lambda_2} \nabla \Gamma_d^n(x) - \frac{\lambda_1^2 \alpha}{2\lambda_2} \Big(\mathbf{U}^{(i)}\Big)^\top e_{d+1} e_{d+1}^\top \mathbf{U}^{(i)} x, \\ \nabla_y f_i(x, y) &= \lambda_1 \Big(\mathbf{V}^{(i)}\Big)^\top B_d \mathbf{U}^{(i)} x - 2\lambda_2 \Big(\mathbf{V}^{(i)}\Big)^\top \mathbf{V}^{(i)} y, \end{aligned} \tag{29}$$

[2] Also known as the "Domino argument" in Ibrahim et al. [2020].

then we discuss the smoothness of $\{\bar{f}_i\}_i$.

Lemma C.2 (Properties of $\bar{f}$) *For $n \in \mathbb{N}^+$, $L \geq 2n\mu > 0$, if we set*

$$\lambda = \lambda^* = (\lambda_1^*, \lambda_2^*) = \left(\sqrt{\frac{n}{40}}L, \frac{n\mu}{2}\right) \quad \text{and} \quad \alpha = \frac{n\mu}{50L}, \tag{30}$$

then the function $\{\bar{f}_i\}_i$ is L-averaged smooth, and $\bar{f}(x, \cdot)$ is μ-strongly concave for any fixed $x \in \mathbb{R}^{d+1}$.

Proof For the strong concavity, note that $\bar{f}$ is $\frac{2\lambda_2}{n}$-strongly concave, so by substitution we have $\bar{f}$ is μ-strongly concave in y. Then for the average smoothness, by definition, we have for any $(x_1, y_1), (x_2, y_2) \in \mathbb{R}^{d+1} \times \mathbb{R}^{d+1}$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x_1, y_1) - \nabla f_i(x_2, y_2)\|^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \Big[\|\nabla_x f_i(x_1, y_1) - \nabla_x f_i(x_2, y_2)\|^2 + \|\nabla_y f_i(x_1, y_1) - \nabla_y f_i(x_2, y_2)\|^2\Big],
\end{aligned} \tag{31}$$

then note that Γ_d^n and Γ_d enjoys the same Lipschitz smoothness parameter as that of Γ, so we have

$$\begin{aligned}
&\|\nabla_x f_i(x_1, y_1) - \nabla_x f_i(x_2, y_2)\|^2 \\
&\leq 4\left\|\lambda_1 \left(\mathbf{U}^{(i)}\right)^\top B_d^\top \mathbf{V}^{(i)}(y_1 - y_2)\right\|^2 + 4\left\|\frac{\lambda_1^2 \alpha}{2n\lambda_2}(\nabla\Gamma_d^n(x_1) - \nabla\Gamma_d^n(x_2))\right\|^2 \\
&\qquad\qquad + 4\left\|\frac{\lambda_1^2\alpha}{2\lambda_2}\left(\mathbf{U}^{(i)}\right)^\top e_{d+1}e_{d+1}^\top \mathbf{U}^{(i)}(x_1 - x_2)\right\|^2 \\
&= 4\lambda_1^2\left\|B_d^\top \mathbf{V}^{(i)}(y_1 - y_2)\right\|^2 + \frac{\lambda_1^4\alpha^2}{n^2\lambda_2^2}\|\nabla\Gamma_d^n(x_1) - \nabla\Gamma_d^n(x_2)\|^2 + \frac{\lambda_1^4\alpha^2}{\lambda_2^2}\left\|e_{d+1}e_{d+1}^\top \mathbf{U}^{(i)}(x_1 - x_2)\right\|^2 \\
&\leq 16\lambda_1^2\left\|\mathbf{V}^{(i)}(y_1 - y_2)\right\|^2 + \frac{C_\gamma^2\lambda_1^4\alpha^2}{n^2\lambda_2^2}\|x_1 - x_2\|^2 + \frac{\lambda_1^4\alpha^2}{\lambda_2^2}\left\|\mathbf{U}^{(i)}(x_1 - x_2)\right\|^2,
\end{aligned} \tag{32}$$

and

$$\begin{aligned}
&\|\nabla_y f_i(x_1, y_1) - \nabla_y f_i(x_2, y_2)\|^2 = \left\|\lambda_1\left(\mathbf{V}^{(i)}\right)^\top B_d \mathbf{U}^{(i)}(x_1 - x_2) - 2\lambda_2\left(\mathbf{V}^{(i)}\right)^\top \mathbf{V}^{(i)}(y_1 - y_2)\right\|^2 \\
&\leq 2\left\|\lambda_1\left(\mathbf{V}^{(i)}\right)^\top B_d\mathbf{U}^{(i)}(x_1 - x_2)\right\|^2 + 2\left\|2\lambda_2\left(\mathbf{V}^{(i)}\right)^\top \mathbf{V}^{(i)}(y_1 - y_2)\right\|^2 \\
&\leq 8\lambda_1^2\left\|\mathbf{U}^{(i)}(x_1 - x_2)\right\|^2 + 8\lambda_2^2\left\|\mathbf{V}^{(i)}(y_1 - y_2)\right\|^2,
\end{aligned} \tag{33}$$

so we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x_1, y_1) - \nabla f_i(x_2, y_2)\|^2 \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left[\left(16\lambda_1^2 + 8\lambda_2^2\right)\left\|\mathbf{V}^{(i)}(y_1 - y_2)\right\|^2 + \left(\frac{\lambda_1^4\alpha^2}{\lambda_2^2} + 8\lambda_1^2\right)\left\|\mathbf{U}^{(i)}(x_1 - x_2)\right\|^2 + \frac{C_\gamma^2\lambda_1^4\alpha^2}{n^2\lambda_2^2}\|x_1 - x_2\|^2\right] \\
&= \frac{1}{n}\left(16\lambda_1^2 + 8\lambda_2^2\right)\sum_{i=1}^{n}\left[\left\|\mathbf{V}^{(i)}(y_1 - y_2)\right\|^2\right] + \frac{1}{n}\left(\frac{\lambda_1^4\alpha^2}{\lambda_2^2} + 8\lambda_1^2\right)\sum_{i=1}^{n}\left[\left\|\mathbf{U}^{(i)}(x_1 - x_2)\right\|^2\right] + \frac{C_\gamma^2\lambda_1^4\alpha^2}{n^2\lambda_2^2}\|x_1 - x_2\|^2 \\
&= \frac{1}{n}\left(16\lambda_1^2 + 8\lambda_2^2\right)\|y_1 - y_2\|^2 + \frac{1}{n}\left(\frac{\lambda_1^4\alpha^2}{\lambda_2^2} + 8\lambda_1^2\right)\|x_1 - x_2\|^2 + \frac{C_\gamma^2\lambda_1^4\alpha^2}{n^2\lambda_2^2}\|x_1 - x_2\|^2 \\
&\leq \frac{1}{n}\max\left\{16\lambda_1^2 + 8\lambda_2^2, \frac{C_\gamma^2\lambda_1^4\alpha^2}{n\lambda_2^2} + \frac{\lambda_1^4\alpha^2}{\lambda_2^2} + 8\lambda_1^2\right\}\left(\|x_1 - x_2\|^2 + \|y_1 - y_2\|^2\right),
\end{aligned} \tag{34}$$

then note that $\alpha \in [0,1]$ because we set $L \geq 2n\mu \geq \frac{1}{50}n\mu$, so substitute parameters into the above, we have

$$\begin{aligned}
&\frac{1}{n}\max\left\{16\lambda_1^2 + 8\lambda_2^2, \frac{C_\gamma^2\lambda_1^4\alpha^2}{n\lambda_2^2} + \frac{\lambda_1^4\alpha^2}{\lambda_2^2} + 8\lambda_1^2\right\} \\
={}& \frac{1}{n}\max\left\{16\lambda_1^2 + 2n^2\mu^2, \frac{4C_\gamma^2\lambda_1^4\alpha^2}{n^3\mu^2} + \frac{4\lambda_1^4\alpha^2}{n^2\mu^2} + 8\lambda_1^2\right\} \\
\leq{}& \frac{1}{n}\max\left\{16\lambda_1^2 + 2n^2\mu^2, 1000000\alpha^2\frac{\lambda_1^4}{n^3\mu^2} + 8\lambda_1^2\right\} \\
={}& \frac{1}{n}\max\left\{\frac{16nL^2}{40} + 2n^2\mu^2, 1000000\cdot\frac{n^2\mu^2}{2500L^2}\cdot\frac{n^2L^4}{1600n^3\mu^2} + \frac{8nL^2}{40}\right\} \\
\leq{}& \max\left\{\frac{2L^2}{5} + \frac{L^2}{2}, \frac{L^2}{4} + \frac{L^2}{5}\right\} \\
\leq{}& \max\left\{\frac{9L^2}{10}, \frac{9L^2}{20}\right\} \leq L^2,
\end{aligned} \tag{35}$$

where the first inequality is attained by the computation with the value of $C_\gamma = 360$, the second inequality comes from the assumption $L \geq 2n\mu \geq 2\sqrt{n}\mu$; the last equality is attained by parameter substitution, which verifies the conclusion. ■

Next we discuss the primal function of the finite-sum hard instance.

Lemma C.3 (Primal of Averaged Smooth Finite-Sum Hard Instance) *For the function $\bar{f} = \frac{1}{n}\sum_{i=1}^n \bar{f}_i$ defined in* (15), *define $\bar{\Phi}(x) \triangleq \max_y \bar{f}(x,y)$, then we have*

$$\bar{\Phi}(x) = \frac{1}{n}\sum_{i=1}^n \bar{\Phi}_i(x), \quad \textit{where} \quad \bar{\Phi}_i(x) \triangleq \Phi_d\Big(\mathbf{U}^{(i)}x\Big), \tag{36}$$

while Φ_d is defined in (9).

