Supplementary Materials for "Robust and Faster Zeroth-Order Minimax Optimization: Complexity and Applications"

In this Appendix, we first provide detailed explanations for some descriptions in the main paper in Section A. Then, we provide the complexity analysis of our ZO-GDEGA algorithm for solving NC-C and NC-SC problems in Section B. Finally, we supplement the experiments in more detail and provide more experimental results in Section C.

A Detailed Explanations for some Descriptions

This section provides some detailed explanations for the main paper.

A.1 Properties of the max function $\Phi(x)$

Compared with existing nonconvex minimax optimization methods such as [31], Assumption 2 can still guarantee the ℓ-weak convexity of the max function $\Phi(x)$. Specifically, existing works rely on the compactness assumption of the maximization domain, thus applying Danskin's theorem for convenience, which implies the weak convexity of the max function $\Phi(x)$. Some machine learning models can satisfy the compactness assumption, e.g., Wasserstein GAN [2] with weight clipping, but some models are difficult to satisfy it, e.g., [17]. Thus, similar to [4], we also use the extension of the classical Danskin's theorem based on Assumption 2, which means that the solution set $Y^*(x) := \left\{y^* | y^* \in \arg\max_{y \in \mathbb{R}^{d_y}} \{f(x,y) - h(y)\}\right\}$ is non-empty for $\forall x \in \mathbb{R}^{d_x}$ and the max function $\Phi(x)$ is ℓ-weakly convex.

A.2 Stochastic ZO-GDEGA for solving NC-C and NC-SC problems

To analyze the stochastic ZO-GDEGA, we first restate the stochastic version of the zeroth-order randomized gradient estimators as follows:

$$\hat{\nabla}_x f(x,y;\mathcal{I}_1) = \frac{1}{b_1}\sum_{j=1}^{b_1} \hat{\nabla}_x f(x,y;\zeta_j) = \frac{1}{b_1}\sum_{j=1}^{b_1} \frac{d_x[f(x+\mu_1 u_j, y;\zeta_j) - f(x,y;\zeta_j)]}{\mu_1} u_j, \tag{A.1a}$$

$$\hat{\nabla}_y f(x,y;\mathcal{I}_2) = \frac{1}{b_2}\sum_{j=2}^{b_2} \hat{\nabla}_y f(x,y;\xi_j) = \frac{1}{b_2}\sum_{j=1}^{b_2} \frac{d_y[f(x,y+\mu_2 v_j;\xi_j) - f(x,y;\xi_j)]}{\mu_2} v_j. \tag{A.1b}$$

According to [19, 47], we know that for given random variables ζ and ξ, $\mathbb{E}_{u,\zeta}[\hat{\nabla}_x f(x,y;\zeta)] = \nabla_x f_{\mu_1}(x,y)$ and $\mathbb{E}_{v,\xi}[\hat{\nabla}_y f(x,y;\xi)] = \nabla_y f_{\mu_2}(x,y)$. The smoothed functions associated to function $f(x,y;\xi)$ can be defined as: $f_{\mu_1}(x,y;\zeta) = \mathbb{E}_u[f(x+\mu_1 u, y;\zeta)]$ and $f_{\mu_2}(x,y;\xi) = \mathbb{E}_v[f(x,y+\mu_2 v;\xi)]$. Then, according to [24, Lemma 5], the ZO estimators (A.1) are unbiased, i.e., $\mathbb{E}_{U,\mathcal{I}_1}[\hat{\nabla}_x f(x,y;\mathcal{I}_1)] = \nabla_x f_{\mu_1}(x,y)$ and $\mathbb{E}_{V,\mathcal{I}_2}[\hat{\nabla}_y f(x,y;\mathcal{I}_2)] = \nabla_y f_{\mu_2}(x,y)$ with $U = \{u_i\}_{i=1}^{b_1}$ and $V = \{v_i\}_{i=1}^{b_2}$. Moreover, we suppose that the variance of zeroth-order stochastic gradient estimation is bounded for any random variables ζ and ξ, i.e.,

$$\begin{aligned} &\mathbb{E}_{u,\zeta}\|\hat{\nabla}_x f(x,y;\zeta) - \nabla_x f_{\mu_1}(x,y)\|^2 \le \sigma_1^2, \\ &\mathbb{E}_{v,\xi}\|\hat{\nabla}_y f(x,y;\xi) - \nabla_y f_{\mu_2}(x,y)\|^2 \le \sigma_1^2. \end{aligned} \tag{A.2}$$

In addition, we also need the following assumptions. Note that these assumptions are common in stochastic optimization [31, 19].

Assumption 6. *We assume that the variance of stochastic gradient is bounded, i.e., there exist a constant σ_2 such that $\mathbb{E}\|\nabla_x f(x,y;\zeta) - \nabla_x f(x,y)\|^2 \le \sigma_2^2$ and $\mathbb{E}\|\nabla_y f(x,y;\xi) - \nabla_y f(x,y)\|^2 \le \sigma_2^2$.*

Assumption 7. *Each component function $f(x,y;\Xi)$ is ℓ-smooth, i.e., for all x, x' and y, y'*

$$\|\nabla f(x,y;\Xi) - \nabla f(x',y';\Xi)\| \le \ell\|(x,y) - (x',y')\|, \tag{A.3}$$

where $\nabla f(x,y;\Xi) = (\nabla_x f(x,y;\zeta), \nabla_y f(x,y;\xi))$.

For notational simplicity, let $\sigma = \max\{\sigma_1, \sigma_2\}$. According to [19], we have

$$\begin{aligned}
&\mathbb{E}\|\hat{\nabla}_x f(x, y; \mathcal{I}_1) - \nabla_x f_{\mu_1}(x, y)\|^2 \le \frac{\sigma^2}{b_1} \\
&\mathbb{E}\|\hat{\nabla}_y f(x, y; \mathcal{I}_2) - \nabla_y f_{\mu_2}(x, y)\|^2 \le \frac{\sigma^2}{b_2}.
\end{aligned} \tag{A.4}$$

Based on the properties mentioned above, we propose our stochastic ZO-GDEGA algorithm, as shown in Algorithm 2, and the complexity of Algorithm 2 can be analyzed successfully.

Algorithm 2 Stochastic Zeroth-Order Gradient Descent Extragradient Ascent Algorithm

```
Initialize: $x_0$, $z_0 = y_0$, step sizes $\eta_x$ and $\eta_y$.
for $t = 0, 1, \dots, T-1$ do
    Draw i.i.d. $\mathcal{I}_1 = \{\zeta_j\}_{j=1}^{b_1}$ and $\mathcal{I}_2 = \{\xi_j\}_{j=1}^{b_2}$ stochastic samples, respectively;
    $\hat{\nabla}_x F(x_t; \mathcal{I}_1) = \begin{cases} \hat{\nabla}_x f(x_t, z_t; \mathcal{I}_1) \text{ for NC-C case;} \\ \hat{\nabla}_x f(x_t, y_t; \mathcal{I}_1) \text{ for NC-SC case;} \end{cases}$
    $x_{t+1} = \text{prox}_{\eta_x}^{g}(x_t - \eta_x \hat{\nabla}_x F(x_t; \mathcal{I}_1))$;
    $z_{t+1} = \text{prox}_{\eta_y}^{h}(y_t + \eta_y \hat{\nabla}_y f(x_t, y_t; \mathcal{I}_2))$;
    $y_{t+1} = \text{prox}_{\eta_y}^{h}(y_t + \eta_y \hat{\nabla}_y f(x_{t+1}, z_{t+1}; \mathcal{I}_2))$;
end for
Randomly draw $\hat{x}$ from $x_1, \dots, x_T$ at uniform;
Output: $\hat{x}$.
```

A.3 Proofs of Propositions 1 and 2

For analyzing our ZO-GDEGA algorithm solving Problem (1), we extend the Propositions 4.11 and 4.12 in [31]. Detailed analysis is as follows.

Proposition 3 (A detailed description of Proposition 1 - the generalized version of Proposition 4.12 in [31])**.** *Under Assumptions 1 and 2, if a point $(\hat{x}, \hat{y})$ is an $\epsilon^2/\ell D_h$-stationary point in terms of Definition 3, a point $\hat{x}$ is an $\mathcal{O}(\epsilon)$-stationary point in terms of Definition 2. Conversely, if a point $\hat{x}$ is an ϵ-stationary point in terms of Definition 2, an $\mathcal{O}(\epsilon)$-statinary point (x', y') in terms of Definition 3 can be obtained using additional $\mathcal{O}(\epsilon^{-2})$ gradients or $\mathcal{O}(\epsilon^{-4})$ stochastic gradients.*

Proof. We have the facts that the objective function $g(x) + f(x, y) + \ell\|x - \hat{x}\|^2$ is strongly convex in x and concave in y, and $x^*(\hat{x}) = \arg\min_{x \in \mathbb{R}^{d_x}} g(x) + \Phi(x) + \ell\|x - \hat{x}\|^2 = \text{prox}_{1/2\ell}^{g+\Phi}(\hat{x})$ is uniquely defined.

- If a point $(\hat{x}, \hat{y})$ is an $\frac{\epsilon^2}{\ell D_h}$-stationary point in terms of Definition 3, i.e.,

$$\|\ell(\hat{x} - \text{prox}_{1/\ell}^{g}(\hat{x} - \frac{1}{\ell}\nabla_x f(\hat{x}, \hat{y})))\| \le \frac{\epsilon^2}{\ell D_h}, \|\ell(\hat{y} - \text{prox}_{1/\ell}^{h}(\hat{y} + \frac{1}{\ell}\nabla_y f(\hat{x}, \hat{y})))\| \le \frac{\epsilon^2}{\ell D_h}. \tag{A.5}$$

By definition, we have

$$\|\nabla\psi_{1/2\ell}(\hat{x})\|^2 = 4\ell^2\|\hat{x} - x^*(\hat{x})\|^2. \tag{A.6}$$

Since $\psi(\cdot) + \ell\|\cdot - \hat{x}\|^2$ is ℓ-strongly convex, we have

$$\begin{aligned}
&g(\hat{x}) + \max_{y \in dom\ h}\{f(\hat{x}, y) - h(y)\} - g(x^*(\hat{x})) - \max_{y \in dom\ h}\{f(x^*(\hat{x}), y) - h(y)\} - \ell\|\hat{x} - x^*(\hat{x})\|^2 \\
&= \psi(\hat{x}) - \psi(x^*(\hat{x})) - \ell\|x^*(\hat{x}) - \hat{x}\|^2 \ge \frac{\ell\|\hat{x} - x^*(\hat{x})\|^2}{2} = \frac{\|\nabla\psi_{1/2\ell}(\hat{x})\|^2}{8\ell}.
\end{aligned} \tag{A.7}$$

Furthermore, we define $\hat{y}^+ = \text{prox}^h_{1/\ell}(\hat{y} + \frac{1}{\ell}\nabla_y f(\hat{x}, \hat{y}))$ and we have

$$\begin{aligned}
g(\hat{x}) \ &+ \max_{y \in dom\ h}\{f(\hat{x}, y) - h(y)\} - g(x^*(\hat{x})) - \max_{y \in dom\ h}\{f(x^*(\hat{x}), y) - h(y)\} - \ell\|\hat{x} - x^*(\hat{x})\|^2 \\
&= g(\hat{x}) + \max_{y \in dom\ h}\{f(\hat{x}, y) - h(y)\} - [f(\hat{x}, \hat{y}^+) - h(\hat{y}^+)] + [f(\hat{x}, \hat{y}^+) - h(\hat{y}^+)] - g(x^*(\hat{x})) \\
&\quad - \max_{y \in dom\ h}\{f(x^*(\hat{x}), y) - h(y)\} - \ell\|\hat{x} - x^*(\hat{x})\|^2 \\
&\leq g(\hat{x}) + \max_{y \in dom\ h}\{f(\hat{x}, y) - h(y)\} - [f(\hat{x}, \hat{y}^+) - h(\hat{y}^+)] + [f(\hat{x}, \hat{y}^+) - h(\hat{y}^+)] - g(x^*(\hat{x})) \\
&\quad - \{f(x^*(\hat{x}), \hat{y}^+) - h(\hat{y}^+)\} - \ell\|\hat{x} - x^*(\hat{x})\|^2 \\
&\leq \max_{y \in dom\ h}\{f(\hat{x}, y) - h(y)\} - [f(\hat{x}, \hat{y}^+) - h(\hat{y}^+)] \\
&\quad + (\|\hat{x} - x^*(\hat{x})\|\|\nabla_x f(\hat{x}, \hat{y}^+) + w\| - \frac{\ell}{2}\|\hat{x} - x^*(\hat{x})\|^2) \\
&\leq \underbrace{\max_{y \in dom\ h}\{f(\hat{x}, y) - h(y)\} - [f(\hat{x}, \hat{y}^+) - h(\hat{y}^+)]}_{M_0} + \frac{\|\nabla_x f(\hat{x}, \hat{y}^+) + w\|^2}{2\ell},
\end{aligned} \tag{A.8}$$

where $w \in \partial g(\hat{x})$, the second inequality follows from the ℓ-strongly convexity of $g(x) + \Gamma(x, \hat{y}^+) + \ell\|x - \hat{x}\|^2$ and Cauchy-Schwarz inequality, and the last inequality holds due to Young's inequality.

Next, we consider the part M_0 in (A.8). By the definition of $\hat{y}^+$, the first-order optimality condition yields that

$$h(y) - h(\hat{y}^+) + \ell(y - \hat{y}^+)^\top(\hat{y}^+ - \hat{y} - \frac{1}{\ell}\nabla_y f(\hat{x}, \hat{y})) \geq 0 \text{ for all } y \in dom\ h. \tag{A.9}$$

Together with the ℓ-smoothness and concavity of the function $f(\hat{x}, \cdot)$, we have

$$-f(\hat{x}, \hat{y}^+) + f(\hat{x}, y) \ \leq \langle -\nabla_y f(\hat{x}, y), \hat{y}^+ - y\rangle + \frac{\ell}{2}\|y - \hat{y}^+\|^2 \tag{A.10}$$

$$f(\hat{x}, y) - f(\hat{x}, \hat{y}) \ \leq \langle \nabla_y f(\hat{x}, \hat{y}), y - \hat{y}\rangle. \tag{A.11}$$

Letting $y = \hat{y}$ in (A.10), we have

$$-f(\hat{x}, \hat{y}^+) + f(\hat{x}, \hat{y}) \ \leq \langle -\nabla_y f(\hat{x}, \hat{y}), \hat{y}^+ - \hat{y}\rangle + \frac{\ell}{2}\|\hat{y} - \hat{y}^+\|^2. \tag{A.12}$$

Adding inequality (A.11) and (A.12), we have

$$\begin{aligned}
f(\hat{x}, y) - f(\hat{x}, \hat{y}^+) \ &\leq \langle \nabla_y f(\hat{x}, \hat{y}), y - \hat{y}^+\rangle + \frac{\ell}{2}\|\hat{y} - \hat{y}^+\|^2 \\
&\leq h(y) - h(\hat{y}^+) + \ell\langle y - \hat{y}^+, \hat{y}^+ - \hat{y}\rangle + \frac{\ell}{2}\|\hat{y} - \hat{y}^+\|^2 \\
&= h(y) - h(\hat{y}^+) - \frac{\ell}{2}\|y - \hat{y}^+\|^2 + \frac{\ell}{2}\|y - \hat{y}\|^2,
\end{aligned} \tag{A.13}$$

where the second inequality holds due to (A.9) and the last equality holds due to $\langle a - b, a - c\rangle = \frac{1}{2}\|a - b\|^2 + \frac{1}{2}\|a - c\|^2 - \frac{1}{2}\|b - c\|^2$. Thus, together with the boundedness of $dom\ h$, we have

$$f(\hat{x}, y) - h(y) - [f(\hat{x}, \hat{y}^+) - h(\hat{y}^+)] \ \leq \frac{\ell}{2}(\|y - \hat{y}\|^2 - \|y - \hat{y}^+\|^2) \leq \ell D_h\|\hat{y}^+ - \hat{y}\|. \tag{A.14}$$

Putting these pieces together yeilds that

$$
\begin{aligned}
& g(\hat{x}) \; + \max_{y\in\mathbb{R}^{d_y}} \{f(\hat{x},y)-h(y)\} - g(x^*(\hat{x})) - \max_{y\in\mathbb{R}^{d_y}} \{f(x^*(\hat{x}),y)-h(y)\} - \ell\|\hat{x}-x^*(\hat{x})\|^2 \\
& \leq \ell D_h\|\hat{y}^+-\hat{y}\| + \frac{\|\nabla_x f(\hat{x},\hat{y}^+)+w\|^2}{2\ell} \\
& \leq \ell D_h\|\hat{y}^+-\hat{y}\| + \frac{\|\nabla_x f(\hat{x},\hat{y})+w\|^2}{\ell} + \frac{\|\nabla_x f(\hat{x},\hat{y}^+)-\nabla_x f(\hat{x},\hat{y})\|^2}{\ell} \\
& \leq \ell D_h\|\hat{y}^+-\hat{y}\| + \frac{\|\nabla_x f(\hat{x},\hat{y})+w\|^2}{\ell} + \ell\|\hat{y}-\hat{y}^+\|^2 \\
& \leq \frac{\epsilon^2}{\ell} + \mathcal{O}(\epsilon^4) + \frac{\epsilon^4}{\ell^3 D_h^2},
\end{aligned}
\tag{A.15}
$$

where the second inequality holds due to Young's Inequality and the third inequality holds due to ℓ-smoothness. According to [3, Theorem 3.1] and the properties of the subgradient descent method in [5], $\|\ell(\hat{x}-\text{prox}^g_{1/\ell}(\hat{x}-\frac{1}{\ell}\nabla_x f(\hat{x},\hat{y})))\| \leq \frac{\epsilon^2}{\ell D_h}$ means that $\|\nabla_x f(\hat{x},\hat{y})+w\|^2 = \mathcal{O}(\epsilon^4)$. Combining (A.7) and (A.15) yeilds that $\|\nabla\psi_{1/2\ell}(\hat{x})\| = \mathcal{O}(\epsilon)$. Thus, the point $\hat{x}$ is a $\mathcal{O}(\epsilon)$-stationary point in terms of Definition 2.

• Conversely, we let a point $\hat{x}$ satisfy that $\|\nabla\psi_{1/2\ell}(\hat{x})\| \leq \epsilon$. We can apply the proximal extragradient algorithm for solving $x^*(\hat{x})$ and obtain a point (x',y') satisfying that

$$
dist(-\partial g(x^+), \nabla_x f(x',y^+) + 2\ell(x'-\hat{x})) \leq \epsilon, \|y^+-y'\| \leq \epsilon/\ell, \|x'-x^*(\hat{x})\| \leq \epsilon/\ell, \tag{A.16}
$$

where $y^+ = \text{prox}^h_{1/\ell}(y'+\frac{1}{\ell}\nabla_y f(x',y'))$ and $x^+ = \text{prox}^g_{1/\ell}(x'-\frac{1}{\ell}\nabla_x f(x',y'))$. Thus, we obtain $\|\ell(y'-\text{prox}^h_{1/\ell}(y'+\frac{1}{\ell}\nabla_y f(x',y')))\| \leq \epsilon$. Next, we prove the part of x. Since $2\ell\|x^*(\hat{x})-\hat{x}\| = \|\nabla\psi_{1/2\ell}(\hat{x})\| \leq \epsilon$, we have

$$
\begin{aligned}
& \|\ell(x'-\text{prox}^g_{1/\ell}(x'-\frac{1}{\ell}\nabla_x f(x',y')))\| \\
\leq\; & \|\ell(x'-\text{prox}^g_{1/\ell}(x'-\frac{1}{\ell}\nabla_x f(x',y'))) + 2\ell(x'-\hat{x})\| + 2\ell\|x'-\hat{x}\| \\
\leq\; & \|\ell(x'-\text{prox}^g_{1/\ell}(x'-\frac{1}{\ell}\nabla_x f(x',y'))) + 2\ell(x'-\hat{x})\| \\
& + 2\ell\|x'-x^*(\hat{x})\| + 2\ell\|x^*(\hat{x})-\hat{x}\| \\
\leq\; & \|\ell(x'-\text{prox}^g_{1/\ell}(x'-\frac{1}{\ell}\nabla_x f(x',y'))) + 2\ell(x'-\hat{x})\| + 3\epsilon, \\
\leq\; & \|\ell(x'-(I+\frac{1}{\ell}\partial g)^{-1}(x'-\frac{1}{\ell}\nabla_x f(x',y'))) + 2\ell(x'-\hat{x})\| + 3\epsilon,
\end{aligned}
\tag{A.17}
$$

where $(I+\tau\partial g)^{-1}(z) = \arg\min_x\{g(x)+\frac{1}{2\tau}\|x-z\|^2\}$). According to [3, Theorem 3.1] and the properties of the subgradient descent method in [5], there exists a $g_0 \in \partial g(x^+)$ such that

$$
\begin{aligned}
& \|\ell(x'-\text{prox}^g_{1/\ell}(x'-\frac{1}{\ell}\nabla_x f(x',y')))\| \\
& \leq \|\ell[x'-(x'-\frac{1}{\ell}(g_0+\nabla_x f(x',y')))] + 2\ell(x'-\hat{x})\| + 3\epsilon \\
& \leq \|\ell[x'-(x'-\frac{1}{\ell}(g_0+\nabla_x f(x',y^+)))] + 2\ell(x'-\hat{x})\| + \|y'-y^+\| + 3\epsilon \\
& \leq \epsilon + 4\epsilon \\
& = \mathcal{O}(\epsilon),
\end{aligned}
\tag{A.18}
$$

where the second inequality holds due to the Triangular Inequality and the smoothness of $f(x,y)$, and the last inequality holds due to (A.16). Thus, if a generalized ϵ-stationary point of f is obtained, the required number of gradient evaluations is $\mathcal{O}(\epsilon^{-2})$ [36]. This argument holds for applying stochastic mirror-prox algorithm and the required number of stochastic gradient evaluations is $\mathcal{O}(\epsilon^{-4})$ [26].

