Appendix

In the appendix, we provide the missing proofs and derivations from the main manuscript, as well as proposing a general variant of the OGDA algorithm where different learning rates can be employed in primal and dual updates.

Table of Contents

A Proof of Convergence in Nonconvex-Strongly-Concave Setting

A.1 Proof of Convergence of OGDA

Here we present the convergence proof for the OGDA algorithm in the NC-SC setting as detailed in Algorithm 2. Note that it is clear from context we abuse the notation and use $\boldsymbol{y}_t^*$ instead of $\boldsymbol{y}^*(\boldsymbol{x}_t)$. In the following, we provide a proof sketch, making our analysis easier to follow.

Algorithm 2 shows the deterministic and stochastic variants of the OGDA algorithm in detail.

Algorithm 2 (Stochastic) OGDA

```
Input :Initialization ($\boldsymbol{x}_{-1} = \boldsymbol{x}_0, \boldsymbol{y}_{-1} = \boldsymbol{y}_0$), learning rates $\eta_x, \eta_y$
for $t = 1, 2, \ldots, T$ do
    $\boldsymbol{x}_t = \boldsymbol{x}_{t-1} - \eta_x \nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) + \eta_x(\nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_x f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2}))$,
    $\boldsymbol{y}_t = \boldsymbol{y}_{t-1} + \eta_y(\nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \eta_y(\nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_y f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2}))$.   # OGDA
    $\boldsymbol{x}_t = \boldsymbol{x}_{t-1} - \eta_y \boldsymbol{g}_{x,t-1} + \eta_y(\boldsymbol{g}_{x,t-1} - \boldsymbol{g}_{x,t-2})$,
    $\boldsymbol{y}_t = \boldsymbol{y}_{t-1} + \eta_y \boldsymbol{g}_{y,t-1} - \eta_y(\boldsymbol{g}_{y,t-1} - \boldsymbol{g}_{y,t-2})$.   # Stochastic OGDA
end
```

Proof sketch. We provide a sketch of key technical ideas. Specifically, we develop three key lemmas to prove the convergence. First lemma is primal descent, in which we use the $\kappa\ell$-smoothness property of $\Phi(\boldsymbol{x})$ at point $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t-1}$ to find an upper bound for $\mathbb{E}[\Phi(\boldsymbol{x}_t) - \Phi(\boldsymbol{x}_{t-1})]$, and then by taking summation on this upper bound for all $t \in \{1, \ldots, T\}$ we are able to show the following:

$$\begin{aligned} \mathbb{E}[\Phi(\boldsymbol{x}_T)] - \Phi(\boldsymbol{x}_1) \leq -\frac{\eta_x}{2} \sum_{i=1}^{T-1} \mathbb{E}[\|\nabla \Phi(\boldsymbol{x}_i)\|^2] + O(\eta_x \ell^2) \\ + O(\eta_x \ell^2) \left(\sum_{i=1}^{T-1} \|\mathbb{E}[\boldsymbol{y}_i - \boldsymbol{y}_i^*\|^2] + \sum_{i=1}^{T-1} \mathbb{E}[\|\boldsymbol{y}_i - \boldsymbol{y}_{i-1}\|^2] \right) \\ - \frac{\eta_x}{2}(1 - O(\eta_x)) \sum_{i=1}^{T-2} \mathbb{E}[\|\boldsymbol{g}_i\|^2] + O\left(\eta_x \frac{T\sigma^2}{M_x}\right) \end{aligned} \tag{6}$$

where $\boldsymbol{g}_i = 2\nabla_x f(\boldsymbol{x}_i, \boldsymbol{y}_i) - \nabla_x f(\boldsymbol{x}_{i-1}, \boldsymbol{y}_{i-1})$.

The second key lemma is dual descent. To derive this lemma, first note that OGDA alternatively can be written in view of Past Extra-gradient algorithm (PEG) as defined in [23]:

$$\boldsymbol{y}_t = \boldsymbol{z}_t + \eta_y \boldsymbol{g}_{y,t-1} \quad , \quad \boldsymbol{z}_{t+1} = \boldsymbol{z}_t + \eta_y \boldsymbol{g}_{y,t} \tag{Dual update}$$

where $\boldsymbol{z}_t = \boldsymbol{y}_{t-1} + \eta_y(\boldsymbol{g}_{y,t-1} - \boldsymbol{g}_{y,t-2})$. Also, we have the following primal update:

$$\boldsymbol{x}_t = \boldsymbol{w}_t - \eta_x \boldsymbol{g}_{x,t-1} \quad , \quad \boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_x \boldsymbol{g}_{x,t} \tag{Primal update}$$

where $\boldsymbol{w}_t = \boldsymbol{x}_{t-1} - \eta_x(\boldsymbol{g}_{x,t-1} - \boldsymbol{g}_{x,t-2})$. This view of OGDA is presented in [23, 13, 39]. Motivated by this interpretation of the OGDA algorithm, we define the following potential function to derive the dual descent. Let $\boldsymbol{r}_t = \|\boldsymbol{z}_{t+1} - \boldsymbol{y}_t^*\|^2 + \frac{1}{4}\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2$, and $\eta_y = \frac{1}{6\ell}$, then we show that:

$$\mathbb{E}[\boldsymbol{r}_t] \leq (1 - \frac{1}{12\kappa})\mathbb{E}[\boldsymbol{r}_{t-1}] + O(\eta_x^2)\mathbb{E}[\|\boldsymbol{g}_{t-2}\|^2] + O(\eta_x^2\kappa^3)\mathbb{E}[\|\boldsymbol{g}_{t-1}\|^2] + O\left(\frac{\sigma^2}{\ell^2 M_y}\right).$$

We built on the top of OGDA analysis in [39, 23] in strongly-concave-strongly-concave setting to prove the above lemma, which helps us directly find an upper bound for $\sum_{i=1}^{T-1} \|\boldsymbol{y}_i - \boldsymbol{y}_{i-1}\|^2$ in Equation 6.

Our third key lemma aims to upper bound $\sum_{i=1}^{T-1} \mathbb{E}[\|\boldsymbol{y}_i - \boldsymbol{y}_i^*\|^2]$ in terms of $\sum_{i=1}^{T-1} \mathbb{E}[\boldsymbol{r}_i]$. Particularly we show that:

$$\sum_{i=1}^{T-1}\mathbb{E}[\|\boldsymbol{y}_i-\boldsymbol{y}_i^*\|^2]\leq\left(\|\boldsymbol{y}_0-\boldsymbol{y}_0^*\|^2+\sum_{i=2}^{T-1}\mathbb{E}[\boldsymbol{r}_i]+\eta_x^2\kappa^2\sum_{i=1}^{T-2}\mathbb{E}[\|\boldsymbol{g}_i\|^2]+\frac{T\sigma^2}{\ell^2 M_y}\right).$$

Now note that using second, and third lemma both $\sum_{i=1}^{T-1}\mathbb{E}[\|\boldsymbol{y}_i-\boldsymbol{y}_{i-1}\|^2]$, and $\sum_{i=1}^{T-1}\mathbb{E}[\|\boldsymbol{y}_i-\boldsymbol{y}_i^*\|^2]$ terms can be upper bounded in terms of $\sum_{i=1}^{T-2}\mathbb{E}[\|\boldsymbol{g}_i\|^2]$, and by properly choosing η_x we show that $\sum_{i=1}^{T-2}\mathbb{E}[\|\boldsymbol{g}_i\|^2]$ term can be ignored, which entails the desired convergence rate.

A.1.1 Useful lemmas

Lemma A.1 (Lemma 4.3 in [30]). *Let $\Phi(\boldsymbol{x})=\max_{\boldsymbol{y}} f(\boldsymbol{x},\boldsymbol{y})$, and $\boldsymbol{y}^*(\boldsymbol{x})=\arg\max_{\boldsymbol{y}} f(\boldsymbol{x},\boldsymbol{y})$. Then, under Assumption 4.1, $\Phi(\boldsymbol{x})$ is $\kappa\ell+\ell$-smooth, and $\boldsymbol{y}^*(\boldsymbol{x})$ is κ Lipschitz.*

Lemma A.2. *Let $\{a_t\}_{t=0}^{\infty}$, $\{b_t\}_{t=0}^{\infty}$ be the sequence of positive real valued number, and $\gamma\in(2,\infty)$ such that $\forall t\geq 1$:*

$$a_t\leq(1-\frac{1}{\gamma})a_{t-1}+b_t \tag{7}$$

then the following inequality holds for any $t_1>t_2\geq 0$:

$$\sum_{i=t_1}^{t_2}a_i\leq\gamma a_{t_1}+\gamma\sum_{i=t_1+1}^{t_2}b_i \tag{8}$$

Proof of Lemma A.2. Unfolding the recursion in Equation 7 for $t-t_1$ steps we have:

$$a_t\leq(1-\frac{1}{\gamma})^{t-t_1}a_{t_1}+\sum_{i=t_1+1}^{t}(1-\frac{1}{\gamma})^{t-i}b_i \tag{9}$$

Now taking summation of above equation we have:

$$\sum_{t=t_1}^{t_2}a_t\leq\left(\sum_{t=t_1}^{t_2}(1-\frac{1}{\gamma})^{t-t_1}\right)a_{t_1}+\sum_{t=t_1+1}^{t_2}\sum_{i=t_1+1}^{t}(1-\frac{1}{\gamma})^{t-i}b_i \tag{10}$$

However note that, we can write:

$$\begin{aligned}\sum_{t=t_1+1}^{t_2}\sum_{i=t_1+1}^{t}(1-\frac{1}{\gamma})^{t-i}b_i=\sum_{i=t_1+1}^{t_2}\left(b_i\sum_{j=0}^{t_2-i}(1-\frac{1}{\gamma})^{j}\right)&=\sum_{i=t_1+1}^{t_2}b_i\frac{1-(1-\frac{1}{\gamma})^{t_2-i+1}}{1-(1-\frac{1}{\gamma})}\\&\leq\gamma\sum_{i=t_1+1}^{t_2}b_i\end{aligned} \tag{11}$$

□

Plugging this back to Equation 10, and noting that $\sum_{t=t_1}^{t_2}(1-\frac{1}{\gamma})^{t-t_1}=\frac{1-(1-\frac{1}{\gamma})^{t_2-t_1+1}}{1-(1-\frac{1}{\gamma})}\leq\gamma$, we have:

$$\sum_{t=t_1}^{t_2}a_t\leq\gamma a_{t_1}+\gamma\sum_{i=t_1+1}^{t_2}b_i \tag{12}$$

Lemma A.3. *Let $\boldsymbol{y}_{t+1}=\boldsymbol{y}_t+\eta_y\boldsymbol{g}_{y,t}$, where $\boldsymbol{g}_{y,t}$ is the unbiased estimator of $\nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t)$. If $\eta_y\leq\frac{1}{2\ell}$, we have:*

$$\|\boldsymbol{y}_{t+1}-\boldsymbol{y}_t^*\|^2\leq(1-\eta_y\mu)\|\boldsymbol{y}_t-\boldsymbol{y}_t^*\|^2+2\eta_y^2\|\boldsymbol{\delta}_t^y\|^2+2\eta_y\langle\boldsymbol{\delta}_t^y,\boldsymbol{y}_t-\boldsymbol{y}_t^*\rangle \tag{13}$$

where $\boldsymbol{\delta}_t^y=\boldsymbol{g}_{y,t}-\nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t)$.

Proof of Lemma A.3. Using the update rule for $\boldsymbol{y}_{t+1}$, we can write:

$$\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t^*\|^2 = \|\boldsymbol{y}_t - \boldsymbol{y}_t^* + \eta_y \boldsymbol{g}_{y,t}\|^2 = \|\boldsymbol{y}_t - \boldsymbol{y}_t^*\|^2 + \eta_y^2 \|\boldsymbol{g}_{y,t}\|^2 + 2\eta_y \langle \boldsymbol{y}_t - \boldsymbol{y}_t^*, \boldsymbol{g}_{y,t}\rangle \tag{14}$$

Now replacing $\boldsymbol{g}_{y,t} = \boldsymbol{\delta}_t^y + \nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t)$, and using Young's inequality we have:

$$\begin{aligned}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t^*\|^2 \le \|\boldsymbol{y}_t - \boldsymbol{y}_t^*\|^2 &+ 2\eta_y^2 \|\nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t)\|^2 + 2\eta_y \langle \nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t), \boldsymbol{y}_t - \boldsymbol{y}_t^*\rangle \\ &+ 2\eta_y^2 \|\boldsymbol{\delta}_t^y\|^2 + 2\eta_y \langle \boldsymbol{\delta}_t^y, \boldsymbol{y}_t - \boldsymbol{y}_t^*\rangle\end{aligned} \tag{15}$$

However, note that since $f(\boldsymbol{x}, .)$ is μ-strongly-concave, and ℓ-smooth, we have:

$$\begin{aligned}\langle \nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t), \boldsymbol{y}_t - \boldsymbol{y}_t^*\rangle &\le -\frac{1}{\ell+\mu}\|\nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t)\|^2 - \frac{\ell\mu}{\ell+\mu}\|\boldsymbol{y}_t - \boldsymbol{y}_t^*\|^2 \\ &\le -\frac{1}{2\ell}\|\nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t)\|^2 - \frac{\mu}{2}\|\boldsymbol{y}_t - \boldsymbol{y}_t^*\|^2,\end{aligned} \tag{16}$$

where in the last inequality, we used the fact that $\kappa \ge 1$, which means that $\ell \ge \mu$. Plugging Equation 16 back to Equation 15, we have:

$$\begin{aligned}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t^*\|^2 \le (1 - \mu\eta_y)\|\boldsymbol{y}_t - \boldsymbol{y}_t^*\|^2 &- \eta_y\left(\frac{1}{\ell} - 2\eta_y\right)\|\nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t)\|^2 \\ &+ 2\eta_y^2 \|\boldsymbol{\delta}_t^y\|^2 + 2\eta_y \langle \boldsymbol{\delta}_t^y, \boldsymbol{y}_t - \boldsymbol{y}_t^*\rangle\end{aligned} \tag{17}$$

Since $\eta_y \le \frac{1}{2\ell}$, we have:

$$\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t^*\|^2 \le (1 - \mu\eta_y)\|\boldsymbol{y}_t - \boldsymbol{y}_t^*\|^2 + 2\eta_y^2 \|\boldsymbol{\delta}_t^y\|^2 + 2\eta_y \langle \boldsymbol{\delta}_t^y, \boldsymbol{y}_t - \boldsymbol{y}_t^*\rangle \tag{18}$$

□

A.1.2 Key lemmas, and proof of Theorem 4.2, and 4.4 for OGDA

For the sake of brevity, we only present the convergence proof for the stochastic version of OGDA (Theorem 4.4), since by letting $\sigma = 0$, we can recover the proof for the deterministic algorithm (Theorem 4.2). Our proof is built on three key lemmas. First, we prove the following lemma, which we call primal descent:

Lemma A.4. *Let $\Phi(\boldsymbol{x}) = \max_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})$, and $\boldsymbol{y}^*(\boldsymbol{x}) = \arg\max_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})$. Also, let $\boldsymbol{g}_i = 2\boldsymbol{g}_{x,i} - \boldsymbol{g}_{x,i-1}$. Then for Algorithm 2, we have:*

$$\begin{aligned}\mathbb{E}[\Phi(\boldsymbol{x}_t)] \le \mathbb{E}[\Phi(\boldsymbol{x}_{t-1})] &- \frac{\eta_x}{2}\mathbb{E}[\|\nabla\Phi(\boldsymbol{x}_{t-1})\|^2] - \frac{\eta_x}{2}(1 - 2\kappa\ell\eta_x)\mathbb{E}[\|\boldsymbol{g}_{t-1}\|^2] + \frac{3}{2}\eta_x^3\ell^2\mathbb{E}[\|\boldsymbol{g}_{t-2}\|^2] \\ &+ \frac{3}{2}\eta_x\ell^2\mathbb{E}[\|\boldsymbol{y}_{t-1}^* - \boldsymbol{y}_{t-1}\|^2] + \frac{3}{2}\eta_x\ell^2\mathbb{E}[\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2] + 15\eta_x\frac{\sigma^2}{M_x}\end{aligned} \tag{19}$$

Proof of Lemma A.4. First, let $\boldsymbol{\delta}_i^x = \boldsymbol{g}_{x,i} - \nabla_x f(\boldsymbol{x}_i, \boldsymbol{y}_i)$. By definition of $\boldsymbol{g}_{x,i}$, we have $\mathbb{E}[\boldsymbol{\delta}_i^x] = \mathbf{0}$, for all $i \in [T]$.

Using the fact that $\Phi(\boldsymbol{x})$ is $2\kappa\ell$ smooth, we have:

$$\begin{aligned}\Phi(\boldsymbol{x}_t) &\le \Phi(\boldsymbol{x}_{t-1}) + \langle \nabla\Phi(\boldsymbol{x}_{t-1}), \boldsymbol{x}_t - \boldsymbol{x}_{t-1}\rangle + \kappa\ell\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|^2 \\ &= \Phi(\boldsymbol{x}_{t-1}) - \eta_x\langle \nabla\Phi(\boldsymbol{x}_{t-1}), \boldsymbol{g}_{t-1}\rangle + \kappa\ell\eta_x^2\|\boldsymbol{g}_{t-1}\|^2 \\ &= \Phi(\boldsymbol{x}_{t-1}) - \frac{\eta_x}{2}\|\nabla\Phi(\boldsymbol{x}_{t-1})\|^2 - \frac{\eta_x}{2}\|\boldsymbol{g}_{t-1}\|^2 + \frac{\eta_x}{2}\|\nabla\Phi(\boldsymbol{x}_{t-1}) - \boldsymbol{g}_{t-1}\|^2 + \kappa\ell\eta_x^2\|\boldsymbol{g}_{t-1}\|^2 \\ &= \Phi(\boldsymbol{x}_{t-1}) - \frac{\eta_x}{2}\|\nabla\Phi(\boldsymbol{x}_{t-1})\|^2 - \frac{\eta_x}{2}(1 - 2\kappa\ell\eta_x)\|\boldsymbol{g}_{t-1}\|^2 + \frac{\eta_x}{2}\|\nabla\Phi(\boldsymbol{x}_{t-1}) - \boldsymbol{g}_{t-1}\|^2\end{aligned} \tag{20}$$

Now using ℓ-smoothness of f, and κ-Lipschitzness of $\boldsymbol{y}^*(\boldsymbol{x})$ (Lemma A.1) we have:

$$\begin{aligned}
\|\nabla\Phi(\boldsymbol{x}_{t-1}) - \boldsymbol{g}_{t-1}\|^2 &= \|\nabla\Phi(\boldsymbol{x}_{t-1}) - \nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) \\
&\quad - \big(\nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_x f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2})\big) - (2\boldsymbol{\delta}^x_{t-1} - \boldsymbol{\delta}^x_{t-2})\|^2 \\
&\leq 3\|\nabla\Phi(\boldsymbol{x}_{t-1}) - \nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\|^2 + 3\|\nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) \\
&\quad - \nabla_x f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2})\|^2 + 3\|2\boldsymbol{\delta}^x_{t-1} - \boldsymbol{\delta}^x_{t-2}\|^2 \\
&\leq 3\ell^2\|\boldsymbol{y}^*(\boldsymbol{x}_{t-1}) - \boldsymbol{y}_{t-1}\|^2 + 3\ell^2\|\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-2}\|^2 + 3\ell^2\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2 \\
&\qquad + 24\|\boldsymbol{\delta}^x_{t-1}\|^2 + 6\|\boldsymbol{\delta}^x_{t-2}\|^2
\end{aligned} \tag{21}$$

where in the first and second inequalities, we used Young's inequality.

By combining Equations 20 and 21 we have:

$$\begin{aligned}
\Phi(\boldsymbol{x}_t) \leq \Phi(\boldsymbol{x}_{t-1}) &- \frac{\eta_x}{2}\|\nabla\Phi(\boldsymbol{x}_{t-1})\|^2 - \frac{\eta_x}{2}(1 - 2\kappa\ell\eta_x)\|\boldsymbol{g}_{t-1}\|^2 \\
&+ \frac{3}{2}\eta_x\ell^2\|\boldsymbol{y}^*_{t-1} - \boldsymbol{y}_{t-1}\|^2 + \frac{3}{2}\eta_x\ell^2\|\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-2}\|^2 + \frac{3}{2}\eta_x\ell^2\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2 \\
&+ 12\eta_x\|\boldsymbol{\delta}^x_{t-1}\|^2 + 3\eta_x\|\boldsymbol{\delta}^x_{t-2}\|^2 \\
\leq \Phi(\boldsymbol{x}_{t-1}) &- \frac{\eta_x}{2}\|\nabla\Phi(\boldsymbol{x}_{t-1})\|^2 - \frac{\eta_x}{2}(1 - 2\kappa\ell\eta_x)\|\boldsymbol{g}_{t-1}\|^2 + \frac{3}{2}\eta_x^3\ell^2\|\boldsymbol{g}_{t-2}\|^2 \\
&+ \frac{3}{2}\eta_x\ell^2\|\boldsymbol{y}^*_{t-1} - \boldsymbol{y}_{t-1}\|^2 + \frac{3}{2}\eta_x\ell^2\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2 + 12\eta_x\|\boldsymbol{\delta}^x_{t-1}\|^2 + 3\eta_x\|\boldsymbol{\delta}^x_{t-2}\|^2
\end{aligned} \tag{22}$$

We proceed by taking expectations on both sides of Equation 22 to get:

$$\begin{aligned}
\mathbb{E}[\Phi(\boldsymbol{x}_t)] \leq \mathbb{E}[\Phi(\boldsymbol{x}_{t-1})] &- \frac{\eta_x}{2}\mathbb{E}[\|\nabla\Phi(\boldsymbol{x}_{t-1})\|^2] - \frac{\eta_x}{2}(1 - 2\kappa\ell\eta_x)\mathbb{E}[\|\boldsymbol{g}_{t-1}\|^2] + \frac{3}{2}\eta_x^3\ell^2\mathbb{E}[\|\boldsymbol{g}_{t-2}\|^2] \\
&+ \frac{3}{2}\eta_x\ell^2\mathbb{E}[\|\boldsymbol{y}^*_{t-1} - \boldsymbol{y}_{t-1}\|^2] + \frac{3}{2}\eta_x\ell^2\mathbb{E}[\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2] + 15\eta_x\frac{\sigma^2}{M_x}
\end{aligned} \tag{23}$$

where we used the fact that $\mathbb{E}[\|\boldsymbol{\delta}^x_i\|^2] \leq \frac{\sigma^2}{M_x}$ for all $i \in [T]$.