Proof By the expression of $\bar{f}$ in (28), take the gradient over y and set it as 0, denote the maximizer as $y^*(x)$, we have

$$-\frac{2\lambda_2}{n}y^*(x) + \frac{1}{n}\sum_{i=1}^n \lambda_1\Big(\mathbf{V}^{(i)}\Big)^\top B_d\mathbf{U}^{(i)}x = 0 \quad\Longrightarrow\quad y^*(x) = \frac{\lambda_1}{2\lambda_2}\sum_{i=1}^n\Big(\mathbf{V}^{(i)}\Big)^\top B_d\mathbf{U}^{(i)}x, \tag{37}$$

so we have

$$\begin{aligned}
&\bar{\Phi}(x) = \bar{f}(x, y^*(x)) \\
={}& \frac{1}{n}\sum_{i=1}^n\left[\frac{\lambda_1^2}{4\lambda_2}\Big\|B_d\mathbf{U}^{(i)}x\Big\|^2 - \frac{\lambda_1^2\sqrt{\alpha}}{2\lambda_2}\Big\langle e_1, \mathbf{U}^{(i)}x\Big\rangle + \frac{\lambda_1^2\alpha}{2n\lambda_2}\Gamma_d^n(x) - \frac{\lambda_1^2\alpha}{4\lambda_2}\Big(u_{d+1}^{(i)}\Big)^2 + \frac{\lambda_1^2\sqrt{\alpha}}{4\lambda_2}\right] \\
={}& \frac{1}{n}\sum_{i=1}^n\left[\frac{\lambda_1^2}{4\lambda_2}\Big\|B_d\mathbf{U}^{(i)}x\Big\|^2 - \frac{\lambda_1^2\sqrt{\alpha}}{2\lambda_2}\Big\langle e_1, \mathbf{U}^{(i)}x\Big\rangle + \frac{\lambda_1^2\alpha}{2n\lambda_2}\sum_{j=1}^n\Gamma_d\Big(\mathbf{U}^{(j)}x\Big) - \frac{\lambda_1^2\alpha}{4\lambda_2}\Big(u_{d+1}^{(i)}\Big)^2 + \frac{\lambda_1^2\sqrt{\alpha}}{4\lambda_2}\right] \\
={}& \frac{1}{n}\sum_{i=1}^n\left[\frac{\lambda_1^2}{4\lambda_2}\Big\|B_d\mathbf{U}^{(i)}x\Big\|^2 - \frac{\lambda_1^2\sqrt{\alpha}}{2\lambda_2}\Big\langle e_1, \mathbf{U}^{(i)}x\Big\rangle + \frac{\lambda_1^2\alpha}{2\lambda_2}\Gamma_d\Big(\mathbf{U}^{(i)}x\Big) - \frac{\lambda_1^2\alpha}{4\lambda_2}\Big(u_{d+1}^{(i)}\Big)^2 + \frac{\lambda_1^2\sqrt{\alpha}}{4\lambda_2}\right] \\
={}& \frac{1}{n}\sum_{i=1}^n\left[\frac{\lambda_1^2}{2\lambda_2}\left(\frac{1}{2}\Big(\mathbf{U}^{(i)}x\Big)^\top A_d\mathbf{U}^{(i)}x - \sqrt{\alpha}\Big\langle e_1, \mathbf{U}^{(i)}x\Big\rangle + \alpha\Gamma_d\Big(\mathbf{U}^{(i)}x\Big) + \frac{1-\alpha}{2}\Big(u_{d+1}^{(i)}\Big)^2 + \frac{\sqrt{\alpha}}{2}\right)\right] \\
={}& \frac{1}{n}\sum_{i=1}^n\Phi_d\Big(\mathbf{U}^{(i)}x\Big),
\end{aligned} \tag{38}$$

where the third equality follows from (16), and A_d and Φ_d are defined in (10) and (9), which concludes the proof. ■

The above two lemmas proves the statements in Lemma 3.3. Before we present the main theorem, we first discuss the behavior of the scaled hard instance, which will be used in the final lower bound analysis.

Lemma C.4 (Primal of the Scaled Finite-Sum Hard Instance) *With the function $\bar{f}(x, y)$ and $\bar{f}_i(x, y)$ defined in* (15), $\bar{\Phi}(x) \triangleq \max_y \bar{f}(x, y)$ *, then for any $\eta \in \mathbb{R}$ and the following function:*

$$f(x, y) = \frac{1}{n}\sum_{i=1}^{n} f_i(x, y) = \frac{1}{n}\sum_{i=1}^{n} \eta^2 \bar{f}_i\left(\frac{x}{\eta}, \frac{y}{\eta}\right), \tag{39}$$

then for its primal function $\Phi(x) \triangleq \max_{y\in\mathbb{R}^{d+1}} f(x, y)$, we have

$$\Phi(x) = \frac{1}{n}\sum_{i=1}^{n} \Phi_i(x), \quad \textit{where} \quad \Phi_i(x) = \eta^2 \bar{\Phi}_i\left(\frac{x}{\eta}\right) = \eta^2 \Phi_d\left(\frac{1}{\eta}\mathbf{U}^{(i)} x\right). \tag{40}$$

Proof Based on (28), we can write out the formulation of f:

$$\begin{aligned}
f(x, y) &= \eta^2 \bar{f}\left(\frac{x}{\eta}, \frac{y}{\eta}\right) \\
&= \eta^2\Bigg(- \frac{\lambda_2}{n}\left\|\frac{y}{\eta}\right\|^2 + \frac{1}{n}\sum_{i=1}^{n}\Bigg[\lambda_1\left\langle B_d \mathbf{U}^{(i)}\frac{x}{\eta}, \mathbf{V}^{(i)}\frac{y}{\eta}\right\rangle - \frac{\lambda_1^2\sqrt{\alpha}}{2\lambda_2}\left\langle e_1, \mathbf{U}^{(i)}\frac{x}{\eta}\right\rangle + \frac{\lambda_1^2\alpha}{2n\lambda_2}\Gamma_d^n\left(\frac{x}{\eta}\right) \\
&\qquad\qquad - \frac{\lambda_1^2\alpha}{4\lambda_2}\left(\frac{u_{d+1}^{(i)}}{\eta}\right)^2 + \frac{\lambda_1^2\sqrt{\alpha}}{4\lambda_2}\Bigg]\Bigg) \\
&= - \frac{\lambda_2}{n}\|y\|^2 + \frac{1}{n}\sum_{i=1}^{n}\lambda_1\left\langle B_d\mathbf{U}^{(i)}x, \mathbf{V}^{(i)}y\right\rangle + \eta^2\Bigg(\frac{1}{n}\sum_{i=1}^{n}\Bigg[- \frac{\lambda_1^2\sqrt{\alpha}}{2\lambda_2}\left\langle e_1, \mathbf{U}^{(i)}\frac{x}{\eta}\right\rangle + \frac{\lambda_1^2\alpha}{2n\lambda_2}\Gamma_d^n\left(\frac{x}{\eta}\right) \\
&\qquad\qquad - \frac{\lambda_1^2\alpha}{4\lambda_2}\left(\frac{u_{d+1}^{(i)}}{\eta}\right)^2 + \frac{\lambda_1^2\sqrt{\alpha}}{4\lambda_2}\Bigg]\Bigg),
\end{aligned} \tag{41}$$

check the gradient over y and set it to be 0 to solve for $y^*(x)$, we have

$$\nabla_y f(x, y^*(x)) = -\frac{2\lambda_2}{n}y^*(x) + \frac{\lambda_1}{n}\sum_{i=1}^{n}\left(\mathbf{V}^{(i)}\right)^\top B_d\mathbf{U}^{(i)}x = 0 \quad\Longrightarrow\quad y^*(x) = \frac{\lambda_1}{2\lambda_2}\sum_{i=1}^{n}\left(\mathbf{V}^{(i)}\right)^\top B_d\mathbf{U}^{(i)}x, \tag{42}$$

which implies that

$$\begin{aligned}
\mathbf{V}^{(i)}y^*(x) &= \frac{\lambda_1}{2\lambda_2}\sum_{j=1}^{n}\mathbf{V}^{(i)}\left(\mathbf{V}^{(j)}\right)^\top B_d\mathbf{U}^{(j)}x = \frac{\lambda_1}{2\lambda_2}B_d\mathbf{U}^{(i)}x \\
\|y^*(x)\|^2 &= \frac{\lambda_1^2}{4\lambda_2^2}\sum_{i=1}^{n}\left\|B_d\mathbf{U}^{(i)}x\right\|^2,
\end{aligned} \tag{43}$$

so the primal function is

$$\begin{aligned}
\Phi(x) &= f(x, y^*(x)) = \eta^2\bar{f}\left(\frac{x}{\eta}, \frac{y^*(x)}{\eta}\right) \\
&= \frac{\lambda_1^2}{4\lambda_2 n}\sum_{i=1}^{n}\left\|B_d\mathbf{U}^{(i)}x\right\|^2 + \frac{\eta^2}{n}\sum_{i=1}^{n}\left[- \frac{\lambda_1^2\sqrt{\alpha}}{2\lambda_2}\left\langle e_1, \mathbf{U}^{(i)}\frac{x}{\eta}\right\rangle + \frac{\lambda_1^2\alpha}{2n\lambda_2}\Gamma_d^n\left(\frac{x}{\eta}\right) - \frac{\lambda_1^2\alpha}{4\lambda_2}\left(\frac{u_{d+1}^{(i)}}{\eta}\right)^2 + \frac{\lambda_1^2\sqrt{\alpha}}{4\lambda_2}\right] \\
&= \frac{\eta^2}{n}\sum_{i=1}^{n}\left[\frac{\lambda_1^2}{4\lambda_2}\left\|B_d\mathbf{U}^{(i)}\frac{x}{\eta}\right\|^2 - \frac{\lambda_1^2\sqrt{\alpha}}{2\lambda_2}\left\langle e_1, \mathbf{U}^{(i)}\frac{x}{\eta}\right\rangle + \frac{\lambda_1^2\alpha}{2n\lambda_2}\Gamma_d^n\left(\frac{x}{\eta}\right) - \frac{\lambda_1^2\alpha}{4\lambda_2}\left(\frac{u_{d+1}^{(i)}}{\eta}\right)^2 + \frac{\lambda_1^2\sqrt{\alpha}}{4\lambda_2}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\eta^2\Phi_d\left(\frac{1}{\eta}\mathbf{U}^{(i)}x\right)\right),
\end{aligned} \tag{44}$$

where the last equality directly applies the conclusion in Lemma C.3, which concludes the proof. ■

C.2.2 Proof of Theorem 3.2

Recall that the complexity for averaged smooth finite-sum nonconvex-strongly-concave problems is defined as

$$\begin{aligned}\mathrm{Compl}_\epsilon\left(\mathcal{F}^{L,\mu,\Delta}_{\mathrm{NCSC}},\mathcal{A},\mathbb{O}^{L,\mathrm{AS}}_{\mathrm{IFO}}\right) &\triangleq \sup_{f\in\mathcal{F}^{L,\mu,\Delta}_{\mathrm{NCSC}}}\ \inf_{\mathtt{A}\in\mathcal{A}\left(\mathbb{O}^{L,\mathrm{AS}}_{\mathrm{IFO}}\right)}\ \mathbb{E}\,T_\epsilon(f,\mathtt{A})\\ &= \sup_{f\in\mathcal{F}^{L,\mu,\Delta}_{\mathrm{NCSC}}}\ \inf_{\mathtt{A}\in\mathcal{A}\left(\mathbb{O}^{L,\mathrm{AS}}_{\mathrm{IFO}}\right)}\ \mathbb{E}\,\inf\left\{T\in\mathbb{N}\,\Big|\,\left\|\nabla\Phi\left(x^T\right)\right\|\le\epsilon\right\}.\end{aligned} \tag{45}$$

Based on the discussion of the properties of the hard instance, we come to the final statement and proof of the theorem.