In summary, the ϵ-stationary point definition in terms of ψ is stronger than the ϵ-stationary point definition in terms of f. This completes the proof. □

Proposition 4 (A detailed description of Proposition 2 - the generalized version of Proposition 4.11 in [31])**.** *Under Assumption 4, if a point $(\hat{x}, \hat{y})$ is an ϵ/κ-stationary point in terms of Definition 3, a point $\hat{x}$ is an $\mathcal{O}(\epsilon)$-stationary point in terms of Definition 2. Conversely, if a point $\hat{x}$ is an ϵ-stationary point in terms of Definition 2, an $\mathcal{O}(\epsilon)$-stationary point $(\hat{x}, y')$ in terms of Definition 3 can be obtained using additional $\mathcal{O}(\kappa \log(1/\epsilon))$ gradients or $\mathcal{O}(1/\epsilon^2)$ stochastic gradients.*

Proof. • If a point $(\hat{x}, \hat{y})$ is an $\frac{\epsilon}{\kappa}$-stationary point in terms of Definition 3, i.e.,

$$\|\ell(\hat{x} - \text{prox}^g_{1/\ell}(\hat{x} - \frac{1}{\ell}\nabla_x f(\hat{x}, \hat{y})))\| \leq \frac{\epsilon}{\kappa}, \|\ell(\hat{y} - \text{prox}^h_{1/\ell}(\hat{y} + \frac{1}{\ell}\nabla_y f(\hat{x}, \hat{y})))\| \leq \frac{\epsilon}{\kappa}. \tag{A.19}$$

Then, there exists $w \in \partial\psi(\hat{x})$ such that

$$\begin{aligned} \|w\| &\leq \|w - \ell(\hat{x} - \text{prox}^g_{1/\ell}(\hat{x} - \frac{1}{\ell}\nabla_x f(\hat{x}, \hat{y})))\| + \frac{\epsilon}{\kappa} \\ &\leq \|\nabla_x f(\hat{x}, y^*(\hat{x})) - \nabla_x f(\hat{x}, \hat{y})\| + \frac{\epsilon}{\kappa} \\ &\leq \ell\|\hat{y} - y^*(\hat{x})\| + \frac{\epsilon}{\kappa}. \end{aligned} \tag{A.20}$$

Similar to [31, Proposition 4.11], since $f(\hat{x}, \cdot)$ is μ-strongly concave over $dom\ h$, the global error bound condition [10] holds true and we have

$$\|w\| \leq \ell\kappa\|\hat{y} - \text{prox}^h_{1/\ell}(\hat{y} + \frac{1}{\ell}\nabla_y f(\hat{x}, \hat{y}))\| + \frac{\epsilon}{\kappa} \leq \epsilon + \frac{\epsilon}{\kappa} = \mathcal{O}(\epsilon). \tag{A.21}$$

Thus, the point $\hat{x}$ is an $\mathcal{O}(\epsilon)$-stationary point in terms of Definition 2.

• Conversely, if $\hat{x}$ is an ϵ-stationary point in terms of ψ, thus $dist(0, \partial\psi(\hat{x})) \leq \epsilon$, where $dist(x, C) = \min_{c \in C} \|x - c\|$. Thus, there exists $w \in \partial\psi(\hat{x})$ such that $\|w\| \leq \epsilon$. The optimization problem $\max_{y \in dom\ h} f(\hat{x}, y) - h(y)$ is strongly concave and $y^*(\hat{x})$ is uniquely defined. We apply proximal gradient descent for solving such problem and obtain a point $y' \in dom\ h$ satisfying that

$$y^+ = \text{prox}^h_{1/\ell}(y' + \frac{1}{\ell}\nabla_y f(\hat{x}, y')), \|\ell(y' - y^+)\| \leq \epsilon, \|y^+ - y^*(\hat{x})\| \leq \epsilon. \tag{A.22}$$

Then, for x, we have

$$\begin{aligned} &\|\ell(\hat{x} - \text{prox}^g_{1/\ell}(\hat{x} - \frac{1}{\ell}\nabla_x f(\hat{x}, y^+)))\| \\ &\quad\leq \|\ell(\hat{x} - \text{prox}^g_{1/\ell}(\hat{x} - \frac{1}{\ell}\nabla_x f(\hat{x}, y^+))) - w\| + \|w\| \\ &\quad= \|\ell(\hat{x} - \text{prox}^g_{1/\ell}(\hat{x} - \frac{1}{\ell}\nabla_x f(\hat{x}, y^+))) - w\| + \epsilon \\ &\quad= \|\ell(\hat{x} - \arg\min_x\{g(x) + \frac{\ell}{2}\|x - \hat{x} + \frac{1}{\ell}\nabla_x f(\hat{x}, y^+)\|^2\}) - w\| + \epsilon. \end{aligned} \tag{A.23}$$

Thus, according to the properties of the subgradient descent method in [5], we have $\|\ell(\hat{x} - \text{prox}^g_{1/\ell}(\hat{x} - \frac{1}{\ell}\nabla_x f(\hat{x}, y^+)))\| \leq \|\nabla_x f(\hat{x}, y^+) - \nabla_x f(\hat{x}, y^*(\hat{x}))\| + \epsilon$. According to the ℓ-smoothness of f, we have

$$\begin{aligned} \|\ell(\hat{x} - \text{prox}^g_{1/\ell}(\hat{x} - \frac{1}{\ell}\nabla_x f(\hat{x}, y^+)))\| &\leq \|\nabla_x f(\hat{x}, y^+) - \nabla_x f(\hat{x}, y^*(\hat{x}))\| + \epsilon \\ &\leq \ell\|y^+ - y^*(\hat{x})\| + \epsilon = \mathcal{O}(\epsilon). \end{aligned} \tag{A.24}$$

Thus, if a generalized ϵ-stationary point of f is obatined, the required number of gradient evaluations is $\mathcal{O}(\kappa \log(1/\epsilon))$ [48]. This argument holds for applying proximal stochastic gradient with proper stepsize and the required number of stochastic gradient evaluations is $\mathcal{O}(1/\epsilon^2)$ [11]. This completes the proof.

In summary, the ϵ-stationary point definition in terms of ψ is stronger than the ϵ-stationary point definition in terms of f. □

A.4 First-Order Gradient Descent Extragradient Ascent Algorithm

As by-products, we also provide the first-order variants of our ZO-GDEGA algorithm, as shown in Algorithms 3 and 4, respectively. The two algorithms can also reduce the per-iteration complexity of the standard first-order EG algorithms in [35] while maintaining their theoretical advantages.

Algorithm 3 Deterministic FO-GDEGA: First-order Gradient Descent Extragradient Ascent Algorithm (the first-order variant of Algorithm 1)

```
Initialize: $x_0$, $z_0 = y_0$, step sizes $\eta_x$ and $\eta_y$.
for $t = 0, 1, \ldots, T-1$ do
    $\nabla_x F(x_t) = \begin{cases} \nabla_x f(x_t, z_t) \text{ for NC-C case;} \\ \nabla_x f(x_t, y_t) \text{ for NC-SC case;} \end{cases}$
    $x_{t+1} = \text{prox}^g_{\eta_x}(x_t - \eta_x \nabla_x F(x_t))$;
    $z_{t+1} = \text{prox}^h_{\eta_y}(y_t + \eta_y \nabla_y f(x_t, y_t))$;
    $y_{t+1} = \text{prox}^h_{\eta_y}(y_t + \eta_y \nabla_y f(x_{t+1}, z_{t+1}))$;
end for
Randomly draw $\hat{x}$ from $x_1, \ldots, x_T$ at uniform;
Output: $\hat{x}$.
```

Algorithm 4 Stochastic FO-GDEGA: Stochastic First-order Gradient Descent Extragradient Ascent Algorithm (the first-order variant of Algorithm 2)

```
Initialize: $x_0$, $z_0 = y_0$, step sizes $\eta_x$ and $\eta_y$.
for $t = 0, 1, \ldots, T-1$ do
    Draw i.i.d. $\mathcal{I}_1 = \{\zeta_j\}_{j=1}^{b_1}$ and $\mathcal{I}_2 = \{\xi_j\}_{j=1}^{b_2}$ stochastic samples, respectively;
    $\nabla_x F(x_t; \mathcal{I}_1) = \begin{cases} \nabla_x f(x_t, z_t; \mathcal{I}_1) \text{ for NC-C case;} \\ \nabla_x f(x_t, y_t; \mathcal{I}_1) \text{ for NC-SC case;} \end{cases}$
    $x_{t+1} = \text{prox}^g_{\eta_x}(x_t - \eta_x \nabla_x F(x_t; \mathcal{I}_1))$;
    $z_{t+1} = \text{prox}^h_{\eta_y}(y_t + \eta_y \nabla_y f(x_t, y_t; \mathcal{I}_2))$;
    $y_{t+1} = \text{prox}^h_{\eta_y}(y_t + \eta_y \nabla_y f(x_{t+1}, z_{t+1}; \mathcal{I}_2))$;
end for
Randomly draw $\hat{x}$ from $x_1, \ldots, x_T$ at uniform;
Output: $\hat{x}$.
```

B Proofs of Overall Complexities for Our ZO-GDEGA Algorithm

This section provides detailed proofs of the complexity results for our ZO-GDEGA in various settings. We first provide several key lemmas and propositions as follows.

Definition 4 (Another form of smoothness)**.** *The smoothness of f means that $f(x, y)$ has Lipschitz continuous gradients, i.e., there also exist L_x and L_y such that*

$$\begin{aligned} \|\nabla_x f(x_1, y) - \nabla_x f(x_2, y)\| &\le L_x \|x_1 - x_2\| \\ \|\nabla_x f(x, y_1) - \nabla_x f(x, y_2)\| &\le L_x \|y_1 - y_2\| \\ \|\nabla_y f(x_1, y) - \nabla_y f(x_2, y)\| &\le L_y \|x_1 - x_2\| \\ \|\nabla_y f(x, y_1) - \nabla_y f(x, y_2)\| &\le L_y \|y_1 - y_2\|. \end{aligned} \tag{B.25}$$

Lemma B.1. *[31] Definition 1 means that f is also ℓ-weakly convex in the first component x, i.e.,*

$$f(\cdot, y) + \frac{\ell}{2}\|\cdot\|^2 \text{ is convex for all } y \in \text{dom } h. \tag{B.26}$$

Lemma B.2. *[14, Lemma 4.1(c)] If $f(x, y)$ is concave on y, then $f_{\mu_2}(x, y)$ is concave on y. If $f(x, y)$ has Lipschitz continuous gradients with constant ℓ, then both $f_{\mu_1}(x, y)$ and $f_{\mu_2}(x.y)$ have Lipschitz continuous gradients with constant $L_{\mu_1} \le \ell$ and $L_{\mu_2} \le \ell$, respectively.*

Proposition 5. *[8, Proposition 4.1] For all $r, \zeta \in \mathbb{R}^n$, if $w = prox^J_\eta(r - \zeta)$, where ζ is a stochastic gradient, then for all $z \in \mathbb{R}^n$, we have*

$$\mathbb{E}\langle \frac{1}{\eta}\zeta, w - z\rangle + J(w) - J(z) \le \frac{1}{2\eta}\mathbb{E}\|r - z\|^2 - \frac{1}{2\eta}\mathbb{E}\|r - w\|^2 - \frac{1}{2\eta}\mathbb{E}\|w - z\|^2. \tag{B.27}$$

Proposition 6. *If $p = prox^J_\eta(r - u)$, $q = prox^J_\eta(r - v)$, and $\mathbb{E}\|u - v\|^2 \le C_1^2 \mathbb{E}\|p - r\|^2 + C_2^2$, then for any $z \in \mathbb{R}^n$ we have*

$$\mathbb{E}\langle \frac{1}{\eta}v, p - z\rangle + J(p) - J(z) \le \frac{1}{2\eta}\mathbb{E}\|r - z\|^2 - \frac{1}{2\eta}\mathbb{E}\|q - z\|^2 - \left(\frac{1}{2\eta} - \frac{C_1^2}{2\eta}\right)\mathbb{E}\|r - p\|^2 + \frac{C_2^2}{2\eta}. \tag{B.28}$$

Proof. Applying Proposition 5 to p and q, respectively, for any $z \in \mathbb{R}^n$ we have

$$\mathbb{E}\langle \frac{1}{\eta}u, p - z\rangle + J(p) - J(z) \leq \frac{1}{2\eta}\mathbb{E}\|r - z\|^2 - \frac{1}{2\eta}\mathbb{E}\|r - p\|^2 - \frac{1}{2\eta}\mathbb{E}\|p - z\|^2, \tag{B.29}$$

$$\mathbb{E}\langle \frac{1}{\eta}v, q - z\rangle + J(q) - J(z) \leq \frac{1}{2\eta}\mathbb{E}\|r - z\|^2 - \frac{1}{2\eta}\mathbb{E}\|r - q\|^2 - \frac{1}{2\eta}\mathbb{E}\|q - z\|^2. \tag{B.30}$$

Let $z = q$ in (B.29), and we have

$$\mathbb{E}\langle \frac{1}{\eta}u, p - q\rangle + J(p) - J(q) \leq \frac{1}{2\eta}\mathbb{E}\|r - q\|^2 - \frac{1}{2\eta}\mathbb{E}\|r - p\|^2 - \frac{1}{2\eta}\mathbb{E}\|p - q\|^2. \tag{B.31}$$

Combining (B.30) and (B.31), then

$$\begin{aligned} &\mathbb{E}\langle \frac{1}{\eta}v, q - z\rangle + \mathbb{E}\langle \frac{1}{\eta}u, p - q\rangle + J(p) - J(z) \\ &\leq \frac{1}{2\eta}\mathbb{E}\|r - z\|^2 - \frac{1}{2\eta}\mathbb{E}\|q - z\|^2 - \frac{1}{2\eta}\mathbb{E}\|r - p\|^2 - \frac{1}{2\eta}\mathbb{E}\|p - q\|^2, \end{aligned} \tag{B.32}$$

which is equivalent to

$$\begin{aligned} &\mathbb{E}\langle \frac{1}{\eta}v, p - z\rangle + J(p) - J(z) \\ &\leq \mathbb{E}\langle \frac{1}{\eta}(v - u), p - q\rangle + \frac{1}{2\eta}\mathbb{E}\|r - z\|^2 - \frac{1}{2\eta}\mathbb{E}\|q - z\|^2 - \frac{1}{2\eta}\mathbb{E}\|r - p\|^2 - \frac{1}{2\eta}\mathbb{E}\|p - q\|^2 \\ &\leq \mathbb{E}\|\frac{1}{\eta}(v - u)\|\|p - q\| + \frac{1}{2\eta}\mathbb{E}\|r - z\|^2 - \frac{1}{2\eta}\mathbb{E}\|q - z\|^2 - \frac{1}{2\eta}\mathbb{E}\|r - p\|^2 - \frac{1}{2\eta}\mathbb{E}\|p - q\|^2 \\ &\leq \frac{1}{2\eta}\mathbb{E}\|v - u\|^2 + \frac{1}{2\eta}\mathbb{E}\|p - q\|^2 + \frac{1}{2\eta}\mathbb{E}\|r - z\|^2 - \frac{1}{2\eta}\mathbb{E}\|q - z\|^2 - \frac{1}{2\eta}\mathbb{E}\|r - p\|^2 - \frac{1}{2\eta}\mathbb{E}\|p - q\|^2 \\ &= \frac{1}{2\eta}\mathbb{E}\|v - u\|^2 + \frac{1}{2\eta}\mathbb{E}\|r - z\|^2 - \frac{1}{2\eta}\mathbb{E}\|q - z\|^2 - \frac{1}{2\eta}\mathbb{E}\|r - p\|^2 \\ &\leq \frac{1}{2\eta}\mathbb{E}\|r - z\|^2 - \frac{1}{2\eta}\mathbb{E}\|q - z\|^2 - (\frac{1}{2\eta} - \frac{C_1^2}{2\eta})\mathbb{E}\|r - p\|^2 + \frac{C_2^2}{2\eta}, \end{aligned} \tag{B.33}$$

where the second inequality holds due to the Schwartz inequality and the third inequality holds due to the Young's inequality. □

Lemma B.3. *[51, Lemma 2.3] For ℓ-smooth function $f(x, y)$, let $q_1 = d_x, q_2 = d_y$. Then we have*

$$\|\hat{\nabla}_x f(x, y) - \nabla_x f(x, y)\|^2 \leq \frac{\mu_1^2 L_x^2 d_x}{4}, \quad \|\hat{\nabla}_y f(x, y) - \nabla_y f(x, y)\|^2 \leq \frac{\mu_2^2 L_y^2 d_y}{4}. \tag{B.34}$$

Lemma B.4. *[14, Lemma 4.1(b)] Suppose that $f(x, y)$ is smooth. In general, it holds that $\|\nabla_x f_{\mu_1}(x, y) - \nabla_x f(x, y)\|^2 \leq \frac{\mu_1^2 d_x^2 L_x^2}{4}$ and $\|\nabla_y f_{\mu_2}(x, y) - \nabla_y f(x, y)\|^2 \leq \frac{\mu_2^2 d_y^2 L_y^2}{4}$.*

Lemma B.5. *$f_{\mu_1}(x, y)$ is weakly-convex.*

Proof. According to [31], smoothness ensures weak convexity of $f(x, y)$ w.r.t. x. Moreover, $F(x, y) := f(x, y) + \frac{\|x\|^2}{2}$ is convex w.r.t. x, so $F_{\mu_1}(x, y) := \mathbb{E}_u F(x + \mu_1 u, y) = \mathbb{E}_u f(x + \mu_1 u, y) + \frac{\|x+\mu_1 u\|^2}{2}$ is convex. Let $z = x + \mu_1 u$, so $\mathbb{E}_u F(z, y) = \mathbb{E}_u f(z, y) + \frac{\|z\|^2}{2}$ is convex w.r.t. z, thus, $\mathbb{E}_u f(z, y)$ is weakly-convex. Thus, $\mathbb{E}_u f(x + \mu_1 u, y)$ is weakly-convex and $\mathbb{E}_u f(x + \mu_1 u, y) = f_{\mu_1}(x, y)$. In fact, the smoothness of $f_{\mu_1}(x, y)$ can directly ensure weak convexity of $f_{\mu_1}(x, y)$ w.r.t. x [31, Lemma A.1]. □

Based on the lemmas and propositions above, we provide the complexity analysis for our ZO-GDEGA algorithm as follows.

All analyzes are organized as follows. We first provide a complexity analysis of our ZO-GDEGA algorithm for solving NC-SC problems in Subsection B.1. Then based on this analysis, we provide a continuity-agnostic analysis (more relaxed condition) in Subsection B.2 for ZO-GDEGA solving NC-C problems. Finally, we provide tighter results for the NC-C setting in Subsections B.3 and B.4 under the Lipschitz continuity assumption.

B.1 Complexity Analysis of ZO-GDEGA for Solving Nonconvex-Strongly Concave Problems

We provide detailed derivation for the complexity results of the deterministic ZO-GDEGA to solve NC-SC problems. Because the derivation of the stochastic ZO-GDEGA is similar to that of the deterministic ZO-GDEGA, we only give some key results for the stochastic setting.

Lemma B.6 (Lipschitz continuity of the solution mapping)**.** *[4, Lemma 4.1] The solution map $y^*(x)$ which fulfills $\Gamma(x, y^*(x)) = \max_{y\in\mathbb{R}^{d_y}} \Gamma(x, y)$ for all $x \in \mathbb{R}^{d_x}$ is κ-Lipschitz.*

Lemma B.7. *For determinisitic ZO-GDEGA algorithm solving the NC-SC problems, the iterates $\{x_t\}_{t=1}^T$ satisfies the following inequality when $q_1 = d_x$,*

$$\frac{\eta_x}{2}\mathbb{E}\|w_{t+1}\|^2 \le \mathbb{E}[\psi(x_t)-\psi(x_{t+1})]+\big(\frac{\ell+\kappa\ell}{2}+(\kappa+1)^2\ell^2\eta_x-\frac{1}{2\eta_x}\big)\mathbb{E}\|x_{t+1}-x_t\|^2+2\eta_x\ell^2\delta_t+\frac{\eta_x\mu_1^2 d_x L_x^2}{2}. \tag{B.35}$$

Similarly, for stochastic ZO-GDEGA algorithm solving the NC-SC problems, the iterates $\{x_t\}_{t=1}^T$ satisfies the following inequality,

$$\frac{\eta_x}{2}\mathbb{E}\|w_{t+1}\|^2 \le \mathbb{E}[\psi(x_t)-\psi(x_{t+1})]+\big(\frac{\ell+\kappa\ell}{2}+(\kappa+1)^2\ell^2\eta_x-\frac{1}{2\eta_x}\big)\mathbb{E}\|x_{t+1}-x_t\|^2+2\ell^2\eta_x\delta_t+\mu_1^2 d_x^2 L_x^2\eta_x+\frac{\eta_x\sigma^2}{2b_1}, \tag{B.36}$$

where $w_t \in (\partial g + \nabla\Phi)(x_t)$.