□

Lemma A.5. *Let* $\eta_y = \frac{1}{6\ell}$*, then the following inequality holds true for OGDA iterates:*

$$\sum_{i=1}^{t+1}\mathbb{E}[\|\boldsymbol{y}_i - \boldsymbol{y}^*_i\|^2] \leq \frac{9}{7}\mathbb{E}[\|\boldsymbol{y}_1 - \boldsymbol{y}^*_1\|^2] + \frac{36}{7}\sum_{i=2}^{t+1}\mathbb{E}[\|\boldsymbol{z}_i - \boldsymbol{y}^*_i\|^2] + \frac{18}{7}\eta_x^2\kappa^2\sum_{i=1}^{t}\mathbb{E}[\|\boldsymbol{g}_i\|^2] + \frac{2T\sigma^2}{7\ell^2 M_y} \tag{24}$$

Proof of Lemma A.5. Using Young's inequality and κ-Lipschitzness of $\boldsymbol{y}^*(\boldsymbol{x})$, we have:

$$\begin{aligned}
\|\boldsymbol{y}_{t+1} - \boldsymbol{y}^*_{t+1}\|^2 &\leq 2\|\boldsymbol{y}_{t+1} - \boldsymbol{y}^*_t\|^2 + 2\|\boldsymbol{y}^*_{t+1} - \boldsymbol{y}^*_t\|^2 \\
&\leq 2\|\boldsymbol{y}_{t+1} - \boldsymbol{y}^*_t\|^2 + 2\kappa^2\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2
\end{aligned} \tag{25}$$

Now, we try to find an upper bound for $\|\boldsymbol{y}_{t+1} - \boldsymbol{y}^*_t\|^2$. Let $\boldsymbol{z}_{t+1} = \boldsymbol{y}_t + \eta_y(\boldsymbol{g}_{y,t} - \boldsymbol{g}_{y,t-1})$, and $\boldsymbol{\delta}^y_i = \boldsymbol{g}_{y,i} - \nabla_y f(\boldsymbol{x}_i, \boldsymbol{y}_i)$. Then we have:

$$\begin{aligned}
\|\boldsymbol{y}_{t+1} - \boldsymbol{y}^*_t\|^2 &= \|\boldsymbol{z}_{t+1} - \boldsymbol{y}^*_t + \eta_y\boldsymbol{g}_{y,t}\|^2 \\
&\leq 2\|\boldsymbol{z}_{t+1} - \boldsymbol{y}^*_t\|^2 + 2\eta_y^2\|\boldsymbol{g}_{y,t}\|^2 \\
&\leq 2\|\boldsymbol{z}_{t+1} - \boldsymbol{y}^*_t\|^2 + 4\eta_y^2\|\nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t)\|^2 + 4\eta_y^2\|\boldsymbol{\delta}^y_t\|^2 \\
&\leq 2\|\boldsymbol{z}_{t+1} - \boldsymbol{y}^*_t\|^2 + 4\eta_y^2\ell^2\|\boldsymbol{y}_t - \boldsymbol{y}^*_t\|^2 + 4\eta_y^2\|\boldsymbol{\delta}^y_t\|^2
\end{aligned} \tag{26}$$

where in the first and second inequality, we used Young's inequality, and for the last inequality, we used smoothness of f. Now, replacing replacing the choice $\eta_y = \frac{1}{6\ell}$ in Equation 26 yields:

$$\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_t^*\|^2 \leq \frac{1}{9}\|\boldsymbol{y}_t - \boldsymbol{y}_t^*\|^2 + 2\|\boldsymbol{z}_{t+1} - \boldsymbol{y}_t^*\|^2 + \frac{1}{9\ell^2}\|\boldsymbol{\delta}_t^y\|^2 \tag{27}$$

Now plugging Equation 27 in Equation 25 we have:

$$\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_{t+1}^*\|^2 \leq \frac{2}{9}\|\boldsymbol{y}_t - \boldsymbol{y}_t^*\|^2 + 4\|\boldsymbol{z}_{t+1} - \boldsymbol{y}_t^*\|^2 + 2\kappa^2\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2 + \frac{2}{9\ell^2}\|\boldsymbol{\delta}_t^y\|^2 \tag{28}$$

Now taking expectations from both sides of Equation 28, we have:

$$\mathbb{E}[\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_{t+1}^*\|^2] \leq \frac{2}{9}\mathbb{E}[\|\boldsymbol{y}_t - \boldsymbol{y}_t^*\|^2] + 4\mathbb{E}[\|\boldsymbol{z}_{t+1} - \boldsymbol{y}_t^*\|^2] + 2\kappa^2\mathbb{E}[\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2] + \frac{2\sigma^2}{9\ell^2 M_y} \tag{29}$$

Using Lemma A.2, it can be easily shown that:

$$\sum_{i=1}^{t+1}\mathbb{E}[\|\boldsymbol{y}_i - \boldsymbol{y}_i^*\|^2] \leq \frac{9}{7}\mathbb{E}[\|\boldsymbol{y}_1 - \boldsymbol{y}_1^*\|^2] + \frac{36}{7}\sum_{i=2}^{t+1}\mathbb{E}[\|\boldsymbol{z}_i - \boldsymbol{y}_i^*\|^2] + \frac{18}{7}\eta_x^2\kappa^2\sum_{i=1}^{t}\mathbb{E}[\|\boldsymbol{g}_i\|^2] + \frac{2T\sigma^2}{7\ell^2 M_y} \tag{30}$$

□

By extending the analysis in [39] for OGDA from SC-SC to NC-SC, we derive the following lemma:

Lemma A.6. *Let* $\boldsymbol{z}_{t+1} = \boldsymbol{y}_t + \eta_y(\boldsymbol{g}_{y,t} - \boldsymbol{g}_{y,t-1})$*,* $\boldsymbol{r}_t = \|\boldsymbol{z}_{t+1} - \boldsymbol{y}_t^*\|^2 + \frac{1}{4}\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2$ *and* $\eta_y = \frac{1}{6\ell}$*. Then OGDA iterates satisfy the following inequalities:*

$$\mathbb{E}[\boldsymbol{r}_t] \leq \left(1 - \frac{1}{12\kappa}\right)\mathbb{E}[\boldsymbol{r}_{t-1}] + 12\eta_x^2\kappa^3\mathbb{E}[\|\boldsymbol{g}_{t-1}\|^2] + \frac{\eta_x^2}{18}\mathbb{E}[\|\boldsymbol{g}_{t-2}\|^2] + \frac{\sigma^2}{3\ell^2 M_y} \tag{31}$$

and

$$\sum_{i=1}^{t}\mathbb{E}[\boldsymbol{r}_i] \leq 12\kappa\mathbb{E}[\boldsymbol{r}_1] + \frac{2}{3}\kappa\mathbb{E}[\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2] + 145\eta_x^2\kappa^4\sum_{i=1}^{t-1}\mathbb{E}[\|\boldsymbol{g}_i\|^2] + \frac{4\kappa\sigma^2(t-1)}{\ell^2 M_y}. \tag{32}$$

Proof of Lemma A.6. Let $\boldsymbol{\delta}_i^y = \boldsymbol{g}_{y,i} - \nabla_y f(\boldsymbol{x}_i, \boldsymbol{y}_i)$, and note that we have $\boldsymbol{z}_{t+1} - \boldsymbol{z}_t = \eta_y \boldsymbol{g}_{y,t}$. We have:

$$\begin{aligned}
\|\boldsymbol{z}_{t+1} - \boldsymbol{y}_t^*\|^2 &= \|\boldsymbol{z}_t - \boldsymbol{y}_t^* + \eta_y \boldsymbol{g}_{y,t}\|^2 \\
&= \|\boldsymbol{z}_t - \boldsymbol{y}_t^*\|^2 + 2\eta_y\langle \boldsymbol{g}_{y,t}, \boldsymbol{z}_t - \boldsymbol{y}_t^*\rangle + \eta_y^2\|\boldsymbol{g}_{y,t}\|^2 \\
&= \|\boldsymbol{z}_t - \boldsymbol{y}_t^*\|^2 - 2\eta_y^2\langle \boldsymbol{g}_{y,t}, \boldsymbol{g}_{y,t-1}\rangle + 2\eta_y\langle \boldsymbol{g}_{y,t}, \boldsymbol{y}_t - \boldsymbol{y}_t^*\rangle + \eta_y^2\|\boldsymbol{g}_{y,t}\|^2 \\
&= \|\boldsymbol{z}_t - \boldsymbol{y}_t^*\|^2 + \eta_y^2\|\boldsymbol{g}_{y,t} - \boldsymbol{g}_{y,t-1}\|^2 + 2\eta_y\langle \boldsymbol{g}_{y,t}, \boldsymbol{y}_t - \boldsymbol{y}_t^*\rangle - \eta_y^2\|\boldsymbol{g}_{y,t-1}\|^2 \\
&\leq \|\boldsymbol{z}_t - \boldsymbol{y}_t^*\|^2 + 3\eta_y^2\|\nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\|^2 \\
&\quad + 2\eta_y\langle \nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t), \boldsymbol{y}_t - \boldsymbol{y}_t^*\rangle - \eta_y^2\|\boldsymbol{g}_{y,t-1}\|^2 \\
&\quad + 3\eta_y^2\|\boldsymbol{\delta}_t^y\|^2 + 3\eta_y^2\|\boldsymbol{\delta}_{t-1}^y\|^2 + 2\eta_y\langle \boldsymbol{\delta}_t^y, \boldsymbol{y}_t - \boldsymbol{y}_t^*\rangle \\
&\leq \|\boldsymbol{z}_t - \boldsymbol{y}_t^*\|^2 + 3\eta_y^2\ell^2\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|^2 + 3\eta_y^2\ell^2\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 - 2\eta_y\mu\|\boldsymbol{y}_t - \boldsymbol{y}_t^*\|^2 \\
&\quad - \eta_y^2\|\boldsymbol{g}_{y,t-1}\|^2 + 3\eta_y^2\|\boldsymbol{\delta}_t^y\|^2 + 3\eta_y^2\|\boldsymbol{\delta}_{t-1}^y\|^2 + 2\eta_y\langle \boldsymbol{\delta}_t^y, \boldsymbol{y}_t - \boldsymbol{y}_t^*\rangle
\end{aligned} \tag{33}$$

where the last inequality follows from the smoothness of f and strong concavity of $f(\boldsymbol{x}_t, .)$. Now note that using Young's inequality, we can write:

$$\|\boldsymbol{y}_t - \boldsymbol{y}_t^*\|^2 \geq \frac{1}{2}\|\boldsymbol{z}_t - \boldsymbol{y}_t^*\|^2 - \eta_y^2\|\boldsymbol{g}_{y,t-1}\|^2 \tag{34}$$

Now plugging Equation 34 back to Equation 33, we have:

$$\begin{aligned}
\|\boldsymbol{z}_{t+1} - \boldsymbol{y}_t^*\|^2 &\leq (1 - \eta_y\mu)\|\boldsymbol{z}_t - \boldsymbol{y}_t^*\|^2 + 3\eta_y^2\ell^2\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|^2 + 3\eta_y^2\ell^2\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 \\
&\quad - \eta_y^2(1 - 2\eta_y\mu)\|\boldsymbol{g}_{y,t-1}\|^2 + 3\eta_y^2\|\boldsymbol{\delta}_t^y\|^2 + 3\eta_y^2\|\boldsymbol{\delta}_{t-1}^y\|^2 + 2\eta_y\langle \boldsymbol{\delta}_t^y, \boldsymbol{y}_t - \boldsymbol{y}_t^*\rangle
\end{aligned} \tag{35}$$

Now note that we have the following:

$$\begin{aligned}
\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 &= \eta_y^2 \|\boldsymbol{g}_{y,t-1} + \boldsymbol{g}_{y,t-1} - \boldsymbol{g}_{y,t-2}\|^2 \\
&\le 2\eta_y^2 \|\boldsymbol{g}_{y,t-1}\|^2 + 2\eta_y^2 \|\boldsymbol{g}_{y,t-1} - \boldsymbol{g}_{y,t-2}\|^2 \\
&\le 2\eta_y^2 \|\boldsymbol{g}_{y,t-1}\|^2 + 6\eta_y^2 \|\nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_y f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2})\|^2 \\
&\quad + 6\eta_y^2 \|\boldsymbol{\delta}^y_{t-1}\|^2 + 6\eta_y^2 \|\boldsymbol{\delta}^y_{t-2}\|^2 \\
&\le 2\eta_y^2 \|\boldsymbol{g}_{y,t-1}\|^2 + 6\eta_y^2 \ell^2 \|\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-2}\|^2 + 6\eta_y^2 \ell^2 \|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2 \\
&\quad + 6\eta_y^2 \|\boldsymbol{\delta}^y_{t-1}\|^2 + 6\eta_y^2 \|\boldsymbol{\delta}^y_{t-2}\|^2
\end{aligned} \tag{36}$$

Now adding $9\eta_y^2 \ell^2 \|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2$ to both side of Equation 35, and using Equation 36 we have:

$$\begin{aligned}
\|\boldsymbol{z}_{t+1} - \boldsymbol{y}_t^*\|^2 + 9\eta_y^2 \ell^2 \|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 \le\ & (1 - \eta_y \mu) \|\boldsymbol{z}_t - \boldsymbol{y}_t^*\|^2 + 3\eta_y^2 \ell^2 \|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|^2 \\
& - \eta_y^2 (1 - 2\eta_y \mu - 24\eta_y^2 \ell^2) \|\boldsymbol{g}_{y,t-1}\|^2 \\
& + 72\eta_y^4 \ell^4 \|\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-2}\|^2 + 72\eta_y^4 \ell^4 \|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2 \\
& + 3\eta_y^2 (1 + 24\eta_y^2 \ell^2) \|\boldsymbol{\delta}^y_t\|^2 + 3\eta_y^2 (1 + 24\eta_y^2 \ell^2) \|\boldsymbol{\delta}^y_{t-1}\|^2 \\
& + 2\eta_y \langle \boldsymbol{\delta}^y_t, \boldsymbol{y}_t - \boldsymbol{y}_t^* \rangle
\end{aligned} \tag{37}$$

We proceed by plugging $\eta_y = \frac{1}{6\ell}$ into Equation 37:

$$\begin{aligned}
\|\boldsymbol{z}_{t+1} - \boldsymbol{y}_t^*\|^2 + \frac{1}{4} \|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 \le\ & \left(1 - \frac{1}{6\kappa}\right) \left(\|\boldsymbol{z}_t - \boldsymbol{y}_t^*\|^2\right) + \frac{1}{18} \|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2 \\
& + \frac{1}{12} \|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|^2 + \frac{1}{18} \|\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-2}\|^2 \\
& + \frac{1}{6\ell^2} \|\boldsymbol{\delta}^y_t\|^2 + \frac{1}{6\ell^2} \|\boldsymbol{\delta}^y_{t-1}\|^2 + \frac{2}{6\ell} \langle \boldsymbol{\delta}^y_t, \boldsymbol{y}_t - \boldsymbol{y}_t^* \rangle
\end{aligned} \tag{38}$$

Taking expectations from both sides of Equation 38, we have:

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{z}_{t+1} - \boldsymbol{y}_t^*\|^2 + \frac{1}{4} \|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2\right] \le\ & \left(1 - \frac{1}{6\kappa}\right) \mathbb{E}\left[\|\boldsymbol{z}_t - \boldsymbol{y}_t^*\|^2\right] + \frac{1}{18} \mathbb{E}[\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2] \\
& + \frac{1}{12} \mathbb{E}[\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|^2] + \frac{1}{18} \mathbb{E}[\|\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-2}\|^2] \\
& + \frac{\sigma^2}{3\ell^2 M_y}
\end{aligned} \tag{39}$$

Also, using Young's inequality, we have:

$$\|\boldsymbol{z}_t - \boldsymbol{y}_t^*\|^2 \le (1 + \frac{1}{12\kappa}) \|\boldsymbol{z}_t - \boldsymbol{y}_{t-1}^*\|^2 + (1 + 12\kappa) \kappa^2 \|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|^2, \tag{40}$$

where we used the fact that for any $\alpha > 0$, $\|\boldsymbol{x} + \boldsymbol{y}\|^2 \le (1 + \alpha) \|\boldsymbol{x}\|^2 + (1 + \frac{1}{\alpha}) \|\boldsymbol{y}\|^2$, and κ-lipschitzness of $\boldsymbol{y}^*(\boldsymbol{x})$. Plugging Equation 40 back to Equation 39, we have:

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{z}_{t+1} - \boldsymbol{y}_t^*\|^2 + \frac{1}{4} \|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2\right] \le\ & \left(1 - \frac{1}{12\kappa}\right) \mathbb{E}\left[\|\boldsymbol{z}_t - \boldsymbol{y}_{t-1}^*\|^2 + \frac{1}{4} \mathbb{E}[\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2]\right] \\
& + 12\kappa^3 \mathbb{E}[\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|^2] + \frac{1}{18} \mathbb{E}[\|\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-2}\|^2] \\
& + \frac{\sigma^2}{3\ell^2 M_y}
\end{aligned} \tag{41}$$

Therefore, if we let $\boldsymbol{r}_t = \|\boldsymbol{z}_{t+1} - \boldsymbol{y}_t^*\|^2 + \frac{1}{4} \|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2$, then we have:

$$\mathbb{E}[\boldsymbol{r}_t] \le \left(1 - \frac{1}{12\kappa}\right) \mathbb{E}[\boldsymbol{r}_{t-1}] + 12\eta_x^2 \kappa^3 \mathbb{E}[\|\boldsymbol{g}_{t-1}\|^2] + \frac{\eta_x^2}{18} \mathbb{E}[\|\boldsymbol{g}_{t-2}\|^2] + \frac{\sigma^2}{3\ell^2 M_y} \tag{42}$$

We can derive the following equation by applying Lemma A.2.

$$\begin{aligned}\sum_{i=1}^{t}\mathbb{E}[\boldsymbol{r}_i] \le 12\kappa\mathbb{E}[\boldsymbol{r}_1] + 144\eta_x^2\kappa^4\sum_{i=1}^{t-1}\mathbb{E}[\|\boldsymbol{g}_i\|^2] + \frac{2}{3}\eta_x^2\kappa\sum_{i=1}^{t-2}\mathbb{E}[\|\boldsymbol{g}_i\|^2] + \frac{2}{3}\kappa\mathbb{E}[\|\boldsymbol{x}_1 - \boldsymbol{x}_0]\|^2 \\ + \frac{4\kappa\sigma^2(t-1)}{\ell^2 M_y}\end{aligned} \tag{43}$$

Or equivalently, we have:

$$\sum_{i=1}^{t}\mathbb{E}[\boldsymbol{r}_i] \le 12\kappa\mathbb{E}[\boldsymbol{r}_1] + \frac{2}{3}\kappa\mathbb{E}[\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2] + 145\eta_x^2\kappa^4\sum_{i=1}^{t-1}\mathbb{E}[\|\boldsymbol{g}_i\|^2] + \frac{4\kappa\sigma^2(t-1)}{\ell^2 M_y} \tag{44}$$

□

Proof of Theorem 4.2, and Theorem 4.4 for OGDA. We begin by taking summation of Equation 19 (Lemma A.4) from $t = 2$ to $t = T$ which yields:

$$\begin{aligned}\frac{\eta_x}{2}\sum_{i=1}^{T-1}\mathbb{E}[\|\nabla\Phi(\boldsymbol{x}_i)\|^2] \le \Phi(\boldsymbol{x}_1) - \mathbb{E}[\Phi(\boldsymbol{x}_T)] + \frac{3}{2}\eta_x\ell^2\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2 \\ - \frac{\eta_x}{2}(1 - 2\kappa\ell\eta_x)\sum_{i=1}^{T-1}\mathbb{E}[\|\boldsymbol{g}_i\|^2] + \frac{3}{2}\eta_x^3\ell^2\sum_{i=1}^{T-2}\mathbb{E}[\|\boldsymbol{g}_i\|^2] \\ + \frac{3}{2}\eta_x\ell^2\sum_{i=1}^{T-1}\|\boldsymbol{y}_i - \boldsymbol{y}_i^*\|^2 + \frac{3}{2}\eta_x\ell^2\sum_{i=1}^{T-1}\mathbb{E}[\|\boldsymbol{y}_i - \boldsymbol{y}_{i-1}\|^2] \\ + 15\eta_x\frac{(T-1)\sigma^2}{M_x}\end{aligned} \tag{45}$$

We proceed by noting that if $\eta_x \le \frac{1}{2\kappa\ell}$, then we can drop $\|\boldsymbol{g}_{T-1}\|^2$ term in above equation. By considering this, and multiplying both sides by $\frac{2}{\eta_x}$ we get (also let $\Delta_\Phi = \max(\Phi(\boldsymbol{x}_0), \Phi(\boldsymbol{x}_1)) - \min_{\boldsymbol{x}}\Phi(\boldsymbol{x})$) :

$$\begin{aligned}\sum_{i=1}^{T-1}\mathbb{E}[\|\nabla\Phi(\boldsymbol{x}_i)\|^2] \le \frac{2\Delta_\Phi}{\eta_x} + 3\ell^2\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2 \\ - (1 - 2\kappa\ell\eta_x - 3\eta_x^2\ell^2)\sum_{i=1}^{T-2}\mathbb{E}[\|\boldsymbol{g}_i\|^2] \\ + 3\ell^2\sum_{i=1}^{T-1}\mathbb{E}[\|\boldsymbol{y}_i^* - \boldsymbol{y}_i\|^2] + 3\ell^2\sum_{i=1}^{T-1}\mathbb{E}[\|\boldsymbol{y}_i - \boldsymbol{y}_{i-1}\|^2] + 30\frac{(T-1)\sigma^2}{M_x}\end{aligned} \tag{46}$$

We can replace $\sum_{i=1}^{T-1}\|\boldsymbol{y}_i^* - \boldsymbol{y}_i\|^2$ with its upper bound obtained in Lemma A.5 to get:

$$\begin{aligned}\sum_{i=1}^{T-1}\|\nabla\Phi(\boldsymbol{x}_i)\|^2 \le \frac{2\Delta_\Phi}{\eta_x} + 3\ell^2\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2 + \frac{27}{7}\ell^2\|\boldsymbol{y}_1 - \boldsymbol{y}_1^*\|^2 \\ - (1 - 2\kappa\ell\eta_x - 3\eta_x^2\ell^2 - \frac{54}{7}\eta_x^2\kappa^2\ell^2)\sum_{i=1}^{T-2}\mathbb{E}[\|\boldsymbol{g}_i\|^2] \\ + \frac{108}{7}\ell^2\sum_{i=2}^{T-1}\mathbb{E}[\|\boldsymbol{z}_i - \boldsymbol{y}_{i-1}^*\|^2] + 3\ell^2\sum_{i=1}^{T-1}\mathbb{E}[\|\boldsymbol{y}_i - \boldsymbol{y}_{i-1}\|^2] + 30\frac{(T-1)\sigma^2}{M_x} \\ + \frac{6}{7}\frac{(T-2)\sigma^2}{M_y}\end{aligned} \tag{47}$$

Now note that $\frac{108}{7}\mathbb{E}[\|\boldsymbol{z}_{i+1}-\boldsymbol{y}_i^*\|^2]+3\sum_{i=2}^{T-1}\mathbb{E}[\|\boldsymbol{y}_i-\boldsymbol{y}_{i-1}\|^2]\leq 15.5\mathbb{E}[\boldsymbol{r}_i]$. Therefore we have:

$$\begin{aligned}\sum_{i=1}^{T-1}\|\nabla\Phi(\boldsymbol{x}_i)\|^2\leq\frac{2\Delta_\Phi}{\eta_x}&+3\ell^2\|\boldsymbol{x}_1-\boldsymbol{x}_0\|^2+\frac{27}{7}\ell^2\|\boldsymbol{y}_1-\boldsymbol{y}_1^*\|^2\\&-(1-2\kappa\ell\eta_x-3\eta_x^2\ell^2-\frac{54}{7}\eta_x^2\kappa^2\ell^2)\sum_{i=1}^{T-2}\mathbb{E}[\|\boldsymbol{g}_i\|^2]\\&+15.5\ell^2\sum_{i=1}^{T-1}\mathbb{E}[\boldsymbol{r}_i]+30\frac{(T-1)\sigma^2}{M_x}+\frac{6}{7}\frac{(T-2)\sigma^2}{M_y}\end{aligned}\tag{48}$$

Furthermore, using Lemma A.6, we can find an upper bound on $\sum_{i=1}^{T-1}\mathbb{E}[\boldsymbol{r}_i]$, and replacing it in above equation yields:

$$\begin{aligned}\sum_{i=1}^{T-1}\|\nabla\Phi(\boldsymbol{x}_i)\|^2\leq\frac{2\Delta_\Phi}{\eta_x}&+186\kappa\ell^2\mathbb{E}[\boldsymbol{r}_1]+11\kappa\ell^2\|\boldsymbol{x}_1-\boldsymbol{x}_0\|^2+3\ell^2\|\boldsymbol{x}_1-\boldsymbol{x}_0\|^2+\frac{27}{7}\ell^2\|\boldsymbol{y}_1-\boldsymbol{y}_1^*\|^2\\&-(1-2\kappa\ell\eta_x-3\eta_x^2\ell^2-\frac{54}{7}\eta_x^2\kappa^2\ell^2-2248\eta_x^2\kappa^4\ell^2)\sum_{i=1}^{T-2}\mathbb{E}[\|\boldsymbol{g}_i\|^2]\\&+\frac{62\kappa\sigma^2(T-2)}{M_y}+30\frac{(T-1)\sigma^2}{M_x}+\frac{6}{7}\frac{(T-2)\sigma^2}{M_y}\end{aligned}\tag{49}$$

By letting $\eta_x=\frac{1}{50\kappa^2\ell}$, it holds that $-(1-2\kappa\ell\eta_x-3\eta_x^2\ell^2-\frac{54}{7}\eta_x^2\kappa^2\ell^2-2248\eta_x^2\kappa^4\ell^2)\sum_{i=1}^{T-2}\mathbb{E}[\|\boldsymbol{g}_i\|^2]\leq 0$. Therefore, with the choice of letting rate $\eta_x=\frac{1}{50\kappa^2\ell}$ and simplifying the terms, we have:

$$\begin{aligned}\frac{1}{T-1}\sum_{i=1}^{T-1}\mathbb{E}[\|\nabla\Phi(\boldsymbol{x}_i)\|^2]\leq 100\frac{\kappa^2\ell\Delta_\Phi}{T-1}&+186\frac{\kappa\ell^2}{T-1}\|\boldsymbol{y}_1-\boldsymbol{y}_1^*+\eta_y(\boldsymbol{g}_{y,1}-\boldsymbol{g}_{y,0})\|^2\\&+47\frac{\kappa\ell^2}{T-1}\|\boldsymbol{y}_1-\boldsymbol{y}_0\|^2+14\frac{\kappa\ell^2}{T-1}\|\boldsymbol{x}_1-\boldsymbol{x}_0\|^2\\&+\frac{27}{7}\frac{\ell^2}{T-1}\|\boldsymbol{y}_1-\boldsymbol{y}_1^*\|^2+\frac{63\kappa\sigma^2}{M_y}+30\frac{\sigma^2}{M_x}\end{aligned}\tag{50}$$

Using Young's inequality and ℓ-smoothness of f, we have:

$$\|\boldsymbol{y}_1-\boldsymbol{y}_1^*+\eta_y(\boldsymbol{g}_{y,1}-\boldsymbol{g}_{y,0})\|^2\leq 2\|\boldsymbol{y}_1-\boldsymbol{y}_1^*\|^2+\frac{1}{18}\|\boldsymbol{y}_1-\boldsymbol{y}_0\|^2+\frac{1}{18}\|\boldsymbol{x}_1-\boldsymbol{x}_0\|^2\tag{51}$$

Plugging this into Equation 50, we have:

$$\begin{aligned}\frac{1}{T-1}\sum_{i=1}^{T-1}\mathbb{E}[\|\nabla\Phi(\boldsymbol{x}_i)\|^2]\leq 100\frac{\kappa^2\ell\Delta_\Phi}{T-1}&+376\frac{\kappa\ell^2}{T-1}\|\boldsymbol{y}_1-\boldsymbol{y}_1^*\|^2\\&+58\frac{\kappa\ell^2}{T-1}\|\boldsymbol{y}_1-\boldsymbol{y}_0\|^2+25\frac{\kappa\ell^2}{T-1}\|\boldsymbol{x}_1-\boldsymbol{x}_0\|^2\\&+\frac{63\kappa\sigma^2}{M_y}+30\frac{\sigma^2}{M_x}\end{aligned}\tag{52}$$

Now by letting $M_x=\frac{\sigma^2}{\epsilon^2}$, $M_y=\frac{\kappa\sigma^2}{\epsilon^2}$ and $D_0=\max(\|\boldsymbol{y}_1-\boldsymbol{y}_1^*\|^2,\|\boldsymbol{x}_1-\boldsymbol{x}_0\|^2,\|\boldsymbol{y}_1-\boldsymbol{y}_0\|^2)$, we have:

$$\frac{1}{T-1}\sum_{i=1}^{T-1}\mathbb{E}[\|\nabla\Phi(\boldsymbol{x}_i)\|^2]\leq O(\frac{\kappa^2\ell\Delta_\Phi+\kappa\ell^2D_0}{T-1})+O(\epsilon^2)\tag{53}$$

which completes the proof as stated. □

A.2 Proof of Convergence of EG

In this section, we present the convergence proof of the EG algorithm as detailed in Algorithm 3. We start by providing the proof sketch.