Theorem C.2 (Lower Bound for Finite-Sum AS NC-SC, Restate Theorem 3.2) *For any linear-span first-order algorithm* $\mathtt{A}\in\mathcal{A}$*, and parameters* $L,\mu,\Delta>0$ *with a desired accuracy* $\epsilon>0$*, for the following function* $f:\mathbb{R}^{(d+1)}\times\mathbb{R}^{(d+1)}\to\mathbb{R}$*:*

$$f_i(x,y)=\eta^2\bar{f}_i\left(\frac{x}{\eta},\frac{y}{\eta}\right),\quad f(x,y)=\frac{1}{n}\sum_{i=1}^n f_i(x,y) \tag{46}$$

where $\bar{f}_i$ *is defined as* (15) *and* $\left\{\mathbf{U}^{(i)}\right\}_{i=1}^n\in\mathbf{Orth}(d+1,(d+1)n,n)$ *is defined in* (13)*, with its primal function* $\Phi(x)\triangleq\max_{y\in\mathbb{R}^{d+1}}f(x,y)$*, for small enough* $\epsilon>0$ *satisfying*

$$\epsilon^2\le\min\left(\frac{\sqrt{\alpha}L^2\Delta}{76800n\mu},\frac{\alpha L^2\Delta}{1280n\mu},\frac{L^2\Delta}{\mu}\right), \tag{47}$$

if we set $L\ge 2n\mu>0$ *and*

$$\lambda^*=\left(\sqrt{\frac{n}{40}}L,\frac{n\mu}{2}\right),\quad \eta=\frac{160\sqrt{2n}\mu}{L^2}\alpha^{-\frac{3}{4}}\epsilon,\quad \alpha=\frac{n\mu}{50L},\quad d=\left\lfloor\frac{\sqrt{\alpha}L^2\Delta}{25600n\mu}\epsilon^{-2}\right\rfloor\ge 3, \tag{48}$$

we have

- *The function* $f\in\mathcal{F}^{L,\mu,\Delta}_{\mathrm{NCSC}}$*,* $\{f_i\}_{i=1}^n$ *is L-averaged smooth.*
- *In the worst case, the algorithm* $\mathcal{A}$ *requires at least* $\Omega\left(n+\sqrt{n\kappa}\Delta L\epsilon^{-2}\right)$ *IFO calls to attain a point* $\hat{x}\in\mathbb{R}^{d+1}$ *such that* $\mathbb{E}\|\nabla\Phi(\hat{x})\|\le\epsilon$*, i.e.,*

$$\mathrm{Compl}_\epsilon\left(\mathcal{F}^{L,\mu,\Delta}_{\mathrm{NCSC}},\mathcal{A},\mathbb{O}^{L,\mathrm{AS}}_{\mathrm{IFO}}\right)=\Omega\left(n+\sqrt{n\kappa}\Delta L\epsilon^{-2}\right). \tag{49}$$

Proof We divide our proof into two cases.

Case 1 The first case builds an $\Omega(n)$ lower bound from a special case of NC-SC function. Consider the following function: $x,y\in\mathbb{R}^d$ and

$$h_i(x,y)\triangleq\theta\langle v_i,x\rangle+L\langle x,y\rangle-\frac{\mu}{2}\|y\|^2,\quad h(x,y)\triangleq\frac{1}{n}\sum_{i=1}^n h_i(x,y), \tag{50}$$

where $\theta\le\sqrt{\frac{2L^2n^2\Delta}{\mu d}}$, $0<\mu\le L$, the dimension number d is set as a multiple of n, and $v_i\in\mathbb{R}^d$ is defined as

$$v_i\triangleq\begin{bmatrix}0&\cdots&0&1&\cdots&1&0&\cdots&0\end{bmatrix}^\top, \tag{51}$$

such that elements with indices from $\frac{i-1}{n}d+1$ to $\frac{i}{n}d$ are 1 and the others are all 0, namely, there are $\frac{d}{n}$ non-zero elements. It is easy to see that the function h_i is μ-strongly convex and L-smooth in both x and y. For the initial value gap, denote $\varphi\triangleq\max_y h$. We have

$$\varphi(x)=\frac{1}{n}\sum_{i=1}^n\left(\theta\langle v_i,x\rangle+\frac{L^2}{2\mu}\|x\|^2\right)=\frac{L^2}{2\mu}\|x\|^2+\frac{\theta}{n}\sum_{i=1}^n\langle v_i,x\rangle, \tag{52}$$

which is a strongly convex function, and its optimal point x^* is

$$x^* = -\frac{\mu\theta}{L^2 n}\sum_{i=1}^{n} v_i, \quad \varphi^* = -\frac{\mu\theta^2}{2L^2n^2}\left\|\sum_{i=1}^{n} v_i\right\|^2. \tag{53}$$

Based on the setting of θ,

$$\varphi(0) - \varphi^* = \frac{\mu\theta^2}{2L^2n^2}\left\|\sum_{i=1}^{n} v_i\right\|^2 = \frac{\mu\theta^2 d}{2L^2n^2} \le \Delta. \tag{54}$$

Hence, we have $h \in \mathcal{F}^{L,\mu,\Delta}_{\mathrm{NCSC}}$. Then based on the expression of $\nabla_x h_i$ and $\nabla_y h_i$, we have that, starting from $(x, y) = (0, 0)$ and denoting $\{i_1, i_2, \cdots, i_t\}$ as the index of IFO sequence for t queries, then the output $(\hat{x}_t, \hat{y}_t)$ will be

$$\hat{x}_t, \hat{y}_t \in \mathrm{Span}\{v_{i_1}, v_{i_2}, \cdots, v_{i_t}\}. \tag{55}$$

then note that each v_i contains only $\frac{d}{n}$ non-zero elements, by the expression of the gradient of the primal function $\nabla\varphi$, we have that if $t \le n/2$, then there must be at least $\frac{n}{2} \times \frac{d}{n} = \frac{d}{2}$ zero elements in $\hat{x}_t$, which implies that for $\epsilon^2 \le \frac{L^2\Delta}{\mu}$,

$$\|\nabla\varphi(\hat{x}_t)\| = \left\|\frac{L^2}{\mu}\hat{x}_t + \frac{\theta}{n}\sum_{i=1}^{n} v_i\right\| \ge \frac{\theta}{n}\sqrt{\frac{d}{2}} \ge \epsilon, \tag{56}$$

where we follow the setting of θ above. So we proved that it requires $\Omega(n)$ IFO calls to find an ϵ-stationary point.

Case 2 The second case provides an $\Omega(\sqrt{n\kappa}\Delta L\epsilon^{-2})$ lower bound concerning the second term in the result. Throughout the case, we assume $L \ge 2n\mu > 0$ as that in Lemma C.2.

Here we still use the hard instance constructed in (15), note that $\nabla f_i(x, y) = \eta\nabla\bar{f}_i\left(\frac{x}{\eta}, \frac{y}{\eta}\right)$ is a scaled version of $\bar{f}_i$, which is L-averaged smooth by Lemma C.2, so by Lemma B.3 we have $\{f_i\}_i$ is also L-average smooth. The for the strong concavity, note that $\bar{f}$ is μ-strongly concave on y, so as the scaled version, f is also μ-strongly concave on y.

Then for the primal function of f, let $\Phi(x) \triangleq \max_y f(x, y)$, by Lemma C.3 and Lemma C.4, we have

$$\Phi(x) = \eta^2\bar{\Phi}\left(\frac{x}{\eta}\right) = \frac{1}{n}\sum_{i=1}^{n}\eta^2\bar{\Phi}_i\left(\frac{x}{\eta}\right), \tag{57}$$

where $\bar{\Phi}$ and $\bar{\Phi}_i$ follow the definition in Lemma C.3,

We first justify the lower bound argument by lower bounding the norm of the gradient. Recall the definition of $\mathcal{I}$ (see (18)), which is the index set such that $u_d^{(i)} = u_{d+1}^{(i)} = 0$, $\forall i \in \mathcal{I}$ while $u^{(i)} = \mathbf{U}^{(i)}x$. By substituting the parameters in the statement above into (18) and Lemma 3.2, we have that when the size of the set $\mathcal{I}$, i.e., $|\mathcal{I}| > n/2$ (note that scaling does not affect the activation status),

$$\begin{aligned}\|\nabla\Phi(x)\|^2 &= \left\|\eta\nabla\bar{\Phi}\left(\frac{x}{\eta}\right)\right\|^2 = \eta^2\left\|\nabla\bar{\Phi}\left(\frac{x}{\eta}\right)\right\|^2 \\ &\ge \frac{51200n\mu^2}{L^4}\alpha^{-\frac{3}{2}}\epsilon^2 \cdot \frac{\lambda_1^4}{128n\lambda_2^2}\alpha^{\frac{3}{2}} \\ &= \frac{51200n\mu^2}{L^4}\alpha^{-\frac{3}{2}}\epsilon^2 \cdot \frac{L^4}{51200n\mu^2}\alpha^{\frac{3}{2}} = \epsilon^2.\end{aligned} \tag{58}$$

Next, we upper bound the starting optimality gap. By substitution of parameter settings and the initial gap of $\bar{\Phi}$ in (17), also recall the setting of ϵ, we have

$$\begin{aligned}\Phi(0) - \Phi^* &= \eta^2\left(\bar{\Phi}(0) - \inf_{x\in\mathbb{R}^{d+1}}\bar{\Phi}(x)\right) = \frac{51200n\mu^2}{L^4}\alpha^{-\frac{3}{2}}\epsilon^2 \cdot \frac{nL^2}{40n\mu}\left(\frac{\sqrt{\alpha}}{2} + 10\alpha d\right) \\ &= \frac{1280n\mu}{L^2}\left(\frac{1}{2\alpha} + \frac{10d}{\sqrt{\alpha}}\right)\epsilon^2 = \frac{640n\mu\epsilon^2}{\alpha L^2} + \frac{12800n\mu d\epsilon^2}{L^2\sqrt{\alpha}} \\ &\le \frac{640n\mu}{\alpha L^2} \cdot \frac{\alpha L^2\Delta}{1280n\mu} + \frac{12800n\mu\epsilon^2}{L^2\sqrt{\alpha}} \cdot \frac{\sqrt{\alpha}L^2\Delta}{25600n\mu}\epsilon^{-2} \\ &\le \frac{\Delta}{2} + \frac{\Delta}{2} = \Delta,\end{aligned} \tag{59}$$

so we conclude that $f \in \mathcal{F}^{L,\mu,\Delta}_{\text{NCSC}}$, i.e. the function class requirement is satisfied.