Proof. From the optimality condition of the proximal operator under the NC-SC setting in Algorithm 1, we deduce that

$$0 \in \partial g(x_{t+1}) + \hat{\nabla}_x f(x_t, y_t) + \frac{1}{\eta_x}(x_{t+1} - x_t). \tag{B.37}$$

Then, we let $w_{t+1} := \frac{1}{\eta_x}(x_t - x_{t+1}) + \nabla\Phi(x_{t+1}) - \hat{\nabla}_x f(x_t, y_t) \in \partial g(x_{t+1}) + \nabla\Phi(x_{t+1})$ and bound the $\|w_{t+1}\|^2$ as follows:

$$\|w_{t+1}\|^2 = \frac{1}{\eta_x^2}\|x_t-x_{t+1}\|^2+\frac{2}{\eta_x}\langle x_t-x_{t+1}, \nabla\Phi(x_{t+1})-\hat{\nabla}_x f(x_t, y_t)\rangle+\|\nabla\Phi(x_{t+1})-\hat{\nabla}_x f(x_t, y_t)\|^2. \tag{B.38}$$

The $(\ell + \kappa\ell)$-smoothness [31, Lemma 4.3] of $\Phi(x)$ implies that

$$\langle\nabla\Phi(x_{t+1}), x_t - x_{t+1}\rangle - \frac{\ell + \kappa\ell}{2}\|x_{t+1} - x_t\|^2 \le \Phi(x_t) - \Phi(x_{t+1}). \tag{B.39}$$

Since the proximal operator minimizes a $\frac{1}{\eta_x}$-strongly convex function, we have that

$$\begin{aligned} &g(x_{t+1}) + \langle\hat{\nabla}_x f(x_t, y_t), x_{t+1} - x_t\rangle + \frac{1}{2\eta_x}\|x_{t+1} - x_t\|^2 + \frac{1}{2\eta_x}\|x_{t+1} - x\|^2 \\ &\le g(x) + \langle\hat{\nabla}_x f(x_t, y_t), x - x_t\rangle + \frac{1}{2\eta_x}\|x - x_t\|^2 \end{aligned} \tag{B.40}$$

for $\forall x \in \mathbb{R}^{d_x}$. Combining (B.39) and (B.40) and letting $x = x_t$, taking the expectation of both sides yields that

$$\mathbb{E}\langle\nabla\Phi(x_{t+1}) - \hat{\nabla}_x f(x_t, y_t), x_t - x_{t+1}\rangle \le \mathbb{E}[\psi(x_t) - \psi(x_{t+1})] + \left(\frac{\ell + \kappa\ell}{2} - \frac{1}{\eta_x}\right)\mathbb{E}\|x_{t+1} - x_t\|^2. \tag{B.41}$$

Lastly, by the Young's inequality, we deduce that

$$\begin{aligned} &\mathbb{E}\|\nabla\Phi(x_{t+1}) - \hat{\nabla}_x f(x_t, y_t)\|^2 \\ &= \mathbb{E}\|\nabla\Phi(x_{t+1}) - \nabla\Phi(x_t) + \nabla\Phi(x_t) - \hat{\nabla}_x f(x_t, y_t)\|^2 \\ &\le 2(\kappa + 1)^2\ell^2\mathbb{E}\|x_{t+1} - x_t\|^2 + 2\mathbb{E}\|\nabla\Phi(x_t) - \hat{\nabla}_x f(x_t, y_t)\|^2 \\ &\le 2(\kappa + 1)^2\ell^2\mathbb{E}\|x_{t+1} - x_t\|^2 + 4\|\nabla\Phi(x_t) - \nabla_x f(x_t, y_t)\|^2 + 4\mathbb{E}\|\nabla_x f(x_t, y_t) - \hat{\nabla}_x f(x_t, y_t)\|^2 \\ &\le 2(\kappa + 1)^2\ell^2\mathbb{E}\|x_{t+1} - x_t\|^2 + 4\ell^2\delta_t + \mu_1^2 d_x L_x^2, \end{aligned} \tag{B.42}$$

where $\delta_t := \|y^*(x_t) - y_t\|^2$. Plugging (B.42) and (B.41) into (B.38) yeilds the desired result. □

Lemma B.8. *For deterministic ZO-GDEGA algorithm solving NC-SC problems, the following statement holds for the generated sequence $\{z_{t+1}\}, \{y_{t+1}\}$ when we choose $q_2 = d_y$:*

$$\begin{aligned}
&\mathbb{E}\langle -\hat{\nabla}_y f(x_{t+1}, z_{t+1}), z_{t+1} - y\rangle + h(z_{t+1}) - h(y) \\
\le\, &\frac{1}{2\eta_y}\mathbb{E}\|y_t - y\|^2 - \frac{1}{2\eta_y}\mathbb{E}\|y_{t+1} - y\|^2 - (\frac{1}{2\eta_y} - \frac{3\eta_y \ell^2}{2})\mathbb{E}\|y_t - z_{t+1}\|^2 \\
&+ \frac{3\ell^2\eta_y\|x_{t+1} - x_t\|^2}{2} + \frac{3\eta_y\mu_2^2 L_y^2 d_y}{4}.
\end{aligned} \tag{B.43}$$

Similarly, for stochastic ZO-GDEGA algorithm solving NC-SC problems, the following statement holds for the generated sequence $\{z_{t+1}\}, \{y_{t+1}\}$ during algorithm proceeding:

$$\begin{aligned}
&\mathbb{E}\langle -\hat{\nabla}_y f(x_{t+1}, z_{t+1}), z_{t+1} - y\rangle + h(z_{t+1}) - h(y) \\
\le\, &\frac{1}{2\eta_y}\mathbb{E}\|y_t - y\|^2 - \frac{1}{2\eta_y}\mathbb{E}\|y_{t+1} - y\|^2 - (\frac{1}{2\eta_y} - \frac{3\eta_y \ell^2}{2})\mathbb{E}\|y_t - z_{t+1}\|^2 \\
&+ \frac{3\ell^2\eta_y\|x_{t+1} - x_t\|^2}{2} + \frac{3\eta_y\mu_2^2 L_y^2 d_y^2}{2} + \frac{3\sigma^2}{b_2}.
\end{aligned} \tag{B.44}$$

Proof. According to Proposition 6, we set $r = y_t$, $q = y_{t+1}$, $p = z_{t+1}$, $\eta = \eta_y$ and $v = -\eta_y\hat{\nabla}_y f(x_{t+1}, z_{t+1})$, $u = -\eta_y\hat{\nabla}_y f(x_t, y_t)$. We can verify that:

$$\begin{aligned}
\mathbb{E}\|u - v\|^2 &= \mathbb{E}\|\eta_y\hat{\nabla}_y f(x_{t+1}, z_{t+1}) - \eta_y\hat{\nabla}_y f(x_t, y_t)\|^2 \\
&\le \eta_y^2\mathbb{E}\big[3\|\hat{\nabla}_y f(x_{t+1}, z_{t+1}) - \nabla_y f(x_{t+1}, z_{t+1})\|^2 + 3\|\hat{\nabla}_y f(x_t, y_t) - \nabla_y f(x_t, y_t)\|^2 \\
&\quad + 3\|\nabla_y f(x_{t+1}, z_{t+1}) - \nabla_y f(x_t, y_t)\|^2\big] \\
&\le \eta_y^2\left(\frac{3\mu_2^2 L_y^2 d_y}{2} + 3\|\nabla_y f(x_{t+1}, z_{t+1}) - \nabla_y f(x_t, y_t)\|^2\right) \\
&\le 3\ell^2\eta_y^2\|p - r\|^2 + 3\ell^2\eta_y^2\|x_{t+1} - x_t\|^2 + \frac{3\eta_y^2\mu_2^2 L_y^2 d_y}{2},
\end{aligned} \tag{B.45}$$

where the second inequality holds due to Lemma B.3, and the third inequality holds due to the ℓ-smoothness of f. Thus, if we set $C_1^2 = 3\eta_y^2\ell^2$, and $C_2^2 = 3\ell^2\eta_y^2\|x_{t+1} - x_t\|^2 + \frac{3\eta_y^2\mu_2^2 L_y^2 d_y}{2}$, $J = h$, according to Proposition 6, we have the following inequality holding for any y:

$$\begin{aligned}
&\mathbb{E}\langle -\hat{\nabla}_y f(x_{t+1}, z_{t+1}), z_{t+1} - y\rangle + h(z_{t+1}) - h(y) \\
\le\, &\frac{1}{2\eta_y}\mathbb{E}\|y_t - y\|^2 - \frac{1}{2\eta_y}\mathbb{E}\|y_{t+1} - y\|^2 - (\frac{1}{2\eta_y} - \frac{3\eta_y \ell^2}{2})\mathbb{E}\|y_t - z_{t+1}\|^2 \\
&+ \frac{3\ell^2\eta_y\|x_{t+1} - x_t\|^2 + \frac{3\eta_y\mu_2^2 L_y^2 d_y}{2}}{2}.
\end{aligned} \tag{B.46}$$

□

Lemma B.9. *Let $\delta_t = \|y^*(x_t) - y_t\|^2$ and $\eta_y \le \min\{\frac{4}{\mu}, \frac{1}{2\ell}\}$, the following statement holds true,*

$$\sum_{t=0}^{T-1}\delta_t \le 8a\delta_0 + 8a\sum_{i=0}^{T-1}[((1 - \frac{1}{4a})(1 + 8a)\kappa^2 + 3\ell^2\eta_y^2)\mathbb{E}\|x_i - x_{i+1}\|^2 + M], \tag{B.47}$$

where $a = \frac{1}{\mu\eta_y}$, $M \triangleq 2\eta_y L_y\mu_2^2 + \frac{3\eta_y^2\mu_2^2 L_y^2 d_y}{2}$ for deterministic setting and $M \triangleq 2\eta_y L_y\mu_2^2 + 3\eta_y^2\mu_2^2 L_y^2 d_y^2 + \frac{6\eta_y\sigma^2}{b_2}$ for stochastic setting.

Proof. About deterministic ZO-GDEGA for solving nonconvex-strongly concave problems, according to Lemma B.8, the following statement holds for the generated sequence $\{z_{t+1}\}, \{y_{t+1}\}$ during algorithm proceeding:

$$\begin{aligned}
&\mathbb{E}\langle -\hat{\nabla}_y f(x_{t+1}, z_{t+1}), z_{t+1} - y\rangle + h(z_{t+1}) - h(y) \\
&\le \frac{1}{2\eta_y}\mathbb{E}\|y_t - y\|^2 - \frac{1}{2\eta_y}\mathbb{E}\|y_{t+1} - y\|^2 - (\frac{1}{2\eta_y} - \frac{3\eta_y \ell^2}{2})\mathbb{E}\|y_t - z_{t+1}\|^2 \\
&\quad + \frac{3\ell^2\eta_y\|x_{t+1} - x_t\|^2 + \frac{3\eta_y\mu_2^2 L_y^2 d_y}{2}}{2}
\end{aligned} \tag{B.48}$$

for $\forall y \in dom\ h$. The strong concavity of $f(x,\cdot)$ ensures that the solution set $Y^*(x) := \big\{y^*|y^* \in \arg\max_{y\in dom\ h}\{f(x,y) - h(y)\}\big\}$ is a singleton and consists of a single element $y^*(x)$ for a given x. Letting $y = y^*(x_{t+1})$, we have

$$\begin{aligned}
&\mathbb{E}\|y_{t+1} - y^*(x_{t+1})\|^2 \\
&\le \mathbb{E}\|y_t - y^*(x_{t+1})\|^2 + 2\eta_y[\langle \hat{\nabla}_y f(x_{t+1}, z_{t+1}), z_{t+1} - y^*(x_{t+1})\rangle - h(z_{t+1}) + h(y^*(x_{t+1}))] \\
&\quad - (1 - 3\eta_y^2\ell^2)\mathbb{E}\|y_t - z_{t+1}\|^2 + 3\ell^2\eta_y^2\|x_{t+1} - x_t\|^2 + \frac{3\eta_y^2\mu_2^2 L_y^2 d_y}{2}.
\end{aligned} \tag{B.49}$$

Taking the expectation of both sides of the above inequality conditioned on (x_{t+1}, z_{t+1}) yields that

$$\begin{aligned}
&\mathbb{E}[\|y_{t+1} - y^*(x_{t+1})\|^2|x_{t+1}, z_{t+1}] \\
&\le \mathbb{E}[\|y_t - y^*(x_{t+1})\|^2|x_{t+1}, z_{t+1}] + 2\eta_y[\langle \nabla_y f_{\mu_2}(x_{t+1}, z_{t+1}), z_{t+1} - y^*(x_{t+1})\rangle - h(z_{t+1}) + h(y^*(x_{t+1}))] \\
&\quad - (1 - 3\eta_y^2\ell^2)\mathbb{E}[\|y_t - z_{t+1}\|^2|x_{t+1}, z_{t+1}] + 3\ell^2\eta_y^2\mathbb{E}[\|x_{t+1} - x_t\|^2|x_{t+1}, z_{t+1}] + \frac{3\eta_y^2\mu_2^2 L_y^2 d_y}{2}.
\end{aligned} \tag{B.50}$$

Taking the expectation of both sides deduces that

$$\begin{aligned}
&\mathbb{E}\|y_{t+1} - y^*(x_{t+1})\|^2 \\
&\le \mathbb{E}\underbrace{\|y_t - y^*(x_{t+1})\|^2}_{J_1} + 2\eta_y\mathbb{E}\underbrace{[\langle \nabla_y f_{\mu_2}(x_{t+1}, z_{t+1}), z_{t+1} - y^*(x_{t+1})\rangle - h(z_{t+1}) + h(y^*(x_{t+1}))]}_{J_2} \\
&\quad - (1 - 3\eta_y^2\ell^2)\mathbb{E}\|y_t - z_{t+1}\|^2 + 3\ell^2\eta_y^2\mathbb{E}\|x_{t+1} - x_t\|^2 + \frac{3\eta_y^2\mu_2^2 L_y^2 d_y}{2}.
\end{aligned} \tag{B.51}$$

Now, we bound J_1 and J_2 as follows:

$$\begin{aligned}
J_1 &= (1 - \frac{1}{4a})\|y_t - y^*(x_{t+1})\|^2 + \frac{1}{4a}\|y_t - y^*(x_{t+1})\|^2 \\
&\le (1 - \frac{1}{4a})[(1 + \frac{1}{8a})\|y_t - y^*(x_t)\|^2 + (1 + 8a)\|y^*(x_t) - y^*(x_{t+1})\|^2] \\
&\quad + \frac{1}{2a}[\|y_t - z_{t+1}\|^2 + \|z_{t+1} - y^*(x_{t+1})\|^2] \\
&\le (1 - \frac{1}{8a})\|y_t - y^*(x_t)\|^2 + (1 - \frac{1}{4a})(1 + 8a)\kappa^2\|x_t - x_{t+1}\|^2 \\
&\quad + \frac{1}{2a}[\|y_t - z_{t+1}\|^2 + \|z_{t+1} - y^*(x_{t+1})\|^2],
\end{aligned} \tag{B.52}$$

where we let $a = \frac{1}{\mu\eta_y}$, the first inequality holds due to the Young's inequality, and the last inequality holds due to the κ-Lipschitz continuity of $y^*(\cdot)$ in Lemma B.6.

$$\begin{aligned}
J_2 &= \langle z_{t+1} - y^*(x_{t+1}), \nabla_y f_{\mu_2}(x_{t+1}, z_{t+1})\rangle - h(z_{t+1}) + h(y^*(x_{t+1})) \\
&\le\ -\frac{\mu}{2}\|z_{t+1} - y^*(x_{t+1})\|^2 + f_{\mu_2}(x_{t+1}, z_{t+1}) - h(z_{t+1}) - [f_{\mu_2}(x_{t+1}, y^*(x_{t+1})) - h(y^*(x_{t+1}))] \\
&\le\ -\frac{\mu}{2}\|z_{t+1} - y^*(x_{t+1})\|^2 + f(x_{t+1}, z_{t+1}) - h(z_{t+1}) - [f(x_{t+1}, y^*(x_{t+1})) - h(y^*(x_{t+1}))] + L_y\mu_2^2 \\
&\le\ -\frac{\mu}{2}\|z_{t+1} - y^*(x_{t+1})\|^2 + L_y\mu_2^2,
\end{aligned} \tag{B.53}$$

where the first inequality holds because the μ strong concavity of $f(x,\cdot)$ implies μ strong concavity of $f_{\mu_2}(x,\cdot)$ [39], the second inequality holds due to [14, Lemma 4.1 (b)], and the third inequality holds due to the definition of $y^*(x)$, i.e., $y^*(x) = \arg\max_y\{f(x,y) - h(y)\}$. Plugging (B.52) and (B.53) into (B.51), we have

$$\begin{aligned}
&\mathbb{E}\|y_{t+1} - y^*(x_{t+1})\|^2 \\
&\leq (1-\frac{1}{8a})\mathbb{E}\|y_t - y^*(x_t)\|^2 + (1-\frac{1}{4a})(1+8a)\kappa^2\mathbb{E}\|x_t - x_{t+1}\|^2 \\
&\quad - \left(\mu\eta_y - \frac{1}{2a}\right)\mathbb{E}\|z_{t+1} - y^*(x_{t+1})\|^2 - (1 - 3\eta_y^2\ell^2 - \frac{1}{2a})\mathbb{E}\|y_t - z_{t+1}\|^2 \\
&\quad + 2\eta_y L_y \mu_2^2 + 3\ell^2\eta_y^2\mathbb{E}\|x_{t+1} - x_t\|^2 + \frac{3\eta_y^2\mu_2^2L_y^2 d_y}{2}.
\end{aligned} \tag{B.54}$$

According to the setting of $a = \frac{1}{\eta_y\mu}$, we choose $\eta_y = \frac{1}{2\ell} < \frac{\sqrt{\mu^2+48\ell^2}-\mu}{12\ell^2}$, so we have

$$\begin{aligned}
&\mathbb{E}\|y_{t+1} - y^*(x_{t+1})\|^2 \\
&\leq (1-\frac{1}{8a})\mathbb{E}\|y_t - y^*(x_t)\|^2 + [8a\kappa^2 + 3\ell^2\eta_y^2]\mathbb{E}\|x_t - x_{t+1}\|^2 + 2\eta_y L_y\mu_2^2 + \frac{3\eta_y^2\mu_2^2L_y^2d_y}{2}.
\end{aligned} \tag{B.55}$$

To simplify the analysis, we let $M \triangleq 2\eta_y L_y \mu_2^2 + \frac{3\eta_y^2\mu_2^2L_y^2d_y}{2}$. By recursively applying (B.55), we obtain for $\forall t \geq 1$

$$\delta_t \leq \left(1-\frac{1}{8a}\right)^t\delta_0 + \sum_{j=1}^{t}\left(1-\frac{1}{8a}\right)^{t-j}[(8a\kappa^2 + 3\ell^2\eta_y^2)\mathbb{E}\|x_j - x_{j+1}\|^2 + M]. \tag{B.56}$$

Summing this inequality from $t = 1$ to $T-1$ deduces that

$$\sum_{t=1}^{T-1}\delta_t \leq \sum_{t=1}^{T-1}\left(1-\frac{1}{8a}\right)^t\delta_0 + \sum_{t=1}^{T-1}\sum_{i=1}^{t}\left(1-\frac{1}{8a}\right)^{t-j}[(8a\kappa^2 + 3\ell^2\eta_y^2)\mathbb{E}\|x_j - x_{j+1}\|^2 + M]. \tag{B.57}$$

We can write that

$$\begin{aligned}
&\sum_{t=1}^{T-1}\sum_{j=1}^{t}\left(1-\frac{1}{8a}\right)^{t-j}[(8a\kappa^2 + 3\ell^2\eta_y^2)\mathbb{E}\|x_j - x_{j+1}\|^2 + M]. \\
&= \sum_{i=1}^{T-1}[(8a\kappa^2 + 3\ell^2\eta_y^2)\mathbb{E}\|x_i - x_{i+1}\|^2 + M]\sum_{j=0}^{T-i-1}\left(1-\frac{1}{8a}\right)^j \\
&\leq 8a\sum_{i=1}^{T-1}[(8a\kappa^2 + 3\ell^2\eta_y^2)\mathbb{E}\|x_i - x_{i+1}\|^2 + M],
\end{aligned} \tag{B.58}$$

and

$$\sum_{t=0}^{T-1}\left(1-\frac{1}{8a}\right)^t = \frac{1-(1-\frac{1}{8a})^T}{1-(1-\frac{1}{8a})} \leq 8a \tag{B.59}$$

with $8a \geq 2 \iff \eta_y \leq \frac{4}{\mu}$. Adding δ_0 on both sides of (B.57) and plugging (B.58) and (B.59) into (B.57) yeilds that

$$\sum_{t=0}^{T-1}\delta_t \leq 8a\delta_0 + 8a\sum_{i=1}^{T-1}[(8a\kappa^2 + 3\ell^2\eta_y^2)\mathbb{E}\|x_i - x_{i+1}\|^2 + M]. \tag{B.60}$$

□

Theorem B.5 (Detailed description of Theorem 1)**.** *We choose* $\eta_x \leq \frac{1}{256\kappa^2\ell}$ *(we hope that* $\frac{\ell+\kappa\ell}{2} + (\kappa+1)^2\ell^2\eta_x - \frac{1}{2\eta_x} + 16a\ell^2\eta_x[8a\kappa^2 + 3\ell^2\eta_y] \leq 0$*, thus we can get* $\eta_x \leq \frac{1}{256\kappa^2\ell} \leq \frac{\sqrt{9(\ell+\kappa\ell)^2+128a\ell^2(8a\kappa^2+3\ell^2\eta_y)}-(\ell+\kappa\ell)}{4(\kappa\ell+\ell)^2+64a\ell^2(8a\kappa^2+3\ell^2\eta_y)}$*),* $\eta_y = \frac{1}{2\ell}$ *(due to* $\eta_y \leq \min\{\frac{4}{\mu}, \frac{1}{2\ell}\}$ *and* $\frac{4}{\mu} > \frac{1}{2\ell}$*). Under*

Assumptions 3, 4, 6 and 7, our ZO-GDEGA converges to an ϵ-stationary point for solving Problem (1), i.e., $\min_{1\le t\le T}$ *dist*$(-\partial g(x_t), \nabla\Phi(x_t)) \le \epsilon$*, with iteration number T bounded by*

$$\mathcal{O}\left(\frac{\kappa^2\ell\Delta_\psi + \kappa\ell^2\delta_0}{\epsilon^2}\right), \tag{B.61}$$

for both deterministic and stochastic settings, where $\Delta_\psi = \psi(x_0) - \min_x \psi(x)$. Thus, the overall complexity is bounded by $\mathcal{O}(\kappa^2(d_x + d_y)\epsilon^{-2})$ for the deterministic setting and $\mathcal{O}(\kappa^2(d_x + \kappa d_y)\epsilon^{-4})$ for the stochastic setting.