Algorithm 3 (Stochastic) EG

```
Input :Initialization $(\boldsymbol{x}_{-1} = \boldsymbol{x}_0, \boldsymbol{y}_{-1} = \boldsymbol{y}_0)$, learning rates $\eta_x, \eta_y$
for $t = 1, 2, \ldots, T$ do
    $\boldsymbol{x}_{t+1/2} = \boldsymbol{x}_t - \eta_x \nabla_x f(\boldsymbol{x}_t, \boldsymbol{y}_t)$ ;    $\boldsymbol{y}_{t+1/2} = \boldsymbol{y}_t + \eta_y \nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t)$
    $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_x \nabla_x f(\boldsymbol{x}_{t+1/2}, \boldsymbol{y}_{t+1/2})$ ; $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \eta_y \nabla_y f(\boldsymbol{x}_{t+1/2}, \boldsymbol{y}_{t+1/2})$ ;    #EG
    $\boldsymbol{x}_{t+1/2} = \boldsymbol{x}_t - \eta_x \boldsymbol{g}_{x,t}$ ;    $\boldsymbol{y}_{t+1/2} = \boldsymbol{y}_t + \eta_y \boldsymbol{g}_{y,t}$ ;
    $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_x \boldsymbol{g}_{x,t+1/2}$ ;    $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \eta_y \boldsymbol{g}_{y,t+1/2}$ ;    # Stochastic EG
end
```

Proof sketch. We highlight the key ideas here. The first step is to derive to find an upper bound on $\Phi(\boldsymbol{x}_{t+1}) - \Phi(\boldsymbol{x}_t)$. Using $\kappa\ell$-smoothness property of $\Phi(\boldsymbol{x})$ at point $\boldsymbol{x}_{t+1}$, and $\boldsymbol{x}_t$ we bound the $\Phi(\boldsymbol{x}_{t+1}) - \Phi(\boldsymbol{x}_t)$ term, and then taking summation over all iterates, we derive the following primal descent lemma:

$$\begin{aligned}\mathbb{E}[\Phi(\boldsymbol{x}_T)] - \Phi(\boldsymbol{x}_0) \leq -\frac{\eta_x}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\Phi(\boldsymbol{x}_t)\|^2] - \frac{\eta_x}{4}(1 - O(\eta_x))\sum_{t=0}^{T-1}\mathbb{E}[\|\boldsymbol{g}_{x,t}\|^2] \\ + O(\eta_x \ell^2)\sum_{t=0}^{T-1}\mathbb{E}[\|\boldsymbol{y}_t - \boldsymbol{y}_t^*\|^2] O(\eta_x)\frac{\sigma^2 T}{M}.\end{aligned} \tag{54}$$

We also show the following dual descent lemma to directly bound $\sum_{t=0}^{T-1}\|\boldsymbol{y}_t - \boldsymbol{y}_t^*\|^2$ term in above inequality:

$$\mathbb{E}[\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_{t+1}^*\|^2] \leq (1 - \frac{1}{12\kappa})\mathbb{E}[\|\boldsymbol{y}_t - \boldsymbol{y}_t^*\|^2] + O(\kappa^3\eta_x^2)\mathbb{E}[\|\boldsymbol{g}_{x,t}\|^2] + \frac{2\sigma^2}{M\ell^2}$$

where we assumed $\eta_y = \frac{1}{4\ell}$. Combining the primal and dual descent lemmas yields the desired result on the convergence of EG to an ϵ-stationary point.

In what follows, we provide the formal key lemmas, and the complete proof of Theorem 4.2, and Theorem 4.4 for EG algorithm. Similar to OGDA, for the sake of brevity, we only present the convergence proof for stochastic version of EG (Theorem 4.4), since by letting $\sigma = 0$, we can recover the proof for deterministic algorithm (Theorem 4.2).

Lemma A.7. *Let* $\eta_y = \frac{1}{4\ell}$*, and* $M = \max(M_x, M_y)$*. Also assume* $\eta_x \leq \frac{1}{64\kappa^2\ell}$*, then the iterates of Algorithm 3 satisfy the following inequalities:*

$$\mathbb{E}[\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_{t+1}^*\|^2] \leq (1 - \frac{1}{12\kappa})\mathbb{E}[\|\boldsymbol{y}_t - \boldsymbol{y}_t^*\|^2] + 18\eta_x^2\kappa^3\mathbb{E}[\|\boldsymbol{g}_{x,t}\|^2] + 2\frac{\sigma^2}{M\ell^2} \tag{55}$$

$$\sum_{i=0}^{T-1}\mathbb{E}[\|\boldsymbol{y}_i - \boldsymbol{y}_i^*\|^2] \leq 12\kappa\|\boldsymbol{y}_0 - \boldsymbol{y}_0^*\|^2 + 216\eta_x^2\kappa^4\sum_{i=0}^{T-2}\mathbb{E}[\|\boldsymbol{g}_{x,i}\|^2] + \frac{24\kappa\sigma^2(T-1)}{M\ell^2} \tag{56}$$

Proof of Lemma A.7. Now we turn to convergence analysis for EG. The deterministic and stochastic variants of the EG algorithm are detailed in Algorithm 3.

To prove this lemma, we built on top of analysis in [39]. We start by noting that:

$$\begin{aligned}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}_{t+\frac{1}{2}}^*\|^2 = \|\boldsymbol{y}_t - \boldsymbol{y}_{t+\frac{1}{2}}^*\|^2 \\ - \|\boldsymbol{y}_{t+1} - \boldsymbol{y}_{t+\frac{1}{2}}\|^2 \\ - \|\boldsymbol{y}_{t+\frac{1}{2}} - \boldsymbol{y}_t\|^2 \\ + 2\eta_y\langle \boldsymbol{g}_{y,t}, \boldsymbol{y}_{t+\frac{1}{2}} - \boldsymbol{y}_{t+1}\rangle \\ + 2\eta_y\langle \boldsymbol{g}_{y,t+\frac{1}{2}}, \boldsymbol{y}_{t+1} - \boldsymbol{y}_{t+\frac{1}{2}}^*\rangle\end{aligned} \tag{57}$$

Let $\boldsymbol{\delta}_i^y = \boldsymbol{g}_{y,i} - \nabla_y f(\boldsymbol{x}_i, \boldsymbol{y}_i)$. We have:

$$\begin{aligned}
&2\eta_y \langle \boldsymbol{g}_{y,t}, \boldsymbol{y}_{t+1/2} - \boldsymbol{y}_{t+1} \rangle + 2\eta_y \langle \boldsymbol{g}_{y,t+\frac{1}{2}}, \boldsymbol{y}_{t+1} - \boldsymbol{y}^*_{t+\frac{1}{2}} \rangle \\
&= 2\eta_y \langle \boldsymbol{g}_{y,t} - \boldsymbol{g}_{y,t+\frac{1}{2}}, \boldsymbol{y}_{t+1/2} - \boldsymbol{y}_{t+1} \rangle + 2\eta_y \langle \nabla_y f(\boldsymbol{x}_{t+1/2}, \boldsymbol{y}_{t+1/2}), \boldsymbol{y}_{t+1/2} - \boldsymbol{y}^*_{t+1/2} \rangle \\
&\quad + \langle \boldsymbol{\delta}^y_{t+\frac{1}{2}}, \boldsymbol{y}_{t+1/2} - \boldsymbol{y}^*_{t+1/2} \rangle \\
&\leq \|\boldsymbol{y}_{t+1/2} - \boldsymbol{y}_{t+1}\|^2 + \eta_y^2 \|\boldsymbol{g}_{y,t} - \boldsymbol{g}_{y,t+\frac{1}{2}}\|^2 - 2\eta_y \mu \|\boldsymbol{y}_{t+1/2} - \boldsymbol{y}^*_{t+1/2}\|^2 \\
&\quad + \langle \boldsymbol{\delta}^y_{t+\frac{1}{2}}, \boldsymbol{y}_{t+1/2} - \boldsymbol{y}^*_{t+1/2} \rangle \\
&\leq \|\boldsymbol{y}_{t+1/2} - \boldsymbol{y}_{t+1}\|^2 + 2\eta_y^2 \|\nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \nabla_y f(\boldsymbol{x}_{t+\frac{1}{2}}, \boldsymbol{y}_{t+\frac{1}{2}})\|^2 - 2\eta_y \mu \|\boldsymbol{y}_{t+1/2} - \boldsymbol{y}^*_{t+1/2}\|^2 \\
&\quad + \langle \boldsymbol{\delta}^y_{t+\frac{1}{2}}, \boldsymbol{y}_{t+1/2} - \boldsymbol{y}^*_{t+1/2} \rangle + 4\eta_y^2 \|\boldsymbol{\delta}^y_t\|^2 + 4\eta_y^2 \|\boldsymbol{\delta}^y_{t+\frac{1}{2}}\|^2 \\
&\leq \|\boldsymbol{y}_{t+1/2} - \boldsymbol{y}_{t+1}\|^2 + 2\eta_y^2 \ell^2 \|\boldsymbol{x}_{t+\frac{1}{2}} - \boldsymbol{x}_t\|^2 + 2\eta_y^2 \ell^2 \|\boldsymbol{y}_{t+\frac{1}{2}} - \boldsymbol{y}_t\|^2 - 2\eta_y \mu \|\boldsymbol{y}_{t+1/2} - \boldsymbol{y}^*_{t+1/2}\|^2 \\
&\quad + \langle \boldsymbol{\delta}^y_{t+\frac{1}{2}}, \boldsymbol{y}_{t+1/2} - \boldsymbol{y}^*_{t+1/2} \rangle + 4\eta_y^2 \|\boldsymbol{\delta}^y_t\|^2 + 4\eta_y^2 \|\boldsymbol{\delta}^y_{t+\frac{1}{2}}\|^2
\end{aligned} \tag{58}$$

where in the first inequality, we used μ-strong-concavity of $f(\boldsymbol{x}, .)$, and in the second inequality, we used Young's inequality, and in the last one, we used the smoothness property. Now plugging Equation 58 back to Equation 57, we have:

$$\begin{aligned}
\|\boldsymbol{y}_{t+1} - \boldsymbol{y}^*_{t+1/2}\|^2 \leq \|\boldsymbol{y}_t - \boldsymbol{y}^*_{t+1/2}\|^2 &- (1 - 2\eta_y^2 \ell^2) \|\boldsymbol{y}_{t+1/2} - \boldsymbol{y}_t\|^2 \\
&+ 2\eta_y^2 \ell^2 \|\boldsymbol{x}_{t+1/2} - \boldsymbol{x}_t\|^2 - 2\eta_y \mu \|\boldsymbol{y}_{t+1/2} - \boldsymbol{y}^*_{t+1/2}\|^2 \\
&+ \langle \boldsymbol{\delta}^y_{t+\frac{1}{2}}, \boldsymbol{y}_{t+1/2} - \boldsymbol{y}^*_{t+1/2} \rangle + 4\eta_y^2 \|\boldsymbol{\delta}^y_t\|^2 + 4\eta_y^2 \|\boldsymbol{\delta}^y_{t+\frac{1}{2}}\|^2
\end{aligned} \tag{59}$$

Using Young's inequality, we can rewrite Equation 59 as follows:

$$\begin{aligned}
\|\boldsymbol{y}_{t+1} - \boldsymbol{y}^*_{t+1/2}\|^2 \leq (1 - \eta_y \mu) \|\boldsymbol{y}_t - \boldsymbol{y}^*_{t+1/2}\|^2 &- (1 - 2\eta_y \mu - 2\eta_y^2 \ell^2) \|\boldsymbol{y}_{t+1/2} - \boldsymbol{y}_t\|^2 \\
&+ 2\eta_y^2 \ell^2 \|\boldsymbol{x}_{t+1/2} - \boldsymbol{x}_t\|^2 + \langle \boldsymbol{\delta}^y_{t+\frac{1}{2}}, \boldsymbol{y}_{t+1/2} - \boldsymbol{y}^*_{t+1/2} \rangle + 4\eta_y^2 \|\boldsymbol{\delta}^y_t\|^2 \\
&+ 4\eta_y^2 \|\boldsymbol{\delta}^y_{t+\frac{1}{2}}\|^2
\end{aligned} \tag{60}$$

Assuming $\eta_y = \frac{1}{4\ell}$, using Young's inequality, we have the following equation:

$$\|\boldsymbol{y}_t - \boldsymbol{y}^*_{t+1/2}\|^2 \leq (1 + \frac{1}{16\kappa}) \|\boldsymbol{y}_t - \boldsymbol{y}^*_t\|^2 + (1 + 16\kappa) \|\boldsymbol{y}^*_{t+1/2} - \boldsymbol{y}^*_t\|^2 \tag{61}$$

$$\|\boldsymbol{y}_{t+1} - \boldsymbol{y}^*_{t+1}\|^2 \leq (1 + \frac{1}{16\kappa}) \|\boldsymbol{y}_{t+1} - \boldsymbol{y}^*_{t+1/2}\|^2 + (1 + 16\kappa) \|\boldsymbol{y}^*_{t+1} - \boldsymbol{y}^*_{t+1/2}\|^2 \tag{62}$$

Combining Equations 60, 61, 62 and using the κ Lipschitzness of $\boldsymbol{y}^*(.)$, and noting that $1 - 2\eta_y \mu - 2\eta_y^2 \ell^2 > 0$, we get:

$$\begin{aligned}
\|\boldsymbol{y}_{t+1} - \boldsymbol{y}^*_{t+1}\|^2 \leq (1 - \frac{1}{8\kappa}) \|\boldsymbol{y}_t - \boldsymbol{y}^*_t\|^2 &+ 17\kappa^3 \|\boldsymbol{x}_{t+1/2} - \boldsymbol{x}_t\|^2 + 17\kappa^3 \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t+1/2}\|^2 \\
&+ 2\langle \boldsymbol{\delta}^y_{t+\frac{1}{2}}, \boldsymbol{y}_{t+1/2} - \boldsymbol{y}^*_{t+1/2} \rangle + \frac{1}{2\ell^2} \|\boldsymbol{\delta}^y_t\|^2 + \frac{1}{2\ell^2} \|\boldsymbol{\delta}^y_{t+\frac{1}{2}}\|^2
\end{aligned} \tag{63}$$

Using Young's inequality, we have:

$$\begin{aligned}
\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t+\frac{1}{2}}\|^2 &= \eta_x^2 \|\boldsymbol{g}_{x,t+\frac{1}{2}} - \boldsymbol{g}_{x,t}\|^2 \\
&\leq 2\eta_x^2 \|\nabla_x f(\boldsymbol{x}_{t+\frac{1}{2}}, \boldsymbol{y}_{t+\frac{1}{2}}) - \nabla_x f(\boldsymbol{x}_t, \boldsymbol{y}_t)\|^2 + 4\eta_x^2 \|\boldsymbol{\delta}^x_{t+\frac{1}{2}}\|^2 + 4\eta_x^2 \|\boldsymbol{\delta}^x_t\|^2 \\
&\leq 2\eta_x^2 \ell^2 \|\boldsymbol{x}_{t+\frac{1}{2}} - \boldsymbol{x}_t\|^2 + 2\eta_x^2 \ell^2 \|\boldsymbol{y}_{t+\frac{1}{2}} - \boldsymbol{y}_t\|^2 + 4\eta_x^2 \|\boldsymbol{\delta}^x_{t+\frac{1}{2}}\|^2 + 4\eta_x^2 \|\boldsymbol{\delta}^x_t\|^2 \\
&\leq 2\eta_x^2 \ell^2 \|\boldsymbol{x}_{t+\frac{1}{2}} - \boldsymbol{x}_t\|^2 + 4\eta_x^2 \ell^2 \|\boldsymbol{y}_{t+\frac{1}{2}} - \boldsymbol{y}^*_t\|^2 + 4\eta_x^2 \ell^2 \|\boldsymbol{y}_t - \boldsymbol{y}^*_t\|^2 \\
&\quad + 4\eta_x^2 \|\boldsymbol{\delta}^x_{t+\frac{1}{2}}\|^2 + 4\eta_x^2 \|\boldsymbol{\delta}^x_t\|^2 \\
&\leq 2\eta_x^2 \ell^2 \|\boldsymbol{x}_{t+\frac{1}{2}} - \boldsymbol{x}_t\|^2 + 8\eta_x^2 \ell^2 \|\boldsymbol{y}_t - \boldsymbol{y}^*_t\|^2 + \frac{\eta_x^2}{2} \|\boldsymbol{\delta}^y_t\|^2 + 4\eta_x^2 \|\boldsymbol{\delta}^x_{t+\frac{1}{2}}\|^2 \\
&\quad + 4\eta_x^2 \|\boldsymbol{\delta}^x_t\|^2 + 2\eta_x^2 \ell \langle \boldsymbol{\delta}^y_t, \boldsymbol{y}_t - \boldsymbol{y}^*_t \rangle
\end{aligned} \tag{64}$$

where in the last inequality, we used Lemma A.3. Plugging Equation 64, in Equation 63, and assuming $\eta_x \le \frac{1}{64\kappa^2\ell}$ gives:

$$\begin{aligned}\|\boldsymbol{y}_{t+1}-\boldsymbol{y}^*_{t+1}\|^2 \le{}& (1-\frac{1}{12\kappa})\|\boldsymbol{y}_t-\boldsymbol{y}^*_t\|^2 + 18\kappa^3\|\boldsymbol{x}_{t+1/2}-\boldsymbol{x}_t\|^2\\ &+ 2\langle \boldsymbol{\delta}^y_{t+\frac{1}{2}}, \boldsymbol{y}_{t+1/2}-\boldsymbol{y}^*_{t+\frac{1}{2}}\rangle + \frac{1}{64\kappa\ell}\langle \boldsymbol{\delta}^y_t, \boldsymbol{y}_t-\boldsymbol{y}^*_t\rangle\\ &+ \frac{1}{\ell^2}\|\boldsymbol{\delta}^y_t\|^2 + \frac{1}{2\ell^2}\|\boldsymbol{\delta}^y_{t+\frac{1}{2}}\|^2 + \frac{1}{4\ell^2}\|\boldsymbol{\delta}^x_{t+\frac{1}{2}}\|^2 + \frac{1}{4\ell^2}\|\boldsymbol{\delta}^x_t\|^2\end{aligned} \tag{65}$$

or equivalently:

$$\begin{aligned}\|\boldsymbol{y}_{t+1}-\boldsymbol{y}^*_{t+1}\|^2 \le{}& (1-\frac{1}{12\kappa})\|\boldsymbol{y}_t-\boldsymbol{y}^*_t\|^2 + 18\eta_x^2\kappa^3\|\boldsymbol{g}_{x,t}\|^2\\ &+ 2\langle \boldsymbol{\delta}^y_{t+\frac{1}{2}}, \boldsymbol{y}_{t+\frac{1}{2}}-\boldsymbol{y}^*_{t+\frac{1}{2}}\rangle + \frac{1}{64\kappa\ell}\langle \boldsymbol{\delta}^y_t, \boldsymbol{y}_t-\boldsymbol{y}^*_t\rangle\\ &+ \frac{1}{\ell^2}\|\boldsymbol{\delta}^y_t\|^2 + \frac{1}{2\ell^2}\|\boldsymbol{\delta}^y_{t+\frac{1}{2}}\|^2 + \frac{1}{4\ell^2}\|\boldsymbol{\delta}^x_{t+\frac{1}{2}}\|^2 + \frac{1}{4\ell^2}\|\boldsymbol{\delta}^x_t\|^2\end{aligned} \tag{66}$$

Taking expectation from both sides of Equation 66 yields:

$$\mathbb{E}[\|\boldsymbol{y}_{t+1}-\boldsymbol{y}^*_{t+1}\|^2] \le \left(1-\frac{1}{12\kappa}\right)\mathbb{E}[\|\boldsymbol{y}_t-\boldsymbol{y}^*_t\|^2] + 18\eta_x^2\kappa^3\mathbb{E}[\|\boldsymbol{g}_{x,t}\|^2] + 2\frac{\sigma^2}{M\ell^2} \tag{67}$$

Now using Lemma A.2 we get

$$\sum_{i=0}^{T-1}\mathbb{E}[\|\boldsymbol{y}_i-\boldsymbol{y}^*_i\|^2] \le 12\kappa\|\boldsymbol{y}_0-\boldsymbol{y}^*_0\|^2 + 216\eta_x^2\kappa^4\sum_{i=0}^{T-2}\mathbb{E}[\|\boldsymbol{g}_{x,i}\|^2] + \frac{24\kappa\sigma^2(T-1)}{M\ell^2} \tag{68}$$

as stated in the lemma. □

Lemma A.8. *Let* $\Phi(\boldsymbol{x}) = \max_y f(\boldsymbol{x},\boldsymbol{y})$*, and* $\eta_y = \frac{1}{4\ell}$*. Then the iterates of Algorithm 3 satisfy the following inequality:*

$$\begin{aligned}\mathbb{E}[\Phi(\boldsymbol{x}_{t+1})] \le{}& \mathbb{E}[\Phi(\boldsymbol{x}_t)] - \frac{\eta_x}{2}\mathbb{E}[\|\nabla\Phi(\boldsymbol{x}_t)\|^2] - \frac{\eta_x}{4}(1-2\eta_x\kappa\ell-8\eta_x^2\ell^2)\mathbb{E}[\|\boldsymbol{g}_{x,t}\|^2]\\ &+ 5\eta_x\ell^2\mathbb{E}[\|\boldsymbol{y}_t-\boldsymbol{y}^*_t\|^2] + 7\eta_x\frac{\sigma^2}{M}\end{aligned} \tag{69}$$

Proof of Lemma A.8. Let $\boldsymbol{\delta}^x_i = \boldsymbol{g}_{x,i} - \nabla_x f(\boldsymbol{x}_i,\boldsymbol{y}_i)$.Using smoothness property at $\boldsymbol{x}_{t+1}$ and $\boldsymbol{x}_t$, we have:

$$\begin{aligned}\Phi(\boldsymbol{x}_{t+1}) &\le \Phi(\boldsymbol{x}_t) - \eta_x\langle\nabla\Phi(\boldsymbol{x}_t), \boldsymbol{g}_{x,t+\frac{1}{2}}\rangle + \eta_x^2\kappa\ell\|\boldsymbol{g}_{x,t+\frac{1}{2}}\|^2\\ &= \Phi(\boldsymbol{x}_t) - \frac{\eta_x}{2}\|\nabla\Phi(\boldsymbol{x}_t)\|^2 - \frac{\eta_x}{2}(1-2\eta_x\kappa\ell)\|\boldsymbol{g}_{x,t+\frac{1}{2}}\|^2 + \frac{\eta_x}{2}\|\nabla\Phi(\boldsymbol{x}_t)-\boldsymbol{g}_{x,t+\frac{1}{2}}\|^2\\ &\le \Phi(\boldsymbol{x}_t) - \frac{\eta_x}{2}\|\nabla\Phi(\boldsymbol{x}_t)\|^2 - \frac{\eta_x}{2}(1-2\eta_x\kappa\ell)\|\boldsymbol{g}_{x,t+\frac{1}{2}}\|^2\\ &\quad + \eta_x\|\nabla\Phi(\boldsymbol{x}_t)-\nabla_x f(\boldsymbol{x}_{t+\frac{1}{2}},\boldsymbol{y}_{t+\frac{1}{2}})\|^2 + \eta_x\|\boldsymbol{\delta}^x_{t+\frac{1}{2}}\|^2\\ &\le \Phi(\boldsymbol{x}_t) - \frac{\eta_x}{2}\|\nabla\Phi(\boldsymbol{x}_t)\|^2 - \frac{\eta_x}{2}(1-2\eta_x\kappa\ell)\|\boldsymbol{g}_{x,t+\frac{1}{2}}\|^2 + \eta_x\ell^2\|\boldsymbol{x}_{t+\frac{1}{2}}-\boldsymbol{x}_t\|^2\\ &\quad + \eta_x\ell^2\|\boldsymbol{y}_{t+\frac{1}{2}}-\boldsymbol{y}^*_t\|^2 + \eta_x\|\boldsymbol{\delta}^x_{t+\frac{1}{2}}\|^2\end{aligned} \tag{70}$$

Using Young's inequality, we have:

$$\|\boldsymbol{g}_{x,t+\frac{1}{2}}\|^2 \ge \frac{1}{2}\|\boldsymbol{g}_{x,t}\|^2 - \|\boldsymbol{g}_{x,t+\frac{1}{2}}-\boldsymbol{g}_{x,t}\|^2 \tag{71}$$