To show the lower bound, by previous analysis and the choice of (13), the activation process for each component will also mimic the "alternating zero-chain" mechanism (see Lemma 3.1) independently. So we have, by the lower bound argument (18), it requires to activate at least half of the components through until their d-th elements (or at least half of $\{u^{(i)}\}_i$ are not activated through until the d-th element, note that each $u^{(i)}$ corresponds to an unique part of x with length $(d+1)$) for the primal stationarity convergence of the objective function, which takes (note that $2\lfloor x \rfloor - 1 \geq x$ when $x \geq 3$)

$$T = \frac{n}{2}(2d-1) \geq \frac{n}{2} \cdot \frac{\sqrt{\alpha} L^2 \Delta}{25600 n\mu}\epsilon^{-2} = \Omega\big(\sqrt{\alpha}\Delta L\kappa\epsilon^{-2}\big) = \Omega\big(\sqrt{n\kappa}\Delta L\epsilon^{-2}\big) \tag{60}$$

IFO oracle queries. So we found that for any fixed index sequence $\{i_t\}_{t=1}^{T}$, the output z^{T+1} from a randomized algorithm[3] must not be an approximate stationary point, which verifies the $\Omega\big(n \vee \sqrt{n\kappa}\Delta L\epsilon^{-2}\big)$ or $\Omega\big(n + \sqrt{n\kappa}\Delta L\epsilon^{-2}\big)$ lower bound by combining the two cases discussed above together. We conclude the proof by applying Yao's minimax theorem [Yao, 1977], the lower bound will also hold for a randomized index sequence incurred by IFOs. ■

D PROOF OF NC-SC CATALYST

D.1 OUTER-LOOP COMPLEXITY

In this section, we first introduce a few useful definitions. The Moreau envelop of a function F with a positive parameter $\lambda > 0$ is:

$$F_\lambda(x) = \min_{z\in\mathbb{R}^{d_1}} F(z) + \frac{1}{2\lambda}\|z-x\|^2.$$

We also define the proximal point of x:

$$\text{prox}_{\lambda F}(x) = \operatorname*{arg\,min}_{z\in\mathbb{R}^{d_1}} \left\{ F(z) + \frac{1}{2\lambda}\|z-x\|^2 \right\}.$$

When F is differentiable and ℓ-weakly convex, for $\lambda \in (0, 1/\ell)$ we have

$$\nabla F(\text{prox}_{\lambda F}(x)) = \nabla F_\lambda(x) = \lambda^{-1}(x - \text{prox}_{\lambda F}(x)). \tag{61}$$

Thus a small gradient $\|\nabla F_\lambda(x)\|$ implies that x is near a point $\text{prox}_{\lambda F}(x)$ that is nearly stationary for F. Therefore $\|\nabla F_\lambda(x)\|$ is also commonly used as a measure of stationarity. We refer readers to [Drusvyatskiy and Paquette, 2019] for more discussion on Moreau envelop.

In this subsection, we use (x^t, y^t) as a shorthand for (x_0^t, y_0^t). We will denote $(\hat{x}^t, \hat{y}^t)$ as the optimal solution to the auxiliary problem $(\star)$ at t-th iteration: $\min_{x\in\mathbb{R}^{d_1}}\max_{y\in\mathbb{R}^{d_2}}\left[\hat{f}_t(x,y) \triangleq f(x,y) + L\|x-x^t\|^2\right]$. It is easy to observe that $\hat{x}^t = \text{prox}_{\Phi/2L}(x^t)$. Define $\hat{\Phi}_t(x) = \max_y f(x,y) + L\|x-x^t\|^2$. In the following theorem, we show the convergence of the Moreau envelop $\|\nabla\Phi_{1/2L}(x)\|^2$ when we replace the inexactness measure (20) by another inexactness measure $\text{gap}_{\hat{f}_t}(x^{t+1}, y^{t+1}) \leq \beta_t(\|x^t - \hat{x}^t\| + \|y^t - \hat{y}^t\|^2)$. Later we will show this inexactness measure can be implied by (20) with our choice of β_t and α_t.

Theorem D.1 *Suppose function f is NC-SC with strong convexity parameter μ and L-Lipschitz smooth. If we replace the stopping criterion (20) by $\text{gap}_{\hat{f}_t}(x^{t+1}, y^{t+1}) \leq \beta_t(\|x^t - \hat{x}^t\|^2 + \|y^t - \hat{y}^t\|^2)$ with $\beta_t = \frac{\mu^4}{28L^3}$ for $t > 0$ and $\beta_0 = \frac{\mu^4}{32L^4\max\{1,L\}}$, then iterates from Algorithm 1 satisfy*

$$\sum_{t=0}^{T-1}\|\nabla\Phi_{1/2L}(x^t)\|^2 \leq \frac{87L}{5}\Delta_0 + \frac{7L}{5}D_y^0, \tag{62}$$

where $D_y^0 = \|y^0 - y^*(x^0)\|^2$ and $\Delta_0 = \Phi(x^0) - \inf_x\Phi(x)$.

[3] Note that randomization does not affect the lower bound, as long as the algorithm satisfies the linear-span assumption.

Proof Define $b_{t+1} = \mathrm{gap}_{\hat{f}_t}(x^{t+1}, y^{t+1})$. By Lemma 4.3 in [Drusvyatskiy and Paquette, 2019],

$$\begin{aligned}\|\nabla \Phi_{1/2L}(x^t)\|^2 = 4L^2\|x^t - \mathrm{prox}_{\Phi/2L}(x^t)\|^2 \leq& 8L[\hat{\Phi}_t(x^t) - \hat{\Phi}_t(\mathrm{prox}_{\Phi/2L}(x^t))] \\ \leq& 8L[\hat{\Phi}_t(x^t) - \hat{\Phi}_t(x^{t+1}) + b_{t+1}] \\ =& 8L\big\{\Phi(x^t) - \big[\Phi(x^{t+1}) + L\|x^{t+1} - x^t\|^2\big] + b_{t+1}\big\} \\ \leq& 8L[\Phi(x^t) - \Phi(x^{t+1}) + b_{t+1}], \end{aligned} \tag{63}$$

where in the first inequality we use L-strongly convexity of $\hat{\Phi}_t$. Then, for $t \geq 1$

$$\begin{aligned}\|y^t - \hat{y}^t\|^2 \leq& 2\|y^t - \hat{y}^{t-1}\|^2 + 2\|y^*(\hat{x}^{t-1}) - y^*(\hat{x}^t)\|^2 \\ \leq& 2\|y^t - \hat{y}^{t-1}\|^2 + 2\left(\frac{L}{\mu}\right)^2\|\hat{x}^t - \hat{x}^{t-1}\|^2 \\ \leq& 2\|y^t - \hat{y}^{t-1}\|^2 + 4\left(\frac{L}{\mu}\right)^2\|\hat{x}^t - x^t\|^2 + 4\left(\frac{L}{\mu}\right)^2\|x^t - \hat{x}^{t-1}\|^2 \\ \leq& \frac{8L}{\mu^2}b_t + 4\left(\frac{L}{\mu}\right)^2\|\hat{x}^t - x^t\|^2, \end{aligned}$$

where we use Lemma B.1 in the second inequality, and (L, μ)-SC-SC of $\tilde{f}_{t-1}(x, y)$ and Lemma B.2 in the last inequality. Therefore,

$$\|x^t - \hat{x}^t\|^2 + \|y^t - \hat{y}^t\|^2 \leq \frac{8L}{\mu^2}b_t + \left(\frac{4L^2}{\mu^2} + 1\right)\|\hat{x}^t - x^t\|^2. \tag{64}$$

By our stopping criterion and $\|\nabla\Phi_{1/2L}(x^t)\|^2 = 4L^2\|x^t - \hat{x}^t\|^2$, for $t \geq 1$

$$b_{t+1} \leq \beta_t\left[\|x_t - \hat{x}^t\|^2 + \|y_t - \hat{y}^t\|^2\right] \leq \frac{8L\beta_t}{\mu^2}b_t + \beta_t\left(\frac{1}{\mu^2} + \frac{1}{4L^2}\right)\|\nabla\Phi_{1/2L}(x^t)\|^2.$$

Define $\theta = \frac{2}{7}$ and $w = \frac{5\mu^2}{112L^3}$. It is easy to verify that as $\beta_t = \frac{\mu^4}{28L^3}$, then $\frac{8L\beta_t}{\mu^2} \leq \theta$ and $\beta_t\left(\frac{1}{\mu^2} + \frac{1}{4L^2}\right) \leq w$. We conclude the following recursive bound

$$b_{t+1} \leq \theta b_t + w\|\nabla\Phi_{1/2L}(x^t)\|^2. \tag{65}$$

For $t = 0$,

$$\|y^0 - \hat{y}^0\|^2 \leq 2\|y^0 - y^*(x^0)\|^2 + 2\|\hat{y}^0 - y^*(x^0)\|^2 \leq 2\|y^0 - y^*(x^0)\|^2 + 2\left(\frac{L}{\mu}\right)^2\|x^0 - \hat{x}^0\|^2. \tag{66}$$