Proof. Taking summation from $t = 0$ to $t = T - 1$ of (B.35) in Lemma B.7, we get that

$$\begin{aligned}
&\frac{\eta_x}{2}\sum_{t=0}^{T-1}\mathbb{E}\|w_{t+1}\|^2 \\
&\le \mathbb{E}[\psi(x_0) - \psi(x_T)] + 16a\ell^2\eta_x\delta_0 + \frac{T\eta_x\mu_1^2 d_x L_x^2}{2} \\
&\quad + \big(\frac{\ell+\kappa\ell}{2} + (\kappa+1)^2\ell^2\eta_x - \frac{1}{2\eta_x} + 16a\ell^2\eta_x[8a\kappa^2 + 3\ell^2\eta_y^2]\big)\sum_{t=0}^{T-1}\mathbb{E}\|x_{t+1} - x_t\|^2 \\
&\quad + 16a\ell^2\eta_x T(2\eta_y L_y\mu_2^2 + \frac{3\eta_y^2\mu_2^2L_y^2d_y}{2}) \\
&\le \mathbb{E}[\psi(x_0) - \psi(x_T)] + 32\kappa\ell^2\eta_x\delta_0 + \frac{T\eta_x\mu_1^2 d_x L_x^2}{2} \\
&\quad + \kappa\ell^2\eta_x T(\frac{32\mu_2^2L_y}{\ell} + \frac{12\mu_2^2L_y^2d_y}{\ell^2}),
\end{aligned} \tag{B.62}$$

where the second inequality holds due to the choice of η_x, and $a = \frac{1}{\mu\eta_y} = \frac{2\ell}{\mu} = 2\kappa$. Thus

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|w_{t+1}\|^2 &\le \frac{2\mathbb{E}[\psi(x_0) - \psi(x_T)]}{T\eta_x} + \frac{64\kappa\ell^2\delta_0}{T} + \mu_1^2 d_x L_x^2 + 8\kappa\ell^2(\frac{8\mu_2^2L_y}{\ell} + \frac{3\mu_2^2L_y^2d_y}{\ell^2}) \\
&\le \mathcal{O}(\frac{\kappa^2\ell(\psi(x_0) - \psi(x_T)) + \kappa\ell^2\delta_0}{T}) + \mathcal{O}(\epsilon^2)
\end{aligned} \tag{B.63}$$

for deterministic settting.

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|w_{t+1}\|^2 \\
&\le \frac{2\mathbb{E}[\psi(x_0) - \psi(x_T)]}{T\eta_x} + \frac{64\kappa\ell^2\delta_0}{T} + 2\mu_1^2d_x^2L_x^2 + \frac{\sigma^2}{2b_1} + 8\kappa\ell^2(\frac{8\mu_2^2L_y}{\ell} + \frac{3\mu_2^2L_y^2d_y^2}{\ell^2} + \frac{12\sigma^2}{b_2\ell}) \\
&\le \mathcal{O}(\frac{\kappa^2\ell(\psi(x_0) - \psi(x_T)) + \kappa\ell^2\delta_0}{T}) + \mathcal{O}(\epsilon^2)
\end{aligned} \tag{B.64}$$

for stochastic settting. Choosing $\mu_1 = \mathcal{O}(\epsilon)$, $\mu_2 = \mathcal{O}(\epsilon\kappa^{-1/2})$, $q_1 = d_x$, and $q_2 = d_y$ for deterministic setting, and $\mu_1 \le \frac{\epsilon}{\sqrt{2}d_xL_x}$, $\mu_2 \le \frac{\epsilon}{4\sqrt{3\kappa}d_yL_y}$, $b_1 = \mathcal{O}(d_x\epsilon^{-2})$, and $b_2 = \mathcal{O}(d_y\kappa\epsilon^{-2})$ for stochastic setting, we can get the desired results. □

B.2 Continuity-Agnostic analysis for Our ZO-GDEGA Solving Nonconvex-Concave Problems

Theorem B.6 (A detailed description of Theorem 2)**.** *Under Assumptions 1 and 2, for any given $\epsilon > 0$ and arbitrary $\hat{y} \in dom\ h$, letting x_ϵ be such that dist*$(-\partial g(x_\epsilon), \nabla\hat{\Phi}(x_\epsilon)) \le \frac{\epsilon}{2\sqrt{6}}$, $\hat{f}(x,y) = f(x,y) - \frac{\hat{\mu}}{2}\|y - \hat{y}\|^2$, $\hat{\Phi}(x) \triangleq \max_y\{\hat{f}(x,y) - h(y)\}$*, our ZO-GDEGA algorithm applied to solve approximate NC-SC model:* $\min_x\max_y \hat{\Psi}(x,y) = g(x) + \hat{f}(x,y) - h(y)$ *with* $\hat{\mu} = \min\left\{\frac{\epsilon^2}{24\ell D_h^2},\ \frac{L_y}{L_x}\frac{\epsilon}{2\sqrt{6}D_h}\right\} = \mathcal{O}(\frac{\epsilon^2}{\ell D_h^2})$*, is guaranteed to generate a point x_ϵ such that* $\mathbb{E}[\|\nabla\psi_{1/2\ell}(x_\epsilon)\|] \le \epsilon$ *for solving NC-C* *problem (1) with total complexity* $\mathcal{O}(\frac{\ell^2}{\hat{\mu}^2}(d_x + d_y)\epsilon^{-2}) = \mathcal{O}((d_x + d_y)\epsilon^{-6})$ *for the deterministic setting and* $\mathcal{O}(\frac{\ell^2}{\hat{\mu}^2}(d_x + \frac{\ell}{\hat{\mu}} d_y)\epsilon^{-4}) = \mathcal{O}(d_x\epsilon^{-8} + d_y\epsilon^{-10})$ *for the stochastic setting (in this setting, we also need Assumptions 3, 6 and 7).*

Proof. Below we state some useful relations that will be used later in the proof. The definition of $\hat{\Psi}$ implies that for all (x, y), we have

$$\nabla_x \Gamma(x, y) = \nabla_x \hat{\Gamma}(x, y), \quad \|\nabla_y f(x, y) - \nabla_y \hat{f}(x, y)\| \le \hat{\mu} D_h, \tag{B.65}$$

where $\hat{\Gamma}(x, y) = \hat{f}(x, y) - h(y)$. We define $\hat{y}^*(x) \triangleq \arg\max_y \hat{\Gamma}(x, y)$. Following [60, Lemma 11], we have that

$$\hat{y}^*(x_\epsilon) = \mathrm{prox}_\alpha^h(\hat{y}^*(x_\epsilon) + \alpha \nabla_y \hat{f}(x_\epsilon, \hat{y}^*(x_\epsilon))). \tag{B.66}$$

Now we are ready for the proof of Theorem B.6. Let $y^+ \triangleq \mathrm{prox}_\alpha^h(\hat{y}^*(x_\epsilon) + \alpha \nabla_y f(x_\epsilon, \hat{y}^*(x_\epsilon)))$, then we have

$$\begin{aligned}
&\|y^+ - \hat{y}^*(x_\epsilon)\| \\
&= \|\mathrm{prox}_\alpha^h(\hat{y}^*(x_\epsilon) + \alpha \nabla_y \hat{f}(x_\epsilon, \hat{y}^*(x_\epsilon))) - \mathrm{prox}_\alpha^h(\hat{y}^*(x_\epsilon) + \alpha \nabla_y f(x_\epsilon, \hat{y}^*(x_\epsilon)))\| \\
&\le \alpha \|\nabla_y \hat{f}(x_\epsilon, \hat{y}^*(x_\epsilon)) - \nabla_y f(x_\epsilon, \hat{y}^*(x_\epsilon))\| \\
&\le \alpha \hat{\mu} D_h,
\end{aligned} \tag{B.67}$$

where the first equality is by the definitions of $\hat{y}^*(x)$ and y^+; the first inequality holds due to the non-expansiveness of proximal operators, and the last inequality holds due to (B.65). Recall that our ultimate goal is to show that $\|\nabla \psi_\lambda(x_\epsilon)\| \le \epsilon$. Now, considering $\mathrm{prox}_\lambda^\psi(x_\epsilon) = \arg\min_v \psi(v) + \frac{1}{2\lambda}\|v - x_\epsilon\|^2$, where $\lambda = \frac{1}{2\ell}$. According to [31, Lemma A.1], we have

$$\|\nabla \psi_\lambda(x_\epsilon)\|^2 = \frac{1}{\lambda^2}\|x_\epsilon - \mathrm{prox}_\lambda^\psi(x_\epsilon)\|^2. \tag{B.68}$$

Since $\psi(x)$ is weakly convex and $\lambda = \frac{1}{2\ell}$, $\psi(x) + \frac{1}{2\lambda}\|x - x_\epsilon\|^2$ is ℓ-strongly convex with the unique minimizer $\mathrm{prox}_\lambda^\psi(x_\epsilon)$ (the definition of proximal operator); we have

$$\begin{aligned}
&g(x_\epsilon) + \max_y \Gamma(x_\epsilon, y) - g(\mathrm{prox}_\lambda^\psi(x_\epsilon)) - \max_y \Gamma(\mathrm{prox}_\lambda^\psi(x_\epsilon), y) - \frac{1}{2\lambda}\|\mathrm{prox}_\lambda^\psi(x_\epsilon) - x_\epsilon\|^2 \\
&= \psi(x_\epsilon) - \psi(\mathrm{prox}_\lambda^\psi(x_\epsilon)) - \frac{1}{2\lambda}\|\mathrm{prox}_\lambda^\psi(x_\epsilon) - x_\epsilon\|^2 \\
&\ge \frac{\ell}{2}\|x_\epsilon - \mathrm{prox}_\lambda^\psi(x_\epsilon)\|^2 = \frac{\lambda^2 \ell}{2}\|\nabla \psi_\lambda(x_\epsilon)\|^2,
\end{aligned} \tag{B.69}$$

where the last inequality holds due to the ℓ-strongly convex $\psi(x) + \frac{1}{2\lambda}\|x - x_\epsilon\|^2$. In the following analysis, we will continue to polish the upper bound on $\|\nabla \psi_\lambda(x_\epsilon)\|^2$ on the left hand side of Eq. (B.69). Indeed,

$$\begin{aligned}
&g(x_\epsilon) + \max_y \Gamma(x_\epsilon, y) - g(\mathrm{prox}_\lambda^\psi(x_\epsilon)) - \max_y \Gamma(\mathrm{prox}_\lambda^\psi(x_\epsilon), y) - \frac{1}{2\lambda}\|\mathrm{prox}_\lambda^\psi(x_\epsilon) - x_\epsilon\|^2 \\
&= \max_y \Gamma(x_\epsilon, y) - \Gamma(x_\epsilon, y^+) + g(x_\epsilon) + \Gamma(x_\epsilon, y^+) - g(\mathrm{prox}_\lambda^\psi(x_\epsilon)) - \max_y \Gamma(\mathrm{prox}_\lambda^\psi(x_\epsilon), y) - \frac{1}{2\lambda}\|\mathrm{prox}_\lambda^\psi(x_\epsilon) - x_\epsilon\|^2 \\
&\le \max_y \Gamma(x_\epsilon, y) - \Gamma(x_\epsilon, y^+) + g(x_\epsilon) + \Gamma(x_\epsilon, y^+) - g(\mathrm{prox}_\lambda^\psi(x_\epsilon)) - \Gamma(\mathrm{prox}_\lambda^\psi(x_\epsilon), y^+) - \frac{1}{2\lambda}\|\mathrm{prox}_\lambda^\psi(x_\epsilon) - x_\epsilon\|^2 \\
&\le \max_y \Gamma(x_\epsilon, y) - \Gamma(x_\epsilon, y^+) + \|\mathrm{prox}_\lambda^\psi(x_\epsilon) - x_\epsilon\|\|\nabla_x \Gamma(x_\epsilon, y^+) + w_\epsilon\| - \frac{\ell}{2}\|\mathrm{prox}_\lambda^\psi(x_\epsilon) - x_\epsilon\|^2 \\
&\le \max_y \Gamma(x_\epsilon, y) - \Gamma(x_\epsilon, y^+) + \frac{\|\nabla_x \Gamma(x_\epsilon, y^+) + w_\epsilon\|^2}{2\ell},
\end{aligned} \tag{B.70}$$

where $w_\epsilon \in \partial g(x_\epsilon)$. Since $dist(-\partial g(x_\epsilon), \nabla \hat{\Phi}(x_\epsilon)) \le \frac{\epsilon}{2\sqrt{6}}$, we can choose $w_\epsilon = \arg\min_{w \in \partial g(x_\epsilon)} \|w + \nabla \hat{\Phi}(x_\epsilon)\|^2$, the second inequality follows from the ℓ-strongly convexity of

$g(x) + \Gamma(x, y^+) + \frac{1}{2\lambda}\|x - x_\epsilon\|^2$ and Cauchy-Schwarz inequality, the last inequality holds due to Young's inequality. Then, using the smoothness of $f(x, y)$ and Eq. (B.67), we have

$$\begin{aligned}
&\|\nabla_x \Gamma(x_\epsilon, y^+) + w_\epsilon\| \\
\le\ &\|\nabla_x \Gamma(x_\epsilon, y^+) + w_\epsilon - (w_\epsilon + \nabla_x \hat{\Phi}(x_\epsilon))\| + \|w_\epsilon + \nabla_x \hat{\Phi}(x_\epsilon)\| \\
\le\ &\|\nabla_x \Gamma(x_\epsilon, y^+) - \nabla_x \Gamma(x_\epsilon, y^*(x_\epsilon))\| + \|\nabla_x \hat{\Phi}(x_\epsilon) + w_\epsilon\| \\
\le\ &L_x \alpha \hat{\mu} D_h + \|\nabla_x \hat{\Phi}(x_\epsilon) + w_\epsilon\|,
\end{aligned} \tag{B.71}$$

where the second inequality holds due to the Danskin's theorem and (B.65), and last inequality holds due to (B.67). Thus, using $(a + b)^2 \le 2(a^2 + b^2)$ for any $a, b \in \mathbb{R}$, we get

$$\|\nabla_x \Gamma(x_\epsilon, y^+) + w_\epsilon\|^2 \le 2L_x^2 \alpha^2 \hat{\mu}^2 D_h^2 + \|\nabla_x \hat{\Phi}(x_\epsilon) + w_\epsilon\|^2 \le 2L_x^2 \alpha^2 \hat{\mu}^2 D_h^2 + \frac{\epsilon^2}{12}, \tag{B.72}$$

where the last inequality holds due to the fact that $dist(-\partial g(x_\epsilon), \nabla \hat{\Phi}(x_\epsilon)) \le \frac{\epsilon}{2\sqrt{6}}$. Next, we bound $\max_y \Gamma(x_\epsilon, y) - \Gamma(x_\epsilon, y^+)$ in (B.70). Recall that $y^+ = \text{prox}_\alpha^h(\hat{y}^*(x_\epsilon) + \alpha \nabla_y f(x_\epsilon, \hat{y}^*(x_\epsilon)))$, the first-order optimality condition yields that

$$-\frac{1}{\alpha}[y^+ - \hat{y}^*(x_\epsilon) - \alpha \nabla_y f(x_\epsilon, \hat{y}^*(x_\epsilon))] \in \partial h(y^+),$$

Therefore, for $\forall y \in dom\ h$, we have that

$$h(y) - h(y^+) \ge \langle y - y^+, -\frac{1}{\alpha}(y^+ - \hat{y}^*(x_\epsilon) - \alpha \nabla_y f(x_\epsilon, \hat{y}^*(x_\epsilon)))\rangle, \tag{B.73}$$

which is equivalent to

$$h(y^+) - h(y) \le \frac{1}{\alpha}\langle y - y^+, y^+ - \hat{y}^*(x_\epsilon)\rangle - \langle y - y^+, \nabla_y f(x_\epsilon, \hat{y}^*(x_\epsilon))\rangle. \tag{B.74}$$

Now, we are ready to provide an upper bound on $\max_{y \in dom\ h} \Gamma(x_\epsilon, y) - \Gamma(x_\epsilon, y^+)$. Indeed, given any $\widetilde{y} \in \arg\max_{y \in dom\ h} \Gamma(x_\epsilon, y)$, we have

$$\begin{aligned}
&\max_{y \in dom\ h} \Gamma(x_\epsilon, y) - \Gamma(x_\epsilon, y^+) = \Gamma(x_\epsilon, \widetilde{y}) - \Gamma(x_\epsilon, \hat{y}^*(x_\epsilon)) + \Gamma(x_\epsilon, \hat{y}^*(x_\epsilon)) - \Gamma(x_\epsilon, y^+) \\
=\ &\underbrace{f(x_\epsilon, \widetilde{y}) - f(x_\epsilon, \hat{y}^*(x_\epsilon))}_{M_4} - h(\widetilde{y}) + h(\hat{y}^*(x_\epsilon)) + \underbrace{f(x_\epsilon, \hat{y}^*(x_\epsilon)) - f(x_\epsilon, y^+)}_{M_5} - h(\hat{y}^*(x_\epsilon)) + h(y^+)
\end{aligned} \tag{B.75}$$

We use concavity and smoothness of $f(x_\epsilon, \cdot)$ for M_4 and M_5, respectively. Thus,

$$\begin{aligned}
&\max_{y \in dom\ h} \Gamma(x_\epsilon, y) - \Gamma(x_\epsilon, y^+) \\
\le\ &\langle \nabla_y f(x_\epsilon, \hat{y}^*(x_\epsilon)), \widetilde{y} - \hat{y}^*(x_\epsilon)\rangle - h(\widetilde{y}) + h(y^+) + \langle \nabla_y f(x_\epsilon, \hat{y}_*(x_\epsilon)), \hat{y}_*(x_\epsilon) - y^+\rangle + \frac{L_y}{2}\|\hat{y}_*(x_\epsilon) - y^+\|^2 \\
=\ &\langle \nabla_y f(x_\epsilon, \hat{y}_*(x_\epsilon)), \tilde{y} - y^+\rangle - h(\tilde{y}) + h(y^+) + \frac{L_y}{2}\|\hat{y}_*(x_\epsilon) - y^+\|^2 \\
\le\ &\frac{1}{\alpha}\langle \tilde{y} - y^+, y^+ - \hat{y}_*(x_\epsilon)\rangle + \frac{L_y}{2}\|\hat{y}_*(x_\epsilon) - y^+\|^2 \\
=\ &-\frac{L_y}{2}\|\hat{y}_*(x_\epsilon) - y^+\|^2 + L_y\langle \tilde{y} - \hat{y}_*(x_\epsilon), y^+ - \hat{y}_*(x_\epsilon)\rangle \\
\le\ &L_y D_h \|y^+ - \hat{y}_*(x_\epsilon)\|,
\end{aligned} \tag{B.76}$$

where the second inequality holds due to the above optimality condition (B.74); in the last equality, we set $\alpha = L_y^{-1}$; the last inequality holds due to Cauchy-Schwarz inequality and the fact that $\max_{y_1, y_2 \in dom\ h} \|y_1 - y_2\| \le D_h$. Next, we plug (B.76) into (B.70) and it follows that

$$\begin{aligned}
&g(x_\epsilon) + \max_{y \in dom\ h} \Gamma(x_\epsilon, y) - g(\text{prox}_\lambda^\psi(x_\epsilon)) - \max_{y \in dom\ h} \Gamma(\text{prox}_\lambda^\psi(x_\epsilon), y) - \frac{1}{2\lambda}\|\text{prox}_\lambda^\psi(x_\epsilon) - x_\epsilon\|^2 \\
\le\ &L_y D_h \|y^+ - \hat{y}^*(x_\epsilon)\| + \frac{\|\nabla_x \Gamma(x_\epsilon, y^+) + w_\epsilon\|^2}{2\ell} \\
\le\ &\hat{\mu} D_h^2 + \frac{\epsilon^2}{24\ell} + \frac{L_x^2}{L_y^2} \cdot \frac{\hat{\mu}^2}{\ell} \cdot D_h^2,
\end{aligned} \tag{B.77}$$

where the last inequality follows from (B.67) and (B.72) with $\alpha = L_y^{-1}$. Combining (B.69) and (B.77), we have

$$\frac{1}{8\ell}\|\nabla\psi_\lambda(x_\epsilon)\|^2 \le \hat{\mu}D_h^2 + \frac{\epsilon^2}{24\ell} + \frac{L_x^2}{L_y^2}\frac{\hat{\mu}^2}{\ell}D_h^2. \tag{B.78}$$

Thus,

$$\|\nabla\psi_\lambda(x_\epsilon)\|^2 \le 8\ell\hat{\mu}D_h^2 + \frac{\epsilon^2}{3} + \frac{8L_x^2}{L_y^2}\hat{\mu}^2 D_h^2. \tag{B.79}$$

Thus, we get$\|\nabla\psi_{1/2\ell}(x_\epsilon)\| \le \epsilon$ when setting $\hat{\mu} = \min\left\{\frac{\epsilon^2}{24\ell D_h^2},\ \frac{L_y}{L_x}\frac{\epsilon}{2\sqrt{6}D_h}\right\}$. □

B.3 Continuity-dependent Analysis of Deterministic ZO-GDEGA for Solving Nonconvex-Concave Problems

Lemma B.10. *For NC-C problems, we bound $\|x_{t+1} - x_t\|^2$ in our deterministic ZO-GDEGA algorithm as follows:*

Proof.

$$\begin{aligned}\|x_{t+1} - x_t\|^2 &\le 2\|\text{prox}_{\eta_x}^g(x_t - \eta_x\hat{\nabla}f(x_t,z_t)) - (x_t - \eta_x\hat{\nabla}f(x_t,z_t))\|^2 + 2\eta_x^2\|\hat{\nabla}f(x_t,z_t)\|^2 \\ &\le 2\eta_x^2L_g^2 + 2\eta_x^2\|\hat{\nabla}_x f(x_t,z_t)\|^2,\end{aligned} \tag{B.80}$$

where the second inequality holds due to Assumption 2. □

Lemma B.11. *We bound $\mathbb{E}_u\|\hat{\nabla}_x f(x,y)\|$ when $q_1 = d_x$.*

Proof.

$$\begin{aligned}\mathbb{E}_u\|\hat{\nabla}_x f(x,y)\| = \mathbb{E}_u\sqrt{\|\hat{\nabla}_x f(x,y)\|^2} &\overset{\text{Jensen's Inequality}}{\le} \sqrt{\mathbb{E}_u\|\hat{\nabla}_x f(x,y)\|^2} \\ &\le \sqrt{2\mathbb{E}_u\|\hat{\nabla}_x f(x,y) - \nabla_x f(x,y)\|^2 + 2\|\nabla_x f(x,y)\|^2} \\ &\overset{\text{Lemma B.3 and Assumption 5}}{\le} \sqrt{\frac{\mu_1^2 d_x L_x^2}{2} + 2G^2}.\end{aligned} \tag{B.81}$$

□

Lemma B.12 ([4])**.** *The Lipschitz continuity of $f(x,y)$ in its first component implies that $\Phi(x)$ is Lipschitz as well with the same constant G.*

Lemma B.13. *We suppose that Assumptions 2 and 5 hold. For the deterministic ZO-GDEGA algorithm solving the NC-C problem (1), the iterates $\{x_t\}_{t=0}^T$ satisfies the following inequality:*

$$\mathbb{E}[\psi_{1/2\ell}(x_{t+1})] \le \mathbb{E}[\psi_{1/2\ell}(x_t)] + 2\eta_x\beta\ell\mathbb{E}[\Delta_t] - \frac{\eta_x}{2}\mathbb{E}\|\nabla\psi_{1/2\ell}(x_t)\|^2 + 2\eta_x\beta\ell^2\mu_1^2 + \eta_x^2\mu_1^2 d_x L_x^2\ell + 6\eta_x^2G^2\ell, \tag{B.82}$$

where $\beta = 1 - 2\ell\eta_x$ and $\Delta_t = \Phi(x_t) - \Gamma(x_t,z_t)$.