Plugging Equation 71 back to Equation 70, and assuming $\eta_x \leq \frac{1}{2\kappa\ell}$ results in:

$$\begin{aligned}
\Phi(\boldsymbol{x}_{t+1}) &\leq \Phi(\boldsymbol{x}_t) - \frac{\eta_x}{2}\|\nabla\Phi(\boldsymbol{x}_t)\|^2 - \frac{\eta_x}{4}(1-2\eta_x\kappa\ell)\|\boldsymbol{g}_{x,t}\|^2 + \eta_x\ell^2\|\boldsymbol{x}_{t+\frac{1}{2}} - \boldsymbol{x}_t\|^2 \\
&\quad + \eta_x\ell^2\|\boldsymbol{y}_{t+\frac{1}{2}} - \boldsymbol{y}_t^*\|^2 + \frac{\eta_x}{2}\|\boldsymbol{g}_{x,t+\frac{1}{2}} - \boldsymbol{g}_{x,t}\|^2 + \eta_x\|\boldsymbol{\delta}_{t+\frac{1}{2}}^x\|^2 \\
&\leq \Phi(\boldsymbol{x}_t) - \frac{\eta_x}{2}\|\nabla\Phi(\boldsymbol{x}_t)\|^2 - \frac{\eta_x}{4}(1-2\eta_x\kappa\ell)\|\boldsymbol{g}_{x,t}\|^2 + \eta_x\ell^2\|\boldsymbol{x}_{t+\frac{1}{2}} - \boldsymbol{x}_t\|^2 \\
&\quad + \eta_x\ell^2\|\boldsymbol{y}_{t+\frac{1}{2}} - \boldsymbol{y}_t^*\|^2 + \eta_x\|\nabla_x f(\boldsymbol{x}_{t+\frac{1}{2}}, \boldsymbol{y}_{t+\frac{1}{2}}) - \nabla_x f(\boldsymbol{x}_t, \boldsymbol{y}_t)\|^2 \\
&\quad + 2\eta_x\|\boldsymbol{\delta}_{t+\frac{1}{2}}^x\|^2 + 2\eta_x\|\boldsymbol{\delta}_t^x\|^2 + \eta_x\|\boldsymbol{\delta}_{t+\frac{1}{2}}^x\|^2 \\
&\leq \Phi(\boldsymbol{x}_t) - \frac{\eta_x}{2}\|\nabla\Phi(\boldsymbol{x}_t)\|^2 - \frac{\eta_x}{4}(1-2\eta_x\kappa\ell)\|\boldsymbol{g}_{x,t}\|^2 + \eta_x\ell^2\|\boldsymbol{x}_{t+\frac{1}{2}} - \boldsymbol{x}_t\|^2 \\
&\quad + \eta_x\ell^2\|\boldsymbol{y}_{t+\frac{1}{2}} - \boldsymbol{y}_t^*\|^2 + \eta_x\ell^2\|\boldsymbol{x}_{t+\frac{1}{2}} - \boldsymbol{x}_t\|^2 + \eta_x\ell^2\|\boldsymbol{y}_{t+\frac{1}{2}} - \boldsymbol{y}_t\|^2 \\
&\quad + 2\eta_x\|\boldsymbol{\delta}_{t+\frac{1}{2}}^x\|^2 + 2\eta_x\|\boldsymbol{\delta}_t^x\|^2 + \eta_x\|\boldsymbol{\delta}_{t+\frac{1}{2}}^x\|^2
\end{aligned} \tag{72}$$

Using Lemma A.3 and Young's inequality, we have:

$$\begin{aligned}
\|\boldsymbol{y}_{t+\frac{1}{2}} - \boldsymbol{y}_t^*\|^2 + \|\boldsymbol{y}_{t+\frac{1}{2}} - \boldsymbol{y}_t\|^2 &\leq 3\|\boldsymbol{y}_{t+\frac{1}{2}} - \boldsymbol{y}_t^*\|^2 + 2\|\boldsymbol{y}_t - \boldsymbol{y}_t^*\|^2 \\
&\leq 5\|\boldsymbol{y}_t - \boldsymbol{y}_t^*\|^2 + \frac{3}{8\ell^2}\|\boldsymbol{\delta}_t^y\|^2 + \frac{3}{2\ell}\langle\boldsymbol{\delta}_t^y, \boldsymbol{y}_t - \boldsymbol{y}_t^*\rangle
\end{aligned} \tag{73}$$

Plugging Equation 73 in Equation 72, we get:

$$\begin{aligned}
\Phi(\boldsymbol{x}_{t+1}) \leq \Phi(\boldsymbol{x}_t) &- \frac{\eta_x}{2}\|\nabla\Phi(\boldsymbol{x}_t)\|^2 - \frac{\eta_x}{4}(1-2\eta_x\kappa\ell - 8\eta_x^2\ell^2)\|\boldsymbol{g}_{x,t}\|^2 + 5\eta_x\ell^2\|\boldsymbol{y}_t - \boldsymbol{y}_t^*\|^2 \\
&+ \frac{3}{8}\eta_x\|\boldsymbol{\delta}_t^y\|^2 + \frac{3}{2}\eta_x\ell\langle\boldsymbol{\delta}_t^y, \boldsymbol{y}_t - \boldsymbol{y}_t^*\rangle + 2\eta_x\|\boldsymbol{\delta}_{t+\frac{1}{2}}^x\|^2 + 2\eta_x\|\boldsymbol{\delta}_t^x\|^2 + \eta_x\|\boldsymbol{\delta}_{t+\frac{1}{2}}^x\|^2
\end{aligned} \tag{74}$$

Taking expectations from both sides of Equation 74, we have:

$$\begin{aligned}
\mathbb{E}[\Phi(\boldsymbol{x}_{t+1})] \leq \mathbb{E}[\Phi(\boldsymbol{x}_t)] &- \frac{\eta_x}{2}\mathbb{E}[\|\nabla\Phi(\boldsymbol{x}_t)\|^2] - \frac{\eta_x}{4}(1-2\eta_x\kappa\ell - 8\eta_x^2\ell^2)\mathbb{E}[\|\boldsymbol{g}_{x,t}\|^2] \\
&+ 5\eta_x\ell^2\mathbb{E}[\|\boldsymbol{y}_t - \boldsymbol{y}_t^*\|^2] + 7\eta_x\frac{\sigma^2}{M}
\end{aligned} \tag{75}$$

□

Proof of Theorem 4.2, and Theorem 4.4 for EG. Equipped with the above lemmas, we can prove the theorem as follows. We start by taking summation from $t = 0$ to $t = T-1$ of Equation 69 in Lemma A.8, to get:

$$\begin{aligned}
\mathbb{E}[\Phi(\boldsymbol{x}_T)] \leq \Phi(\boldsymbol{x}_0) &- \frac{\eta_x}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\Phi(\boldsymbol{x}_t)\|^2] - \frac{\eta_x}{4}(1-2\eta_x\kappa\ell - 8\eta_x^2\ell^2)\sum_{t=0}^{T-1}\mathbb{E}[\|\boldsymbol{g}_{x,t}\|^2] \\
&+ 5\eta_x\ell^2\sum_{t=0}^{T-1}\mathbb{E}[\|\boldsymbol{y}_t - \boldsymbol{y}_t^*\|^2] + 7\eta_x\frac{\sigma^2 T}{M}
\end{aligned} \tag{76}$$

Replacing $\sum_{t=0}^{T-1}\mathbb{E}[\|\boldsymbol{y}_t - \boldsymbol{y}_t^*\|^2]$ with the upper bound in Lemma A.7, we have:

$$\begin{aligned}
\mathbb{E}[\Phi(\boldsymbol{x}_T)] \leq 60\eta_x\kappa\ell^2\|\boldsymbol{y}_0 - \boldsymbol{y}_0^*\|^2 + \Phi(\boldsymbol{x}_0) &- \frac{\eta_x}{2}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\Phi(\boldsymbol{x}_t)\|^2] \\
&- \frac{\eta_x}{4}(1-2\eta_x\kappa\ell - 8\eta_x^2\ell^2 - 4320\eta_x^2\kappa^4\ell^2)\sum_{t=0}^{T-1}\mathbb{E}[\|\boldsymbol{g}_{x,t}\|^2] \\
&+ \frac{120\eta_x\kappa\sigma^2(T-1)}{M} + 7\eta_x\frac{\sigma^2 T}{M}
\end{aligned} \tag{77}$$

Let $\eta_x = \frac{1}{75\kappa^2\ell}$. Then $1 - 2\eta_x\kappa\ell - 8\eta_x^2\ell^2 - 4320\eta_x^2\kappa^4\ell^2 > 0$. After rearranging and simplifying the terms of Equation 77, we have:

$$\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla\Phi(\boldsymbol{x}_t)\|^2] \leq \frac{2\Delta_\Phi}{\eta_x} + 120\kappa\ell^2\|\boldsymbol{y}_0 - \boldsymbol{y}_0^*\|^2 + \frac{240\kappa\sigma^2 T}{M} + \frac{14\sigma^2 T}{M} \tag{78}$$

Replacing $\eta_x = \frac{1}{75\kappa^2\ell}$ in Equation 78, we have:

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla\Phi(\boldsymbol{x}_t)\|^2] \leq \frac{150\kappa^2\ell\Delta_\Phi + 120\kappa\ell^2\|\boldsymbol{y}_0 - \boldsymbol{y}_0^*\|^2}{T} + \frac{240\kappa\sigma^2}{M} + \frac{14\sigma^2}{M}. \tag{79}$$

Now by letting, $M = \frac{\kappa\sigma^2}{\epsilon^2}$, and $D_0 = \|\boldsymbol{y}_0 - \boldsymbol{y}_0^*\|^2$, we have:

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla\Phi(\boldsymbol{x}_t)\|^2] \leq O(\frac{\kappa^2\ell\Delta_\Phi + \kappa\ell^2 D_0}{T}) + O(\epsilon^2) \tag{80}$$

□

A.3 Tightness Analysis

In this section we provide the complete proofs for Theorem 4.5 (Subsection A.3.1), and Theorem 4.6 (Subsection A.3.2), showing the tightness of the obtained upper bounds given our choice of learning rates.

A.3.1 GDA

Proof of Theorem 4.5. Recall that we consider the following quadratic NC-SC function $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$

$$f(x, y) := -\tfrac{1}{4}\ell x^2 + bxy - \tfrac{1}{2}\mu y^2.$$

We know f is nonconvex in x (it is actually concave in x) and μ strongly concave in y. Assume $\kappa := \ell/\mu \geq 4$ and choose $b = \sqrt{\mu(\ell + 2\mu_x)/2}$ for some $0 < \mu_x \leq \ell/2$ to be chosen later. Then we know $b \leq \ell/2$ and it is easy to verify f is ℓ smooth. Note that the primal function

$$\Phi(x) = \max_y f(x, y) = \tfrac{1}{2}\mu_x x^2$$

is actually strongly convex. This also justifies the symbol for μ_x. We use GDA to find the solution for $\min_x \max_y f(x, y)$. Actually for this problem the optimal solution is achieved at the origin. The stepsizes are chosen as $\eta_x = \frac{c_1}{\kappa^2\ell}$ and $\eta_y = \frac{c_2}{\ell}$ for some small enough numerical constants c_1 and c_2 such that $c = c_2/c_1 \geq 1$. Also denote $r = \eta_y/\eta_x = c\kappa^2$ as the stepsize ratio. Then the GDA update rule can be written as

$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} = (I + \eta_x \mathbf{M}) \cdot \begin{pmatrix} x_k \\ y_k \end{pmatrix}, \tag{81}$$

where

$$\mathbf{M} := \begin{pmatrix} \ell/2 & -b \\ rb & -\mu r \end{pmatrix}.$$

We note that the above update is a linear time invariant system. We need to analyze its eigenvalues. Let λ_1 and λ_2 be the two eigenvalues of $\mathbf{M}$, we have

$$\lambda_{1,2} = -\frac{1}{2}\left(\mu r - \frac{1}{2}\ell\right) \pm \frac{1}{2}\sqrt{\left(\mu r - \frac{1}{2}\ell\right)^2 - 4r\mu\mu_x}.$$

Note that if we choose $\mu_x < \ell/8$, plugging into $r = c\kappa^2$, we can bound

$$\begin{aligned} 0 \geq \lambda_1 &= -\frac{(2c\kappa - 1)\ell}{4}\left(1 - \sqrt{1 - \frac{4c\kappa\mu_x}{(c\kappa - 1/2)^2\ell}}\right) \\ &\geq -\frac{2c\kappa\mu_x}{c\kappa - 1/2} \geq -4\mu_x. \end{aligned}$$

Let s_1 be the corresponding eigenvalue of $I + \eta_x M$, for small enough $c_1 \leq 1$, it satisfies

$$0 \leq 1 - \frac{4c_1\mu_x}{\kappa^2} \leq s_1 = 1 + \eta_x\lambda_1 \leq 1.$$

We adversarially choose the initial point (x_0, y_0) such that it is parallel to the eigenvector of $I + \eta_x M$ corresponding to s_1. We can always choose $x_0 \geq 0$ for simplicity. Then we have

$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} = (I + \eta_x \mathbf{M})^T \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = s_1^T \begin{pmatrix} x_0 \\ y_0 \end{pmatrix},$$

so we can compute the magnitude of x_T as $x_T = s_1^T x_0$. Also note that $\Delta_\Phi = \Phi(x_0) = \frac{1}{2}\mu_x x_0^2$. Note that if $\Delta_\Phi = 0$, this lemma is trivially true. Therefore we can assume $\Delta_\Phi > 0$. Choosing $\mu_x = \epsilon^2/\Delta_\Phi$, we have

$$\begin{aligned} |\nabla\Phi(\bar{x})| = \mu_x\bar{x} \geq& \mu_x x_T \geq \mu_x x_0 \left(1 - \frac{4c_1\mu_x}{\kappa^2}\right)^T \\ =& \sqrt{2}\epsilon\left(1 - \frac{4c_1\epsilon^2}{\kappa^2\Delta_\Phi}\right)^T, \end{aligned}$$

where $\bar{x} \geq x_T$ because $x_0 \geq x_1 \geq \cdots \geq x_T$ and $\bar{x}$ is sampled from this sequence. Then we know that to achieve $|\nabla\Phi(\bar{x})| \leq \epsilon$, we must have $T = \Omega\left(\frac{\kappa^2\Delta_\Phi}{\epsilon^2}\right)$ as stated.

□

A.3.2 EG/OGDA

Proof of Theorem 4.6 for EG. We consider the same quadratic hard example f and notation used in the proof of Theorem 4.5. For simplicity, denote $\boldsymbol{w} = (x, y)$. Then EG satisfies

$$\begin{aligned} \boldsymbol{w}_{k+1/2} =& (I + \eta_x\mathbf{M})\boldsymbol{w}_k, \\ \boldsymbol{w}_{k+1} =& \boldsymbol{w}_k + \eta_x\mathbf{M}\boldsymbol{w}_{k+1/2} \\ =& (I + \eta_x\mathbf{M} + \eta_x^2\mathbf{M}^2)\boldsymbol{w}_k. \end{aligned}$$

Therefore, similar to GDA, EG is also a linear time invariant system. The transition matrix for EG is $(I + \eta_x\mathbf{M} + \eta_x^2\mathbf{M}^2)$. Its eigenvalues are

$$s_i = 1 + \eta_x\lambda_i + \eta_x^2\lambda_i^2 \geq 1 + \eta_x\lambda_i, \quad i = 1, 2.$$

The rest of analysis is the same as that of GDA.

□

Proof of Theorem 4.6 for OGDA. We consider the same quadratic hard example f and the notation used in the proofs of Theorems 5.1 and 5.2. The dynamics of OGDA is

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + 2\eta_x\mathbf{M}\boldsymbol{w}_k - \eta_x\mathbf{M}\boldsymbol{w}_{k-1}.$$

If we initialize $\boldsymbol{w}_0$ parallel to the eigenvector of $\mathbf{M}$ corresponding to λ_1 and let $\boldsymbol{w}_1 = \boldsymbol{w}_0$, we know every $\boldsymbol{w}_k$ is parallel to it, i.e., $\boldsymbol{w}_k = z_k\boldsymbol{w}_0$ for some scalar z_k which satisfies

$$z_{k+1} = z_k + 2\eta_x\lambda_1 z_k - \eta_x\lambda_1 z_{k-1}.$$

The general solution of the above recurrence relation is

$$z_k = a\alpha^k + b\beta^k$$

for some constant a, b and

$$\begin{aligned} \alpha =& \frac{1}{2}\left(1 + 2\eta_x\lambda_1 + \sqrt{1 + 4\eta_x^2\lambda_1^2}\right), \\ \beta =& \frac{1}{2}\left(1 + 2\eta_x\lambda_1 - \sqrt{1 + 4\eta_x^2\lambda_1^2}\right). \end{aligned}$$

We have

$$1+\eta_x\lambda_1 \le \alpha \le 1, \quad \eta_x\lambda_1 \le \beta \le 0.$$

Using the initial condition $z_{-1}=z_0=1$, we can get the constants

$$\begin{aligned} a =& \frac{\alpha(1-\beta)}{\alpha-\beta} = \frac{1}{2} + \frac{1}{2\sqrt{1+4\eta_x^2\lambda_1^2}} \ge 1/2, \\ b =& -\frac{\beta(1-\alpha)}{\alpha-\beta} = \frac{\sqrt{1+4\eta_x^2\lambda_1^2}-1}{2\sqrt{1+4\eta_x^2\lambda_1^2}} \le \eta_x^2\lambda_1^2. \end{aligned}$$

We can bound

$$\begin{aligned} |z_T| &\ge \frac{1}{2}\left(1+\eta_x\lambda_2\right)^T - |\eta_x\lambda_1|^{k+2} \\ &\ge \frac{1}{2}\left(1-\frac{4c_1\mu_x}{\kappa^2}\right)^T - \frac{1}{4}, \end{aligned}$$

where we use the fact $|\eta_x\lambda_1| \le 1/2$. Similar to the analysis for GDA, choosing $\mu_x = 50\epsilon^2/\Delta_\Phi$, we have

$$\begin{aligned} |\nabla\Phi(\bar{x})| = \mu_x\bar{x} \ge& \mu_x x_T \ge \mu_x x_0\left[\frac{1}{2}\left(1-\frac{4c_1\mu_x}{\kappa^2}\right)^T - \frac{1}{4}\right] \\ =& 10\epsilon\left[\frac{1}{2}\left(1-\frac{4c_1\mu_x}{\kappa^2}\right)^T - \frac{1}{4}\right]. \end{aligned}$$

Therefore, if $|\nabla\Phi(\bar{x})| \le \epsilon$, we must have

$$T = \Omega\left(\frac{\kappa^2}{\mu_x}\right) = \Omega\left(\frac{\kappa^2\Delta_\Phi}{\epsilon^2}\right).$$

□

B Proof of Convergence in Nonconvex-Concave Setting

B.1 Proof of convergence of OGDA

In this section, the convergence of OGDA in NC-C setting has been established. Before presenting the complete proofs, here we briefly discuss the proof sketch.

Proof sketch We start from the standard descent analysis on Moreau envelope function [9]. Let $\delta_t = \Phi(\boldsymbol{x}_t) - f(\boldsymbol{x}_t, \boldsymbol{y}_t)$, then we can show:

$$\begin{aligned} \frac{1}{T+1}\sum_{t=0}^{T}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_t)\|^2 \le \frac{\Phi_{1/2\ell}(\boldsymbol{x}_0)-\Phi_{1/2\ell}(\boldsymbol{x}_{T+1})}{T+1} + O\left(\frac{1}{T+1}\sum_{t=0}^{T}\ell\delta_t\right) + O(\ell\eta_x^2G^2) \\ + \frac{1}{T+1}\sum_{t=0}^{T}O(\|\nabla_x f(\boldsymbol{x}_t,\boldsymbol{y}_t) - \nabla_x f(\boldsymbol{x}_{t-1},\boldsymbol{y}_{t-1})\|^2). \end{aligned}$$

It turns out that the gradient norm depends on two terms, difference between gradient at time t and $t-1$ and δ_t: primal function gap at iteration t. To bound the first term, we can utilize smoothness of ∇f and reduce the problem to bounding $\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2$:

$$\sum_{t=0}^{T}\|\boldsymbol{y}_t-\boldsymbol{y}_{t-1}\|^2 \le \sum_{t=0}^{T}O\left(\eta_y^2\ell\sum_{j=0}^{T}(2\eta_y^2\ell^2)^j\right)\delta_t + \sum_{t=0}^{T}O\left(\eta_x^2\eta_y^2\ell^2G^2\sum_{j=0}^{T}(2\eta_y^2\ell^2)^j\right).$$

Here we reduce difference between dual iterates to primal function gap δ_t. Now, it remains to bound δ_t. We have the following recursion relation holding for any t and any $s \leq t$:

$$\Phi(\boldsymbol{x}_t) - f(\boldsymbol{x}_t, \boldsymbol{y}_t) \leq O(\eta_x(t-s)G^2) + \frac{1}{2\eta_y}(\|\boldsymbol{y}_{t-1} - y^*(\boldsymbol{x}_s)\|^2 - \|\boldsymbol{y}_t - \boldsymbol{y}^*(\boldsymbol{x}_s)\|^2 + \eta_x^2\eta_y\ell G^2$$
$$+ \frac{1}{2}\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2 - \frac{1}{2}\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2) + \langle \nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_y f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2}), \boldsymbol{y}_{t-1} - \boldsymbol{y}^*(\boldsymbol{x}_s)\rangle$$
$$- \langle \nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}), \boldsymbol{y}_t - \boldsymbol{y}^*(\boldsymbol{x}_s)\rangle. \tag{82}$$

If we let s stay the same for some iterations, $(1/T+1)\sum_{t=0}^{T}\delta_t$ vanishes in a telescoping fashion.

In the following, we present the key lemmas, and complete convergence proof of OGDA. First let us introduce some useful lemmas for deterministic setting.

B.1.1 Useful Lemmas

Lemma B.1. *For OGDA (Algorithm 2), under Theorem 4.9's assumptions, the following statement holds for the generated sequence $\{\boldsymbol{y}_t\}$ during algorithm proceeding and any $\boldsymbol{y} \in \mathcal{Y}$:*

$$\begin{aligned}\|\boldsymbol{y}_t - \boldsymbol{y}\|^2 \leq \|\boldsymbol{y}_{t-1} - \boldsymbol{y}\|^2 &- \frac{1}{2}\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 + \frac{1}{2}\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2 + 2\eta_y\langle \boldsymbol{y}_t - \boldsymbol{y}, \nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t)\rangle \\ &+ \eta_y\eta_x^2\ell G^2 - 2\eta_y\langle \nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}), \boldsymbol{y}_t - \boldsymbol{y}\rangle \\ &+ 2\eta_y\langle \nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_y f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2}), \boldsymbol{y}_{t-1} - \boldsymbol{y}\rangle.\end{aligned} \tag{83}$$

Proof. According to updating rule of $\boldsymbol{y}$:

$$\boldsymbol{y}_t = \mathcal{P}_{\mathcal{Y}}\left(\boldsymbol{y}_{t-1} + 2\eta_y\nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \eta_y\nabla_y f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2})\right).$$

Following the analysis in [40], we let $\varepsilon_{t-1} = \eta_y(\nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})) - \eta_y(\nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_y f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2}))$ and re-write the updating rule as:

$$\boldsymbol{y}_t = \mathcal{P}_{\mathcal{Y}}\left(\boldsymbol{y}_{t-1} + \eta_y\nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \varepsilon_{t-1}\right)$$

Then, due to the property of projection onto convex set we have the following inequality that holds for any $\boldsymbol{y} \in \mathcal{Y}$:

$$(\boldsymbol{y} - \boldsymbol{y}_t)^\top(\boldsymbol{y}_t - \boldsymbol{y}_{t-1} - \eta_y\nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t) + \varepsilon_{t-1}) \geq 0.$$

Using the identity that $\langle \boldsymbol{a}, \boldsymbol{b}\rangle = \frac{1}{2}(\|\boldsymbol{a}+\boldsymbol{b}\|^2 - \|\boldsymbol{a}\|^2 - \|\boldsymbol{b}\|^2)$ we have:

$$\begin{aligned}0 &\leq \|\boldsymbol{y} - \boldsymbol{y}_{t-1} - \eta_y\nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t) + \varepsilon_{t-1}\|^2 - \|\boldsymbol{y} - \boldsymbol{y}_t\|^2 - \|\boldsymbol{y}_t - \boldsymbol{y}_{t-1} - \eta_y\nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t) + \varepsilon_{t-1}\|^2 \\ &\leq \|\boldsymbol{y} - \boldsymbol{y}_{t-1}\|^2 - \|\boldsymbol{y} - \boldsymbol{y}_t\|^2 - \|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 + 2\langle \boldsymbol{y}_t - \boldsymbol{y}, \eta_y\nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t)\rangle - 2\langle \boldsymbol{y}_t - \boldsymbol{y}, \varepsilon_{t-1}\rangle\end{aligned}$$

Now we plug the definition of ε_{t-1} into above inequality to get:

$$\begin{aligned}\|\boldsymbol{y} - \boldsymbol{y}_t\|^2 &\leq \|\boldsymbol{y} - \boldsymbol{y}_{t-1}\|^2 - \|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 + 2\eta_y\langle \boldsymbol{y}_t - \boldsymbol{y}, \nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t)\rangle \\ &\quad - 2\eta_y\langle \boldsymbol{y}_t - \boldsymbol{y}, (\nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}))\rangle \\ &\quad + 2\eta_y\langle \boldsymbol{y}_t - \boldsymbol{y}, \nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_y f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2})\rangle \\ &\leq \|\boldsymbol{y} - \boldsymbol{y}_{t-1}\|^2 - \|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 + 2\eta_y\langle \boldsymbol{y}_t - \boldsymbol{y}, \nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t)\rangle \\ &\quad - 2\eta_y\langle \boldsymbol{y}_t - \boldsymbol{y}, (\nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}))\rangle \\ &\quad + 2\eta_y\langle \boldsymbol{y}_{t-1} - \boldsymbol{y}, \nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_y f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2})\rangle \\ &\quad + \eta_y\ell(\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 + \|\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-2}\|^2 + \|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2) \\ &\leq \|\boldsymbol{y} - \boldsymbol{y}_{t-1}\|^2 - \frac{1}{2}\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 + \frac{1}{2}\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2 + 2\eta_y\langle \boldsymbol{y}_t - \boldsymbol{y}, \nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t)\rangle \\ &\quad - 2\eta_y\langle \boldsymbol{y}_t - \boldsymbol{y}, (\nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}))\rangle \\ &\quad + 2\eta_y\langle \boldsymbol{y}_{t-1} - \boldsymbol{y}, \nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_y f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2})\rangle + \eta_y\eta_x^2\ell G^2,\end{aligned} \tag{84}$$

which concludes the proof.