Because $\Phi(x) + L\|x - x^0\|^2$ is L-strongly convex, we have

$$\left(\Phi(\hat{x}^0) + L\|\hat{x}^0 - x^0\|^2\right) + \frac{L}{2}\|\hat{x}^0 - x^0\|^2 \leq \Phi(x^0) = \Phi^* + (\Phi(x^0) - \Phi^*) \leq \Phi(\hat{x}^0) + (\Phi(x^0) - \Phi^*).$$

This implies $\|\hat{x}^0 - x^0\|^2 \leq \frac{L}{2}(\Phi(x^0) - \Phi^*)$. Then combining with (66)

$$\|y^0 - \hat{y}^0\|^2 + \|x^0 - \hat{x}^0\|^2 \leq \left(\frac{L^3}{\mu^2} + \frac{L}{2}\right)(\Phi(x^0) - \Phi^*) + 2\|y^0 - y^*(x^0)\|^2.$$

Hence, by the stopping criterion,

$$b_1 \leq \beta_0\left(\frac{L^3}{\mu^2} + \frac{L}{2}\right)(\Phi(x^0) - \Phi^*) + 2\beta_0\|y^0 - y^*(x^0)\|^2.$$

Define $\theta_0 = \frac{\mu^2}{16L^2}$. With $\beta_0 = \frac{\mu^4}{32L^4\max\{1,L\}}$, $\beta_0\left(\frac{L^3}{\mu^2} + \frac{L}{2}\right) \leq \theta_0$ and $2\beta_0 \leq \theta_0$. So we can write

$$b_1 \leq \theta_0(\Phi(x^0) - \Phi^*) + \theta_0\|y^0 - y^*(x^0)\|^2.$$

Unravelling (65), we have for $t \geq 1$

$$b_{t+1} \leq \theta^t b_1 + w \sum_{k=1}^{t} \theta^{t-k} \|\nabla \Phi_{1/2L}(x_k)\|^2 \leq \theta^t \theta_0 (\Phi(x^0) - \Phi^*) + \theta^t \theta_0 \|y^0 - y^*(x^0)\|^2 + w \sum_{k=1}^{t} \theta^{t-k} \|\nabla \Phi_{1/2L}(x_k)\|^2. \tag{67}$$

Summing from $t = 0$ to $T - 1$,

$$\begin{aligned} \sum_{t=0}^{T-1} b_{t+1} &= \sum_{t=1}^{T-1} b_t + b_1 \\ &\leq \theta_0 \sum_{t=0}^{T-1} \theta^t [\Phi(x^0) - \Phi^*] + \theta_0 \sum_{t=0}^{T-1} \theta^t \|y^0 - y^*(x^0)\|^2 + w \sum_{t=1}^{T-1} \sum_{k=1}^{t} \theta^{t-k} \|\nabla \Phi_{1/2L}(x_k)\|^2 \\ &\leq \theta_0 \sum_{t=0}^{T-1} \theta^t [\Phi(x^0) - \Phi^*] + \theta_0 \sum_{t=0}^{T-1} \theta^t \|y^0 - y^*(x^0)\|^2 + w \sum_{t=1}^{T-1} \frac{1}{1-\theta} \|\nabla \Phi_{1/2L}(x^t)\|^2, \end{aligned} \tag{68}$$

where we use $\sum_{t=1}^{T-1} \sum_{k=1}^{t} \theta^{t-k} \|\nabla \Phi_{1/2L}(x_k)\|^2 = \sum_{k=1}^{T-1} \sum_{t=k}^{T} \theta^{t-k} \|\nabla \Phi_{1/2L}(x_k)\|^2 \leq \sum_{k=1}^{T-1} \frac{1}{1-\theta} \|\nabla \Phi_{1/2L}(x_k)\|^2$. Now, by telescoping (63),

$$\frac{1}{8L} \sum_{t=0}^{T-1} \|\nabla \Phi_{1/2L}(x^t)\|^2 \leq \Phi(x^0) - \Phi^* + \sum_{t=0}^{T-1} b_{t+1}.$$

Plugging (68) in,

$$\frac{1}{8L} \sum_{t=0}^{T-1} \|\nabla \Phi_{1/2L}(x^t)\|^2 - w \sum_{t=1}^{T-1} \frac{1}{1-\theta} \|\nabla \Phi_{1/2L}(x^t)\|^2 \leq \left(1 + \frac{\theta_0}{1-\theta}\right) [\Phi(x^0) - \Phi^*] + \frac{\theta_0}{1-\theta} \|y^0 - y^*(x^0)\|^2. \tag{69}$$

Plugging in $w \leq \frac{5}{112L}$, $\frac{1}{1-\theta} = \frac{7}{5}$ and $\theta_0 \leq \frac{1}{16}$

$$\frac{1}{16L} \sum_{t=0}^{T-1} \|\nabla \Phi_{1/2L}(x^t)\|^2 \leq \frac{87}{80} [\Phi(x^0) - \Phi^*] + \frac{7}{80} \|y^0 - y^*(x^0)\|^2.$$

■

Proof of Theorem 4.1

Proof We first show that criterion (20) implies the criterion in Theorem D.1. By Lemma B.2, as $\hat{f}_t$ is (L, μ)-SC-SC and $3L$-smooth,

$$2\mu \operatorname{gap}_{\hat{f}_t}(x^{t+1}, y^{t+1}) \leq \|\nabla \hat{f}_t(x^{t+1}, y^{t+1})\|^2 \leq \alpha_t \|\nabla \hat{f}_t(x^t, y^t)\|^2 \leq 36L^2 \alpha_t (\|x^t - \hat{x}^t\|^2 + \|y^t - \hat{y}^t\|^2),$$

therefore,

$$\operatorname{gap}_{\hat{f}_t}(x^{t+1}, y^{t+1}) \leq \frac{18L^2 \alpha_t}{\mu} (\|x^t - \hat{x}^t\|^2 + \|y^t - \hat{y}^t\|^2),$$

which implies $\operatorname{gap}_{\hat{f}_t}(x^{t+1}, y^{t+1}) \leq \beta_t (\|x^t - \hat{x}^t\|^2 + \|y^t - \hat{y}^t\|^2)$ by our choice of $\{\beta_t\}_t$ and $\{\alpha_t\}_t$.

We still use $b_{t+1} = \operatorname{gap}_{\hat{f}_t}(x^{t+1}, y^{t+1})$ as in the proof of Theorem (D.1). First, note that

$$\begin{aligned} \|\nabla \Phi(x^{t+1})\|^2 &\leq 2\|\nabla \Phi(x^{t+1}) - \nabla \Phi(\hat{x}^t)\|^2 + 2\|\nabla \Phi(\hat{x}^t)\|^2 \\ &\leq 2\left(\frac{2L^2}{\mu}\right) \|x^{t+1} - \hat{x}^t\|^2 + 2\|\nabla \Phi_{1/2L}(x^t)\|^2 \\ &\leq \frac{16L^3}{\mu^2} b_{t+1} + 2\|\nabla \Phi_{1/2L}(x^t)\|^2. \end{aligned} \tag{70}$$

where in the second inequality we use Lemma B.1 and Lemma 4.3 in [Drusvyatskiy and Paquette, 2019]. Summing from $t = 0$ to $T - 1$, we have

$$\sum_{t=0}^{T-1} \|\nabla\Phi(x^{t+1})\|^2 \leq \frac{16L^3}{\mu^2} \sum_{t=0}^{T-1} b_{t+1} + 2 \sum_{t=0}^{T-1} \|\nabla\Phi_{1/2L}(x^t)\|^2. \tag{71}$$

Applying (68), we have

$$\frac{16L^3}{\mu^2} \sum_{t=0}^{T-1} b_{t+1} \leq \frac{16L^3\theta_0}{\mu^2} \sum_{t=0}^{T-1} \theta^t [\Phi(x^0) - \Phi^*] + \frac{16L^3\theta_0}{\mu^2} \sum_{t=0}^{T-1} \theta^t \|y^0 - y^*(x^0)\|^2 + \frac{16L^3 w}{\mu^2} \sum_{t=1}^{T-1} \frac{1}{1-\theta} \|\nabla\Phi_{1/2L}(x^t)\|^2.$$

Plugging in $\theta_0 = \frac{\mu^2}{16L^2}$, $\theta = \frac{2}{7}$ and $w = \frac{5\mu^2}{112L^3}$,

$$\frac{16L^3}{\mu^2} \sum_{t=0}^{T-1} b_{t+1} \leq \frac{7L}{5} [\Phi(x^0) - \Phi^*] + \frac{7L}{5} \|y^0 - y^*(x^0)\|^2 + \sum_{t=1}^{T-1} \|\nabla\Phi_{1/2L}(x^t)\|^2.$$

Plugging back into (71),

$$\sum_{t=0}^{T-1} \|\nabla\Phi(x^{t+1})\|^2 \leq \frac{7L}{5} [\Phi(x^0) - \Phi^*] + \frac{7L}{5} \|y^0 - y^*(x^0)\|^2 + 3 \sum_{t=0}^{T-1} \|\nabla\Phi_{1/2L}(x^t)\|^2.$$

Applying Theorem D.1,

$$\frac{1}{T} \sum_{t=1}^{T} \|\nabla\Phi(x^{t+1})\|^2 \leq \frac{268L}{5T} [\Phi(x^0) - \Phi^*] + \frac{28L}{5T} \|y^0 - y^*(x^0)\|^2.$$

■

D.2 COMPLEXITY OF SOLVING AUXILIARY PROBLEM ($\star$) AND PROOF OF THEOREM 4.2

In this layer, we apply an inexact proximal point algorithm to solve the (L, μ)-SC-SC and $3L$-smooth auxiliary problem: $\min_x \max_y \hat{f}_t(x, y)$. Throughout this subsection, we suppress the outer-loop index t without confusion, i.e. we use $\hat{f}$ instead of $\hat{f}_t$ and $\tilde{f}_k = \hat{f}(x, y) - \frac{\tau}{2}\|y - z_k\|^2$ instead of $\tilde{f}_{t,k}$. Accordingly, we also omit the superscript in (x_k^t, y_k^t) and ϵ_k^t.

Before we prove Theorem 4.2, we present a lemma from [Lin et al., 2018]. The inner loop to solve ($\star$) can be considered as applying Catalyst for strongly-convex minimization in [Lin et al., 2018] to the function $-\hat{\Psi}(y) = -\min_{x\in\mathbb{R}^d} \hat{f}(x, y)$. The following lemma captures the convergence of Catalyst framework in minimization, which we present in Algorithm 1.