Proof. We define $\hat{x}_t := \text{prox}_{\frac{1}{2\ell}}^{g+\Phi}(x_t)$. According to the definition of Moreau envelope, we have that

$$\psi_{1/2\ell}(x_{t+1}) = \min_x\{\psi(x) + \ell\|x - x_{t+1}\|^2\} \le \psi(\hat{x}_t) + \ell\|\hat{x}_t - x_{t+1}\|^2, \tag{B.83}$$

and further taking an expectation of both sides of the above inequality,

$$\mathbb{E}\psi_{1/2\ell}(x_{t+1}) \le \mathbb{E}\psi(\hat{x}_t) + \ell\mathbb{E}\|\hat{x}_t - x_{t+1}\|^2. \tag{B.84}$$

According to [4, Lemma 3.2], we have $\hat{x}_t = \mathrm{prox}^g_{\eta_x}\left(2\ell\eta_x x_t - \eta_x v_t + (1-2\ell\eta_x)\hat{x}_t\right)$, where $v_t \in \partial\Phi(\hat{x}_t)$. Thus, with $\beta = (1-2\ell\eta_x)$,

$$\begin{aligned}
&\|\hat{x}_t - x_{t+1}\|^2\\
\overset{(a)}{=}&\|\mathrm{prox}^g_{\eta_x}(2\ell\eta_x x_t - \eta_x v_t + \beta\hat{x}_t) - \mathrm{prox}^g_{\eta_x}(x_t - \eta_x\hat{\nabla}_x f(x_t, z_t))\|^2\\
\overset{(b)}{\leq}&\|\beta(\hat{x}_t - x_t) + \eta_x(\hat{\nabla}_x f(x_t, z_t) - v_t)\|^2\\
=&\beta^2\|\hat{x}_t - x_t\|^2 + 2\eta_x\beta\langle\hat{\nabla}_x f(x_t, z_t) - v_t, \hat{x}_t - x_t\rangle + \eta_x^2\|\hat{\nabla}_x f(x_t, z_t) - v_t\|^2\\
\overset{(c)}{=}&\beta^2\|\hat{x}_t - x_t\|^2 + 2\eta_x\beta\langle\hat{\nabla}_x f(x_t, z_t) - v_t, \hat{x}_t - x_t\rangle + 2\eta_x^2\|\hat{\nabla}_x f(x_t, z_t)\|^2 + 2\eta_x^2 G^2,
\end{aligned} \tag{B.85}$$

where $v_t \in \partial\Phi(\hat{x}_t)$, the equality (a) holds due to the definitions of x_{t+1} and $\hat{x}_t$, the inequality (b) holds due to the non-expansiveness of the proximal operator, and the inequality (c) holds due to Lemma B.12. Then, according to [14, Lemma 4.1 (a)], we have that $\mathbb{E}[\hat{\nabla}_x f(x, y)] = \nabla_x f_{\mu_1}(x, y)$ and $\mathbb{E}[\hat{\nabla}_y f(x, y)] = \nabla_y f_{\mu_2}(x, y)$. Taking an expectation of both sides of the above inequality, conditioning on (x_t, z_t), together with Lemma B.11, yield that

$$\begin{aligned}
&\mathbb{E}\left[\|\hat{x}_t - x_{t+1}\|^2 | x_t, z_t\right]\\
\leq\ &\beta^2\|\hat{x}_t - x_t\|^2 + 2\eta_x\beta\langle\nabla_x f_{\mu_1}(x_t, z_t) - v_t, \hat{x}_t - x_t\rangle\\
&+ 2\eta_x^2\mathbb{E}[\|\hat{\nabla}_x f(x_t, z_t)\|^2 | x_t, z_t] + 2\eta_x^2 G^2.\\
\leq\ &\beta^2\|\hat{x}_t - x_t\|^2 + 2\eta_x\beta\langle\nabla_x f_{\mu_1}(x_t, z_t) - v_t, \hat{x}_t - x_t\rangle + \eta_x^2\mu_1^2 d_x L_x^2 + 6\eta_x^2 G^2.
\end{aligned} \tag{B.86}$$

Taking the expectation of both sides yields that

$$\mathbb{E}\|\hat{x}_t - x_{t+1}\|^2 \leq \beta^2\mathbb{E}\|\hat{x}_t - x_t\|^2 + 2\eta_x\beta\underbrace{\mathbb{E}\langle\nabla_x f_{\mu_1}(x_t, z_t) - v_t, \hat{x}_t - x_t\rangle}_{J_1} + \eta_x^2\mu_1^2 d_x L_x^2 + 6\eta_x^2 G^2. \tag{B.87}$$

According to the ℓ-weak convexity of f_{μ_1} in Lemma B.5, we obtain

$$\begin{aligned}
\langle\nabla_x f_{\mu_1}(x_t, z_t), \hat{x}_t - x_t\rangle &\overset{(a)}{\leq} f_{\mu_1}(\hat{x}_t, z_t) - f_{\mu_1}(x_t, z_t) + \frac{\ell}{2}\|\hat{x}_t - x_t\|^2\\
&\overset{(b)}{\leq} f(\hat{x}_t, z_t) - f(x_t, z_t) + \ell\mu_1^2 + \frac{\ell}{2}\|\hat{x}_t - x_t\|^2\\
&= f(\hat{x}_t, z_t) - h(z_t) - [f(x_t, z_t) - h(z_t)] + \ell\mu_1^2 + \frac{\ell}{2}\|\hat{x}_t - x_t\|^2\\
&\overset{(c)}{\leq} \Phi(\hat{x}_t) - \Gamma(x_t, z_t) + \frac{\ell}{2}\|\hat{x}_t - x_t\|^2 + \ell\mu_1^2,
\end{aligned} \tag{B.88}$$

where the inequality (a) holds due to Lemma B.5, the inequality (b) holds due to [14, Lemma 4.1 (b)], and the inequality (c) holds due to the definitions of $\Phi(\hat{x})$ and $\Gamma(x, y)$. And by the ℓ-weak convexity of $\Phi(x)$, we have

$$-\langle v_t, \hat{x}_t - x_t\rangle \leq \Phi(x_t) - \Phi(\hat{x}_t) + \frac{\ell}{2}\|\hat{x}_t - x_t\|^2. \tag{B.89}$$

Combining (B.88) and (B.89) and taking an expectation of both sides, we have

$$J_1 \leq \mathbb{E}[\Phi(x_t)] - \mathbb{E}[\Gamma(x_t, z_t)] + \ell\mathbb{E}\|\hat{x}_t - x_t\|^2 + \ell\mu_1^2. \tag{B.90}$$

By plugging (B.90) into (B.87), we have

$$\begin{aligned}
&\mathbb{E}\|\hat{x}_t - x_{t+1}\|^2\\
\leq\ &\beta^2\mathbb{E}\|\hat{x}_t - x_t\|^2 + 2\eta_x\beta\underbrace{\mathbb{E}\langle\nabla_x f_{\mu_1}(x_t, z_t) - v_t, \hat{x}_t - x_t\rangle}_{J_1} + \eta_x^2\mu_1^2 d_x L_x^2 + 6\eta_x^2 G^2\\
\leq\ &\beta^2\mathbb{E}\|\hat{x}_t - x_t\|^2 + 2\eta_x\beta\left[\mathbb{E}[\Phi(x_t)] - \mathbb{E}[\Gamma(x_t, z_t)] + \ell\mathbb{E}\|\hat{x}_t - x_t\|^2 + \ell\mu_1^2\right] + \eta_x^2\mu_1^2 d_x L_x^2 + 6\eta_x^2 G^2\\
\leq\ &(1 - 2\eta_x\ell)\mathbb{E}\|\hat{x}_t - x_t\|^2 + 2\eta_x\beta\mathbb{E}[\Delta_t] + 2\eta_x\beta\ell\mu_1^2 + \eta_x^2\mu_1^2 d_x L_x^2 + 6\eta_x^2 G^2,
\end{aligned} \tag{B.91}$$

where $\Delta_t = \Phi(x_t) - \Gamma(x_t, z_t)$ and the last inequality holds due to $\beta^2 + 2\eta_x \ell \beta = \beta(\beta + 2\eta_x \ell) = \beta$. Thus,

$$\begin{aligned}
&\mathbb{E}[\psi_{1/2\ell}(x_{t+1})] \\
&\le \mathbb{E}\psi(\hat{x}_t) + \ell \mathbb{E}\|\hat{x}_t - x_{t+1}\|^2 \\
&\le \mathbb{E}\psi(\hat{x}_t) + \ell \mathbb{E}\|\hat{x}_t - x_t\|^2 + 2\eta_x \beta \ell \mathbb{E}[\Delta_t] - 2\eta_x \ell^2 \mathbb{E}\|\hat{x}_t - x_t\|^2 + 2\eta_x \beta \ell^2 \mu_1^2 + \eta_x^2 \mu_1^2 d_x L_x^2 \ell + 6\eta_x^2 G^2 \ell \\
&= \mathbb{E}\psi_{1/2\ell}(x_t) + 2\eta_x \beta \ell \mathbb{E}[\Delta_t] - \frac{\eta_x}{2} \mathbb{E}\|\nabla \psi_{1/2\ell}(x_t)\|^2 + 2\eta_x \beta \ell^2 \mu_1^2 + \eta_x^2 \mu_1^2 d_x L_x^2 \ell + 6\eta_x^2 G^2 \ell,
\end{aligned} \tag{B.92}$$

where the last equality holds due to the definitions of $\hat{x}_t$ and $\psi_{1/2\ell}(x_t)$ and [31, Lemma A.1]. This completes the proof. □

Lemma B.14. *For deterministic ZO-GDEGA to solve the NC-C problem (1), the following statement holds for the generated sequences $\{z_{t+1}\}, \{y_{t+1}\}$ when $q_1 = d_x$ and $q_2 = d_y$:*

$$\begin{aligned}
&\mathbb{E}\langle -\hat{\nabla}_y f(x_{t+1}, z_{t+1}), z_{t+1} - y\rangle + h(z_{t+1}) - h(y) \\
&\le \frac{1}{2\eta_y} \mathbb{E}\|y_t - y\|^2 - \frac{1}{2\eta_y} \mathbb{E}\|y_{t+1} - y\|^2 - (\frac{1}{2\eta_y} - \frac{3\eta_y \ell^2}{2}) \mathbb{E}\|y_t - z_{t+1}\|^2 \\
&\quad + 3\eta_y \eta_x^2 \ell^2 (\frac{\mu_1^2 d_x L_x^2}{2} + 2G^2 + L_g^2) + \frac{3\mu_2^2 L_y^2 d_y \eta_y}{4}.
\end{aligned} \tag{B.93}$$

Proof. According to Proposition 6, we set $r = y_t$, $q = y_{t+1}$, $p = z_{t+1}$, $\eta = \eta_y$ and $v = -\eta_y \hat{\nabla}_y f(x_{t+1}, z_{t+1})$, $u = -\eta_y \hat{\nabla}_y f(x_t, y_t)$. We can verify that:

$$\begin{aligned}
\mathbb{E}\|u - v\|^2 &= \mathbb{E}\|\eta_y \hat{\nabla}_y f(x_{t+1}, z_{t+1}) - \eta_y \hat{\nabla}_y f(x_t, y_t)\|^2 \\
&\le \eta_y^2 \mathbb{E}\big[3\|\hat{\nabla}_y f(x_{t+1}, z_{t+1}) - \nabla_y f(x_{t+1}, z_{t+1})\|^2 + 3\|\hat{\nabla}_y f(x_t, y_t) - \nabla_y f(x_t, y_t)\|^2 \\
&\qquad + 3\|\nabla_y f(x_{t+1}, z_{t+1}) - \nabla_y f(x_t, y_t)\|^2\big] \\
&\le \eta_y^2 \left(\frac{3\mu_2^2 L_y^2 d_y}{2} + 3\mathbb{E}\|\nabla_y f(x_{t+1}, z_{t+1}) - \nabla_y f(x_t, y_t)\|^2 \right) \\
&\le \eta_y^2 \left(\frac{3\mu_2^2 L_y^2 d_y}{2} + 3\ell^2 \mathbb{E}\|z_{t+1} - y_t\|^2 + 3\ell^2 \mathbb{E}\|x_{t+1} - x_t\|^2 \right) \\
&\le \eta_y^2 \big(3\ell^2 \mathbb{E}\|p - r\|^2 + 6\eta_x^2 \ell^2 \mathbb{E}\|\hat{\nabla}_x f(x_t, z_t)\|^2 + 6\eta_x^2 L_g^2 \ell^2\big) + \frac{3\mu_2^2 L_y^2 d_y \eta_y^2}{2} \\
&\le \eta_y^2 \big(3\ell^2 \mathbb{E}\|p - r\|^2 + 6\eta_x^2 \ell^2 (\frac{\mu_1^2 d_x L_x^2}{2} + 2G^2) + 6\eta_x^2 L_g^2 \ell^2\big) + \frac{3\mu_2^2 L_y^2 d_y \eta_y^2}{2},
\end{aligned} \tag{B.94}$$

where the first inequality holds due to the Young's inequality, the second inequality holds due to Lemma B.3, the third inequality holds due to the ℓ-smoothness of f, the forth inequality holds due to Lemma B.10, and the last inequality holds due to Lemma B.11. Thus, if we set $C_1^2 = 3\eta_y^2 \ell^2$, and $C_2^2 = 6\eta_y^2 \eta_x^2 \ell^2 (\frac{\mu_1^2 d_x L_x^2}{2} + 2G^2 + L_g^2) + \frac{3\mu_2^2 L_y^2 d_y \eta_y^2}{2}$, $J = h$, we have the following inequality holding for any y:

$$\begin{aligned}
&\mathbb{E}\langle -\hat{\nabla}_y f(x_{t+1}, z_{t+1}), z_{t+1} - y\rangle + h(z_{t+1}) - h(y) \\
&\le \frac{1}{2\eta_y} \mathbb{E}\|y_t - y\|^2 - \frac{1}{2\eta_y} \mathbb{E}\|y_{t+1} - y\|^2 - (\frac{1}{2\eta_y} - \frac{3\eta_y \ell^2}{2}) \mathbb{E}\|y_t - z_{t+1}\|^2 \\
&\quad + \frac{6\eta_y \eta_x^2 \ell^2 (\frac{\mu_1^2 d_x L_x^2}{2} + 2G^2 + L_g^2) + \frac{3\mu_2^2 L_y^2 d_y \eta_y}{2}}{2}.
\end{aligned} \tag{B.95}$$

□

Lemma B.15. *We suppose that Assumption 1 holds. For deterministic ZO-GDEGA solving NC-C problems, denoting $R_1 \triangleq \sqrt{\frac{\mu_1^2 d_x L_x^2}{2} + 2G^2}$ and $\Delta_t = \Phi(x_t) - \Gamma(x_t, z_t)$, and letting $\eta_y = \frac{1}{\sqrt{3}\ell}$, the* *following statement holds true for* $\forall s \leq t-1$, $y \in dom\ h$ *and* $q_1 = d_x$,

$$\begin{aligned}\mathbb{E}[\Delta_{t+1}] \leq\ & 2(t-s+1)\eta_x(R_1+L_g)G \\ & + \frac{1}{2\eta_y}\mathbb{E}\|y_t - y^*(x_s)\|^2 - \frac{1}{2\eta_y}\mathbb{E}\|y_{t+1} - y^*(x_s)\|^2 - (\frac{1}{2\eta_y} - \frac{3\eta_y \ell^2}{2})\mathbb{E}\|y_t - z_{t+1}\|^2 \\ & + 3\eta_y\eta_x^2\ell^2(R_1^2 + L_g^2) + \frac{3\eta_y\mu_2^2 d_y L_y^2}{4} + \ell\mu_2^2.\end{aligned} \tag{B.96}$$

Proof.

$$\begin{aligned}\Delta_{t+1} &= \Phi(x_{t+1}) - \Gamma(x_{t+1}, y^*(x_s)) + \Gamma(x_{t+1}, y^*(x_s)) - \Gamma(x_{t+1}, z_{t+1}) \\ &\overset{(a)}{=} f(x_{t+1}, y^*(x_{t+1})) - h(y^*(x_{t+1})) - f(x_{t+1}, y^*(x_s)) \\ &\quad + h(y^*(x_s)) + f(x_{t+1}, y^*(x_s)) - h(y^*(x_s)) - f(x_{t+1}, z_{t+1}) + h(z_{t+1}) \\ &= f(x_{t+1}, y^*(x_{t+1})) - f(x_s, y^*(x_{t+1})) + f(x_s, y^*(x_{t+1})) - h(y^*(x_{t+1})) \\ &\quad - f(x_{t+1}, y^*(x_s)) + h(y^*(x_s)) + f(x_{t+1}, y^*(x_s)) - h(y^*(x_s)) - f(x_{t+1}, z_{t+1}) + h(z_{t+1}) \\ &\overset{(b)}{\leq} f(x_{t+1}, y^*(x_{t+1})) - f(x_s, y^*(x_{t+1})) + f(x_s, y^*(x_s)) - h(y^*(x_s)) \\ &\quad - f(x_{t+1}, y^*(x_s)) + h(y^*(x_s)) + f(x_{t+1}, y^*(x_s)) - h(y^*(x_s)) - f(x_{t+1}, z_{t+1}) + h(z_{t+1}) \\ &= \underbrace{f(x_{t+1}, y^*(x_{t+1})) - f(x_s, y^*(x_{t+1}))}_{M_1} + \underbrace{f(x_s, y^*(x_s)) - f(x_{t+1}, y^*(x_s))}_{M_2} \\ &\quad + \underbrace{f(x_{t+1}, y^*(x_s)) - f(x_{t+1}, z_{t+1}) - h(y^*(x_s)) + h(z_{t+1})}_{M_3},\end{aligned} \tag{B.97}$$

where the equality (a) holds due to the definitions of $\Phi(\cdot)$ and $\Gamma(\cdot,\cdot)$ and the inquality (b) holds due to the definition of $\Phi(x_s)$. Then, we bound M_1, M_2 and M_3 as follows:

$$\begin{aligned}\mathbb{E}[M_1] \leq & G\mathbb{E}\|x_{t+1} - x_s\| \leq G\mathbb{E}\sum_{l=s}^{t}\|x_{l+1} - x_l\| \\ \overset{(a)}{\leq} & G\sum_{l=s}^{t}(\mathbb{E}\|\text{prox}_{\eta_x}^g(x_l - \eta_x\hat{\nabla}_x f(x_l, z_l)) - \text{prox}_{\eta_x}^g(x_l)\| + \mathbb{E}\|\text{prox}_{\eta_x}^g(x_l) - x_l\|) \\ \overset{(b)}{\leq} & G\sum_{l=s}^{t}(\eta_x\mathbb{E}\|\hat{\nabla}_x f(x_l, z_l))\| + \eta_x L_g) \\ \overset{(c)}{\leq} & G\sum_{l=s}^{t}(\eta_x\sqrt{\mathbb{E}\|\hat{\nabla}_x f(x_l, z_l))\|^2} + \eta_x L_g) \\ \overset{(d)}{\leq} & (t-s+1)\eta_x(\sqrt{\frac{\mu_1^2 d_x L_x^2}{2} + 2G^2} + L_g)G,\end{aligned} \tag{B.98}$$

where the inequality (a) holds due to the Triangle inequality, the inequality (b) holds due to the non-expansiveness of the proximal operator and Assumption 2, and the inequality (c) and (d) hold due to Lemma B.11. Similarly, it can be concluded that $\mathbb{E}[M_2] \leq (t-s+1)\eta_x(\sqrt{\frac{\mu_1^2 d_x L_x^2}{2} + 2G^2} + L_g)G$. About M_3, we have the following derivation

$$\begin{aligned}\mathbb{E}[M_3] \leq & f_{\mu_2}(x_{t+1}, y^*(x_s)) - f_{\mu_2}(x_{t+1}, z_{t+1}) - h(y^*(x_s)) + h(z_{t+1}) + \ell\mu_2^2 \\ \overset{(a)}{\leq} & -\langle\nabla_y f_{\mu_2}(x_{t+1}, z_{t+1}), z_{t+1} - y^*(x_s)\rangle - h(y^*(x_s)) + h(z_{t+1}) + \ell\mu_2^2 \\ \overset{(b)}{\leq} & \frac{1}{2\eta_y}\mathbb{E}\|y_t - y^*(x_s)\|^2 - \frac{1}{2\eta_y}\mathbb{E}\|y_{t+1} - y^*(x_s)\|^2 - (\frac{1}{2\eta_y} - \frac{3\eta_y\ell^2}{2})\mathbb{E}\|y_t - z_{t+1}\|^2 \\ & + 3\eta_y\eta_x^2\ell^2(\frac{\mu_1^2 d_x L_x^2}{2} + 2G^2 + L_g^2) + \frac{3\mu_2^2 L_y^2 d_y \eta_y}{4} + \ell\mu_2^2,\end{aligned} \tag{B.99}$$

where the first inequality holds due to [14, Lemma 4.1 (b)], the inequality (a) holds due to the concavity of $f_{\mu_2}(x,\cdot)$ in Lemma B.2 and the inequality (b) holds due to Lemma B.14 with $y=y^*(x_s)$. Plugging M_1, M_2 and M_3 into (B.97) yields the desired result. □