□

Lemma B.2. *For OGDA (Algorithm 2), under the same assumptions made as in Theorem 4.8, the following statement holds for the generated sequence* $\{\boldsymbol{x}_t\}, \{\boldsymbol{y}_t\}$ *during algorithm proceeding:*

$$\begin{aligned}\Phi_{1/2\ell}(\boldsymbol{x}_t) \leq{}& \Phi_{1/2\ell}(\boldsymbol{x}_{t-1}) + 2\eta_x\ell\left(\Phi(\boldsymbol{x}_{t-1}) - f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\right) - \frac{\eta_x}{8}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_{t-1})\|^2 + 3\ell\eta_x^2 G^2 \\ &+ \frac{\eta_x}{2}\|\nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_x f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2})\|^2.\end{aligned}$$

Proof. Let $\hat{\boldsymbol{x}}_{t-1} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} \Phi(\boldsymbol{x}) + \ell\|\boldsymbol{x} - \boldsymbol{x}_{t-1}\|^2$. Notice that:

$$\begin{aligned}\Phi_{1/2\ell}(\boldsymbol{x}_t) &\leq \Phi_{1/2\ell}(\hat{\boldsymbol{x}}_{t-1}) + \ell\|\hat{\boldsymbol{x}}_{t-1} - \boldsymbol{x}_t\|^2 \\ &\leq \Phi_{1/2\ell}(\hat{\boldsymbol{x}}_{t-1}) + \ell(\|\hat{\boldsymbol{x}}_{t-1} - \boldsymbol{x}_{t-1}\|^2 \\ &\quad + 2\eta_x\langle 2\nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_x f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2}), \hat{\boldsymbol{x}}_{t-1} - \boldsymbol{x}_{t-1}\rangle + 3\eta_x^2 G^2)\end{aligned}$$

According to smoothness of $f(\cdot, \boldsymbol{y})$, we have:

$$\begin{aligned}\langle\hat{\boldsymbol{x}}_{t-1} - \boldsymbol{x}_{t-1}, \nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\rangle &\leq f(\hat{\boldsymbol{x}}_{t-1}, \boldsymbol{y}_{t-1}) - f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) + \frac{\ell}{2}\|\hat{\boldsymbol{x}}_{t-1} - \boldsymbol{x}_{t-1}\|^2 \\ &\leq \Phi(\boldsymbol{x}_{t-1}) - f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \frac{\ell}{2}\|\hat{\boldsymbol{x}}_{t-1} - \boldsymbol{x}_{t-1}\|^2.\end{aligned}$$

So we have

$$\begin{aligned}\Phi_{1/2\ell}(\boldsymbol{x}_t) &\leq \Phi_{1/2\ell}(\hat{\boldsymbol{x}}_{t-1}) + \ell\|\hat{\boldsymbol{x}}_{t-1} - \boldsymbol{x}_t\|^2 \\ &\leq \Phi_{1/2\ell}(\hat{\boldsymbol{x}}_{t-1}) + \ell\|\boldsymbol{x}_{t-1} - \hat{\boldsymbol{x}}_{t-1}\|^2 \\ &\quad + 2\eta_x\ell\left(\Phi(\boldsymbol{x}_{t-1}) - f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \frac{\ell}{2}\|\hat{\boldsymbol{x}}_{t-1} - \boldsymbol{x}_{t-1}\|^2\right) + 3\ell\eta_x^2 G^2 \\ &\quad + \eta_x\ell\left(\frac{1}{2\ell}\|\nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_x f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2})\|^2 + \frac{\ell}{2}\|\boldsymbol{x}_{t-1} - \hat{\boldsymbol{x}}_{t-1}\|^2\right) \\ &\leq \Phi_{1/2\ell}(\boldsymbol{x}_{t-1}) + 2\eta_x\ell\left(\Phi(\boldsymbol{x}_{t-1}) - f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\right) - \frac{\eta_x\ell^2}{2}\|\hat{\boldsymbol{x}}_{t-1} - \boldsymbol{x}_{t-1}\|^2 + 3\ell\eta_x^2 G^2 \\ &\quad + \frac{\eta_x}{2}\|\nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_x f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2})\|^2.\end{aligned}$$

Using the fact that $\|\hat{\boldsymbol{x}}_{t-1} - \boldsymbol{x}_{t-1}\| = \frac{1}{2\ell}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_{t-1})\|$ will conclude the proof.

□

Lemma B.3 (Iterates gap)**.** *For OGDA (Algorithm 2), under Theorem 4.8's assumptions, the following statement holds for the generated sequence* $\{\boldsymbol{y}_t\}$ *during algorithm proceeding:*

$$\begin{aligned}\sum_{t=0}^{T}\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 \leq{}& \sum_{t=0}^{T}\left(\sum_{j=0}^{T}(2\eta_y^2\ell^2)^j\right)4\eta_y^2\ell\left(\Phi(\boldsymbol{x}_t) - f(\boldsymbol{x}_t, \boldsymbol{y}_t)\right) \\ &+ \sum_{t=0}^{T}\left(\sum_{j=0}^{T}(2\eta_y^2\ell^2)^j\right)2\eta_x^2\eta_y^2\ell^2 G^2.\end{aligned}$$

Proof. Observe that

$$\begin{aligned}\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 &= \eta_y^2\left\|2\nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_y f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2})\right\|^2 \\ &\leq 2\eta_y^2\left\|\nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\right\|^2 + 2\eta_y^2\left\|\nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_y f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2})\right\|^2 \\ &\leq 4\eta_y^2\ell\left(\Phi(\boldsymbol{x}_{t-1}) - f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\right) + 2\eta_y^2\ell^2\left(\|\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-2}\|^2 + \left\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\right\|^2\right) \\ &\leq 2\eta_y^2\ell^2\left\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\right\|^2 + 4\eta_y^2\ell\left(\Phi(\boldsymbol{x}_{t-1}) - f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\right) + 2\eta_x^2\eta_y^2\ell^2 G^2.\end{aligned}$$

Unrolling the recursion yields:

$$\begin{aligned}\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 \leq (2\eta_y^2\ell^2)^{t-1}\left\|\boldsymbol{y}_0 - \boldsymbol{y}_{-1}\right\|^2 + \sum_{j=1}^{t}(2\eta_y^2\ell^2)^{t-j}4\eta_y^2\ell\left(\Phi(\boldsymbol{x}_{j-1}) - f(\boldsymbol{x}_{j-1}, \boldsymbol{y}_{j-1})\right) \\ + \sum_{j=1}^{t}(2\eta_y^2\ell^2)^{t-j}2\eta_x^2\eta_y^2\ell^2G^2.\end{aligned}$$

Since $\boldsymbol{y}_0 = \boldsymbol{y}_{-1}$, we have:

$$\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 \leq \sum_{j=1}^{t}(2\eta_y^2\ell^2)^{t-j}4\eta_y^2\ell\left(\Phi(\boldsymbol{x}_{j-1}) - f(\boldsymbol{x}_{j-1}, \boldsymbol{y}_{j-1})\right) + \sum_{j=1}^{t}(2\eta_y^2\ell^2)^{t-j}2\eta_x^2\eta_y^2\ell^2G^2.$$

Finally, summing the above inequality over $t = 0$ to T yields:

$$\begin{aligned}\sum_{t=0}^{T}\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 \leq \sum_{t=0}^{T}\left(\sum_{j=0}^{T}(2\eta_y^2\ell^2)^j\right)4\eta_y^2\ell\left(\Phi(\boldsymbol{x}_t) - f(\boldsymbol{x}_t, \boldsymbol{y}_t)\right) \\ + \sum_{t=0}^{T}\left(\sum_{j=0}^{T}(2\eta_y^2\ell^2)^j\right)2\eta_x^2\eta_y^2\ell^2G^2.\end{aligned}$$

□

Lemma B.4. *For OGDA (Algorithm 2), under Theorem 4.8's assumptions, the following statement holds for the generated sequence* $\{\boldsymbol{y}_t\}$ *during algorithm proceeding and* $\forall s \leq t$*:*

$$\begin{aligned}\Phi(\boldsymbol{x}_t) - f(\boldsymbol{x}_t, \boldsymbol{y}_t) \leq 2\eta_x(t-s)G^2 + \frac{1}{2\eta_y}\left(\|\boldsymbol{y}_{t-1} - y^*(\boldsymbol{x}_s)\|^2 - \|\boldsymbol{y}_t - \boldsymbol{y}^*(x_s)\|^2 - \frac{1}{2}\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2\right) \\ + \frac{1}{2\eta_y}\left(\frac{1}{2}\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2 + \eta_y\eta_x^2\ell G^2\right) - \langle\nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}), \\ \boldsymbol{y}_t - \boldsymbol{y}^*(\boldsymbol{x}_s)\rangle + \langle\nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_y f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2}), \boldsymbol{y}_{t-1} - \boldsymbol{y}^*(\boldsymbol{x}_s)\rangle.\end{aligned}$$

Proof. Observe that:

$$\begin{aligned}\Phi(\boldsymbol{x}_t) - f(\boldsymbol{x}_t, \boldsymbol{y}_t) &\leq f(\boldsymbol{x}_t, \boldsymbol{y}^*(\boldsymbol{x}_t)) - f(\boldsymbol{x}_s, \boldsymbol{y}^*(\boldsymbol{x}_t)) + f(\boldsymbol{x}_s, \boldsymbol{y}^*(\boldsymbol{x}_s)) - f(\boldsymbol{x}_t, \boldsymbol{y}^*(\boldsymbol{x}_s)) \\ &\quad + f(\boldsymbol{x}_t, \boldsymbol{y}^*(\boldsymbol{x}_s)) - f(\boldsymbol{x}_t, \boldsymbol{y}_t) \\ &\leq 2(t-s)\eta_x G^2 - \langle\boldsymbol{y}_t - \boldsymbol{y}, \nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t)\rangle,\end{aligned}$$

where in the last step we use the concavity of $f(\boldsymbol{x}_t, \cdot)$.

Plugging in Lemma B.1 will conclude the proof as follows:

$$\begin{aligned}\Phi(\boldsymbol{x}_t) - f(\boldsymbol{x}_t, \boldsymbol{y}_t) \leq\ & 2(t-s)\eta_x G^2 \\ & + \frac{1}{2\eta_y}\left(\|\boldsymbol{y}_{t-1} - \boldsymbol{y}\|^2 - \|\boldsymbol{y}_t - \boldsymbol{y}\|^2 - \frac{1}{2}\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 + \frac{1}{2}\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2\right) \\ & + \frac{1}{2\eta_y}\eta_y\eta_x^2\ell G^2 - \langle\nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}), \boldsymbol{y}_t - \boldsymbol{y}\rangle \\ & + \langle\nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_y f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2}), \boldsymbol{y}_{t-1} - \boldsymbol{y}\rangle.\end{aligned}$$

□

Lemma B.5. *For OGDA (Algorithm 2), under the same assumptions made in Theorem 4.8, the following statement holds for the generated sequence* $\{\boldsymbol{x}_t\}, \{\boldsymbol{y}_t\}$ *during algorithm proceeding:*

$$\frac{1}{T+1}\sum_{t=0}^{T}\Phi(\boldsymbol{x}_t) - f(\boldsymbol{x}_t, \boldsymbol{y}_t) \leq \frac{1}{B}\left(2\eta_x B^2G^2 + \frac{1}{2\eta_y}\left(D^2 + \eta_y\ell D^2\right) + 2(3\eta_x G^2 + D)D\right).$$

Proof. Let $S = (T+1)/B$, and we choose $s = jB$, $j = 0, ..., S$. Then by summing over t on the both side of Lemma B.4 we have:

$$
\begin{aligned}
\frac{1}{T+1}\sum_{t=0}^{T}\Phi(\boldsymbol{x}_t) - f(\boldsymbol{x}_t,\boldsymbol{y}_t) &= \frac{1}{T+1}\sum_{j=0}^{S}\sum_{t=jB}^{(j+1)B-1}\Phi(\boldsymbol{x}_t) - f(\boldsymbol{x}_t,\boldsymbol{y}_t) \\
&\leq \frac{1}{T+1}\sum_{j=0}^{S}\left[2\eta_x B^2G^2 + \frac{1}{2\eta_y}\left(\|\boldsymbol{y}_{jB-1} - \boldsymbol{y}^*(x_{jB})\|^2 + \frac{1}{2}\|\boldsymbol{y}_{jB-1} - \boldsymbol{y}_{jB-2}\|^2\right)\right] \\
&\quad + \frac{1}{T+1}\sum_{j=0}^{S}\big(-\langle\nabla_y f(\boldsymbol{x}_{(j+1)B-1}, y_{(j+1)B-1}) - \nabla_y f(\boldsymbol{x}_{(j+1)B-2}, y_{(j+1)B-2}), \\
&\quad \boldsymbol{y}_{(j+1)B-1} - y^*(\boldsymbol{x}_{jB})\rangle + \langle\nabla_y f(\boldsymbol{x}_{jB-1},\boldsymbol{y}_{jB-1}) - \nabla_y f(\boldsymbol{x}_{jB-2},\boldsymbol{y}_{jB-2}), \boldsymbol{y}_{jB-1} - y^*(\boldsymbol{x}_{jB}\rangle \\
&\leq \frac{1}{T+1}\sum_{j=0}^{S}\left(2\eta_x B^2G^2 + \frac{1}{2\eta_y}\left(D^2 + \frac{1}{2}D^2\right) + 2(3\eta_x G^2 + D)D\right) \\
&\leq \frac{1}{B}\left(2\eta_x B^2G^2 + \frac{1}{2\eta_y}\left(D^2 + \eta_y \ell D^2\right) + 2(3\eta_x G^2 + D)D\right).
\end{aligned}
$$

□

B.1.2 Proof of Theorem 4.8 for OGDA

In this section we are going to provide the proof of Theorem 4.8 on the convergence rate of OGDA in both deterministic and stochastic settings.

We start by establishing the convergence rate in deterministic setting. Before, we first state the formal version of Theorem 4.8 here:

Theorem B.6 (OGDA Deterministic (Theorem 4.8 restated))**.** *Under Assumption 4.7, if we choose* $\eta_x = \Theta\left(\min\left\{\frac{\epsilon}{\ell G}, \frac{\epsilon^2}{\ell G^2}, \frac{\epsilon^4}{D^2G^2\ell^3}\right\}\right)$, $\eta_y = \frac{1}{2\ell}$, *then OGDA (Algorithm 2) guarantees to find ϵ-stationary point, i.e.,* $\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_t)\|^2 \leq \epsilon^2$, *with the gradient complexity bounded by:*

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

Proof. From Lemma B.2 we have:

$$
\begin{aligned}
\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_t)\|^2 &\leq \frac{\Phi_{1/2\ell}(\boldsymbol{x}_0) - \Phi_{1/2\ell}(\boldsymbol{x}_t)}{\eta_x T} + 16\ell\frac{1}{T}\sum_{t=0}^{T-1}(\Phi(\boldsymbol{x}_t) - f(\boldsymbol{x}_t,\boldsymbol{y}_t)) + 24\eta_x\ell G^2 \\
&\quad + 4\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla_x f(\boldsymbol{x}_t,\boldsymbol{y}_t) - \nabla_x f(\boldsymbol{x}_{t-1},\boldsymbol{y}_{t-1})\|^2, \\
&\leq \frac{\Phi_{1/2\ell}(\boldsymbol{x}_0) - \Phi_{1/2\ell}(\boldsymbol{x}_t)}{\eta_x T} + 16\ell\frac{1}{T}\sum_{t=0}^{T-1}(\Phi(\boldsymbol{x}_t) - f(\boldsymbol{x}_t,\boldsymbol{y}_t)) + 24\eta_x\ell G^2 \\
&\quad + 4\frac{1}{T+1}\sum_{t=0}^{T}\ell^2(3\eta_x^2G^2 + \|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2).
\end{aligned}
$$

Plugging in Lemma B.3 yields:

$$\begin{aligned}\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_t)\|^2 \le{}& \frac{\Phi_{1/2\ell}(\boldsymbol{x}_0)-\Phi_{1/2\ell}(\boldsymbol{x}_t)}{\eta_x T}\\ &+16\ell\frac{1}{T}\sum_{t=0}^{T-1}\left(\Phi(\boldsymbol{x}_t)-f(\boldsymbol{x}_t,\boldsymbol{y}_t)\right)+24\eta_x\ell G^2+12\eta_x^2\ell^2G^2\\ &+4\frac{1}{T+1}\ell^2\left(\sum_{t=0}^{T}\left(\sum_{j=0}^{T}(2\eta_y^2\ell^2)^j\right)4\eta_y^2\ell\left(\Phi(\boldsymbol{x}_t)-f(\boldsymbol{x}_t,\boldsymbol{y}_t)\right)+\sum_{t=0}^{T}\left(\sum_{j=0}^{T}(2\eta_y^2\ell^2)^j\right)2\eta_x^2\eta_y^2\ell^2G^2\right),\end{aligned}$$

since we choose $\eta_y\ell\le\frac{1}{2}$, we know that:

$$\sum_{j=0}^{T}\left(2\eta_y^2\ell^2\right)^j\le 2.$$

Hence we have:

$$\begin{aligned}\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_t)\|^2 \le{}& \frac{\Phi_{1/2\ell}(\boldsymbol{x}_0)-\Phi_{1/2\ell}(\boldsymbol{x}_t)}{\eta_x T}+(16\ell+32\eta_y^2\ell^3)\frac{1}{T+1}\sum_{t=0}^{T}\left(\Phi(\boldsymbol{x}_t)-f(\boldsymbol{x}_t,\boldsymbol{y}_t)\right)\\ &+24\eta_x\ell G^2+12\eta_x^2\ell^2G^2+16\eta_x^2\eta_y^2\ell^4G^2.\end{aligned}$$

Now we plug in Lemma B.5 to replace $\Phi(\boldsymbol{x}_t)-f(\boldsymbol{x}_t,\boldsymbol{y}_t)$:

$$\begin{aligned}\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_t)\|^2 \le{}& \frac{\Phi_{1/2\ell}(\boldsymbol{x}_0)-\Phi_{1/2\ell}(\boldsymbol{x}_t)}{\eta_x T}\\ &+(16\ell+32\eta_y^2\ell^3)\frac{1}{B}\left(2\eta_x B^2G^2+\frac{1}{2\eta_y}\left(D^2+\eta_y\ell D^2\right)+2(3\eta_xG^2+D)D\right)\\ &+24\eta_x\ell G^2+12\eta_x^2\ell^2G^2+16\eta_x^2\eta_y^2\ell^4G^2.\end{aligned}$$

Choose $B=O\left(\frac{D}{G\sqrt{\eta_x\eta_y}}\right)$, $\eta_x=O\left(\min\left\{\frac{\epsilon}{\ell G},\frac{\epsilon^2}{\ell G^2},\frac{\epsilon^4}{D^2G^2\ell^3}\right\}\right)$, $\eta_y=\frac{1}{2\ell}$, and then we guarantee that $\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_t)\|^2\le\epsilon^2$ with the gradient complexity is bounded by:

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

□

Stochastic setting.

We now turn to presenting the proof of OGDA in stochastic setting. First let us introduce some useful lemmas.

B.1.3 Useful Lemmas

Lemma B.7. *For Stochastic OGDA (Algorithm 2), under the same assumptions made in Theorem 4.9, if we choose $\eta\le 1/4\ell$ the following statement holds for the generated sequence $\{\boldsymbol{y}_t\}$ during algorithm proceeding and for any $\boldsymbol{y}\in\mathcal{Y}$:*

$$\begin{aligned}\mathbb{E}\|\boldsymbol{y}-\boldsymbol{y}_t\|^2\le{}&\mathbb{E}\|\boldsymbol{y}-\boldsymbol{y}_{t-1}\|^2-\frac{1}{4}\mathbb{E}\|\boldsymbol{y}_t-\boldsymbol{y}_{t-1}\|^2+\frac{1}{4}\mathbb{E}\|\boldsymbol{y}_{t-1}-\boldsymbol{y}_{t-2}\|^2+2\eta_y\langle\boldsymbol{y}_t-\boldsymbol{y},\nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t)\rangle\\ &+\eta_y\eta_x^2\ell(G^2+\sigma^2)+6\eta_y^2\sigma^2-2\eta_y\langle\boldsymbol{y}_t-\boldsymbol{y},\nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t)-\nabla_y f(\boldsymbol{x}_{t-1},\boldsymbol{y}_{t-1})\rangle\\ &+2\eta_y\langle\boldsymbol{y}_{t-1}-\boldsymbol{y},\nabla_y f(\boldsymbol{x}_{t-1},\boldsymbol{y}_{t-1})-\nabla_y f(\boldsymbol{x}_{t-2},\boldsymbol{y}_{t-2})\rangle.\end{aligned}$$

Proof. The proof is similar to deterministic setting. Here we use ξ_{t-1} to denote the random sample at iteration t. According to updating rule of $\boldsymbol{y}$:

$$\boldsymbol{y}_t = \mathcal{P}_{\mathcal{Y}}\left(\boldsymbol{y}_{t-1} + 2\eta_y \nabla_y f(\boldsymbol{x}_{t-1},\boldsymbol{y}_{t-1};\xi_{t-1}) - \eta_y \nabla_y f(\boldsymbol{x}_{t-2},\boldsymbol{y}_{t-2};\xi_{t-1})\right)$$

Similarly to deterministic setting, we let

$$\tilde{\varepsilon}_{t-1} = \eta_y(\nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t) - \nabla_y f(\boldsymbol{x}_{t-1},\boldsymbol{y}_{t-1};\xi_{t-1})) - \eta_y(\nabla_y f(\boldsymbol{x}_{t-1},\boldsymbol{y}_{t-1};\xi_{t-1}) - \nabla_y f(\boldsymbol{x}_{t-2},\boldsymbol{y}_{t-2};\xi_{t-1}))$$
$$\varepsilon_{t-1} = \eta_y(\nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t) - \nabla_y f(\boldsymbol{x}_{t-1},\boldsymbol{y}_{t-1})) - \eta_y(\nabla_y f(\boldsymbol{x}_{t-1},\boldsymbol{y}_{t-1}) - \nabla_y f(\boldsymbol{x}_{t-2},\boldsymbol{y}_{t-2}))$$

and re-write the updating rule as:

$$\boldsymbol{y}_t = \mathcal{P}_{\mathcal{Y}}\left(\boldsymbol{y}_{t-1} + \eta_y \nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t) - \tilde{\varepsilon}_{t-1}\right)$$

Due to the property of projection we have:

$$(\boldsymbol{y}-\boldsymbol{y}_t)^\top(\boldsymbol{y}_t - \boldsymbol{y}_{t-1} - \eta_y \nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t) + \tilde{\varepsilon}_{t-1}) \geq 0$$

Using the identity that $\langle \boldsymbol{a},\boldsymbol{b}\rangle = \frac{1}{2}(\|\boldsymbol{a}+\boldsymbol{b}\|^2 - \|\boldsymbol{a}\|^2 - \|\boldsymbol{b}\|^2)$ we have:

$$\begin{aligned}
0 &\leq \|\boldsymbol{y}-\boldsymbol{y}_{t-1} - \eta_y \nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t) + \tilde{\varepsilon}_{t-1}\|^2 - \|\boldsymbol{y}-\boldsymbol{y}_t\|^2 - \|\boldsymbol{y}_t - \boldsymbol{y}_{t-1} - \eta_y \nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t) + \tilde{\varepsilon}_{t-1}\|^2\\
&= \|\boldsymbol{y}-\boldsymbol{y}_{t-1}\|^2 - \|\boldsymbol{y}-\boldsymbol{y}_t\|^2 - \|\boldsymbol{y}_t-\boldsymbol{y}_{t-1}\|^2 + 2\langle \boldsymbol{y}_t - \boldsymbol{y}, \eta_y \nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t)\rangle\\
&\quad + 2\langle \boldsymbol{y}-\boldsymbol{y}_{t-1}, \tilde{\varepsilon}_{t-1}\rangle - 2\langle \boldsymbol{y}_t - \boldsymbol{y}_{t-1}, \tilde{\varepsilon}_{t-1}\rangle.
\end{aligned}$$

Notice that

$$\begin{aligned}
-2\langle \boldsymbol{y}_t - \boldsymbol{y}_{t-1}, \tilde{\varepsilon}_{t-1}\rangle &= -2\langle \boldsymbol{y}_t - \boldsymbol{y}_{t-1}, \varepsilon_{t-1}\rangle - 2\langle \boldsymbol{y}_t - \boldsymbol{y}_{t-1}, \tilde{\varepsilon}_{t-1} - \varepsilon_{t-1}\rangle\\
&\leq -2\langle \boldsymbol{y}_t - \boldsymbol{y}_{t-1}, \varepsilon_{t-1}\rangle + \frac{1}{2}\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 + 2\|\tilde{\varepsilon}_{t-1} - \varepsilon_{t-1}\|^2
\end{aligned}$$