Lemma D.1 ([Lin et al., 2018]) *Consider the problem $\min_{x\in\mathbb{R}^d} h(x)$. Assume function h is μ-strongly convex. Define $A_k = \prod_{i=1}^k (1 - \alpha_i)$, $\eta_k = \frac{\alpha_k - q}{1-q}$ and a sequence $\{v_t\}_t$ with $v_0 = x_0$ and $v_k = x_{k-1} + \frac{1}{\alpha_k}(x_k - x_{k-1})$ for $k > 1$. Consider the potential function: $S_k = h(x_k) - h(x^*) + \frac{\eta_{k+1}\alpha_{k+1}\tau}{2(1-\alpha_{k+1})}\|x^* - v_k\|^2$, where x^* is the optimal solution. After running Algorithm 1 for K iterations, we have*

$$\frac{1}{A_K} S_K \leq \left(\sqrt{S_0} + 2 \sum_{t=1}^{K} \sqrt{\frac{\epsilon_k}{A_k}} \right)^2. \tag{73}$$

Before we step into the proof of Theorem 4.2, we introduce several notations. We denote the dual function of $\hat{f}$ by $\hat{\Psi}(y) = \min_x \hat{f}(x, y)$. We denote the dual function of $\tilde{f}_k(x, y)$ by $\tilde{\Psi}_k(y) = \min_x \tilde{f}_k(x, y) = \min_x \hat{f}(x, y) - \frac{\tau}{2}\|y - z_k\|^2 = \hat{\Psi}(y) - \frac{\tau}{2}\|y - z_k\|^2$. Let $y_k^* = \arg\max_y \tilde{\Psi}_k(y)$. We also define (x^*, y^*) as the optimal solution to $\min_x \max_y \hat{f}(x, y)$

Algorithm 1 Catalyst for Strongly-Convex Minimization

```
Input: function $h$, initial point $x_0$, strong-convexity constant $\mu$, parameter $\tau > 0$
Initialization: $q = \frac{\mu}{\mu+\tau}$, $z_1 = x_0$, $\alpha_1 = \sqrt{q}$.
for all $k = 1, 2, ..., K$ do
    Find an inexact solution $x_k$ to the following problem with algorithm $\mathcal{M}$
        $$\min_{x \in \mathbb{R}^d} \tilde{h}_k(x) \triangleq \left[ h(x) + \frac{\tau}{2} \|x - z_k\|^2 \right]$$
    such that
        $$\tilde{h}_k(x_k) - \min_{x \in \mathbb{R}^d} \tilde{h}_k(x) \leq \epsilon_k. \qquad (72)$$
    Choose $\alpha_{k+1} \in [0, 1]$ such that $\alpha_{k+1}^2 = (1 - \alpha_{k+1})\alpha_k^2 + q\alpha_{k+1}$.
    $z_{k+1} = x_k + \beta_k(x_k - x_{k-1})$ where $\beta_k = \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2+\alpha_{k+1}}$.
end for
Output: $x_K$.
```

Proof of Theorem 4.2

Proof When the criterion $\|\nabla \tilde{f}_k(x^k, y^k)\|^2 \leq \epsilon_k$ is satisfied, by Lemma B.2,

$$\operatorname{gap}_{\tilde{f}_k}(x_k, y_k) \leq \frac{1}{2\mu} \|\nabla \tilde{f}_k(x^k, y^k)\|^2 \leq \frac{1}{2\mu} \epsilon_k = \frac{\sqrt{2}}{4}(1 - \rho)^k \operatorname{gap}_{\hat{f}}(x_0, y_0) = \hat{\epsilon}_k,$$

where we define $\hat{\epsilon}_k = \frac{\sqrt{2}}{4}(1 - \rho)^k \operatorname{gap}_{\hat{f}}(x_0, y_0)$.

The auxiliary problem ($\star\star$) can be considered as $\max_y \hat{\Psi}(y)$. We see $\operatorname{gap}_{\tilde{f}_k}(x_k, y_k) \leq \hat{\epsilon}_k$ implies $\max_y \tilde{\Psi}_k(y) - \tilde{\Psi}_k(y_k) \leq \hat{\epsilon}_k$. By choosing $\alpha_1 = \sqrt{q}$ in Algorithm 1, it is easy to check that $\alpha_k = \sqrt{q}$ and $\beta_k = \frac{\sqrt{q}-q}{\sqrt{q}+q}$, for all k. So this inner loop can be considered as applying Algorithm 1 to $-\tilde{\Psi}(y)$ and Lemma D.1 can guarantee the convergence of the dual function. Define $S_k = \hat{\Psi}(y^*) - \hat{\Psi}(y_k) + \frac{\eta_{t+1}\alpha_{t+1}\tau}{2(1-\alpha_{t+1})} \|y^* - v_k\|^2$ with $\eta_k = \frac{\alpha_k - q}{1-q}$. Lemma D.1 gives rise to

$$\frac{1}{A_K} S_K \leq \left(\sqrt{S_0} + 2 \sum_{k=1}^{K} \sqrt{\frac{\hat{\epsilon}_k}{A_k}} \right)^2. \qquad (74)$$

Note that $A_k = \prod_{i=1}^{k} (1 - \alpha_i) = (1 - \sqrt{q})^k$ and

$$\frac{\eta_k \alpha_k \tau}{2(1 - \alpha_k)} = \frac{\sqrt{q} - q}{1 - q} \frac{\sqrt{q}\tau}{2(1 - \sqrt{q})} = \frac{\sqrt{q} - q}{\tau/(\mu + \tau)} \frac{\sqrt{q}\tau}{2(1 - \sqrt{q})} = \frac{q(\mu + \tau)}{2} = \frac{\mu}{2}.$$

So $S_0 = \hat{\Psi}(y^*) - \hat{\Psi}(y_0) + \frac{\mu}{2} \|y^* - y_0\|^2 \leq 2(\hat{\Psi}(y^*) - \hat{\Psi}(y_0))$. Then with $\epsilon_k = \frac{\sqrt{2}}{4}(1 - \rho)^k \operatorname{gap}_{\hat{f}}(x_0, y_0)$, and we have

$$\text{Right-hand side of } (74) \leq \left(\sqrt{2(\hat{\Psi}(y^*) - \hat{\Psi}(y_0))} + \sum_{t=1}^{T} \sqrt{2 \left(\frac{1 - \rho}{1 - \sqrt{q}} \right)^t \operatorname{gap}_{\hat{f}}(x_0, y_0)} \right)^2 \qquad (75)$$

$$\leq 2 \left(1 + \sum_{k=1}^{K} \left(\sqrt{\frac{1 - \rho}{1 - \sqrt{q}}} \right)^k \right)^2 \operatorname{gap}_{\hat{f}}(x_0, y_0) \qquad (76)$$

$$\leq 2 \left(\frac{\left(\sqrt{\frac{1-\rho}{1-\sqrt{q}}} \right)^{K+1}}{\sqrt{\frac{1-\rho}{1-\sqrt{q}}} - 1} \right)^2 \operatorname{gap}_{\hat{f}}(x_0, y_0) \leq 2 \left(\frac{\sqrt{\frac{1-\rho}{1-\sqrt{q}}}}{\sqrt{\frac{1-\rho}{1-\sqrt{q}}} - 1} \right)^2 \left(\frac{1 - \rho}{1 - \sqrt{q}} \right)^K \operatorname{gap}_{\hat{f}}(x_0, y_0). \qquad (77)$$

Plugging back into (74),

$$S_K \leq 2\left(\frac{1}{\sqrt{1-\rho}-\sqrt{1-\sqrt{q}}}\right)^2 (1-\rho)^{K+1}\operatorname{gap}_{\hat{f}}(x_0,y_0) \leq \frac{8}{(\sqrt{q}-\rho)^2}(1-\rho)^{K+1}\operatorname{gap}_{\hat{f}}(x_0,y_0), \tag{78}$$

where the second inequality is due to $\sqrt{1-x}+\frac{x}{2}$ is decreasing in $[0,1]$. Note that

$$\begin{aligned}\|x_K-x^*\|^2 \leq& 2\|x_K-x^*(y_K)\|^2+2\|x^*(y_K)-x^*(y^*)\|^2\\ \leq& \frac{4}{L}[\hat{f}(x_K,y_K)-\hat{f}(x^*(y_K),y_K)]+18\|y_K-y^*\|^2\\ \leq& \frac{4}{L}\hat{\epsilon}_K+18\|y_K-y^*\|^2.\end{aligned} \tag{79}$$

where in the second inequality we use Lemma B.1. Then,

$$\|x_K-x^*\|^2+\|y_K-y^*\|^2 \leq 19\|y_K-y^*\|^2+\frac{4}{L}\hat{\epsilon}_K. \tag{80}$$

Because $\|y_K-y^*\|^2 \leq \frac{2}{\mu}[\hat{\Psi}(y^*)-\hat{\Psi}(y_K)] \leq \frac{2}{\mu}S_K$, by plugging in (78) and the definition of $\hat{\epsilon}_k$, we get

$$\|x_K-x^*\|^2+\|y_K-y^*\|^2 \leq \left(\frac{306}{\mu(\sqrt{q}-\rho)^2}+\frac{\sqrt{2}}{L}\right)(1-\rho)^K\operatorname{gap}_{\hat{f}}(x_0,y_0).$$

By Lemma B.2, we have

$$\|x_K-x^*\|^2+\|y_K-y^*\|^2 \geq \frac{1}{36L^2}\|\nabla\hat{f}(x_K,y_K)\|^2 \quad \text{and} \quad \operatorname{gap}_{\hat{f}}(x_0,y_0) \leq \frac{1}{2\mu}\|\nabla\hat{f}(x_0,y_0)\|^2.$$

Then we finish the proof. ■

D.3 COMPLEXITY OF SOLVING SUBPROBLEM (⋆⋆) AND PROOF OF THEOREM 4.3

As in the previous subsection, we suppress the outer-loop index t. Define $\hat{\Phi}(x)=\max_y \hat{f}(x,y)$, $\hat{\Psi}(y)=\min_x \hat{f}(x,y)$ and $\hat{\Phi}^*=\min_x\hat{\Phi}(x)=\max_y\hat{\Psi}(y)=\hat{\Psi}^*$. We still define $\tilde{\Psi}_k(y)=\min_x \tilde{f}_k(x,y)=\min_x \hat{f}(x,y)-\frac{\tau}{2}\|y-z_k\|^2=\hat{\Psi}(y)-\frac{\tau}{2}\|y-z_k\|^2$, and $\tilde{\Phi}_k(x)=\max_y \tilde{f}_k(x,y)$. Let (x^*,y^*) be the optimal solution to $\min_x\max_y \hat{f}(x,y)$ and (x_k^*,y_k^*) the optimal solution to $\min_x\max_y \tilde{f}_k(x,y)$. Also, in this subsection, we denote $x^*(y)=\arg\min_x \hat{f}(x,y)$ and $y^*(x)=\arg\max_y \hat{f}(x,y)$. Recall that we defined a potential function $S_k=\hat{\Psi}(y^*)-\hat{\Psi}(y_k)+\frac{\mu}{2}\|y^*-v_k\|^2$ in the proof of Theorem 4.2.