Lemma B.16. *For deterministic ZO-GDEGA algorithm, letting* $\Delta_{t+1}=\Phi(x_{t+1})-\Gamma(x_{t+1},z_{t+1})$, *the following statement holds:*

$$\begin{aligned}&\frac{1}{T+1}\left(\sum_{t=0}^{T}\mathbb{E}[\Phi(x_{t+1})-\Gamma(x_{t+1},z_{t+1})]\right)\\ \le&\frac{1}{B}[2\eta_xB^2(R_1+L_g)G+\frac{D_h^2}{2\eta_y}+3B\eta_y(\eta_x^2\ell^2R_1^2+\eta_x^2L_g^2\ell^2+\frac{\mu_2^2L_y^2d_y}{4})+\ell\mu_2^2B].\end{aligned}\tag{B.100}$$

Proof. According to Lemma B.15 and $\eta_y\le\frac{1}{\sqrt{3}\ell}$, by splitting the summation into blocks we get that

$$\frac{1}{T+1}\left(\sum_{t=0}^{T}\mathbb{E}[\Delta_{t+1}]\right)=\frac{1}{T+1}\sum_{j=0}^{(T+1)/B-1}\left(\sum_{t=jB}^{(j+1)B-1}\mathbb{E}[\Delta_{t+1}]\right).\tag{B.101}$$

By using (B.96) with $s=-1$ for $j=0$, we get that

$$\begin{aligned}\sum_{t=0}^{B-1}\mathbb{E}[\Delta_{t+1}]\le&B^2\eta_x(\sqrt{\frac{\mu_1^2d_xL_x^2}{2}+2G^2}+L_g)G+\frac{1}{2\eta_y}\mathbb{E}\|y_0-y^*(x_{-1})\|^2-\frac{1}{2\eta_y}\mathbb{E}\|y_B-y^*(x_{-1})\|^2\\ &+3B\eta_y\eta_x^2\ell^2(\frac{\mu_1^2d_xL_x^2}{2}+2G^2+L_g^2)+\frac{3\eta_y\mu_2^2d_yL_y^2B}{4}+\ell\mu_2^2B.\end{aligned}\tag{B.102}$$

Analogously, for $j>0$ and $s=jB$ we have that

$$\begin{aligned}&\sum_{t=jB}^{(j+1)B-1}\mathbb{E}[\Delta_{t+1}]\\ \le&\eta_xB^2(\sqrt{\frac{\mu_1^2d_xL_x^2}{2}+2G^2}+L_g)G+\frac{1}{2\eta_y}\left(\mathbb{E}\|y_{jB}-y^*(x_{jB})\|^2-\|y_{(j+1)B}-y^*(x_{jB})\|^2\right)\\ &+3B\eta_y\eta_x^2\ell^2(\frac{\mu_1^2d_xL_x^2}{2}+2G^2+L_g^2)+\frac{3\eta_yB\mu_2^2L_y^2d_y}{4}+\ell\mu_2^2B.\end{aligned}\tag{B.103}$$

Plugging (B.102) and (B.103) into (B.101) yields that

$$\begin{aligned}&\frac{1}{T+1}\left(\sum_{t=0}^{T}\mathbb{E}[\Delta_{t+1}]\right)\\ \le&\frac{1}{B}\left[2\eta_xB^2(\sqrt{\frac{\mu_1^2d_xL_x^2}{2}+2G^2}+L_g)G+\frac{D_h^2}{2\eta_y}\right.\\ &\left.+3B\eta_y\eta_x^2\ell^2(\frac{\mu_1^2d_xL_x^2}{2}+2G^2+L_g^2)+\frac{3\eta_yB\mu_2^2L_y^2d_y}{4}+\ell\mu_2^2B\right]\\ \triangleq&\frac{1}{B}[2\eta_xB^2(R_1+L_g)G+\frac{D_h^2}{2\eta_y}+3B\eta_y(\eta_x^2\ell^2R_1^2+\eta_x^2L_g^2\ell^2+\frac{\mu_2^2L_y^2d_y}{4})+\ell\mu_2^2B],\end{aligned}\tag{B.104}$$

where $R_1\triangleq\sqrt{\frac{\mu_1^2d_xL_x^2}{2}+2G^2}$. □

Theorem B.7 (Restatement of Theorem 3)**.** *We suppose that Assumptions 1, 2 and 5 hold. If we choose* $\eta_x=\min\{\frac{\epsilon^4}{4096\sqrt{3}\beta^2\ell^3D_h^2G(R_1+L_g)},\frac{\epsilon^2}{16(2\ell\mu_1^2d_xL_x^2+12\ell G^2)},\frac{\epsilon}{16\ell\sqrt{R_1^2+L_g^2}}\}$, $\eta_y=\frac{1}{\sqrt{3}\ell}$, $q_1=d_x$, $q_2=d_y$, $\mu_1=\mathcal{O}(\epsilon)$ *and* $\mu_2=\mathcal{O}(\epsilon)$, *the deterministic ZO-GDEGA algorithm can be guaranteed to* *find ϵ-stationary point, i.e., $\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla\psi_{1/2\ell}(x_t)\|^2 \leq \epsilon^2$, with the iteration number T bounded by:*

$$\mathcal{O}\left(\frac{\ell G(R_1+L_g)\hat{\Delta}_\psi}{\epsilon^4}\max\{1,\frac{\ell^2 D_h^2}{\epsilon^2}\}\right),$$

where $\hat{\Delta}_\psi = \psi_{1/2\ell}(x_0) - \min_x \psi_{1/2\ell}(x)$. Thus, the overall complexity is bounded by $\mathcal{O}((d_x + d_y)\epsilon^{-6})$.

Proof. Summing up the inequality (B.82) in Lemma B.13 from 0 to T, together with Lemma B.16, we have

$$\begin{aligned}
&\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla\psi_{1/2\ell}(x_t)\|^2\\
\leq&\frac{2(\psi_{1/2\ell}(x_0)-\psi_{1/2\ell}(x_{T+1}))}{\eta_x(T+1)}\\
&+4\beta\ell\frac{1}{B}[2\eta_x B^2 G(R_1+L_g)+\frac{D_h^2}{2\eta_y}+3B\eta_y(\eta_x^2\ell^2R_1^2+\eta_x^2L_g^2\ell^2+\frac{\mu_2^2L_y^2d_y}{4})+\ell\mu_2^2B]\\
&+4\beta\ell^2\mu_1^2+2\eta_x\mu_1^2d_xL_x^2\ell+12\eta_xG^2\ell\\
=&\frac{2(\psi_{1/2\ell}(x_0)-\psi_{1/2\ell}(x_{T+1}))}{\eta_x(T+1)}\\
&+4\beta\ell[2\eta_x BG(R_1+L_g)+\frac{D_h^2}{2\eta_y B}+3\eta_y(\eta_x^2\ell^2R_1^2+\eta_x^2L_g^2\ell^2+\frac{\mu_2^2L_y^2d_y}{4})+\ell\mu_2^2]\\
&+4\beta\ell^2\mu_1^2+2\eta_x\mu_1^2d_xL_x^2\ell+12\eta_xG^2\ell.
\end{aligned} \tag{B.105}$$

We choose $B=\frac{D_h}{2\sqrt{\eta_x\eta_y G(R_1+L_g)}}$, thus we have

$$\begin{aligned}
&\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla\psi_{1/2\ell}(x_t)\|^2\\
\leq&\frac{2\hat{\Delta}_\psi}{\eta_x(T+1)}+4\beta\ell^2\mu_1^2+2\eta_x\mu_1^2d_xL_x^2\ell+12\eta_xG^2\ell\\
&+4\beta\ell[\sqrt{\eta_x}\frac{D_hG(R_1+L_g)}{\sqrt{\eta_yG(R_1+L_g)}}+\frac{D_h\sqrt{\eta_x\eta_yG(R_1+L_g)}}{\eta_y}+3\eta_y(\eta_x^2\ell^2R_1^2+\eta_x^2L_g^2\ell^2+\frac{\mu_2^2L_y^2d_y}{4})+\ell\mu_2^2].
\end{aligned} \tag{B.106}$$

Thus, we choose $\eta_x = \min\{\frac{\epsilon^4}{4096\sqrt{3}\ell^3D_h^2G(R_1+L_g)}, \frac{\epsilon^2}{32\ell\mu_1^2d_xL_x^2+192\ell G^2}, \frac{\epsilon}{16\ell\sqrt{R_1^2+L_g^2}}\}$, $\mu_1 \leq \frac{\epsilon}{8\sqrt{d_x}\ell}$, $\mu_2 \leq \frac{\epsilon}{4\sqrt{2L_y^2d_y+4\ell^2}}$ and $\eta_y = \frac{1}{\sqrt{3}\ell}$, we have

$$\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla\psi_{1/2\ell}(x_t)\|^2 \leq \frac{2\hat{\Delta}_\psi}{\eta_x(T+1)}+\frac{\epsilon^2}{2}. \tag{B.107}$$

This implies that the number of iterations required by ZO-GDEGA to find an ϵ-stationary point is bounded by

$$\mathcal{O}\left(\frac{\ell G(R_1+L_g)\hat{\Delta}_\psi}{\epsilon^4}\max\{1,\frac{\ell^2D_h^2}{\epsilon^2}\}\right), \tag{B.108}$$

which means that the overall complexity is bounded by $\mathcal{O}((d_x+d_y)\epsilon^{-6})$. This completes the proof. Note that this bound does not consist of the term $\hat{\Delta}_0 = \Phi(x_0) - (f(x_0,y_0) - h(y_0))$ compared to [31]. □

B.4 Continuity-dependent Analysis of Stochastic ZO-GDEGA for Solving Nonconvex-Concave Problems

Lemma B.17. *For stochastic ZO-GDEGA solving the NC-C problem (1), the iterates $\{x_t\}_{t=0}^{T}$ satisfies the following inequality*

$$\mathbb{E}[\psi_{1/2\ell}(x_{t+1})] \leq \mathbb{E}[\psi_{1/2\ell}(x_t)]+2\eta_x\ell\mathbb{E}[\Delta_t]-\frac{\eta_x}{2}\mathbb{E}\|\nabla\psi_{1/2\ell}(x_t)\|^2+2\eta_x\ell^2\mu_1^2+2\eta_x^2\ell(\frac{\mu_1^2 d_x^2 L_x^2}{2}+\frac{\sigma^2}{b_1})+6\eta_x^2\ell G^2, \tag{B.109}$$

with $\beta = (1-2\ell\eta_x)$, where $\Delta_t = \Phi(x_t) - \Gamma(x_t, z_t)$.

Proof. We first bound the term $\mathbb{E}_{U_1,\mathcal{I}_1}\|\hat{\nabla}_x f(x, y; \mathcal{I}_1)\|^2$ as follows:

$$\begin{aligned}\mathbb{E}_{U,\mathcal{I}_1}\left[\|\hat{\nabla}_x f(x, y; \mathcal{I}_1)\|^2\right] &= \mathbb{E}\left[\|\nabla_x f_{\mu_1}(x, y)\|^2\right] + \frac{\sigma_1^2}{b_1} \\ &\leq 2\mathbb{E}\left[\|\nabla_x f_{\mu_1}(x, y) - \nabla_x f(x, y)\|^2\right] + 2\mathbb{E}\left[\|\nabla_x f(x, y)\|^2\right] + \frac{\sigma_1^2}{b_1} \\ &\leq \frac{\mu_1^2 d_x^2 L_x^2}{2} + 2G^2 + \frac{\sigma^2}{b_1},\end{aligned} \tag{B.110}$$

From the definition of the Moreau envelope we deduce that

$$\mathbb{E}[\psi_{1/2\ell}(x_{t+1})] \leq \mathbb{E}[\psi(\hat{x}_t)] + \ell\mathbb{E}\left[\|\hat{x}_t - x_{t+1}\|^2\right]. \tag{B.111}$$

We yeild that for $\beta = 1 - 2\ell\eta_x$

$$\begin{aligned}&\|\hat{x}_t - x_{t+1}\|^2 \\ &\overset{(a)}{=} \|\text{prox}_{\eta_x}^{g}(2\ell\eta_x x_t - \eta_x v_t + \beta\hat{x}_t) - \text{prox}_{\eta_x}^{g}(x_t - \eta_x\hat{\nabla}_x f(x_t, z_t; \mathcal{I}_1))\|^2 \\ &\overset{(b)}{\leq} \|\beta(\hat{x}_t - x_t) + \eta_x(\hat{\nabla}_x f(x_t, z_t; \mathcal{I}_1) - v_t)\|^2 \\ &= \beta^2\|\hat{x}_t - x_t\|^2 + 2\eta_x\beta\langle\hat{\nabla}_x f(x_t, z_t; \mathcal{I}_1) - v_t, \hat{x}_t - x_t\rangle + \eta_x^2\|\hat{\nabla}_x f(x_t, z_t; \mathcal{I}_1) - v_t\|^2 \\ &\overset{(c)}{=} \beta^2\|\hat{x}_t - x_t\|^2 + 2\eta_x\beta\langle\hat{\nabla}_x f(x_t, z_t; \mathcal{I}_1) - v_t, \hat{x}_t - x_t\rangle + 2\eta_x^2\|\hat{\nabla}_x f(x_t, z_t; \mathcal{I}_1)\|^2 + 2\eta_x^2 G^2,\end{aligned} \tag{B.112}$$

where $v_t \in \partial\Phi(\hat{x}_t)$. Taking an expectation of both sides of the above inequality, conditioning on (x_t, z_t), yields that

$$\begin{aligned}&\mathbb{E}_{U,\mathcal{I}_1}\left[\|\hat{x}_t - x_{t+1}\|^2 | x_t, z_t\right] \\ \leq&\beta^2\|\hat{x}_t - x_t\|^2 + 2\eta_x\beta\langle\nabla_x f_{\mu_1}(x_t, z_t) - v_t, \hat{x}_t - x_t\rangle + 2\eta_x^2\mathbb{E}_{U,\mathcal{I}_1}[\|\hat{\nabla}_x f(x_t, z_t; \mathcal{I}_1)\|^2 | x_t, z_t] + 2\eta_x^2 G^2 \\ \leq&\beta^2\|\hat{x}_t - x_t\|^2 + 2\eta_x\beta\langle\nabla_x f_{\mu_1}(x_t, z_t) - v_t, \hat{x}_t - x_t\rangle + 2\eta_x^2(\frac{\mu_1^2 d_x^2 L_x^2}{2} + \frac{\sigma^2}{b_1}) + 6\eta_x^2 G^2,\end{aligned} \tag{B.113}$$

where the second inequality holds due to (B.110). Taking the expectation of both sides yields that

$$\mathbb{E}\|\hat{x}_t - x_{t+1}\|^2 \leq \beta^2\mathbb{E}\|\hat{x}_t - x_t\|^2 + 2\eta_x\beta\mathbb{E}\langle\nabla_x f_{\mu_1}(x_t, z_t) - v_t, \hat{x}_t - x_t\rangle + 2\eta_x^2(\frac{\mu_1^2 d_x^2 L_x^2}{2} + \frac{\sigma^2}{b_1}) + 6\eta_x^2 G^2. \tag{B.114}$$

According to (B.90), we have that

$$\begin{aligned}\mathbb{E}\|\hat{x}_t - x_{t+1}\|^2 \leq&\beta^2\mathbb{E}\|\hat{x}_t - x_t\|^2 + 2\eta_x\beta\big[\mathbb{E}[\Phi(x_t)] - \mathbb{E}[\Gamma(x_t, z_t)] + \ell\mathbb{E}\|\hat{x}_t - x_t\|^2 + \ell\mu_1^2\big] \\ &+ 2\eta_x^2(\frac{\mu_1^2 d_x^2 L_x^2}{2} + \frac{\sigma^2}{b_1}) + 6\eta_x^2 G^2 \\ =&(1 - 2\eta_x\ell)\mathbb{E}\|\hat{x}_t - x_t\|^2 + 2\eta_x\beta\mathbb{E}[\Delta_t] + 2\eta_x^2(\frac{\mu_1^2 d_x^2 L_x^2}{2} + \frac{\sigma^2}{b_1}) + 6\eta_x^2 G^2,\end{aligned} \tag{B.115}$$

where $\Delta_t = \Phi(x_t) - \Gamma(x_t, z_t)$. Thus,

$$
\begin{aligned}
\mathbb{E}[\psi_{1/2\ell}(x_{t+1})] \le\ & \mathbb{E}[\psi(\hat{x}_t)] + \ell\mathbb{E}\|\hat{x}_t - x_{t+1}\|^2 \\
\le\ & \mathbb{E}[\psi(\hat{x}_t)] + \ell\mathbb{E}\|\hat{x}_t - x_t\|^2 + 2\eta_x\beta\ell\mathbb{E}[\Delta_t] - 2\eta_x\ell^2\mathbb{E}\|\hat{x}_t - x_t\|^2 + 2\eta_x\beta\ell^2\mu_1^2 \\
& + 2\eta_x^2\ell(\frac{\mu_1^2 d_x^2 L_x^2}{2} + \frac{\sigma^2}{b_1}) + 6\eta_x^2\ell G^2. \\
=\ & \mathbb{E}[\psi_{1/2\ell}(x_t)] + 2\eta_x\beta\ell\mathbb{E}[\Delta_t] - \frac{\eta_x}{2}\mathbb{E}\|\nabla\psi_{1/2\ell}(x_t)\|^2 + 2\eta_x\beta\ell^2\mu_1^2 \\
& + 2\eta_x^2\ell(\frac{\mu_1^2 d_x^2 L_x^2}{2} + \frac{\sigma^2}{b_1}) + 6\eta_x^2\ell G^2,
\end{aligned}
\tag{B.116}
$$

where the last equality holds due to the definitions of $\hat{x}_t$ and $\psi_{1/2\ell}(x_t)$. This completes the proof. □

Lemma B.18. *For the NC-C setting of Algorithm 2, we bound $\|x_{t+1} - x_t\|^2$ as follows:*

Proof.

$$
\begin{aligned}
& \|x_{t+1} - x_t\|^2 \\
& \le 2\|\mathrm{prox}^g_{\eta_x}(x_t - \eta_x\hat{\nabla}_x f(x_t, y_t; \mathcal{I}_1)) - (x_t - \eta_x\hat{\nabla}_x f(x_t, y_t; \mathcal{I}_1))\|^2 + 2\eta_x^2\|\hat{\nabla}_x f(x_t, y_t; \mathcal{I}_1)\|^2 \\
& \le 2\eta_x^2 L_g^2 + 2\eta_x^2\|\hat{\nabla}_x f(x_t, y_t; \mathcal{I}_1)\|^2,
\end{aligned}
\tag{B.117}
$$

where the first inequality holds due to Cauchy–Schwarz inequality and the second inequality holds due to Assumption 2. □

Lemma B.19. *For stochastic ZO-GDEGA solving NC-C problems, the following statement holds for the generated sequences $\{z_{t+1}\}, \{y_{t+1}\}$ during algorithm proceeding:*

$$
\begin{aligned}
& \mathbb{E}\langle -\eta_y\hat{\nabla}_y f(x_{t+1}, z_{t+1}; \mathcal{I}_2), z_{t+1} - y\rangle + \eta_y h(z_{t+1}) - \eta_y h(y) \\
\le\ & \frac{1}{2\eta_y}\mathbb{E}\|y_t - y\|^2 - \frac{1}{2\eta_y}\mathbb{E}\|y_{t+1} - y\|^2 - (\frac{1}{2\eta_y} - \frac{3\eta_y\ell^2}{2})\mathbb{E}\|y_t - z_{t+1}\|^2 \\
& + 3\eta_y\eta_x^2\ell^2(\frac{\sigma^2}{b_1} + \frac{\mu_1^2 d_x^2 L_x^2}{2} + 2G^2 + L_g^2) + \frac{3\eta_y\sigma^2}{b_2}.
\end{aligned}
\tag{B.118}
$$