So we have:

$$\begin{aligned}
0 &\leq \|\boldsymbol{y}-\boldsymbol{y}_{t-1} - \eta_y \nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t) + \tilde{\varepsilon}_{t-1}\|^2 - \|\boldsymbol{y}-\boldsymbol{y}_t\|^2 - \|\boldsymbol{y}_t - \boldsymbol{y}_{t-1} - \eta_y \nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t) + \tilde{\varepsilon}_{t-1}\|^2\\
&= \|\boldsymbol{y}-\boldsymbol{y}_{t-1}\|^2 - \|\boldsymbol{y}-\boldsymbol{y}_t\|^2 - \|\boldsymbol{y}_t-\boldsymbol{y}_{t-1}\|^2 + 2\langle \boldsymbol{y}_t - \boldsymbol{y}, \eta_y \nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t)\rangle\\
&\quad + 2\langle \boldsymbol{y}-\boldsymbol{y}_{t-1}, \tilde{\varepsilon}_{t-1}\rangle - 2\langle \boldsymbol{y}_t - \boldsymbol{y}_{t-1}, \varepsilon_{t-1}\rangle + \frac{1}{2}\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 + 2\|\tilde{\varepsilon}_{t-1} - \varepsilon_{t-1}\|^2.
\end{aligned}$$

Taking expectation over ξ_{t-1} yields:

$$\begin{aligned}
0 &\leq \mathbb{E}\|\boldsymbol{y}-\boldsymbol{y}_{t-1}\|^2 - \mathbb{E}\|\boldsymbol{y}-\boldsymbol{y}_t\|^2 - \frac{1}{2}\mathbb{E}\|\boldsymbol{y}_t-\boldsymbol{y}_{t-1}\|^2 + 2\langle \boldsymbol{y}_t - \boldsymbol{y}, \eta_y \nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t)\rangle\\
&\quad - 2\langle \boldsymbol{y}_t - \boldsymbol{y}, \varepsilon_{t-1}\rangle + 6\eta_y^2\sigma^2.
\end{aligned}$$

Now we plug the definition of ε_{t-1} into above inequality:

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{y}-\boldsymbol{y}_t\|^2 &\leq \mathbb{E}\|\boldsymbol{y}-\boldsymbol{y}_{t-1}\|^2 - \mathbb{E}\|\boldsymbol{y}_t-\boldsymbol{y}_{t-1}\|^2 + 2\eta_y\langle \boldsymbol{y}_t - \boldsymbol{y}, \nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t)\rangle\\
&\quad - 2\eta_y\langle \boldsymbol{y}_t - \boldsymbol{y}, \nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t) - \nabla_y f(\boldsymbol{x}_{t-1},\boldsymbol{y}_{t-1})\rangle + 6\eta_y^2\sigma^2\\
&\quad + 2\eta_y\langle \boldsymbol{y}_t - \boldsymbol{y}, \nabla_y f(\boldsymbol{x}_{t-1},\boldsymbol{y}_{t-1}) - \nabla_y f(\boldsymbol{x}_{t-2},\boldsymbol{y}_{t-2})\rangle\\
&\leq \mathbb{E}\|\boldsymbol{y}-\boldsymbol{y}_{t-1}\|^2 - \mathbb{E}\|\boldsymbol{y}_t-\boldsymbol{y}_{t-1}\|^2 + 2\eta_y\langle \boldsymbol{y}_t - \boldsymbol{y}, \nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t)\rangle\\
&\quad - 2\eta_y\langle \boldsymbol{y}_t - \boldsymbol{y}, \nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t) - \nabla_y f(\boldsymbol{x}_{t-1},\boldsymbol{y}_{t-1})\rangle + 6\eta_y^2\sigma^2\\
&\quad + 2\eta_y\langle \boldsymbol{y}_{t-1} - \boldsymbol{y}, \nabla_y f(\boldsymbol{x}_{t-1},\boldsymbol{y}_{t-1}) - \nabla_y f(\boldsymbol{x}_{t-2},\boldsymbol{y}_{t-2})\rangle\\
&\quad + \eta_y\ell(\mathbb{E}\|\boldsymbol{y}_t-\boldsymbol{y}_{t-1}\|^2 + \mathbb{E}\|\boldsymbol{x}_{t-1}-\boldsymbol{x}_{t-2}\|^2 + \mathbb{E}\|\boldsymbol{y}_{t-1}-\boldsymbol{y}_{t-2}\|^2)\\
&\overset{①}{\leq} \mathbb{E}\|\boldsymbol{y}-\boldsymbol{y}_{t-1}\|^2 - \frac{1}{4}\mathbb{E}\|\boldsymbol{y}_t-\boldsymbol{y}_{t-1}\|^2 + \frac{1}{4}\mathbb{E}\|\boldsymbol{y}_{t-1}-\boldsymbol{y}_{t-2}\|^2\\
&\quad + 2\eta_y\langle \boldsymbol{y}_t - \boldsymbol{y}, \nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t)\rangle + \eta_y\eta_x^2\ell(G^2+\sigma^2) + 6\eta_y^2\sigma^2\\
&\quad - 2\eta_y\langle \boldsymbol{y}_t - \boldsymbol{y}, \nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t) - \nabla_y f(\boldsymbol{x}_{t-1},\boldsymbol{y}_{t-1})\rangle\\
&\quad + 2\eta_y\langle \boldsymbol{y}_{t-1} - \boldsymbol{y}, \nabla_y f(\boldsymbol{x}_{t-1},\boldsymbol{y}_{t-1}) - \nabla_y f(\boldsymbol{x}_{t-2},\boldsymbol{y}_{t-2})\rangle,
\end{aligned}$$

where in ① we use the fact that $\eta_y\ell \leq \frac{1}{4}$ and hence can conclude the proof.

□

Lemma B.8. *For Stochastic OGDA (Algorithm 2), under same assumptions as in Theorem 4.9, the following statement holds for the genserated sequence $\{\boldsymbol{x}_t\}, \{\boldsymbol{y}_t\}$ during algorithm proceeding:*

$$\begin{aligned}\mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{x}_t)] &\le \mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{x}_{t-1})] + 2\eta_x\ell\mathbb{E}\left(\Phi(\boldsymbol{x}_{t-1}) - f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\right) - \frac{\eta_x}{8}\mathbb{E}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_{t-1})\|^2 \\ &\quad + 3\ell\eta_x^2(G^2+\sigma^2) + \frac{\eta_x}{2}\mathbb{E}\|\nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_x f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2})\|^2.\end{aligned}$$

Proof. Let $\hat{\boldsymbol{x}}_{t-1} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} \Phi(\boldsymbol{x}) + \ell\|\boldsymbol{x} - \boldsymbol{x}_{t-1}\|^2$. Notice that:

$$\begin{aligned}\mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{x}_t)] &\le \mathbb{E}[\Phi_{1/2\ell}(\hat{\boldsymbol{x}}_{t-1})] + \ell\mathbb{E}\|\hat{\boldsymbol{x}}_{t-1} - \boldsymbol{x}_t\|^2 \\ &\le \mathbb{E}[\Phi_{1/2\ell}(\hat{\boldsymbol{x}}_{t-1})] \\ &\quad + \ell(\mathbb{E}\|\boldsymbol{x}_{t-1} - \hat{\boldsymbol{x}}_{t-1}\|^2 + 2\eta_x\langle 2\nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_x f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2}), \boldsymbol{x}_{t-1} - \hat{\boldsymbol{x}}_{t-1}\rangle \\ &\quad + 3\eta_x^2(G^2+\sigma^2))\end{aligned}$$

According to smoothness of $f(\cdot, \boldsymbol{y})$, we have:

$$\begin{aligned}\langle \hat{\boldsymbol{x}}_{t-1} - \boldsymbol{x}_{t-1}, \nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\rangle &\le f(\hat{\boldsymbol{x}}_{t-1}, \boldsymbol{y}_{t-1}) - f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) + \frac{\ell}{2}\|\hat{\boldsymbol{x}}_{t-1} - \boldsymbol{x}_{t-1}\|^2 \\ &\le \Phi(\boldsymbol{x}_{t-1}) - f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \frac{\ell}{2}\|\hat{\boldsymbol{x}}_{t-1} - \boldsymbol{x}_{t-1}\|^2.\end{aligned}$$

So we have

$$\begin{aligned}\mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{x}_t)] &\le \mathbb{E}[\Phi_{1/2\ell}(\hat{\boldsymbol{x}}_{t-1})] + \ell\mathbb{E}\|\hat{\boldsymbol{x}}_{t-1} - \boldsymbol{x}_t\|^2 \\ &\le \mathbb{E}[\Phi_{1/2\ell}(\hat{\boldsymbol{x}}_{t-1})] + \ell\mathbb{E}\|\boldsymbol{x}_{t-1} - \hat{\boldsymbol{x}}_{t-1}\|^2 \\ &\quad + 2\eta_x\ell\mathbb{E}\left(\Phi(\boldsymbol{x}_{t-1}) - f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \frac{\ell}{2}\mathbb{E}\|\hat{\boldsymbol{x}}_{t-1} - \boldsymbol{x}_{t-1}\|^2\right) + 3\ell\eta_x^2(G^2+\sigma^2) \\ &\quad + \eta_x\ell\left(\frac{1}{2\ell}\mathbb{E}\|\nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_x f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2})\|^2 + \frac{\ell}{2}\mathbb{E}\|\boldsymbol{x}_{t-1} - \hat{\boldsymbol{x}}_{t-1}\|^2\right) \\ &\le \mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{x}_{t-1})] + 2\eta_x\ell\mathbb{E}\left(\Phi(\boldsymbol{x}_{t-1}) - f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\right) - \frac{\eta_x\ell^2}{2}\mathbb{E}\|\hat{\boldsymbol{x}}_{t-1} - \boldsymbol{x}_{t-1}\|^2 \\ &\quad + 3\ell\eta_x^2(G^2+\sigma^2) + \frac{\eta_x}{2}\mathbb{E}\|\nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_x f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2})\|^2.\end{aligned}$$

□

Lemma B.9. *For Stochastic OGDA (Algorithm 2), under Theorem 4.9's assumptions, the following statement holds for the generated sequence $\{\boldsymbol{y}_t\}$ during algorithm proceeding:*

$$\begin{aligned}\sum_{t=0}^{T}\mathbb{E}\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 &\le 4\eta_y^2\ell\sum_{t=0}^{T}\left(\sum_{j=0}^{T}\left(2\eta_y^2\ell^2\right)^j\right)\mathbb{E}[\Phi(\boldsymbol{x}_t) - f(\boldsymbol{x}_t, \boldsymbol{y}_t)] \\ &\quad + \sum_{t=0}^{T}\left(\sum_{j=0}^{T}\left(2\eta_y^2\ell^2\right)^j\right)\left(6\eta_x^2\eta_y^2\ell^2(G^2+\sigma^2) + 6\eta_y^2\sigma^2\right)\end{aligned}$$

Proof. According to updating rule of stochastic OGDA:

$$\begin{aligned}&\mathbb{E}\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 \\ &\le \eta_y^2\mathbb{E}\|2\nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}; \xi_{t-1}) - f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2}; \xi_{t-1})\|^2 \\ &\le 2\eta_y^2\mathbb{E}\|\nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\|^2 + 2\eta_y^2\sigma^2 + 2\eta_y^2\mathbb{E}\|\nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2})\|^2 + 4\eta_y^2\sigma^2 \\ &\le 4\eta_y^2\ell\mathbb{E}[\Phi(\boldsymbol{x}_{t-1}) - f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})] + 2\eta_y^2\ell^2(\mathbb{E}\|\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-2}\|^2 + \mathbb{E}\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2) + 6\eta_y^2\sigma^2 \\ &\le 4\eta_y^2\ell\mathbb{E}[\Phi(\boldsymbol{x}_{t-1}) - f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})] + 2\eta_y^2\ell^2(3\eta_x^2(G^2+\sigma^2) + \mathbb{E}\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2) + 6\eta_y^2\sigma^2.\end{aligned}$$

Unrolling the recursion yields:

$$\begin{aligned}\mathbb{E}\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 \leq{}& 4\eta_y^2\ell \sum_{j=0}^{t-1} \left(2\eta_y^2\ell^2\right)^{t-1-j} \mathbb{E}[\Phi(x_j) - f(x_j, y_j)] \\ &+ \sum_{j=0}^{t-1} \left(2\eta_y^2\ell^2\right)^{t-1-j} \left(6\eta_x^2\eta_y^2\ell^2(G^2+\sigma^2) + 6\eta_y^2\sigma^2\right) + (2\eta_y^2\ell^2)\mathbb{E}\|\boldsymbol{y}_0 - \boldsymbol{y}_{-1}\|^2.\end{aligned}$$

Since $\boldsymbol{y}_0 = \boldsymbol{y}_{-1}$, we can conclude the proof via summing t from 0 to $T-1$:

$$\begin{aligned}\sum_{t=0}^{T} \mathbb{E}\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 \leq{}& 4\eta_y^2\ell \sum_{t=0}^{T} \left(\sum_{j=0}^{T} \left(2\eta_y^2\ell^2\right)^{j} \right) \mathbb{E}[\Phi(x_t) - f(x_t, y_t)] \\ &+ \sum_{t=0}^{T} \left(\sum_{j=0}^{T} \left(2\eta_y^2\ell^2\right)^{j} \right) \left(6\eta_x^2\eta_y^2\ell^2(G^2+\sigma^2) + 6\eta_y^2\sigma^2\right).\end{aligned}$$

□

Lemma B.10. *For Stochastic OGDA (Algorithm 2), under assumptions made in Theorem 4.9, the following statement holds for the generated sequence $\{\boldsymbol{y}_t\}$ during algorithm proceeding and $\forall s \leq t$:*

$$\begin{aligned}\mathbb{E}[\Phi(\boldsymbol{x}_t) - f(\boldsymbol{x}_t, \boldsymbol{y}_t)] \leq{}& 2\eta_x(t-s)G\sqrt{G^2+\sigma^2} + \frac{\eta_y\eta_x^2\ell}{2}(G^2+\sigma^2) + 3\eta_y\sigma^2 \\ &+ \frac{1}{2\eta_y}\left(\mathbb{E}\|\boldsymbol{y}_{t-1} - \boldsymbol{y}^*(\boldsymbol{x}_s)\|^2 - \mathbb{E}\|\boldsymbol{y}_t - \boldsymbol{y}^*(\boldsymbol{x}_s)\|^2 - \frac{1}{4}\mathbb{E}\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 + \frac{1}{4}\mathbb{E}\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2 \right) \\ &+ \langle \nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}), \boldsymbol{y}_t - \boldsymbol{y}^*(\boldsymbol{x}_s)\rangle \\ &- \langle \nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_y f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2}), \boldsymbol{y}_{t-1} - \boldsymbol{y}^*(\boldsymbol{x}_s)\rangle.\end{aligned}$$

Proof. Observe that:

$$\begin{aligned}\mathbb{E}[\Phi(\boldsymbol{x}_t) - f(\boldsymbol{x}_t, \boldsymbol{y}_t)] &\leq \mathbb{E}[f(\boldsymbol{x}_t, \boldsymbol{y}^*(\boldsymbol{x}_t)) - f(\boldsymbol{x}_s, \boldsymbol{y}^*(\boldsymbol{x}_t))] + \mathbb{E}[f(\boldsymbol{x}_s, \boldsymbol{y}^*(\boldsymbol{x}_s)) - f(\boldsymbol{x}_t, \boldsymbol{y}^*(\boldsymbol{x}_s))] \\ &\quad + \mathbb{E}[f(\boldsymbol{x}_t, \boldsymbol{y}^*(\boldsymbol{x}_s)) - f(\boldsymbol{x}_t, \boldsymbol{y}_t)] \\ &\leq 2(t-s)\eta_x G\sqrt{G^2+\sigma^2} - \mathbb{E}\langle \boldsymbol{y}_t - \boldsymbol{y}, \nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t)\rangle.\end{aligned}$$

Plugging in Lemma B.7 will conclude the proof. □

Lemma B.11. *For Stochastic OGDA (Algorithm 2), under Theorem 4.9's assumptions, the following statement holds for the generated sequence $\{\boldsymbol{x}_t\}, \{\boldsymbol{y}_t\}$ during algorithm proceeding:*

$$\begin{aligned}\frac{1}{T+1}\sum_{t=0}^{T} \mathbb{E}[\Phi(\boldsymbol{x}_t) - f(\boldsymbol{x}_t, \boldsymbol{y}_t)] \leq{}& \frac{1}{B}\left(2\eta_x B^2 G\sqrt{G^2+\sigma^2} + \frac{5D^2}{8\eta_y} + 2(3\eta_x G\sqrt{G^2+\sigma^2} + D)D \right) \\ &+ \frac{\eta_y\eta_x^2\ell}{2}(G^2+\sigma^2) + 3\eta_y\sigma^2.\end{aligned}$$

Proof. Let $S = (T+1)/B$, and we choose $s = jB$, $j = 0, ..., S$. Then by summing over t on the both side of Lemma B.11 we have:

$$
\begin{aligned}
&\frac{1}{T+1}\sum_{t=0}^{T-1}\mathbb{E}[\Phi(\boldsymbol{x}_t) - f(\boldsymbol{x}_t, \boldsymbol{y}_t)] = \frac{1}{T}\sum_{j=0}^{S}\sum_{t=jB}^{(j+1)B-1}\mathbb{E}[\Phi(\boldsymbol{x}_t) - f(\boldsymbol{x}_t, \boldsymbol{y}_t)] \\
&\leq \frac{1}{T}\sum_{j=0}^{S}\left[2\eta_x B^2 G\sqrt{G^2+\sigma^2} + \frac{1}{2\eta_y}\left(\|y_{jB} - y^*(x_{jB})\|^2 + \frac{1}{4}\|y_{jB} - y_{jB-1}\|^2\right)\right] \\
&\quad + \frac{\eta_y\eta_x^2\ell}{2}(G^2+\sigma^2) + 3\eta_y\sigma^2 \\
&\quad + \frac{1}{T}\sum_{j=0}^{S}\Big[-\langle\nabla_y f(\boldsymbol{x}_{(j+1)B-1}, y_{(j+1)B-1}) - \nabla_y f(\boldsymbol{x}_{(j+1)B-2}, y_{(j+1)B-2}), \boldsymbol{y}_{(j+1)B-1} - y^*(\boldsymbol{x}_{jB})\rangle \\
&\qquad\qquad + \langle\nabla_y f(\boldsymbol{x}_{jB-1}, \boldsymbol{y}_{jB-1}) - \nabla_y f(\boldsymbol{x}_{jB-2}, \boldsymbol{y}_{jB-2}), \boldsymbol{y}_{jB-1} - y^*(\boldsymbol{x}_{jB}\rangle\Big] \\
&\leq \frac{1}{T}\sum_{j=0}^{S}\left[2\eta_x B^2 G\sqrt{G^2+\sigma^2} + \frac{5D^2}{8\eta_y} + 2(3\eta_x G\sqrt{G^2+\sigma^2} + D)D\right] \\
&\quad + \frac{\eta_y\eta_x^2\ell}{2}(G^2+\sigma^2) + 3\eta_y\sigma^2 \\
&\leq \frac{1}{B}\left[2\eta_x B^2 G\sqrt{G^2+\sigma^2} + \frac{5D^2}{8\eta_y} + 2(3\eta_x G\sqrt{G^2+\sigma^2} + D)D\right] \\
&\quad + \frac{\eta_y\eta_x^2\ell}{2}(G^2+\sigma^2) + 3\eta_y\sigma^2.
\end{aligned}
$$

□

B.1.4 Proof of Theorem 4.9 for OGDA

In this section we are going to provide the proof for Theorem 4.9, the convergence rate of OGDA in stochastic setting. We first introduce the formal version of Theorem 4.9 here:

Theorem B.12 (OGDA Stochastic (Theorem 4.9 restated))**.** *Under Assumption 4.3 and 4.7, if we choose* $\eta_x = O(\min\{\frac{\epsilon^2}{\ell(G^2+\sigma^2)}, \frac{\epsilon^4}{D^2\ell^3 G\sqrt{G^2+\sigma^2}}, \frac{\epsilon^6}{D^2\ell^3\sigma^2 G\sqrt{G^2+\sigma^2}}\})$, $\eta_y = O(\min\{\frac{1}{4\ell}, \frac{\epsilon^2}{\ell\sigma^2}\})$, *then Stochastic OGDA (Algorithm 2) guarantees to find* ϵ*-stationary point, i.e.,* $\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_t)\|^2 \leq \epsilon^2$, *with the gradient complexity bounded by:*

$$
O\left(\frac{D^2\ell^3 G\sqrt{G^2+\sigma^2}}{\epsilon^6}\max\left\{1, \frac{\sigma^2}{\epsilon^2}\right\}\right).
$$

Proof. Similar to the proof in deterministic setting, first according to Lemma B.8 we have:

$$
\begin{aligned}
\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_t)\|^2 \leq\ & \frac{\Phi_{1/2\ell}(x_0) - \Phi_{1/2\ell}(\boldsymbol{x}_{T+1})}{\eta_x(T+1)} \\
& + 16\ell\frac{1}{T+1}\sum_{t=0}^{T}(\Phi(x_t) - f(\boldsymbol{x}_t, \boldsymbol{y}_t)) + 12\eta_x^2\ell^2(G^2+\sigma^2) + 24\ell\eta_x(G^2+\sigma^2) \\
& + 4\ell^2\frac{1}{T+1}\left(4\eta_y^2\ell\sum_{t=0}^{T+1}\left(\sum_{j=0}^{T}(2\eta_y^2\ell^2)^j\right)\mathbb{E}[\Phi(\boldsymbol{x}_t) - f(\boldsymbol{x}_t, \boldsymbol{y}_t)]\right. \\
& + \left.\sum_{t=0}^{T}\left(\sum_{j=0}^{T}(2\eta_y^2\ell^2)^j\right)\left(6\eta_x^2\eta_y^2\ell^2(G^2+\sigma^2) + 6\eta_y^2\sigma^2\right)\right..
\end{aligned}
$$

Since we choose $\eta_y \ell \leq \frac{1}{4}$, it follows that:

$$\sum_{j=0}^{T} \left(2\eta_y^2 \ell^2\right)^j \leq 2.$$

As a result, we can further simplify the bound as:

$$\begin{aligned}\frac{1}{T+1}\sum_{t=0}^{T} \mathbb{E}\|\nabla \Phi_{1/2\ell}(\boldsymbol{x}_t)\|^2 \leq{} & \frac{\Phi_{1/2\ell}(x_0) - \Phi_{1/2\ell}(\boldsymbol{x}_{T+1})}{\eta_x (T+1)} \\ & + (16\ell + 32\eta_y^2\ell^3)\frac{1}{T+1}\sum_{t=0}^{T}\left(\Phi(x_t) - f(\boldsymbol{x}_t, \boldsymbol{y}_t)\right) \\ & + 12\eta_x^2\ell^2(G^2+\sigma^2) + 24\ell\eta_x(G^2+\sigma^2) + 8\ell^2\left(6\eta_x^2\eta_y^2\ell^2(G^2+\sigma^2) + 6\eta_y^2\sigma^2\right).\end{aligned}$$

Plugging in Lemma B.11 yields:

$$\begin{aligned}\frac{1}{T+1}\sum_{t=0}^{T} \mathbb{E}\|\nabla \Phi_{1/2\ell}(\boldsymbol{x}_t)\|^2 \leq{} & \frac{\Phi_{1/2\ell}(x_0) - \Phi_{1/2\ell}(\boldsymbol{x}_{T+1})}{\eta_x (T+1)} \\ & + (16\ell + 32\eta_y^2\ell^3)\frac{1}{B}\left(2\eta_x B^2 G\sqrt{G^2+\sigma^2} + \frac{5D^2}{8\eta_y} + 2(3\eta_x G\sqrt{G^2+\sigma^2} + D)D\right) \\ & + (16\ell + 32\eta_y^2\ell^3)\left(\frac{\eta_y\eta_x^2\ell}{2}(G^2+\sigma^2) + 3\eta_y\sigma^2\right) \\ & + 12\eta_x^2\ell^2(G^2+\sigma^2) + 24\ell\eta_x(G^2+\sigma^2) + 8\ell^2\left(6\eta_x^2\eta_y^2\ell^2(G^2+\sigma^2) + 6\eta_y^2\sigma^2\right).\end{aligned}$$

Choose $B = O(\frac{D}{\sqrt{\eta_x\eta_y G\sqrt{G^2+\sigma^2}}})$, $\eta_x = O(\min\{\frac{\epsilon^2}{\ell(G^2+\sigma^2)}, \frac{\epsilon^4}{D^2\ell^3 G\sqrt{G^2+\sigma^2}}, \frac{\epsilon^6}{D^2\ell^3\sigma^2 G\sqrt{G^2+\sigma^2}}\})$, $\eta_y = O(\min\{\frac{1}{4\ell}, \frac{\epsilon^2}{\ell\sigma^2}\})$, and then it is guaranteed that $\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_t)\|^2 \leq \epsilon^2$ with the gradient complexity is bounded by

$$O\left(\frac{D^2\ell^3 G\sqrt{G^2+\sigma^2}}{\epsilon^6}\max\left\{1, \frac{\sigma^2}{\epsilon^2}\right\}\right).$$

□

B.2 Proof of convergence of EG

In this section, the convergence of EG in NC-C setting has been established. Before presenting the complete proofs, here we briefly discuss the proof sketch.