The following lemma shows that the initial point we choose to solve (⋆⋆) for $\mathcal{M}$ at iteration k is not far from the optimal solution of (⋆⋆) if the stopping criterion is satisfied for every iterations before k.

Lemma D.2 (Initial distance of the warm-start) *Under the same assumptions as Theorem 4.2, with accuracy ϵ_k specified in Theorem 4.2, we assume that for $\forall i<k$, $\|\nabla\tilde{f}_i(x_i,y_i)\|^2\leq\epsilon_i$. At iteration k, solving the subproblem (⋆⋆) from initial point (x_{k-1},y_{k-1}), we have*

$$\|x_{k-1}-x_k^*\|^2+\|y_{k-1}-y_k^*\|^2 \leq C_k\epsilon_k,$$

where $C_1=\left[\frac{72\sqrt{2}}{\mu^2}+\frac{74\sqrt{2}}{(2\tau+\mu)\mu}\right]\frac{1}{1-\rho}$, $C_t=\frac{2}{\mu\min\{L,\mu+\tau\}}\frac{1}{1-\rho}+\frac{288\sqrt{2}\tau^2\max\{40L^2,9\tau^2+4L^2\}}{(\mu+\tau)^2L^2\mu^2(\sqrt{q}-\rho)^2)}\frac{1}{(1-\rho)^2}$ for $t>1$.

Proof We separate the proof into two cases: $k=1$ and $k>1$.
Case $k=1$**:** Note that $z_1=y_0$, and therefore the subproblem at the first iteration is

$$\min_x\max_y\left[\tilde{f}_1(x,y)=\hat{f}(x,y)-\frac{\tau}{2}\|y-y_0\|^2\right]. \tag{81}$$

Since $x_1^* = \arg\min_x \tilde{f}_1(x, y_1^*) = \arg\min_x \hat{f}(x, y_1^*)$ and $x^* = \arg\min_x \hat{f}(x, y^*)$, by Lemma B.1 we have $\|x^* - x_1^*\| \leq 3\|y^* - y_1^*\|$. Furthermore,

$$\begin{aligned} \|x_0 - x_1^*\|^2 + \|y_0 - y_1^*\|^2 \leq & 2\|x_0 - x^*\|^2 + 2\|x^* - x_1^*\|^2 + \|y_0 - y_1^*\|^2 \\ \leq & 2\|x_0 - x^*\|^2 + 18\|y^* - y_1^*\|^2 + \|y_0 - y_1^*\|^2 \\ \leq & 2\|x_0 - x^*\|^2 + 36\|y_0 - y^*\|^2 + 37\|y_0 - y_1^*\|^2 \\ \leq & \frac{72}{\mu} \operatorname{gap}_{\hat{f}}(x_0, y_0) + 37\|y_0 - y_1^*\|^2, \end{aligned} \tag{82}$$

where in the last inequality we use Lemma B.2. It remains to bound $\|y_0 - y_1^*\|$. Since $\hat{\Psi}(y) - \frac{\tau}{2}\|y - y_0\|^2$ is $(\mu + \tau)$ strongly-concave w.r.t. y, we have

$$\left(\hat{\Psi}(y_1^*) - \frac{\tau}{2}\|y_1^* - y_0\|^2\right) - \frac{\tau + \mu}{2}\|y_1^* - y_0\|^2 \geq \hat{\Psi}(y_0) = \hat{\Psi}^* - [\hat{\Psi}^* - \hat{\Psi}(y_0)] \geq \hat{\Psi}(y_1^*) - [\hat{\Psi}^* - \hat{\Psi}(y_0)], \tag{83}$$

It further implies

$$\left(\tau + \frac{\mu}{2}\right) \|y_1^* - y_0\|^2 \leq \hat{\Psi}^* - \hat{\Psi}(y_0) \leq \operatorname{gap}_{\hat{f}}(x_0, y_0). \tag{84}$$

Plugging back into (82), we have

$$\begin{aligned} \|x_0 - x_1^*\|^2 + \|y_0 - y_1^*\|^2 &\leq \left[\frac{72}{\mu} + \frac{74}{2\tau + \mu}\right] \operatorname{gap}_{\hat{f}}(x_0, y_0) \\ &\leq \left[\frac{72\sqrt{2}}{\mu^2} + \frac{74\sqrt{2}}{(2\tau + \mu)\mu}\right] \frac{1}{1 - \rho}\epsilon_1. \end{aligned}$$

Case $k > 1$**:** From the proof of Theorem 4.2, we see that $\|\nabla \tilde{f}_i(x_i^t, y_i)\|^2 \leq \epsilon_i$ implies $\operatorname{gap}_{\tilde{f}_i}(x_i, y_i) \leq \hat{\epsilon}_i$ where $\hat{\epsilon}_i = \frac{\sqrt{2}}{4}(1 - \rho)^i \operatorname{gap}_{\hat{f}}(x_0, y_0)$. Note that $\tilde{f}_k$ is $(L, \mu + \tau)$-SC-SC and $(L + \max\{2L, \tau\})$-smooth. Then

$$\begin{aligned} \|x_{k-1} - x_k^*\|^2 &\leq 2\|x_{k-1} - x^*(y_{k-1}^*)\|^2 + 2\|x^*(y_{k-1}^*) - x^*(y_k^*)\|^2 \\ &\leq 2\|x_{k-1} - x_{k-1}^*\|^2 + 2\left(\frac{L + \max\{2L, \tau\}}{L}\right)^2 \|y_k^* - y_{k-1}^*\|^2. \end{aligned} \tag{85}$$

Furthermore,

$$\begin{aligned} \|x_{k-1} - x_k^*\|^2 + \|y_{k-1} - y_k^*\|^2 \leq \|x_{k-1} - x_k^*\|^2 + 2\|y_{k-1} - y_{k-1}^*\|^2 + 2\|y_{k-1}^* - y_k^*\|^2 \\ \leq 2\|x_{k-1} - x_{k-1}^*\|^2 + 2\|y_{k-1} - y_{k-1}^*\|^2 + 2\left[\left(\frac{L + \max\{2L, \tau\}}{L}\right)^2 + 1\right] \|y_k^* - y_{k-1}^*\|^2 \\ \leq \frac{4\hat{\epsilon}_{k-1}}{\min\{L, \mu + \tau\}} + \max\left\{20, \frac{9\tau^2}{2L^2} + 2\right\} \|y_k^* - y_{k-1}^*\|^2. \end{aligned} \tag{86}$$

Now we want to bound $\|y_{k-1}^* - y_k^*\|$. By optimality condition, we have for $\forall y$,

$$(y - y_k^*)^\top \nabla \tilde{\Psi}_t(y_k^*) \leq 0, \quad (y - y_{k-1}^*)^\top \nabla \tilde{\Psi}_{t-1}(y_{k-1}^*) \leq 0. \tag{87}$$

Choose y in the first inequality to be y_{k-1}^*, y in the second inequality to be y_k^*, and sum them together, we have

$$(y_k^* - y_{k-1}^*)^\top (\nabla \tilde{\Psi}_{k-1}(y_{k-1}^*) - \nabla \tilde{\Psi}_k(y_k^*)) \leq 0. \tag{88}$$

Using $\nabla \tilde{\Psi}_k(y) = \nabla_y \hat{f}(x^*(y), y) - \tau(y - z_k)$, we have

$$(y_k^* - y_{k-1}^*)^\top (\nabla_y \hat{f}(x^*(y_{k-1}^*), y_{k-1}^*) - \tau(y_{k-1}^* - z_{k-1}) - \nabla_y \hat{f}(x^*(y_k^*), y_k^*) + \tau(y_k^* - z_k)) \leq 0. \tag{89}$$

By strong concavity of $\hat{\Psi}(y) = \max_x \hat{f}(x, y)$, we have

$$(y_k^* - y_{k-1}^*)^\top (\nabla \hat{\Psi}(y_k^*) - \nabla \hat{\Psi}(y_{k-1}^*)) \leq -\mu\|y_k^* - y_{k-1}^*\|^2. \tag{90}$$

Adding to (89), we have

$$(y_k^* - y_{k-1}^*)^\top[\tau(y_k^* - z_k) - \tau(y_{k-1}^* - z_{k-1})] \leq -\mu\|y_k^* - y_{k-1}^*\|^2 \tag{91}$$

Rearranging,

$$\frac{\tau}{\mu+\tau}(y_k^* - y_{k-1}^*)^\top(z_{k-1} - z_k) \geq \|y_k^* - y_{k-1}^*\|^2. \tag{92}$$

Further with $(y_k^* - y_{k-1}^*)^\top(z_{k-1} - z_k) \leq \|y_k^* - y_{k-1}^*\|\|z_{k-1} - z_k\|$, we have

$$\|y_k^* - y_{k-1}^*\| \leq \frac{\tau}{\mu+\tau}\|z_{k-1} - z_k\|. \tag{93}$$

From updates of $\{z_k\}_k$, we have for $t > 2$

$$\begin{aligned}\|z_k - z_{k-1}\| &= \left\|y_{k-1} + \frac{\sqrt{q}-q}{\sqrt{q}+q}(y_{k-1} - y_{k-2}) - y_{k-2} - \frac{\sqrt{q}-q}{\sqrt{q}+q}(y_{k-2} - y_{k-3})\right\| \\ &\leq \left(1 + \frac{\sqrt{q}-q}{\sqrt{q}+q}\right)\|y_{k-1} - y_{k-2}\| + \frac{\sqrt{q}-q}{\sqrt{q}+q}\|y_{k-2} - y_{k-3}\| \\ &\leq 2\|y_{k-1} - y_{k-2}\| + \|y_{k-2} - y_{k-3}\| \\ &\leq 6\max\{\|y_{k-1} - y^*\|, \|y_{k-2} - y^*\|, \|y_{k-3} - y^*\|\}\end{aligned} \tag{94}$$