Proof. According to Proposition 6, we set $r = y_t$, $q = y_{t+1}$, $p = z_{t+1}$, $\eta = \eta_y$ and $v = -\eta_y\hat{\nabla}_y f(x_{t+1}, z_{t+1}; \mathcal{I}_2)$, $u = -\eta_y\hat{\nabla}_y f(x_t, y_t; \mathcal{I}_2)$. We can verify that:

$$
\begin{aligned}
\mathbb{E}\|u - v\|^2 &= \mathbb{E}\|\eta_y\hat{\nabla}_y f(x_{t+1}, z_{t+1}; \mathcal{I}_2) - \eta_y\hat{\nabla}_y f(x_t, y_t; \mathcal{I}_2)\|^2 \\
&\le \eta_y^2\mathbb{E}\big[3\|\hat{\nabla}_y f(x_{t+1}, z_{t+1}; \mathcal{I}_2) - \nabla_y f_{\mu_2}(x_{t+1}, z_{t+1})\|^2 + 3\|\hat{\nabla}_y f(x_t, y_t; \mathcal{I}_2) - \nabla_y f_{\mu_2}(x_t, y_t)\|^2 \\
&\quad + 3\|\nabla_y f_{\mu_2}(x_{t+1}, z_{t+1}) - \nabla_y f_{\mu_2}(x_t, y_t)\|^2\big] \\
&\le \eta_y^2\left(\frac{6\sigma^2}{b_2} + 3\mathbb{E}\|\nabla_y f_{\mu_2}(x_{t+1}, z_{t+1}) - \nabla_y f_{\mu_2}(x_t, y_t)\|^2\right) \\
&\le \eta_y^2\left(\frac{6\sigma^2}{b_2} + 3\ell^2\mathbb{E}\|z_{t+1} - y_t\|^2 + 3\ell^2\mathbb{E}\|x_{t+1} - x_t\|^2\right) \\
&\le \eta_y^2\big(3\ell^2\mathbb{E}\|p - r\|^2 + 6\eta_x^2\ell^2\mathbb{E}\|\hat{\nabla}_x f(x_t, y_t; \mathcal{I}_1)\|^2 + 6\eta_x^2 L_g^2\ell^2 + \frac{6\sigma^2}{b_2}\big) \\
&\le \eta_y^2\big(3\ell^2\mathbb{E}\|p - r\|^2 + 6\eta_x^2\ell^2(\frac{\mu_1^2 d_x^2 L_x^2}{2} + 2G^2 + \frac{\sigma^2}{b_1}) + 6\eta_x^2 L_g^2\ell^2 + \frac{6\sigma^2}{b_2}\big),
\end{aligned}
\tag{B.119}
$$

where the second inequality holds due to Eq. (A.4), the third inequality holds due to the L_{μ_2}-smoothness of f_{μ_2} with $L_{\mu_2} \le \ell$, the forth inequality holds due to Lemma B.18, and the last inequality holds due to (B.110). Thus, if we set $C_1^2 = 3\eta_y^2\ell^2$, and $C_2^2 = 3\eta_x^2\eta_y^2\ell^2(\frac{2\sigma^2}{b_1} + \mu_1^2 d_x^2 L_x^2 +$

$4G^2 + 2L_g^2) + \frac{6\eta_y^2\sigma^2}{b_2}$, $J = h$, we have the following inequality holding for any y:

$$
\begin{aligned}
&\mathbb{E}\langle -\hat{\nabla}_y f(x_{t+1}, z_{t+1}; \mathcal{I}_2), z_{t+1} - y\rangle + h(z_{t+1}) - h(y) \\
\leq\, & \frac{1}{2\eta_y}\mathbb{E}\|y_t - y\|^2 - \frac{1}{2\eta_y}\mathbb{E}\|y_{t+1} - y\|^2 - (\frac{1}{2\eta_y} - \frac{3\eta_y\ell^2}{2})\mathbb{E}\|y_t - z_{t+1}\|^2 \\
& + \frac{3\eta_y\eta_x^2\ell^2(\frac{2\sigma^2}{b_1} + \mu_1^2 d_x^2 L_x^2 + 4G^2 + 2L_g^2) + \frac{6\eta_y\sigma^2}{b_2}}{2}.
\end{aligned} \tag{B.120}
$$

□

Lemma B.20. *For stochasitic ZO-GDEGA to solve NC-C problems, letting $\Delta_{t+1} = \Phi(x_{t+1}) - \Gamma(x_{t+1}, z_{t+1})$ and $\eta_y \leq \frac{1}{\sqrt{3}\ell}$, the following statement holds true for $\forall s \leq t-1$ and $y \in dom\ h$,*

$$
\begin{aligned}
\mathbb{E}[\Delta_{t+1}] \leq\, & 2(t - s + 1)\eta_x G(\sqrt{\frac{\mu_1^2 d_x^2 \ell^2}{2} + 2G^2 + \frac{\sigma^2}{b_1}} + L_g) \\
& + \frac{1}{2\eta_y}\mathbb{E}\|y_t - y^*(x_s)\|^2 - \frac{1}{2\eta_y}\mathbb{E}\|y_{t+1} - y^*(x_s)\|^2 - (\frac{1}{2\eta_y} - \frac{3\eta_y\ell^2}{2})\mathbb{E}\|y_t - z_{t+1}\|^2 \\
& + 3\eta_y\eta_x^2\ell^2(\frac{\sigma^2}{b_1} + \frac{\mu_1^2 d_x^2 L_x^2}{2} + 2G^2 + L_g^2) + \frac{3\eta_y\sigma^2}{b_2} + \ell\mu_2^2.
\end{aligned} \tag{B.121}
$$

Proof. We decompose Δ_t into M_1, M_2 and M_3 similar to Lemma B.15. Then we bound M_1 as follows:

$$
\begin{aligned}
\mathbb{E}[M_1] \leq & G\mathbb{E}\|x_{t+1} - x_s\| \leq G\mathbb{E}\sum_{l=s}^{t}\|x_{l+1} - x_l\| \\
\leq & G\sum_{l=s}^{t}(\mathbb{E}\|\text{prox}_{\eta_x}^g(x_l - \eta_x\hat{\nabla}_x f(x_l, z_l; \mathcal{I}_1)) - \text{prox}_{\eta_x}^g(x_l)\| + \mathbb{E}\|\text{prox}_{\eta_x}^g(x_l) - x_l\|) \\
\leq & G\sum_{l=s}^{t}(\eta_x\mathbb{E}\|\hat{\nabla}_x f(x_l, z_l; \mathcal{I}_1))\| + \eta_x L_g) \\
\leq & G\sum_{l=s}^{t}(\eta_x\sqrt{\mathbb{E}\|\hat{\nabla}_x f(x_l, z_l; \mathcal{I}_1))\|^2} + \eta_x L_g) \\
\leq & (t - s + 1)\eta_x G(\sqrt{\frac{\mu_1^2 d_x^2 \ell^2}{2} + 2G^2 + \frac{\sigma^2}{b_1}} + L_g),
\end{aligned} \tag{B.122}
$$

where the last inequality holds due to Equation (B.110). Similarly, it can be concluded that $\mathbb{E}[M_2] \leq (t - s + 1)\eta_x G(\sqrt{\frac{\mu_1^2 d_x^2 \ell^2}{2} + 2G^2 + \frac{\sigma^2}{b_1}} + L_g)$. About M_3, we have the following derivation

$$
\begin{aligned}
\mathbb{E}[M_3] \leq & f_{\mu_2}(x_{t+1}, y^*(x_s)) - f_{\mu_2}(x_{t+1}, z_{t+1}) - h(y^*(x_s)) + h(z_{t+1}) + \ell\mu_2^2 \\
\overset{(a)}{\leq} & -\langle \nabla_y f_{\mu_2}(x_{t+1}, z_{t+1}), z_{t+1} - y^*(x_s)\rangle - h(y^*(x_s)) + h(z_{t+1}) + \ell\mu_2^2 \\
\overset{(b)}{\leq} & \frac{1}{2\eta_y}\mathbb{E}\|y_t - y^*(x_s)\|^2 - \frac{1}{2\eta_y}\mathbb{E}\|y_{t+1} - y^*(x_s)\|^2 - (\frac{1}{2\eta_y} - \frac{3\eta_y\ell^2}{2})\mathbb{E}\|y_t - z_{t+1}\|^2 \\
& + 3\eta_y\eta_x^2\ell^2(\frac{\sigma^2}{b_1} + \frac{\mu_1^2 d_x^2 L_x^2}{2} + 2G^2 + L_g^2) + \frac{3\eta_y\sigma^2}{b_2} + \ell\mu_2^2,
\end{aligned} \tag{B.123}
$$

where the first inequality holds due to [14, Lemma 4.1 (b)], inequality (a) holds due to the concavity of $f(x, \cdot)$ and inequality (b) holds due to Lemma B.19 with $y = y^*(x_s)$. Plugging M_1, M_2 and M_3 into Δ_{t+1} yields the desired result. □

Lemma B.21. *For stochastic ZO-GDEGA solving NC-C problems, letting $\Delta_{t+1} = \Phi(x_{t+1}) - \Gamma(x_{t+1}, z_{t+1})$, the following inequality holds:*

$$\begin{aligned}&\frac{1}{T+1}\left(\sum_{t=0}^{T}\mathbb{E}[\Phi(x_{t+1}) - \Gamma(x_{t+1}, z_{t+1})]\right)\\ \le&\frac{1}{B}[2\eta_x B^2 G(R_2 + L_g) + \frac{D_h^2}{2\eta_y} + B\eta_y(\frac{3\eta_x^2\ell^2\sigma^2}{b_1} + \frac{3}{2}\eta_x^2\ell^2\mu_1^2 d_x^2 L_x^2 + 6\eta_x^2\ell^2 G^2 + 3\eta_x^2 L_g^2\ell^2 + \frac{3\sigma^2}{b_2}) + \ell\mu_2^2 B].\end{aligned} \tag{B.124}$$

Proof. By splitting the summation into blocks we get that

$$\frac{1}{T+1}\left(\sum_{t=0}^{T}\mathbb{E}[\Delta_{t+1}]\right) = \frac{1}{T+1}\sum_{j=0}^{(T+1)/B-1}\left(\sum_{t=jB}^{(j+1)B-1}\mathbb{E}[\Delta_{t+1}]\right). \tag{B.125}$$

According to Lemma B.20 and $\eta_y \le \frac{1}{\sqrt{3}\ell}$, by using (B.121) with $s = -1$ for $j = 0$, we get that

$$\begin{aligned}\sum_{t=0}^{B-1}\mathbb{E}[\Delta_{t+1}] \le& B^2\eta_x G(\sqrt{\frac{\mu_1^2 d_x^2\ell^2}{2} + 2G^2 + \frac{\sigma^2}{b_1}} + L_g) + \frac{1}{2\eta_y}\mathbb{E}\|y_0 - y^*(x_{-1})\|^2 - \frac{1}{2\eta_y}\mathbb{E}\|y_B - y^*(x_{-1})\|^2\\ &+ 3B\eta_y\eta_x^2\ell^2(\frac{\sigma^2}{b_1} + \frac{\mu_1^2 d_x^2 L_x^2}{2} + 2G^2 + L_g^2) + \frac{3B\eta_y\sigma^2}{b_2} + \ell\mu_2^2 B.\end{aligned} \tag{B.126}$$

Analogously, for $j > 0$ and $s = jB$ we have that

$$\begin{aligned}&\sum_{t=jB}^{(j+1)B-1}\mathbb{E}[\Delta_{t+1}]\\ \le&\eta_x B^2 G(\sqrt{\frac{\mu_1^2 d_x^2\ell^2}{2} + 2G^2 + \frac{\sigma^2}{b_1}} + L_g) + \frac{1}{2\eta_y}\left(\mathbb{E}\|y_{jB} - y^*(x_{jB})\|^2 - \|y_{(j+1)B} - y^*(x_{jB})\|^2\right)\\ &+ 3B\eta_y(\frac{\eta_x^2\ell^2\sigma^2}{b_1} + \frac{\eta_x^2\ell^2\mu_1^2 d_x^2 L_x^2}{2} + 2\eta_x^2\ell^2 G^2 + \eta_x^2\ell^2 L_g^2 + \frac{\sigma^2}{b_2}) + \ell\mu_2^2 B.\end{aligned} \tag{B.127}$$

Plugging (B.126) and (B.127) into (B.125) together with Assumption 2 yields that

$$\begin{aligned}&\frac{1}{T+1}\left(\sum_{t=0}^{T}\mathbb{E}[\Delta_{t+1}]\right)\\ \le&\frac{1}{B}\Bigg[2\eta_x B^2 G(\sqrt{\frac{\mu_1^2 d_x^2\ell^2}{2} + 2G^2 + \frac{\sigma^2}{b_1}} + L_g) + \frac{D_h^2}{2\eta_y}\\ &+3B\eta_y\eta_x^2\ell^2(\frac{\sigma^2}{b_1} + \frac{\mu_1^2 d_x^2 L_x^2}{2} + 2G^2 + L_g^2) + \frac{3B\eta_y\sigma^2}{b_2} + \ell\mu_2^2 B\Bigg]\\ \triangleq&\frac{1}{B}[2\eta_x B^2 G(R_2 + L_g) + \frac{D_h^2}{2\eta_y} + B\eta_y(3\eta_x^2\ell^2 R_2^2 + 3\eta_x^2 L_g^2\ell^2 + \frac{3\sigma^2}{b_2}) + \ell\mu_2^2 B],\end{aligned} \tag{B.128}$$

where $R_2 \triangleq \sqrt{\frac{\mu_1^2 d_x^2\ell^2}{2} + 2G^2 + \frac{\sigma^2}{b_1}}$. □

Theorem B.8 (Detailed description of Theorem 4)**.** *We suppose that Assumptions 1, 2, 3, 5, 6 and 7 hold. If we choose the step sizes $\eta_y = \frac{1}{\sqrt{3}\ell}$ and $\eta_x = \min\{\frac{\epsilon^4}{16384\ell^3 D_h^2 G(R_2+L_g)}, \frac{\epsilon^2}{32\ell(\sigma^2+\mu_1^2 d_x^2 L_x^2)+192\ell G^2}, \frac{\epsilon}{4\ell\sqrt{R_2^2+L_g^2}}\}$, and let $b_1 = \mathcal{O}(d_x)$, $b_2 = \mathcal{O}(d_y\epsilon^{-2})$, $\mu_1 \le \mathcal{O}(\epsilon)$, $\mu_2 \le \mathcal{O}(\epsilon)$, the stochastic ZO-GDEGA can be guaranteed to find an ϵ-stationary point, i.e., $\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla\psi_{1/2\ell}(x_t)\|^2 \le \epsilon^2$, with the iteration complexity bounded by:*

$$\mathcal{O}\left(\frac{\ell G(R_2 + L_g)\hat{\Delta}_\psi}{\epsilon^4}\max\{1, \frac{\ell^2 D_h^2}{\epsilon^2}\}\right), \tag{B.129}$$

where $R_2 \triangleq \sqrt{\frac{\mu_1^2 d_x^2 \ell^2}{2} + 2G^2 + \frac{\sigma^2}{b_1}}$, *and* $\hat{\Delta}_\psi = \psi_{1/2\ell}(x_0) - \min_x \psi_{1/2\ell}(x)$. *Thus, the overall complexity is reduced to* $\mathcal{O}(d_x\epsilon^{-6} + d_y\epsilon^{-8})$.

Proof. Summing up the inequality (B.109) in Lemma B.17 from 0 to T, we have

$$\begin{aligned}
&\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla\psi_{1/2\ell}(x_t)\|^2 \\
\leq\ & \frac{2(\psi_{1/2\ell}(x_0) - \psi_{1/2\ell}(x_{T+1}))}{\eta_x(T+1)} + 4\ell^2\mu_1^2 + 2\eta_x\ell(\frac{2\sigma^2}{b_1} + \mu_1^2 d_x^2\ell^2) + 12\eta_x G^2\ell \\
&+ \frac{4\ell}{B}[2\eta_x B^2 G(R_2 + L_g) + \frac{D_h^2}{2\eta_y} + B\eta_y(3\eta_x^2\ell^2R_2^2 + 3\eta_x^2L_g^2\ell^2 + \frac{3\sigma^2}{b_2}) + \ell\mu_2^2 B] \qquad (B.130)\\
=\ & \frac{2(\psi_{1/2\ell}(x_0) - \psi_{1/2\ell}(x_{T+1}))}{\eta_x(T+1)} + 4\ell^2\mu_1^2 + 2\eta_x\ell(\frac{2\sigma^2}{b_1} + \mu_1^2 d_x^2\ell^2) + 12\eta_x G^2\ell \\
&+ 4\ell[2\eta_x BG(R_2 + L_g) + \frac{D_h^2}{2\eta_y B} + \eta_y(3\eta_x^2\ell^2R_2^2 + 3\eta_x^2L_g^2\ell^2 + \frac{3\sigma^2}{b_2}) + \ell\mu_2^2].
\end{aligned}$$

We choose $B = \frac{D_h}{2\sqrt{\eta_x\eta_y G(R_2+L_g)}}$, thus we have

$$\begin{aligned}
&\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla\psi_{1/2\ell}(x_t)\|^2 \\
\leq\ & \frac{2\hat{\Delta}_\psi}{\eta_x(T+1)} + 4\ell^2\mu_1^2 + 2\eta_x\ell(\frac{\sigma^2}{b_1} + \mu_1^2 d_x^2\ell^2) + 12\eta_x G^2\ell \qquad (B.131)\\
&+ 4\ell[\frac{2D_h\sqrt{\eta_x\eta_y G(R_2 + L_g)}}{\eta_y} + \eta_y(3\eta_x^2\ell^2R_2^2 + 3\eta_x^2L_g^2\ell^2 + \frac{3\sigma^2}{b_2}) + \ell\mu_2^2].
\end{aligned}$$

There are three cases to be analyzed.

• When $\eta_y = \frac{1}{\sqrt{3}\ell}$, we choose $\eta_x = \min\{\frac{\epsilon^4}{16384\ell^3 D_h^2 G(R_2+L_g)}, \frac{\epsilon^2}{32\ell(\sigma^2+\mu_1^2 d_x^2 L_x^2)+192\ell G^2}, \frac{\epsilon}{4\ell\sqrt{R_2^2+L_g^2}}\}$, $b_1 = \mathcal{O}(d_x)$, $b_2 = \mathcal{O}(d_y\epsilon^{-2})$, $\mu_1 \leq \frac{\epsilon}{2\ell\sqrt{d_x}}$, and $\mu_2 = \frac{\epsilon}{2\ell}$. Thus, the gradient complexity is bounded by:

$$\mathcal{O}\left(\frac{\ell G(R_2 + L_g)\hat{\Delta}_\psi}{\epsilon^4}\max\{1, \frac{\ell^2 D_h^2}{\epsilon^2}\}\right),$$

which means the overall complexity is bounded by $\mathcal{O}(d_x\epsilon^{-6} + d_y\epsilon^{-8})$.

• When $\eta_y = \frac{\epsilon^p}{12\ell}$, we choose $\eta_x = \min\{\frac{\epsilon^{4+p}}{16384\ell^3 D_h^2 G(R_2+L_g)}, \frac{\epsilon^2}{32\ell(\sigma^2+\mu_1^2 d_x^2 L_x^2)+192\ell G^2}, \frac{\epsilon^{1-\frac{p}{2}}}{16\ell\sqrt{R_2^2+L_g^2}}\}$, $b_1 = \mathcal{O}(d_x)$, $b_2 = \mathcal{O}(d_y\epsilon^{p-2})$, $\mu_1 = \mathcal{O}(\epsilon)$, and $\mu_2 = \mathcal{O}(\epsilon)$, where $0 < p < 2$. Thus, the gradient complexity is bounded by:

$$\mathcal{O}\left(\frac{\ell G(R_2 + L_g)\hat{\Delta}_\psi}{\epsilon^{4+p}}\max\{1, \frac{\ell^2 D_h^2}{\epsilon^2}\}\right),$$

which means the overall complexity is bounded by $\mathcal{O}(d_x\epsilon^{-6-p} + d_y\epsilon^{-8})$.

• When $\eta_y = \frac{\epsilon^2}{192\ell\sigma^2}$, we choose $\eta_x = \min\{\frac{\epsilon^6}{3145728\ell^3\sigma^2 D_h^2 G(R_2+L_g)}, \frac{\epsilon^2}{32\ell(\sigma^2+\mu_1^2 d_x^2 L_x^2)+192\ell G^2}\}$, $b_1 = \mathcal{O}(d_x)$, $b_2 = \mathcal{O}(d_y)$, $\mu_1 = \mathcal{O}(\epsilon)$, and $\mu_2 = \mathcal{O}(\epsilon)$. Thus, the gradient complexity is bounded by:

$$\mathcal{O}\left(\frac{\ell G(R_2 + L_g)\hat{\Delta}_\psi}{\epsilon^6}\max\{1, \frac{\ell^2 D_h^2 \sigma_y^2}{\epsilon^2}\}\right),$$

which means the overall complexity is bounded by $\mathcal{O}(d_x\epsilon^{-8} + d_y\epsilon^{-8})$.

In summary, we choose $\eta_x = \min\{\frac{\epsilon^4}{16384\ell^3 D_h^2 G(R_2+L_g)}, \frac{\epsilon^2}{32\ell(\sigma^2+\mu_1^2 d_x^2 L_x^2)+192\ell G^2}, \frac{\epsilon}{4\ell\sqrt{R_2^2+L_g^2}}\}$, $\mu_1 = \mathcal{O}(\epsilon)$, $\mu_2 = \mathcal{O}(\epsilon)$, $b_1 = \mathcal{O}(d_x)$, $b_2 = \mathcal{O}(d_y\epsilon^{-2})$ and $\eta_y = \frac{1}{\sqrt{3}\ell}$, we have

$$\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla\psi_{1/2\ell}(x_t)\|^2 \leq \frac{2\hat{\Delta}_\psi}{\eta_x(T+1)} + \frac{\epsilon^2}{2}. \tag{B.132}$$

This implies that the overall complexity required by stochastic ZO-GDEGA can be further bounded by

$$\mathcal{O}(d_x\epsilon^{-6} + d_y\epsilon^{-8}) \tag{B.133}$$

to return an ϵ-stationary point. This completes the proofs. □

C More Experimental Details and Results

This section provides more details and results for our experiments. Our code will be publicly available.