Proof sketch Similar to OGDA, we have the following lemma on $\Phi_{1/2\ell}$:

$$\begin{aligned}\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_{t-\frac{1}{2}})\|^2 \leq{} & \Phi_{1/2\ell}(\boldsymbol{x}_{-\frac{1}{2}}) - \Phi_{1/2\ell}(\boldsymbol{x}_{T+\frac{1}{2}}) \\ & + O(\ell + \eta_y^2\ell^3)\frac{1}{T+1}\sum_{t=0}^{T}\delta_{t-\frac{1}{2}} + O(\ell\eta_x^2 G^2).\end{aligned}$$

Now we need to examine $\delta_{t-\frac{1}{2}}$. To bound this term, we have the following recursion:

$$\begin{aligned}\Phi(\boldsymbol{x}_{t+\frac{1}{2}}) - f(\boldsymbol{x}_{t+\frac{1}{2}}, \boldsymbol{y}_{t+\frac{1}{2}}) \leq{} & O((t-s)\eta_x G^2) \\ & + \frac{1}{2\eta_y}\left(\|\boldsymbol{y}_t - \boldsymbol{y}^*(\boldsymbol{x}_s)\|^2 - \|\boldsymbol{y}_{t+1} - \boldsymbol{y}^*(\boldsymbol{x}_s)\|^2 + \frac{\eta_x^2 G^2}{2}\right),\end{aligned}$$

which is derived by the descent property of EG on concave function. Similar to OGDA, here we also obtain neat recursion, which will yield our desired complexity bound.

In the following, we present the key lemmas, and complete convergence proof of EG. First let us introduce some useful lemmas for the deterministic setting.

B.2.1 Useful Lemmas

Proposition B.13 ([5], Proposition 4.2)**.** *If* $\boldsymbol{p} = \mathcal{P}_{\mathcal{Y}}(\boldsymbol{r} - \boldsymbol{u})$, $\boldsymbol{q} = \mathcal{P}_{\mathcal{Y}}(\boldsymbol{r} - \boldsymbol{v})$, *and*

$$\|\boldsymbol{u} - \boldsymbol{v}\|^2 \leq C_1^2 \|\boldsymbol{p} - \boldsymbol{r}\|^2 + C_2^2,$$

then for any $\boldsymbol{z} \in \mathbb{R}^d$ *we have:*

$$\langle \boldsymbol{v}, \boldsymbol{p} - \boldsymbol{z} \rangle \leq \|\boldsymbol{r} - \boldsymbol{z}\|^2 - \|\boldsymbol{q} - \boldsymbol{z}\|^2 - \left(\frac{1}{2} - \frac{C_1^2}{2}\right) \|\boldsymbol{r} - \boldsymbol{p}\|^2 + \frac{C_2^2}{2}.$$

Lemma B.14. *For EG (Algorithm 3), under Theorem 4.9's assumptions, the following statement holds for the generated sequence* $\{\boldsymbol{y}_t\}, \{\boldsymbol{y}_{t+\frac{1}{2}}\}$ *during algorithm proceeding and any* $\boldsymbol{y} \in \mathcal{Y}$*:*

$$\begin{aligned}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}\|^2 \leq \|\boldsymbol{y}_t - \boldsymbol{y}\|^2 + 2\eta_y \langle \boldsymbol{y}_{t+\frac{1}{2}} - \boldsymbol{y}, \nabla_y f(\boldsymbol{x}_{t+\frac{1}{2}}, \boldsymbol{y}_{t+\frac{1}{2}}) \rangle - \left(\frac{1}{2} - \frac{\eta_y^2 \ell^2}{2}\right) \|\boldsymbol{y}_t - \boldsymbol{y}_{t+\frac{1}{2}}\|^2 \\ + \frac{\eta_x^2 \eta_y^2 \ell^2 G^2}{2}.\end{aligned}$$

Proof. According to Proposition B.13, we set $\boldsymbol{r} = \boldsymbol{y}_t$, $\boldsymbol{q} = \boldsymbol{y}_{t+1}$, $\boldsymbol{p} = \boldsymbol{y}_{t+\frac{1}{2}}$ and $\boldsymbol{v} = -\eta_y \nabla_y f(\boldsymbol{x}_{t+\frac{1}{2}}, \boldsymbol{y}_{t+\frac{1}{2}})$, $\boldsymbol{u} = -\eta_y \nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t)$. We can verify that:

$$\begin{aligned}\|\boldsymbol{u} - \boldsymbol{v}\|^2 &= \eta_y^2 \|\nabla_y f(\boldsymbol{x}_{t+\frac{1}{2}}, \boldsymbol{y}_{t+\frac{1}{2}}) - \nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t)\|^2 \\ &\leq \eta_y^2 (\ell^2 \|\boldsymbol{y}_{t+\frac{1}{2}} - \boldsymbol{y}_t\|^2 + \ell^2 \|\boldsymbol{x}_{t+\frac{1}{2}} - \boldsymbol{x}_t\|^2) \\ &\leq \eta_y^2 (\ell^2 \|\boldsymbol{p} - \boldsymbol{r}\|^2 + \ell^2 \eta_x^2 G^2),\end{aligned}$$

so if we set $C_1^2 = \eta_y^2 \ell^2$ and $C_2^2 = \eta_x^2 \eta_y^2 \ell^2 G^2$, we have the following inequality holding for any $\boldsymbol{y} \in \mathcal{Y}$:

$$\begin{aligned}\langle -\eta_y \nabla_y f(\boldsymbol{x}_{t+\frac{1}{2}}, \boldsymbol{y}_{t+\frac{1}{2}}), \boldsymbol{y}_{t+\frac{1}{2}} - \boldsymbol{y} \rangle \leq \|\boldsymbol{y}_t - \boldsymbol{y}\|^2 - \|\boldsymbol{y}_{t+1} - \boldsymbol{y}\|^2 - \left(\frac{1}{2} - \frac{\eta_y^2 \ell^2}{2}\right) \|\boldsymbol{y}_t - \boldsymbol{y}_{t+\frac{1}{2}}\|^2 \\ + \frac{\eta_x^2 \eta_y^2 \ell^2 G^2}{2}.\end{aligned}$$

□

Lemma B.15. *For EG (Algorithm 3), under Theorem 4.9's assumptions, the following statement holds for the generated sequence* $\{\boldsymbol{x}_t\}, \{\boldsymbol{y}_t\}, \{\boldsymbol{x}_{t+\frac{1}{2}}\}, \{\boldsymbol{y}_{t+\frac{1}{2}}\}$ *during algorithm proceeding:*

$$\begin{aligned}\Phi_{1/2\ell}(\boldsymbol{x}_{t+\frac{1}{2}}) \leq \Phi_{1/2\ell}(\boldsymbol{x}_{t-\frac{1}{2}}) + 2\eta_x \ell \left(\Phi(\boldsymbol{x}_{t-\frac{1}{2}}) - f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}})\right) - \frac{\eta_x}{8} \|\nabla \Phi_{1/2\ell}(\boldsymbol{x}_{t-\frac{1}{2}})\|^2 + 3\ell \eta_x^2 G^2 \\ + \frac{\eta_x}{2} \|\nabla_x f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\|^2.\end{aligned}$$

Proof. Let $\hat{\boldsymbol{x}}_{t-\frac{1}{2}} = \arg\min_{\boldsymbol{x} \in \mathbb{R}^d} \Phi(\boldsymbol{x}) + \ell \|\boldsymbol{x} - \boldsymbol{x}_{t-\frac{1}{2}}\|^2$. Notice that:

$$\begin{aligned}\Phi_{1/2\ell}(\boldsymbol{x}_{t+\frac{1}{2}}) &\leq \Phi_{1/2\ell}(\hat{\boldsymbol{x}}_{t-\frac{1}{2}}) + \ell \|\hat{\boldsymbol{x}}_{t-\frac{1}{2}} - \boldsymbol{x}_{t+\frac{1}{2}}\|^2 \\ &\leq \Phi_{1/2\ell}(\hat{\boldsymbol{x}}_{t-\frac{1}{2}}) + \ell \|\hat{\boldsymbol{x}}_{t-\frac{1}{2}} - \boldsymbol{x}_{t+\frac{1}{2}}\|^2 \\ &\quad + \ell(2\eta_x \langle \nabla_x f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}}) + (\nabla_x f(\boldsymbol{x}_t, \boldsymbol{y}_t - \nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}), \hat{\boldsymbol{x}}_{t-\frac{1}{2}} - \boldsymbol{x}_{t-\frac{1}{2}} \rangle + \eta_x^2 G^2) \\ &= \Phi_{1/2\ell}(\hat{\boldsymbol{x}}_{t-\frac{1}{2}}) + \ell(\|\hat{\boldsymbol{x}}_{t-\frac{1}{2}} - \boldsymbol{x}_{t+\frac{1}{2}}\|^2 + 2\eta_x \langle \nabla_x f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}}), \hat{\boldsymbol{x}}_{t-\frac{1}{2}} - \boldsymbol{x}_{t-\frac{1}{2}} \rangle) \\ &\quad + 2\ell \eta_x \langle \nabla_x f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}), \hat{\boldsymbol{x}}_{t-\frac{1}{2}} - \boldsymbol{x}_{t-\frac{1}{2}} \rangle + \eta_x^2 \ell G^2\end{aligned}$$

According to smoothness of $f(\cdot, \boldsymbol{y})$, we have:

$$\begin{aligned}\langle \hat{\boldsymbol{x}}_{t-\frac{1}{2}} - \boldsymbol{x}_{t-\frac{1}{2}}, \nabla_x f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}}) \rangle &\leq f(\hat{\boldsymbol{x}}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}}) - f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}}) + \frac{\ell}{2} \|\hat{\boldsymbol{x}}_{t-\frac{1}{2}} - \boldsymbol{x}_{t-\frac{1}{2}}\|^2 \\ &\leq \Phi(\boldsymbol{x}_{t-\frac{1}{2}}) - f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}}) - \frac{\ell}{2} \|\hat{\boldsymbol{x}}_{t-\frac{1}{2}} - \boldsymbol{x}_{t-\frac{1}{2}}\|^2.\end{aligned}$$

So we have

$$\begin{aligned}
\Phi_{1/2\ell}(\boldsymbol{x}_{t+\frac{1}{2}}) \leq{} & \Phi_{1/2\ell}(\boldsymbol{x}_{t-\frac{1}{2}}) + \ell\|\boldsymbol{x}_{t-\frac{1}{2}} - \hat{\boldsymbol{x}}_{t-\frac{1}{2}}\|^2 \\
& + 2\eta_x\ell\left(\Phi(\boldsymbol{x}_{t-\frac{1}{2}}) - f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}}) - \frac{\ell}{2}\|\hat{\boldsymbol{x}}_{t-\frac{1}{2}} - \boldsymbol{x}_{t-\frac{1}{2}}\|^2\right) + 3\ell\eta_x^2G^2 \\
& + \eta_x\ell\left(\frac{1}{2\ell}\|\nabla_x f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\|^2 + \frac{\ell}{2}\|\boldsymbol{x}_{t-\frac{1}{2}} - \hat{\boldsymbol{x}}_{t-\frac{1}{2}}\|^2\right) \\
\leq{} & \Phi_{1/2\ell}(\boldsymbol{x}_{t-\frac{1}{2}}) + 2\eta_x\ell\left(\Phi(\boldsymbol{x}_{t-\frac{1}{2}}) - f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}})\right) - \frac{\eta_x\ell^2}{2}\|\hat{\boldsymbol{x}}_{t-\frac{1}{2}} - \boldsymbol{x}_{t-\frac{1}{2}}\|^2 + 3\ell\eta_x^2G^2 \\
& + \frac{\eta_x}{2}\|\nabla_x f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\|^2.
\end{aligned}$$

$\square$

Lemma B.16. *For EG (Algorithm 3), under Theorem 4.9's assumptions, the following statement holds for the generated sequence $\{\boldsymbol{x}_t\}, \{\boldsymbol{y}_t\}, \{\boldsymbol{x}_{t+\frac{1}{2}}\}, \{\boldsymbol{y}_{t+\frac{1}{2}}\}$ during algorithm proceeding and $\forall s \leq t$:*

$$\begin{aligned}
\Phi(\boldsymbol{x}_{t+\frac{1}{2}}) - f(\boldsymbol{x}_{t+\frac{1}{2}}, \boldsymbol{y}_{t+\frac{1}{2}}) \leq{} & 2(t-s+1)\eta_x G^2 \\
& + \frac{1}{2\eta_y}\left(\|\boldsymbol{y}_t - \boldsymbol{y}^*(\boldsymbol{x}_s)\|^2 - \|\boldsymbol{y}_{t+1} - \boldsymbol{y}^*(\boldsymbol{x}_s)\|^2 + \frac{\eta_x^2G^2}{2}\right).
\end{aligned}$$

Proof. Observe that:

$$\begin{aligned}
\Phi(\boldsymbol{x}_{t+\frac{1}{2}}) - f(\boldsymbol{x}_{t+\frac{1}{2}}, \boldsymbol{y}_{t+\frac{1}{2}}) \leq{} & f(\boldsymbol{x}_{t+\frac{1}{2}}, \boldsymbol{y}^*(\boldsymbol{x}_{t+\frac{1}{2}})) - f(\boldsymbol{x}_s, \boldsymbol{y}^*(\boldsymbol{x}_{t+\frac{1}{2}})) + f(\boldsymbol{x}_s, \boldsymbol{y}^*(\boldsymbol{x}_s)) \\
& - f(\boldsymbol{x}_{t+\frac{1}{2}}, \boldsymbol{y}^*(\boldsymbol{x}_s)) + f(\boldsymbol{x}_{t+\frac{1}{2}}, \boldsymbol{y}^*(\boldsymbol{x}_s)) - f(\boldsymbol{x}_{t+\frac{1}{2}}, \boldsymbol{y}_{t+\frac{1}{2}}) \\
\leq{} & 2(t-s+1)\eta_x G^2 - \langle \boldsymbol{y}_{t+\frac{1}{2}} - \boldsymbol{y}, \nabla_y f(\boldsymbol{x}_{t+\frac{1}{2}}, \boldsymbol{y}_{t+\frac{1}{2}})\rangle
\end{aligned}$$

Plugging in Lemma B.14 will conclude the proof:

$$\begin{aligned}
\Phi(\boldsymbol{x}_{t+\frac{1}{2}}) - f(\boldsymbol{x}_{t+\frac{1}{2}}, \boldsymbol{y}_{t+\frac{1}{2}}) \leq{} & 2(t-s+1)\eta_x G^2 \\
& + \frac{1}{2\eta_y}\left(\|\boldsymbol{y}_t - \boldsymbol{y}^*(\boldsymbol{x}_s)\|^2 - \|\boldsymbol{y}_{t+1} - \boldsymbol{y}^*(\boldsymbol{x}_s)\|^2 + \frac{\eta_x^2G^2}{2}\right).
\end{aligned}$$

$\square$

Lemma B.17. *For EG (Algorithm 3), under Theorem 4.9's assumptions, the following statement holds for the generated sequence $\{\boldsymbol{x}_t\}, \{\boldsymbol{y}_t\}, \{\boldsymbol{x}_{t+\frac{1}{2}}\}, \{\boldsymbol{y}_{t+\frac{1}{2}}\}$ during algorithm proceeding:*

$$\frac{1}{T+1}\sum_{t=0}^{T}\Phi(\boldsymbol{x}_{t-\frac{1}{2}}) - f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}}) \leq \frac{1}{B}\left(2\eta_x B^2G^2 + \frac{D^2}{2\eta_y} + \frac{B\eta_x^2G^2}{2}\right)$$

Proof. According to Lemma B.16:

$$\begin{aligned}
& \frac{1}{T+1}\sum_{t=0}^{T}\Phi(\boldsymbol{x}_{t-\frac{1}{2}}) - f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}}) \\
={} & \frac{1}{T+1}\sum_{j=0}^{S}\sum_{t=kB}^{(k+1)B-1}\Phi(\boldsymbol{x}_{t-\frac{1}{2}}) - f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}}) \\
\leq{} & \frac{1}{T+1}\sum_{j=0}^{S}\left[2B^2\eta_x G^2 + \frac{1}{2\eta_y}\left(\|\boldsymbol{y}_{kB} - \boldsymbol{y}^*(\boldsymbol{x}_s)\|^2 - \|\boldsymbol{y}_{(k+1)B-1} - \boldsymbol{y}^*(\boldsymbol{x}_s)\|^2 + \frac{\eta_x^2G^2}{2}\right)\right] \\
\leq{} & \frac{1}{B}\left[2\eta_x B^2G^2 + \frac{D^2}{2\eta_y} + \frac{B\eta_x^2G^2}{2}\right].
\end{aligned}$$

$\square$

B.2.2 Proof of Theorem 4.8 for EG

In this section we are going to provide the proof for Theorem 4.8, EG part, the convergence rate of EG in deterministic setting. We first introduce the formal version of Theorem 4.8, EG part here:

Theorem B.18 (EG Deterministic, formal)**.** *Under Assumption 4.7, if we choose* $\eta_x = O\left(\min\left\{\frac{\epsilon}{\ell G}, \frac{\epsilon^2}{\ell G^2}, \frac{\epsilon^4}{D^2G^2\ell^3}\right\}\right)$, $\eta_y = \frac{1}{2\ell}$, *then EG (Algorithm 3) guarantees to find* ϵ*-stationary point, i.e.,* $\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_t)\|^2 \leq \epsilon^2$, *with the gradient complexity bounded by:*

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

Proof. According to Lemma B.15:

$$\begin{aligned}\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_{t-\frac{1}{2}})\|^2 \leq \frac{\Phi_{1/2\ell}(\boldsymbol{x}_{-\frac{1}{2}})}{\eta_x(T+1)} + \frac{1}{T+1}\sum_{t=0}^{T}8\ell\left(\Phi(x_{t-\frac{1}{2}}) - f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}})\right) + 12\eta_x\ell G^2 \\ + 8\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla_x f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\|^2.\end{aligned}$$

For $\|\nabla_x f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\|^2$, notice that:

$$\begin{aligned}\|\nabla_x f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\|^2 &\leq \ell^2\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|^2 + \ell^2\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 \\ &\leq \eta_x^2\ell^2G^2 + \eta_y^2\ell^2\|\nabla_y f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}})\|^2 \\ &\leq \eta_x^2\ell^2G^2 + 2\eta_y^2\ell^3\left(\Phi(\boldsymbol{x}_{t-\frac{1}{2}}) - f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}})\right)\end{aligned}$$

So we have:

$$\begin{aligned}\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_{t-\frac{1}{2}})\|^2 &\leq \frac{\Phi_{1/2\ell}(\boldsymbol{x}_{-\frac{1}{2}})}{\eta_x(T+1)} \\ &+ \frac{1}{T+1}\sum_{t=0}^{T}(8\ell + 2\eta_y^2\ell^3)\left(\Phi(x_{t-\frac{1}{2}}) - f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}})\right) + 12\eta_x G^2 + 8\eta_x^2\ell^2G^2\end{aligned}$$

Now we plug in Lemma B.17:

$$\begin{aligned}\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_{t-\frac{1}{2}})\|^2 &\leq \frac{\Phi_{1/2\ell}(\boldsymbol{x}_{-\frac{1}{2}}) - \Phi_{1/2\ell}(\boldsymbol{x}_{T-\frac{1}{2}})}{\eta_x(T+1)} \\ &+ (8\ell + 2\eta_y^2\ell^3)\left(2\eta_x BG^2 + \frac{D^2}{2\eta_y B} + \frac{\eta_x^2G^2}{2}\right) + 12\ell\eta_x G^2 + 8\eta_x^2\ell^2G^2\end{aligned}$$

Choose $B = O\left(\frac{D}{G\sqrt{\eta_x\eta_y}}\right)$, $\eta_x = O\left(\min\left\{\frac{\epsilon}{\ell G}, \frac{\epsilon^2}{\ell G^2}, \frac{\epsilon^4}{D^2G^2\ell^3}\right\}\right)$, $\eta_y = \frac{1}{2\ell}$, and then we guarantee that $\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_{t-\frac{1}{2}})\|^2 \leq \epsilon^2$ with the gradient complexity is bounded by:

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

□

Stochastic setting.

In this part, we are going to present proof of EG in stochastic setting. First let us introduce some useful lemmas.

B.2.3 Useful Lemmas

Lemma B.19. *For Stochastic EG (Algorithm 3), under Theorem 4.9's assumptions, the following statement holds for the generated sequence $\{\boldsymbol{y}_t\}, \{\boldsymbol{y}_{t+\frac{1}{2}}\}$ during algorithm proceeding and any $\boldsymbol{y} \in \mathcal{Y}$:*

$$\|\boldsymbol{y}_{t+1}-\boldsymbol{y}\|^2 \le \|\boldsymbol{y}_t-\boldsymbol{y}\|^2 + 2\eta_y\langle \boldsymbol{y}_{t+\frac{1}{2}}-\boldsymbol{y}, \nabla_y f(\boldsymbol{x}_{t+\frac{1}{2}},\boldsymbol{y}_{t+\frac{1}{2}})\rangle - \left(\frac{1}{2} - \frac{3\eta_x^2L^2}{2}\right)\|\boldsymbol{y}_t-\boldsymbol{y}_{t+\frac{1}{2}}\|^2 \\ + \frac{1}{2}(3\eta_x^2\eta_y^2\ell^2(G^2+\sigma^2)+6\eta_y^2\sigma^2).$$

Proof. According to Proposition B.13, we set $\boldsymbol{r} = \boldsymbol{y}_t$, $\boldsymbol{q} = \boldsymbol{y}_{t+1}$, $\boldsymbol{p} = \boldsymbol{y}_{t+\frac{1}{2}}$ and $\boldsymbol{v} = -\eta_y\nabla_y f(\boldsymbol{x}_{t+\frac{1}{2}},\boldsymbol{y}_{t+\frac{1}{2}};\xi)$, $\boldsymbol{u} = -\eta_y\nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t;\xi)$. We can verify that:

$$\begin{aligned}
&\|\boldsymbol{u}-\boldsymbol{v}\|^2 = \eta_y^2\|\nabla_y f(\boldsymbol{x}_{t+\frac{1}{2}},\boldsymbol{y}_{t+\frac{1}{2}};\xi) - \nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t;\xi)\|^2\\
&\le 3\eta_y^2\|\nabla_y f(\boldsymbol{x}_{t+\frac{1}{2}},\boldsymbol{y}_{t+\frac{1}{2}}) - \nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t)\|^2 + 3\eta_y^2\|\nabla_y f(\boldsymbol{x}_{t+\frac{1}{2}},\boldsymbol{y}_{t+\frac{1}{2}};\xi) - \nabla_y f(\boldsymbol{x}_{t+\frac{1}{2}},\boldsymbol{y}_{t+\frac{1}{2}})\|^2\\
&\quad + 3\eta_y^2\|\nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t;\xi) - \nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t)\|^2\\
&\le 3(\ell^2\|\boldsymbol{y}_{t+\frac{1}{2}}-\boldsymbol{y}_t\|^2 + \ell^2\|\boldsymbol{x}_{t+\frac{1}{2}}-\boldsymbol{x}_t\|^2) + 3\eta_y^2 Var(\nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t;\xi))\\
&\quad + 3\eta_y^2 Var(\nabla_y f(\boldsymbol{x}_{t+\frac{1}{2}},\boldsymbol{y}_{t+\frac{1}{2}};\xi))\\
&\le 3\eta_y^2(\ell^2\|\boldsymbol{y}_{t+\frac{1}{2}}-\boldsymbol{y}_t\|^2 + \eta_x^2\ell^2(G^2+\sigma^2)) + 3\eta_y^2 Var(\nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t;\xi))\\
&\quad + 3\eta_y^2 Var(\nabla_y f(\boldsymbol{x}_{t+\frac{1}{2}},\boldsymbol{y}_{t+\frac{1}{2}};\xi))
\end{aligned}$$

so if we set $C_1^2 = 3\eta_y^2\ell^2$ and $C_2^2 = 3\eta_x^2\eta_y^2\ell^2(G^2+\sigma^2) + 3\eta_y^2 Var(\nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t;\xi)) + 3\eta_y^2 Var(\nabla_y f(\boldsymbol{x}_{t+\frac{1}{2}},\boldsymbol{y}_{t+\frac{1}{2}};\xi))$, we have the following inequality holding for any $\boldsymbol{y} \in \mathcal{Y}$:

$$\langle -\eta_y\nabla_y f(\boldsymbol{x}_{t+\frac{1}{2}},\boldsymbol{y}_{t+\frac{1}{2}};\xi), \boldsymbol{y}_{t+\frac{1}{2}}-\boldsymbol{y}\rangle \le \|\boldsymbol{y}_t-\boldsymbol{y}\|^2 - \|\boldsymbol{y}_{t+1}-\boldsymbol{y}\|^2 \\ - \left(\frac{1}{2}-\frac{C_1^2}{2}\right)\|\boldsymbol{y}_t-\boldsymbol{y}_{t+\frac{1}{2}}\|^2 + \frac{C_2^2}{2}.$$

Taking expectation on both sides yields:

$$\langle -\eta_y\nabla_y f(\boldsymbol{x}_{t+\frac{1}{2}},\boldsymbol{y}_{t+\frac{1}{2}}), \boldsymbol{y}_{t+\frac{1}{2}}-\boldsymbol{y}\rangle \le \mathbb{E}\|\boldsymbol{y}_t-\boldsymbol{y}\|^2 - \mathbb{E}\|\boldsymbol{y}_{t+1}-\boldsymbol{y}\|^2 \\ - \left(\frac{1}{2}-\frac{3\eta_y^2\ell^2}{2}\right)\mathbb{E}\|\boldsymbol{y}_t-\boldsymbol{y}_{t+\frac{1}{2}}\|^2 + \frac{1}{2}(3\eta_x^2\eta_y^2\ell^2(G^2+\sigma^2)+6\eta_y^2\sigma^2).$$