Therefore,

$$\begin{aligned}\|z_k - z_{k-1}\|^2 &\leq 36\max\{\|y_{k-1} - y^*\|^2, \|y_{k-2} - y^*\|^2, \|y_{k-3} - y^*\|^2\} \\ &\leq \frac{72}{\mu}\max\{\hat{\Psi}(y_{k-1}) - \hat{\Psi}^*, \hat{\Psi}(y_{k-2}) - \hat{\Psi}^*, \hat{\Psi}(y_{k-3}) - \hat{\Psi}^*\} \\ &\leq \frac{72}{\mu}\max\{S_{k-1}, S_{k-2}, S_{k-3}\},\end{aligned}$$

where in the second inequality we use strongly concavity of $\hat{\Psi}$ and in the last we use $\hat{\Psi}(y_k) - \hat{\Psi}^* \leq S_k$. Combining with (93) and (86), we have

$$\|x_{k-1} - x_k^*\|^2 + \|y_{k-1} - y_k^*\|^2 \leq \frac{4\hat{\epsilon}_{k-1}}{\min\{L, \mu+\tau\}} + \frac{36\tau^2\max\{40L^2, 9\tau^2 + 4L^2\}}{(\mu+\tau)^2L^2\mu}\max\{S_{k-1}, S_{k-2}, S_{k-3}\}. \tag{95}$$

Plugging in $S_k \leq \frac{8}{(\sqrt{q}-\rho)^2}(1-\rho)^{k+1}\operatorname{gap}_{\hat{f}}(x_0, y_0)$ from the proof of Theorem 4.2 and from definition of ϵ_k and $\hat{\epsilon}_k$, we have

$$\|x_{k-1} - x_k^*\|^2 + \|y_{k-1} - y_k^*\|^2 \leq \left\{\frac{2}{\mu\min\{L, \mu+\tau\}}\frac{1}{1-\rho} + \frac{288\sqrt{2}\tau^2\max\{40L^2, 9\tau^2 + 4L^2\}}{(\mu+\tau)^2L^2\mu^2(\sqrt{q}-\rho)^2)}\frac{1}{(1-\rho)^2}\right\}\epsilon_k. \tag{96}$$

It is left to discuss the case $t = 2$. Similarly, we have

$$\|z_2 - z_1\| = \left\|y_1 + \frac{\sqrt{q}-q}{\sqrt{q}+q}(y_1 - y_0) - y_0\right\| = \left(1 + \frac{\sqrt{q}-q}{\sqrt{q}+q}\right)\|y_1 - y_0\| \leq 4\max\{\|y_1 - y^*\|, \|y_0 - y^*\|\}$$

Then

$$\begin{aligned}\|z_2 - z_1\|^2 &\leq 16\max\{\|y_1 - y^*\|^2, \|y_0 - y^*\|^2\} \\ \leq \frac{32}{\mu}&\max\{\hat{\Psi}(y_1) - \hat{\Psi}^*, \hat{\Psi}(y_0) - \hat{\Psi}^*\} \leq \frac{32}{\mu}\max\{S_1, \operatorname{gap}_{\hat{f}}(x_0, y_0)\},\end{aligned}$$

Combining with (93) and (86), we have

$$\|x_1 - x_2^*\|^2 + \|y_1 - y_2^*\|^2 \leq \frac{4\hat{\epsilon}_1}{\min\{L, \mu+\tau\}} + \frac{16\tau^2\max\{40L^2, 9\tau^2 + 4L^2\}}{(\mu+\tau)^2L^2\mu}\max\{S_1, \operatorname{gap}_{\hat{f}}(x_0, y_0)\}. \tag{97}$$

Plugging in $S_1 \leq \frac{8}{(\sqrt{q}-\rho)^2}(1-\rho)^2 \operatorname{gap}_{\hat{f}}(x_0, y_0)$ and definition of ϵ_2 and $\hat{\epsilon}_1$, we have

$$\|x_1 - x_2^*\|^2 + \|y_1 - y_2^*\|^2 \leq \left\{ \frac{2}{\mu \min\{L, \mu+\tau\}} \frac{1}{1-\rho} + \frac{128\sqrt{2}\tau^2 \max\{40L^2, 9\tau^2 + 4L^2\}}{(\mu+\tau)^2 L^2 \mu^2 (\sqrt{q}-\rho)^2)} \right\} \epsilon_2. \tag{98}$$

■

Proof of Theorem 4.3

Proof We separate our arguments for the deterministic and stochastic settings. Inside this proof, $(x_{(i)}, y_{(i)})$ denotes the i-th iterate of $\mathcal{M}$ in solving the subproblem: $\min_x \max_y \tilde{f}_k(x, y)$. We use (x_k^*, y_k^*) to denote the optimal solution as before. We pick $(x_{(0)}, y_{(0)})$ to be (x_{k-1}, y_{k-1}).

Deterministic setting. The subproblem is $(L + \max\{2L, \tau\})$-Lipschitz smooth and $(L, \mu+\tau)$-SC-SC. By Lemma B.2, after N iterations of algorithm $\mathcal{M}$,

$$\begin{aligned}\|\nabla \tilde{f}_k(x_{(N)}, y_{(N)})\|^2 &\leq 4(L + \max\{2L, \tau\})^2 [\|x_{(N)} - x_k^*\|^2 + \|y_{(N)} - y_k^*\|^2] \\ &\leq 4(L + \max\{2L, \tau\})^2 \left(1 - \frac{1}{\Lambda_{\mu,L}^{\mathcal{M}}(\tau)}\right)^N [\|x_{k-1} - x_k^*\|^2 + \|y_{k-1} - y_k^*\|^2].\end{aligned}$$

Choosing

$$\begin{aligned}N =& \Lambda_{\mu,L}^{\mathcal{M}}(\tau) \log \frac{4(L + \max\{2L, \tau\})^2 (\|x_{k-1} - x_k^*\|^2 + \|y_{k-1} - y_k^*\|^2)}{\epsilon_k} \\ \leq& \Lambda_{\mu,L}^{\mathcal{M}}(\tau) \log \frac{4(L + \max\{2L, \tau\})^2 C_t \epsilon_k}{\epsilon_k} = \Lambda_{\mu_x,\mu,L}^{\mathcal{M}}(\tau) \log\left(4(L + \max\{2L, \tau\})^2 C_t\right),\end{aligned}$$

where C_t is specified in Lemma D.2, we have $\|\nabla \tilde{f}_k(x_{(N)}, y_{(N)})\|^2 \leq \epsilon_k$.

Stochastic setting. With the same reasoning as in deterministic setting and applying Appendix B.4 of [Lin et al., 2018], after

$$N = \Lambda_{\mu,L}^{\mathcal{M}}(\tau) \log \frac{4(L + \max\{2L, \tau\})^2 (\|x_{k-1} - x_k^*\|^2 + \|y_{k-1} - y_k^*\|^2)}{\epsilon_k} + 1$$

iterations of $\mathcal{M}$, we have $\|\nabla \tilde{f}_k(x_{(N)}, y_{(N)})\|^2 \leq \epsilon_k$.

■

D.4 TOTAL COMPLEXITY

Proof of Corollary 4.1

Proof From Theorem 4.1, the number of outer-loop calls to find an ϵ-stationary point of Φ is $T = O\left(L(\Delta + D_y^0)\epsilon^{-2}\right)$. From Theorem 4.2, by picking $\rho = 0.9\sqrt{q} = 0.9\sqrt{\mu/(\mu+\tau)}$, we have

$$\|\nabla \hat{f}_t(x_k^t, y_k^t)\|^2 \leq \left[\frac{5508L^2}{\mu^2(\sqrt{q}-\rho)^2} + \frac{18\sqrt{2}L^2}{\mu}\right](1-\rho)^k \|\nabla \hat{f}_t(x_0^t, y_0^t)\|^2. \tag{99}$$

Therefore, to achieve $\|\nabla \hat{f}_t(x_K^t, y_K^t)\|^2 \leq \alpha_t \|\nabla \hat{f}_t(x_0^t, y_0^t)\|^2$, we need to solve $(\star\star)$

$$K = 0.9\sqrt{(\tau+\mu)/\mu} \log \frac{\left[\frac{5508L^2}{\mu^2(\sqrt{q}-\rho)^2} + \frac{18\sqrt{2}L^2}{\mu}\right]}{\alpha_t} = O\left(\sqrt{(\tau+\mu)/\mu} \log\left(\frac{\max\{1, L, \tau\}}{\min\{1, \mu\}}\right)\right)$$

times, where α_t is defined as in Theorem 4.1. Finally, Theorem 4.3 implies that solving ($\star\star$) needs $N = O\left(\Lambda_{\mu,L}^{\mathcal{M}}(\tau)\log\left(\frac{\max\{1,L,\tau\}}{\min\{1,\mu\}}\right)\right)$ gradient oracles. The total complexity is

$$T \cdot K \cdot N = O\left(\frac{\Lambda_{\mu,L}^{\mathcal{M}}(\tau)L(\Delta + D_y^0)}{\epsilon^2}\sqrt{\frac{\mu+\tau}{\mu}}\log^2\left(\frac{\max\{1,L,\tau\}}{\min\{1,\mu\}}\right)\right). \tag{100}$$

■

References

Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points ii: first-order methods. *Mathematical Programming*, pages 1–41, 2019.

D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1):503–558, 2019.

A. Ibrahim, W. Azizian, G. Gidel, and I. Mitliagkas. Linear lower bounds and conditioning of differentiable games. In *ICML*, pages 4583–4593. PMLR, 2020.

H. Lin, J. Mairal, and Z. Harchaoui. Catalyst acceleration for first-order convex optimization: from theory to practice. *JMLR*, 18(1):7854–7907, 2018.

T. Lin, C. Jin, and M. I. Jordan. Near-optimal algorithms for minimax optimization. In *COLT*, 2020.

A. C.-C. Yao. Probabilistic computations: Toward a unified measure of complexity. In *18th Annual Symposium on Foundations of Computer Science (sfcs 1977)*, pages 222–227. IEEE Computer Society, 1977.

D. Zhou and Q. Gu. Lower bounds for smooth nonconvex finite-sum optimization. In *International Conference on Machine Learning*, pages 7574–7583. PMLR, 2019.