C.1 Data Poisoning Attack

This part provides more experimental details for solving the data poisoning attack problem. All the experiments were performed on Intel Core i5-11260H 2.6GHz CPU and 16GB RAM platform.

As is well known, when training deep neural networks, attention must be paid to adversarial examples that may lead to the misclassification of deep models. But the black-box model architecture is unknown to the adversary, it is necessary and key to solve Problems (5) with only the (more often noisy) evaluations of the objective functional values. About the model (5), we set $\mathcal{D}_{tr} \triangleq \mathcal{D}_{tr,p}\cup\mathcal{D}_{tr,c}$ denotes the training dataset including n samples, $\mathcal{D}_{tr,p}$ and $\mathcal{D}_{tr,c}$ denote the poisoned and clean subsets of $\mathcal{D}_{tr}$, respectively. We choose the loss function as $F_{tr}(\delta, w; \mathcal{D}_{tr}) := \mathcal{L}(\delta, w; \mathcal{D}_{tr,p}) + \mathcal{L}(0, w; \mathcal{D}_{tr,c})$, where $\mathcal{L}(\delta, w; \mathcal{D}) = -\frac{1}{|\mathcal{D}|}\sum_{(s_i,t_i)\in\mathcal{D}_p}[t_i\log(\mathcal{L}(\delta, w; s_i)) + (1-t_i)\log(1-\mathcal{L}(\delta, w; s_i))]$, and $\mathcal{L}(\delta, w; s_i) = \frac{1}{1+e^{-(s_i+\delta)^\top w}}$. In the data poisoning attack problem, the lower accuracy of the attacked model means the more effective the attack method.

Hyperparameter selection. For solving the NC-C problem (5), we choose the batch size guided by theory and considering the trade-off between time consumption and accuracy, while observing reasonably good performance. We set mini-batch size $b_1 = b_2 = 100$ for the `synthetic` dataset and $b_1 = b_2 = 10$ for the `epsilon_test` dataset and train all the methods for $T = 50,000$ iterations. Besides, we also choose the same step sizes $\eta_x = 0.02$, $\eta_y = 0.05$ for all the cases in the main paper. Note that ZO-AGP [51] requires that the step size of variable x is monotonically decreasing. Thus, we set its step sizes to be $\eta_x^t = \frac{2}{100+\sqrt{t}}$, $\eta_y = 0.05$ for a fair comparison. By the way, to test the accuracy of the zeroth-order gradient estimators, we set different numbers of random direction vectors $q_1 = q_2 = \{5, 20\}$ denoted by q to train all the methods. We set the poisoning ratio $|\mathcal{D}_{tr,p}|/|\mathcal{D}_{tr}| = 0.1$, smoothing parameters $\mu_1 = \mu_2 = 2\times 10^{-5}$ and the range of perturbation $r_x = 2$ as default values in the data poisoning attack experiments. The above hyperparameters are the same in other experiments unless explicitly stated. After training to get poisoned data, we retrain the logistic regression model 1000 times each using clean data and adversarial examples generated at each iteration.

C.1.1 Data Poisoning Attack for the NC-C Problem (5) on More Real-World Datasets

This part provides more experimental results for solving the NC-C problem (5) on the `w8a`, `a9a` and `HIGGS` datasets[4] as shown in Figs. C.4-C.6. **Note that the lower the accuracy is, the stronger the generated attack is, which means better performance.**

- The `w8a` dataset: It contains 49,749 samples of 300 dimensions and we split it into 70% training samples and 30% test samples. We set the batch size to 10.

- The `a9a` dataset: It contains 32,561 samples of 123 dimensions and we split it into 70% training samples and 30% test samples. We set the batch size to 10.

[4] https://www.csie.ntu.edu.tw/∼cjlin/libsvmtools/datasets/

• The HIGGS dataset: It contains 11,000,000 samples of 28 dimensions and we split it into 70% training samples and 30% test samples. We set the batch size to 512.

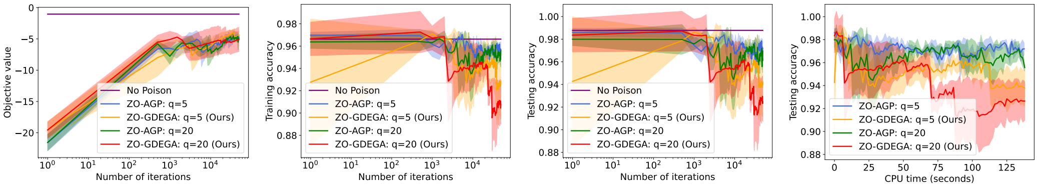


Figure C.4: Comparison of the experimental results for solving the data poisoning attack problem (5) on the w8a dataset.

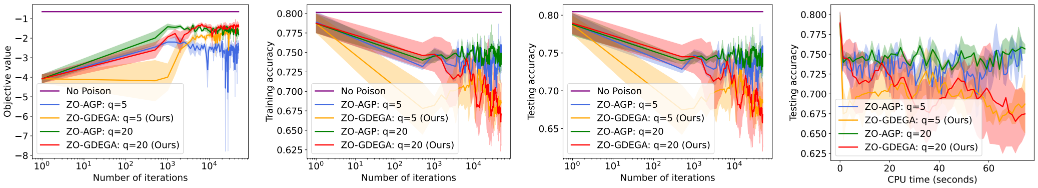


Figure C.5: Comparison of the experimental results for solving the data poisoning attack problem (5) on the a9a dataset.

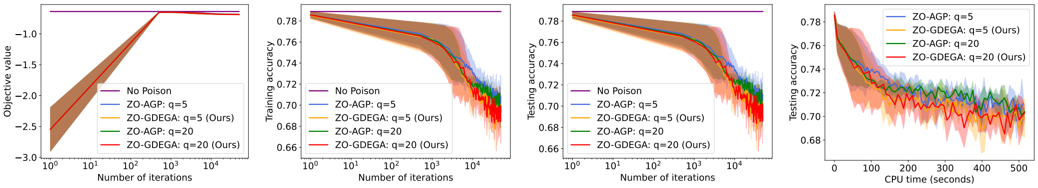


Figure C.6: Comparison of the experimental results for solving the data poisoning attack problem (5) on the HIGGS dataset.

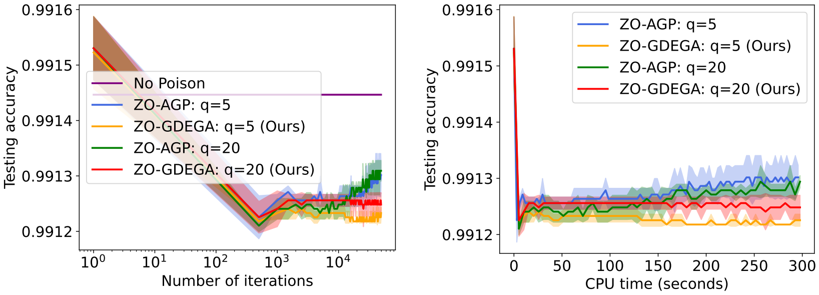


Figure C.7: Comparison of the experimental results for solving the data poisoning attack problem (5) on the bio_train dataset.

From Figs. C.4 -C.6, it can be seen that our ZO-GDEGA algorithm still performs better than baselines in reducing classification accuracy. This part also provides experimental results for solving the NC-C problem (5) on an unbalanced dataset, bio_train[5], where the proportion of positive samples is only 0.89%. The hyperparameter choices are the same as the experiments on the epsilon_test dataset. Experimental results are shown in Fig. C.7. It can be seen that our ZO-GDEGA algorithm performs better and the ZO-AGP algorithm even diverges.

[5] https://osmot.cs.cornell.edu/kddcup/datasets.html

C.1.2 Robustness Study for the NC-C Data Poisoning Attack Problem (5)

To verify the robustness of our ZO-GDEGA for solving NC-C problems, we conduct the data poisoning attack experiment under different smoothing parameters μ_1 and μ_2. We choose the same hyperparameters as in ***"Hyperparameter selection"*** above. The experimental results are shown in Fig. C.8. **Note that the lower the accuracy is, the stronger the generated attack is, which means better performance.** It can be found that the attack performance of our ZO-GDEGA algorithm performs always better than the ZO-AGP algorithm under different smoothing parameters μ_1 and μ_2.

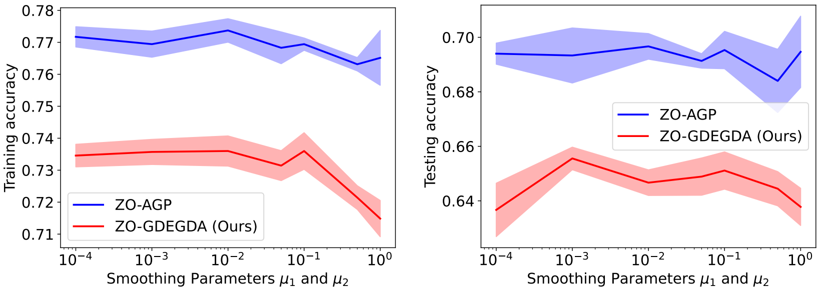


Figure C.8: The performance of ZO-GDEGA and ZO-AGP for solving NC-C Problem (5) with $q_1 = q_2 = 5$ and different μ_1 and μ_2 on the `synthetic` dataset, where $\mu_1 = \mu_2$. We can observe that our ZO-GDEGA algorithm is more robust compared with the baseline method, which is reflected in two aspects. On the one hand, under different smoothing parameter settings, the testing accuracies of the models attacked by our ZO-GDEGA have almost the same and small standard deviation. Whereas the model attacked by the poisoned data generated by ZO-AGP produces a different performance, i.e., the standard deviation of its test accuracy varies dramatically with the smoothing parameter. On the other hand, under large smoothing parameter settings (e.g., $\mu_1 = \mu_2 = 1$), our ZO-GDEGA can still reduce the testing classification accuracy more effectively than ZO-AGP.

C.1.3 Data Poisoning Attack against Sparse Logistic Regression

Sparse models are playing an increasingly important role in the fields such as machine learning and image processing. They have variable selection capabilities and can solve problems such as overfitting in modeling. In order to verify the universality of our ZO-GDEGA, we also conducted black-box attacks on sparse logistic regression models. The minimax formulation of this problem can be expressed as:

$$\max_{\|\delta\|_\infty \le r_x} \min_{w} F_{tr}(\delta, w; \mathcal{D}_{tr}) + \lambda \|w\|_1. \tag{C.134}$$

We test our ZO-GDEGA and ZO-AGP for solving Problem (C.134) on the `synthetic` dataset. We also choose the same hyperparameters as in ***"Hyperparameter selection"*** above and the experimental results are shown in Fig. C.9. **Note that the lower the accuracy is, the stronger the generated attack is, which means better performance.** It can be seen that our ZO-GDEGA still performs better than ZO-AGP in reducing accuracy.

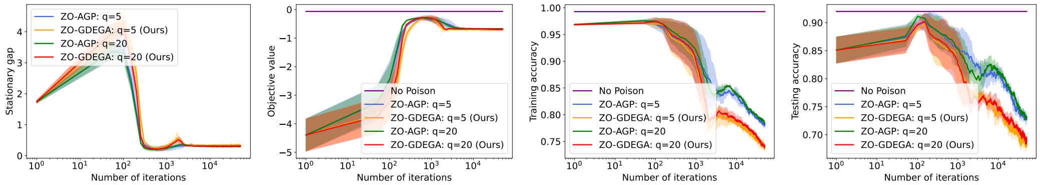


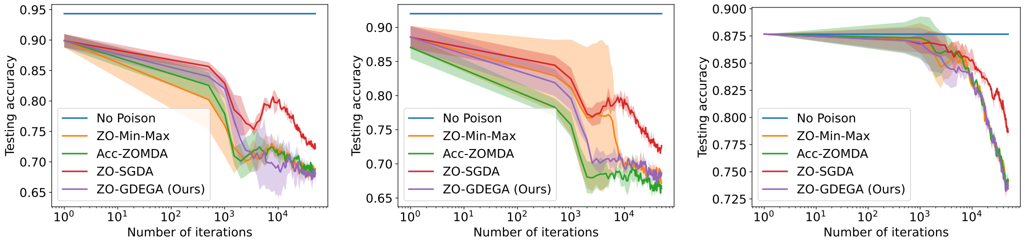
Figure C.9: The performance of all algorithms for solving sparse NC-C Problem (C.134) with $\eta_x = 0.02$, $\eta_y = 0.05$, $\mu_1 = \mu_2 = 2 \times 10^{-5}$, and $\lambda = 10^{-3}$ on the `synthetic` dataset.

C.1.4 Data Poisoning Attack for NC-SC Problems

We also consider the following model (C.135) for data poisoning attack problems. Note that Problem (C.135) can be also rewritten as the form (1) by setting $g(\cdot) = \mathcal{I}_{\|\delta\|_\infty \le \epsilon}(\cdot)$, $h(\cdot) \equiv 0$, $f = -f_2$, and thus Problem (C.135) becomes a NC-SC problem.

$$\max_{\|\delta\|_\infty \le r_x} \min_{w} f_2(\delta, w) := F_{tr}(\delta, w; \mathcal{D}_{tr}) + \lambda \|w\|^2. \tag{C.135}$$

For solving Problem (C.135), we perform our ZO-GDEGA algorithm on the `synthetic` dataset. The baseline methods for Problem (C.135) are ZO-SGDA [47], ZO-Min-Max [33] and Acc-ZOMDA [19]. For NC-SC problems, we also set mini-batch size $b_1 = b_2 = 100$ for ZO-Min-Max, ZO-SGDA, Acc-ZOMDA, and our ZO-GDEGA and train all the methods for $T = 50,000$ iterations.



(a) NC-SC, $\lambda = 1 \times 10^{-5}$ (b) NC-SC, $\lambda = 1 \times 10^{-3}$ (c) NC-SC, $\lambda = 1 \times 10^{-1}$

Figure C.10: The performance of single-loop algorithms for solving NC-SC problems with different λ when $\eta_x = 0.02$, $\eta_y = 0.05$, $\mu_1 = \mu_2 = 2 \times 10^{-5}$ and $q_1 = q_2 = 20$ on the `synthetic` dataset, we can find that our ZO-GDEGA algorithm is competitive with the baseline methods under different λ settings.

• Hyperparameter λ: To verify the performance of our ZO-GDEGA under different hyperparameter λ, we plot Fig. C.10. **Note that the lower the accuracy is, the stronger the generated attack is, which means better performance.** Fig. C.10 shows that the weaker the strong convexity (i.e., $\lambda = 1 \times 10^{-5}$), the better our ZO-GDEGA algorithm performs; on the contrary, our ZO-GDEGA performs competitively to the state-of-the-art zeroth-order stochastic algorithms.

C.2 AUC Maximization

This part provides more experimental details and results for solving AUC maximization problems. All the experiments were performed on the GeForce RTX 2080Ti platform with the PyTorch framework.

For a more detailed explanation, we restate the AUC maximization problem

$$\min_{\|\theta\|, \|a\|, \|b\| \le r_x,} \max_{v \le r_y} \mathbb{E}_{\mathbf{s} \sim \mathbb{P}}[f(\theta, a, b, v; \mathbf{s})], \tag{C.136}$$

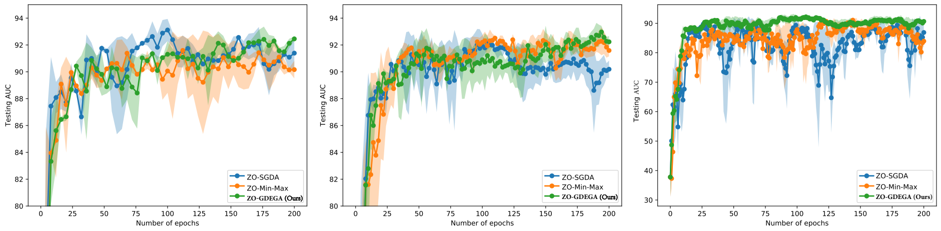
where r_x and r_y are the radii of the projection balls, $\mathbf{s} = (s, t)$ is drawn independently from the distribution $\mathbb{P}$, and $f(\theta, a, b, v; \mathbf{s}) = (1-p)(h(\theta; s) - a)^2 \mathbb{I}_{[t=1]} + p(h(\theta; s) - b)^2 \mathbb{I}_{[t=-1]} + 2(1 + v)(ph(\theta; s)\mathbb{I}_{[y=-1]} - (1-p)h(\theta; s)\mathbb{I}_{[y=1]}) - p(1-p)v^2$, $p = \mathbb{E}_t[\mathbb{I}_{t=1}]$, where $\mathbb{I}(\cdot)$ is an indicator function. When h is a multilayer perception (MLP), Problem (6) becomes a NC-SC problem.

Hyperparameter selection. We choose the hyperparameters according to theoretical guidance and considering the balance between time consumption and accuracy while observing reasonably good performance. We train our ZO-GDEGA and baseline methods with mini-batch size $b_1 = b_2 = 256$ on the `MNIST`, `Fashion-MNIST` and `ijcnn1` datasets for 200 epochs. We set $\eta_x = \eta_y = 0.1$ and $q_1 = q_2 = 10$ for all the methods. The testing accuracy versus the number of epochs on the `Fashion-MNIST` dataset is detailed shown in Fig. C.11.

We also compare our ZO-GDEGA and state-of-the-art methods as shown in Table C.3. From Table C.3, at small smoothing parameter $\mu_1, \mu_2 \le 0.01$, the Acc-ZOMDA [19] algorithm with lower complexity show clear advantages, but at large smoothing parameters $\mu_1, \mu_2 \ge 0.05$, our ZO-GDEGA algorithm performs better than other methods, which verifies that our ZO-GDEGA algorithm can tolerate rougher gradient estimations and provides promising insights into the robustness of zeroth-order minimax optimization.

C.3 Robust Neural Network Training

To verify that our ZO-GDEGA algorithm can solve extensive applications, we conduct robust network training experiments. The purpose of robust network training is against adversarial attacks.



(a) $\mu_1 = \mu_2 = 0.001$ (b) $\mu_1 = \mu_2 = 0.01$ (c) $\mu_1 = \mu_2 = 0.05$

Figure C.11: The testing accuracy vs the number of epochs of ZO-SGDA, ZO-Min-Max, and our ZO-GDEGA algorithms for solving the NC-SC problem (6) with different μ_1 and μ_2 on the `Fashion-MNIST` dataset. We can observe that our ZO-GDEGA algorithm is more robust in smoothing parameters μ_1 and μ_2 compared with the baseline methods.

Table C.3: The average AUC performance with different μ_1 and μ_2 on the `MNIST`, `Fashion-MNIST`, and `ijcnn1` datasets.

Datasets	`MNIST`				`Fashion-MNIST`				`ijcnn1`			
$\mu_1(\mu_2)$	0.001	0.01	0.05	0.1	0.001	0.01	0.05	0.1	0.001	0.01	0.05	0.1
ZO-SGDA	91.67	91.81	88.12	82.32	91.62	90.19	87.27	80.62	78.65	79.02	74.33	69.56
ZO-Min-Max	92.25	92.01	88.56	83.12	90.80	91.58	83.23	78.38	79.56	80.31	76.66	72.23
Acc-ZOMDA	92.45	92.58	89.35	89.48	92.97	91.65	87.75	87.65	82.42	82.45	80.35	76.56
ZO-GDEGA	91.60	92.60	89.70	89.76	91.97	92.23	88.66	88.01	81.01	82.30	80.99	78.01

Although neural networks are widely used for image classification, they are vulnerable to adversarial attacks such as FGSM [16]. For example, a small perturbation can greatly destroy classification performance. Thus, robust network training has been paid more attention by researchers. The optimization formulation of robust networks is

$$\min_{w} \sum_{i=1}^{N} \max_{\delta_i, \text{s.t. } |\delta_i|_\infty \leq \epsilon} \ell(f_3(x_i + \delta_i; w), y_i), \tag{C.137}$$

where w is the parameter of the neural network, the pair (x_i, y_i) denotes the i-th data point, δ_i is the perturbation added to data point i. Referring to [57], we give the following nonconvex-concave minimax problem to reformulate the robust training process.

$$\min_{w} \sum_{i=1}^{N} \max_{t \in \mathcal{T}} \Big\{ \sum_{j=0}^{9} t_j \ell(f_3(x_{i,j}^K; w), y_i) + \lambda \|w\|_1 \Big\}, \text{s.t. } \mathcal{T} = \{(t_1, \cdots, t_9) | \sum_{j=0}^{9} t_j = 1, t_j \geq 0\}, \tag{C.138}$$

where $\lambda > 0$, $x_{i,j}^K$ is an approximated attack on sample x_i by changing the output of the network to label j. We use the same convolutional neural network as in [57]. In our ZO-GDEGA, we set $\mu_1 = \mu_2 = 0.0001$ and $q_1 = q_2 = 5$. we apply our ZO-GDEGA algorithm to train a robust neural network on the `MNIST` dataset against the adversarial attack, FGSM. The experiments in this part were performed on the GeForce RTX 2080Ti platform with the PyTorch framework. The experimental results are shown in Table C.4. Note that the ZO-AGP algorithm can not solve this problem. Although an excellent robust neural network has not been achieved by our ZO-GDEGA algorithm, our ZO-GDEGA algorithm opens the way for ZO algorithms to solve this application. Performance improvement is our future research direction.

Table C.4: Test accuracies under FGSM attack.

	FGSM		
	$\epsilon = 0.02$	$\epsilon = 0.03$	$\epsilon = 0.05$
ZO-GDEGA	79.67%	78.81%	76.22%