□

Lemma B.20. *For Stochastic EG (Algorithm 3), under Theorem 4.9's assumptions, the following statement holds for the generated sequence $\{\boldsymbol{x}_t\}, \{\boldsymbol{y}_t\}, \{\boldsymbol{x}_{t+\frac{1}{2}}\}, \{\boldsymbol{y}_{t+\frac{1}{2}}\}$ during algorithm proceeding:*

$$\mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{x}_{t+\frac{1}{2}})] \le \mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{x}_{t-\frac{1}{2}})] + 2\eta\ell\mathbb{E}[\Phi(x_{t-\frac{1}{2}}) - f(\boldsymbol{x}_{t-\frac{1}{2}},\boldsymbol{y}_{t-\frac{1}{2}})] - \frac{\eta_x}{8}\mathbb{E}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_{t-\frac{1}{2}})\|^2 \\ + 3\eta_x^2\ell(G^2+\sigma^2) + 2\eta_x\mathbb{E}\|\nabla_x f(\boldsymbol{x}_t,\boldsymbol{y}_t) - \nabla_x f(\boldsymbol{x}_{t-1},\boldsymbol{y}_{t-1})\|^2.$$

Proof. Let $\hat{\boldsymbol{x}}_{t-\frac{1}{2}} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} \Phi(\boldsymbol{x}) + \ell\|\boldsymbol{x}-\boldsymbol{x}_{t-\frac{1}{2}}\|^2$. Notice that:

$$\begin{aligned}
&\mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{x}_{t+\frac{1}{2}})] \le \mathbb{E}[\Phi(\hat{\boldsymbol{x}}_{t-\frac{1}{2}})] + \ell\mathbb{E}\|\hat{\boldsymbol{x}}_{t-\frac{1}{2}} - \boldsymbol{x}_{t+\frac{1}{2}}\|^2\\
&\le \Phi_{1/2\ell}(\hat{\boldsymbol{x}}_{t-\frac{1}{2}}) + 3\eta_x^2\ell(\sigma^2+G^2) + \ell\mathbb{E}\|\hat{\boldsymbol{x}}_{t-\frac{1}{2}} - \boldsymbol{x}_{t+\frac{1}{2}}\|^2\\
&\quad + 2\eta_x\ell\mathbb{E}\langle \nabla_x f(\boldsymbol{x}_{t-\frac{1}{2}},\boldsymbol{y}_{t-\frac{1}{2}}) + \nabla_x f(\boldsymbol{x}_t,\boldsymbol{y}_t) - \nabla_x f(\boldsymbol{x}_{t-1},\boldsymbol{y}_{t-1}), \hat{\boldsymbol{x}}_{t-\frac{1}{2}} - \boldsymbol{x}_{t-\frac{1}{2}}\rangle)\\
&= \mathbb{E}[\Phi(\hat{\boldsymbol{x}}_{t-\frac{1}{2}})] + \ell(\mathbb{E}\|\hat{\boldsymbol{x}}_{t-\frac{1}{2}} - \boldsymbol{x}_{t+\frac{1}{2}}\|^2 + 2\eta_x\mathbb{E}\langle \nabla_x f(\boldsymbol{x}_{t-\frac{1}{2}},\boldsymbol{y}_{t-\frac{1}{2}}), \hat{\boldsymbol{x}}_{t-\frac{1}{2}} - \boldsymbol{x}_{t-\frac{1}{2}}\rangle)\\
&\quad + 2\ell\eta_x\mathbb{E}\langle \nabla_x f(\boldsymbol{x}_t,\boldsymbol{y}_t) - \nabla_x f(\boldsymbol{x}_{t-1},\boldsymbol{y}_{t-1}), \hat{\boldsymbol{x}}_{t-\frac{1}{2}} - \boldsymbol{x}_{t-\frac{1}{2}}\rangle + 3\eta_x^2\ell(\sigma^2+G^2)
\end{aligned}$$

According to smoothness of $f(\cdot, \boldsymbol{y})$, we have:

$$\begin{aligned}\langle \hat{\boldsymbol{x}}_{t-\frac{1}{2}} - \boldsymbol{x}_{t-\frac{1}{2}}, \nabla_x f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}})\rangle &\le f(\hat{\boldsymbol{x}}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}}) - f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}}) + \frac{\ell}{2}\|\hat{\boldsymbol{x}}_{t-\frac{1}{2}} - \boldsymbol{x}_{t-\frac{1}{2}}\|^2 \\ &\le \Phi(\boldsymbol{x}_{t-\frac{1}{2}}) - f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}}) - \frac{\ell}{2}\|\hat{\boldsymbol{x}}_{t-\frac{1}{2}} - \boldsymbol{x}_{t-\frac{1}{2}}\|^2.\end{aligned}$$

So we have:

$$\begin{aligned}&\mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{x}_{t+\frac{1}{2}})] \le \mathbb{E}[\Phi(\hat{\boldsymbol{x}}_{t-\frac{1}{2}})] + \ell\mathbb{E}\|\boldsymbol{x}_{t-\frac{1}{2}} - \hat{\boldsymbol{x}}_{t-\frac{1}{2}}\|^2 \\ &\quad + 2\eta_x\ell\mathbb{E}\left(\Phi(\boldsymbol{x}_{t-\frac{1}{2}}) - f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}}) - \frac{\ell}{2}\|\hat{\boldsymbol{x}}_{t-\frac{1}{2}} - \boldsymbol{x}_{t-\frac{1}{2}}\|^2\right) + 3\eta_x^2\ell(G^2+\sigma^2) \\ &\quad + \eta_x\ell\left(\frac{1}{2\ell}\mathbb{E}\|\nabla_x f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\|^2 + \frac{\ell}{2}\mathbb{E}\|\boldsymbol{x}_{t-\frac{1}{2}} - \hat{\boldsymbol{x}}_{t-\frac{1}{2}}\|^2\right) \\ &\le \mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{x}_{t-\frac{1}{2}})] + 2\eta_x\ell\left(\Phi(\boldsymbol{x}_{t-\frac{1}{2}}) - f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}})\right) - \frac{\eta_x\ell^2}{2}\mathbb{E}\|\hat{\boldsymbol{x}}_{t-\frac{1}{2}} - \boldsymbol{x}_{t-\frac{1}{2}}\|^2 \\ &\quad + 3\eta_x^2\ell(G^2+\sigma^2) + \frac{\eta_x}{2}\mathbb{E}\|\nabla_x f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\|^2.\end{aligned}$$

Using the fact that $\|\boldsymbol{x}_{t-\frac{1}{2}} - \hat{\boldsymbol{x}}_{t-\frac{1}{2}}\| = \frac{1}{2\ell}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_{t-\frac{1}{2}})\|$ will conclude the proof.

□

Lemma B.21. *For Stochastic EG (Algorithm 3), under Theorem 4.9's assumptions, the following statement holds for the generated sequence* $\{\boldsymbol{x}_t\}, \{\boldsymbol{y}_t\}, \{\boldsymbol{x}_{t+\frac{1}{2}}\}, \{\boldsymbol{y}_{t+\frac{1}{2}}\}$ *during algorithm proceeding and* $\forall s \le t$*:*

$$\begin{aligned}&\Phi(\boldsymbol{x}_{t+\frac{1}{2}}) - f(\boldsymbol{x}_{t+\frac{1}{2}}, \boldsymbol{y}_{t+\frac{1}{2}}) \le 2(t-s+1)\eta_x G^2 \\ &\quad + \frac{1}{2\eta_y}\left(\|\boldsymbol{y}_t - \boldsymbol{y}^*(\boldsymbol{x}_s)\|^2 - \|\boldsymbol{y}_{t+1} - \boldsymbol{y}^*(\boldsymbol{x}_s)\|^2 + \frac{1}{2}(3\eta_x^2\eta_y^2\ell^2(G^2+\sigma^2) + 6\eta_y^2\sigma^2)\right).\end{aligned}$$

Proof. According to Lemma B.21:

$$\begin{aligned}\Phi(\boldsymbol{x}_{t+\frac{1}{2}}) - f(\boldsymbol{x}_{t+\frac{1}{2}}, \boldsymbol{y}_{t+\frac{1}{2}}) &\le f(\boldsymbol{x}_{t+\frac{1}{2}}, \boldsymbol{y}^*(\boldsymbol{x}_{t+\frac{1}{2}})) - f(\boldsymbol{x}_s, \boldsymbol{y}^*(\boldsymbol{x}_{t+\frac{1}{2}})) + f(\boldsymbol{x}_s, \boldsymbol{y}^*(\boldsymbol{x}_s)) \\ &\quad - f(\boldsymbol{x}_{t+\frac{1}{2}}, \boldsymbol{y}^*(\boldsymbol{x}_s)) + f(\boldsymbol{x}_{t+\frac{1}{2}}, \boldsymbol{y}^*(\boldsymbol{x}_s)) - f(\boldsymbol{x}_{t+\frac{1}{2}}, \boldsymbol{y}_{t+\frac{1}{2}}) \\ &\le 2(t-s+1)\eta_x G^2 - \langle \boldsymbol{y}_{t+\frac{1}{2}} - \boldsymbol{y}, \nabla_y f(\boldsymbol{x}_{t+\frac{1}{2}}, \boldsymbol{y}_{t+\frac{1}{2}})\rangle\end{aligned}$$

Plugging in Lemma B.19 will conclude the proof:

$$\begin{aligned}&\Phi(\boldsymbol{x}_{t+\frac{1}{2}}) - f(\boldsymbol{x}_{t+\frac{1}{2}}, \boldsymbol{y}_{t+\frac{1}{2}}) \le 2(t-s+1)\eta_x G^2 \\ &\quad + \frac{1}{2\eta_y}\left(\|\boldsymbol{y}_t - \boldsymbol{y}^*(\boldsymbol{x}_s)\|^2 - \|\boldsymbol{y}_{t+1} - \boldsymbol{y}^*(\boldsymbol{x}_s)\|^2 + \frac{1}{2}(3\eta_x^2\eta_y^2\ell^2(G^2+\sigma^2) + 6\eta_y^2\sigma^2)\right).\end{aligned}$$

□

Lemma B.22. *For Stochastic EG (Algorithm 3), under Theorem 4.9's assumptions, the following statement holds for the generated sequence* $\{\boldsymbol{x}_t\}, \{\boldsymbol{y}_t\}, \{\boldsymbol{x}_{t+\frac{1}{2}}\}, \{\boldsymbol{y}_{t+\frac{1}{2}}\}$ *during algorithm proceeding:*

$$\frac{1}{T+1}\sum_{t=0}^{T}\Phi(\boldsymbol{x}_{t-\frac{1}{2}}) - f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}}) \le \frac{1}{B}\left(2\eta_x B^2 G^2 + \frac{D^2}{2\eta_y} + \frac{B\eta_x^2 G^2}{2}\right).$$

Proof. Summing over $t = 0$ to T on both side of Lemma B.21 yields:

$$
\begin{aligned}
&\frac{1}{T+1}\sum_{t=0}^{T}\Phi(\boldsymbol{x}_{t-\frac{1}{2}}) - f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}}) \\
&= \frac{1}{T+1}\sum_{j=0}^{S}\sum_{t=kB}^{(k+1)B-1}\Phi(\boldsymbol{x}_{t-\frac{1}{2}}) - f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}}) \\
&\leq \frac{1}{T+1}\sum_{j=0}^{S}\Bigg[2B^2\eta_x G^2 + \frac{1}{2\eta_y}\Big(\|\boldsymbol{y}_{kB} - \boldsymbol{y}^*(\boldsymbol{x}_s)\|^2 - \|\boldsymbol{y}_{(k+1)B-1} - \boldsymbol{y}^*(\boldsymbol{x}_s)\|^2 \\
&\quad + \frac{1}{2}(3\eta_x^2\eta_y^2\ell^2(G^2+\sigma^2) + 6\eta_y^2\sigma^2)\Big)\Bigg] \\
&\leq \frac{1}{B}\left(2\eta_x B^2 G^2 + \frac{D^2}{2\eta_y} + \frac{1}{2}(3\eta_x^2\eta_y^2\ell^2(G^2+\sigma^2) + 6\eta_y^2\sigma^2)\right),
\end{aligned}
$$

which concludes the proof. □

B.2.4 Proof of Theorem 4.9 for EG

In this section we provide the proof for Theorem 4.9 on the convergence rate of EG in stochastic setting. We first introduce the formal version of theorem here:

Theorem B.23 (EG Stochastic, formal)**.** *Under Assumption 4.3, and 4.7, if we choose* $\eta_x = O(\min\{\frac{\epsilon^2}{\ell(G^2+\sigma^2)}, \frac{\epsilon^4}{D^2\ell^3 G\sqrt{G^2+\sigma^2}}, \frac{\epsilon^6}{D^2\ell^3\sigma^2 G\sqrt{G^2+\sigma^2}}\})$, $\eta_y = O(\min\{\frac{1}{2\ell}, \frac{\epsilon^2}{\ell\sigma^2}\})$, *then Stochastic EG (Algorithm 3) guarantees to find* ϵ*-stationary point, i.e.,* $\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_t)\|^2 \leq \epsilon^2$*, with the gradient complexity bounded by:*

$$
O\left(\frac{D^2\ell^3 G\sqrt{G^2+\sigma^2}}{\epsilon^6}\max\left\{1, \frac{\sigma^2}{\epsilon^2}\right\}\right).
$$

Proof. According to Lemma B.20:

$$
\begin{aligned}
\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_{t-\frac{1}{2}})\|^2 \leq{}& \frac{\mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{x}_{-\frac{1}{2}}) - \Phi_{1/2\ell}(\boldsymbol{x}_{T+\frac{1}{2}})]}{T} \\
&+ 16\ell\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\Phi(x_{t-\frac{1}{2}}) - f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}})] + 24\eta_x\ell(G^2+\sigma^2) \\
&+ 16\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla_x f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\|^2.
\end{aligned}
$$

Observe that:

$$
\begin{aligned}
\mathbb{E}\|\nabla_x f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\|^2 &\leq \ell^2\mathbb{E}\|(\boldsymbol{x}_t, \boldsymbol{y}_t) - (\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\|^2 \\
&= \ell^2\mathbb{E}\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|^2 + \ell^2\mathbb{E}\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 \\
&\leq \ell^2\eta_x^2(G^2+\sigma^2) + \ell^2\eta_y^2\mathbb{E}\left\|\nabla_y f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}})\right\|^2 \\
&\leq \ell^2\eta_x^2(G^2+\sigma^2) + \ell^2\eta_y^2\mathbb{E}\left[\Phi(\boldsymbol{x}_{t-\frac{1}{2}}) - f(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}})\right].
\end{aligned}
$$

So we have:

$$
\begin{aligned}
\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_{t-\frac{1}{2}})\|^2 &\leq \frac{\mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{x}_{-\frac{1}{2}})-\Phi_{1/2\ell}(\boldsymbol{x}_{T+\frac{1}{2}})]}{T+1} \\
&\quad + 16\ell\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\Phi(x_{t-\frac{1}{2}})-f(\boldsymbol{x}_{t-\frac{1}{2}},\boldsymbol{y}_{t-\frac{1}{2}})] + 24\eta_x\ell(G^2+\sigma^2) \\
&\quad + 16\ell^2\eta_y^2\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left[\Phi(\boldsymbol{x}_{t-\frac{1}{2}})-f(\boldsymbol{x}_{t-\frac{1}{2}},\boldsymbol{y}_{t-\frac{1}{2}})\right] + 16\ell^2\eta_x^2(G^2+\sigma^2) \\
&\leq \frac{\mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{x}_{-\frac{1}{2}})-\Phi_{1/2\ell}(\boldsymbol{x}_{T+\frac{1}{2}})]}{T+1} + 16(\ell+\ell^2\eta_y^2)\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\Phi(x_{t-\frac{1}{2}})-f(\boldsymbol{x}_{t-\frac{1}{2}},\boldsymbol{y}_{t-\frac{1}{2}})] \\
&\quad + 16\ell^2\eta_x^2(G^2+\sigma^2) + 24\eta_x\ell(G^2+\sigma^2)
\end{aligned}
$$

Plugging in Lemma B.22 yields:

$$
\begin{aligned}
\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_{t-\frac{1}{2}})\|^2 \leq{}& \frac{\mathbb{E}[\Phi_{1/2\ell}(\boldsymbol{x}_{-\frac{1}{2}})-\Phi_{1/2\ell}(\boldsymbol{x}_{T+\frac{1}{2}})]}{T+1} \\
&+ 16(\ell+\ell^2\eta_y^2)\frac{1}{B}\left(2\eta_x B^2G^2 + \frac{D^2}{2\eta_y} + \frac{B\eta_x^2G^2}{2}\right) \\
&+ 16\ell^2\eta_x^2(G^2+\sigma^2) + 24\eta_x\ell(G^2+\sigma^2).
\end{aligned}
$$

Choosing $B = O(\frac{D}{\sqrt{\eta_x\eta_y G\sqrt{G^2+\sigma^2}}})$, $\eta_x = O(\min\{\frac{\epsilon^2}{\ell(G^2+\sigma^2)}, \frac{\epsilon^4}{D^2\ell^3G\sqrt{G^2+\sigma^2}}, \frac{\epsilon^6}{D^2\ell^3\sigma^2G\sqrt{G^2+\sigma^2}}\})$, $\eta_y = O(\min\{\frac{1}{\ell}, \frac{\epsilon^2}{\ell\sigma^2}\})$, guarantees that $\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\|\nabla\Phi_{1/2\ell}(\boldsymbol{x}_t)\|^2 \leq \epsilon^2$ holds with the gradient complexity is bounded by:

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

which completes the proof. □

B.3 Tightness Analysis

In this section, we provide our tightness analysis showing our obtained upper bound is tight given our choice of learning rates. In subsection B.3.1, we introduce our hard example, and show the lower bound on convergence of this example, and then in subsection B.3.2, we extend the tightness result to EG/OGDA using the same hard example.

B.3.1 GDA

Proof of Theorem 4.10. Let $L \geq 0$ be some constants to be chosen later. Inspired by [11], we consider the following function $f : \mathbb{R}\times[-D,D]\to\mathbb{R}$:

$$
f(x,y) = h(x)y
$$

where

$$
h(x) = \begin{cases} \frac{L}{2}x^2 & |x|\leq 1 \\ L - \frac{L}{2}(|x|-2)^2 & 1\leq|x|\leq 2 \\ L & |x|\geq 2. \end{cases}
$$

It is easy to verify that f is nonconvex, $2LD$ smooth, and LD-Lipschitz. We choose $L = \frac{1}{D}\min\{\ell/2, G\}$ to guarantee that f is ℓ smooth and G-Lipschitz with respect to x. The primal function is $\Phi(x) = Dh(x)$ attained when $y = D$. After standard calculations, we know that when $|x| \leq 1$, the Moreau envelope $\Phi_{1/2\ell}(x)$ satisfies

$$\Phi_{1/2\ell}(x) = \frac{LD\ell}{LD + 2\ell} x^2, \quad |x| \leq 1.$$

By definition, we also know $\Phi_{1/2\ell}(x) \geq 0$ for any $x \in \mathbb{R}$.

We first claim that if we choose $|x_0| \leq 1$, $y_0 \geq 0$, we have for any $t \geq 0$, $|x_t| \leq 1$ and $y_t \geq 0$. We verify this claim by induction. First note that when $t = 0$, the claim holds for sure. Let us assume it holds for $t = k$. Then for $t = k + 1$,

$$x_{k+1} = x_k - \eta_x L x_k y_k = (1 - \eta_x L y_k) x_k.$$

Since $0 \leq y_k \leq D$, we have $0 \leq 1 - \eta_x L y_k \leq 1$. Therefore $|x_{k+1}| \leq 1$. For y_{k+1}, we have

$$y_{k+1} = \mathcal{P}_{[-D,D]}(y_k + \eta_y h(x_k)).$$

Since $h(x_k) \geq 0$, we know that $y_{k+1} \geq 0$, which verifies the claim.

We can also bound

$$|x_T| = \left| \prod_{t=0}^{T-1} (1 - \eta_x L y_t) x_0 \right| \geq (1 - \eta_x LD)^T |x_0|.$$

Since $\nabla \Phi_{1/2\ell}(x) = \frac{2LD\ell}{LD+2\ell} x$, choosing $x_0 = \frac{LD+2\ell}{LD\ell}\epsilon$, we have $\epsilon \geq |\nabla \Phi_{1/2\ell}(x_T)| \geq 2\epsilon(1 - \eta_x LD)^T$. Also noting $\hat{\Delta}_\Phi = \frac{LD+2\ell}{LD\ell}\epsilon^2$, we have

$$\begin{aligned} T = \Omega\left(\frac{1}{\eta_x LD}\right) &= \Omega\left(\frac{\hat{\Delta}_\Phi}{\eta_x LD\epsilon^2} \cdot \frac{LD\ell}{LD + 2\ell}\right) \\ &= \Omega\left(\frac{\ell^3 G^2 D^2 \hat{\Delta}_\Phi}{\epsilon^6}\right). \end{aligned}$$

□

B.3.2 EG/OGDA

Proof of Theorem 4.11 for OGDA. We use the same hard example $f(x, y) = h(x)y$ as in proof of Theorem 4.10. Similarly, we first claim that if we choose $0 \leq x_0 \leq 1$ and $y_0 = D$, the following statements hold for any $t \geq 0$:

$$(a)\ 0 \leq x_t \leq 1, \text{and } x_t \geq x_{t-1}/\sqrt{2}, (b)\ y_t = D,$$

where we define $x_{-1} = x_0$ and $y_{-1} = y_0$.

Now we prove the above claim by induction. First, when $t = 0$, the claim holds for sure. Then, let us assume it holds for $t \leq k$. Then for $t = k + 1$, we have

$$\begin{aligned} x_{k+1} &= x_k - 2\eta_x LD x_k y_k + \eta_x LD x_{k-1} y_{k-1} \\ &= (1 - 2\eta_x LD) x_k + \eta_x LD x_{k-1}. \end{aligned}$$

Since $0 \leq x_k, x_{k-1} \leq 1$ and $0 \leq \eta_x LD \leq 0.1$, we have

$$(1 - 2\eta_x LD) x_k \leq x_{k+1} \leq (1 - \eta_x LD) x_k + \eta_x LD x_{k-1},$$

which implies $0 \leq x_k/\sqrt{2} \leq 0.8 x_k \leq x_{k+1} \leq 1$. For y_{k+1}, we know

$$y_{k+1} = \mathcal{P}_{[-D,D]}(y_k + 2\eta_y h(x_k) - \eta_y h(x_{k-1})).$$

Since $h(x) = \frac{L}{2}x^2$ when $|x| \leq 1$, and $x_k \geq \frac{1}{\sqrt{2}} x_{k-1}$, we know that $2\eta_y h(x_k) - \eta_y h(x_{k-1}) \geq 0$ so $y_{k+1} = 1$. Till now, we have proved the claim.

Then, we are going to bound the magnitude of x_T. According to the updating rule we have:

$$x_{t+1} = x_t - 2\eta_x LD x_t + \eta_x LD x_{t-1}.$$

Solving the above recursion we get the solution for x_t as follows:

$$\begin{aligned} x_t = &\left(\frac{1}{2} + \frac{1}{2\sqrt{\Delta}}\right)\left(\frac{1-2\eta_x LD+\sqrt{\Delta}}{2}\right)^t x_0 \\ &+ \left(\frac{1}{2} - \frac{1}{2\sqrt{\Delta}}\right)\left(\frac{1-2\eta_x LD-\sqrt{\Delta}}{2}\right)^t x_0, \end{aligned}$$

where $\Delta = (1-2\eta_x LD)^2 + 4\eta_x LD$.

Let $a_1 = \left(\frac{1}{2} + \frac{1}{2\sqrt{\Delta}}\right)$, $a_2 = \left(\frac{1}{2} - \frac{1}{2\sqrt{\Delta}}\right)$, and $\lambda_1 = \left(\frac{1-2\eta_x LD+\sqrt{\Delta}}{2}\right)$, $\lambda_2 = \left(\frac{1-2\eta_x LD-\sqrt{\Delta}}{2}\right)$. We observe the following facts:

$$\begin{aligned} &a_1 \geq \frac{1}{2}, a_2 \leq \eta_x^2 L^2 D^2, \\ &1-\eta_x LD \leq \lambda_1 \leq 1, -\eta_x LD \leq \lambda_2 \leq 0. \end{aligned}$$

Now, we can bound the magnitude of x_T

$$\begin{aligned} |x_T| = \left|a_1\lambda_1^T + a_2\lambda_2^T\right| x_0 &\geq \left||a_1\lambda_1^T| - |a_2\lambda_2^T|\right| x_0 \\ &\geq \left(\frac{1}{2}(1-2\eta_x LD)^T - (\eta_x LD)^{T+2}\right) x_0. \end{aligned}$$

Since $\nabla\Phi_{1/2\ell}(x) = \frac{2LD\ell}{LD+2\ell}x$, by choosing $x_0 = \frac{LD+2\ell}{LD\ell}\cdot 4\epsilon$, we have

$$\epsilon \geq |\nabla\Phi_{1/2\ell}(x_T)| \geq 8\epsilon\left(\frac{1}{2}(1-2\eta_x LD)^T - \frac{1}{4}\right),$$

which yields $(1-2\eta_x LD)^T \leq 3/4$. The rest of proof is similar to that of Theorem 4.10. □

Proof of Theorem 4.11 for EG. We use the same hard example $f(x,y) = h(x)y$ as in proof of Theorem 4.10. Similarly to our previous proofs for GDA and OGDA, we first claim that if we choose $0 \leq x_0 \leq 1$ and $y_0 = D$, the following statements hold for any $t \geq 0$:

$$(a)\ 0 \leq x_t \leq 1;\ (b)\ y_t = D, y_{t+1/2} = D.$$

We prove this claim by induction. First, when $t=0$, the claim holds for sure. Then, let us assume it holds for $t \leq k$. Then for $t = k+1$, we have

$$\begin{aligned} x_{k+1} &= x_k - \eta_x L y_{k+1/2} x_{k+1/2} \\ &= x_k - \eta_x L y_{k+1/2}\left(1-\eta_x L y_k\right) x_k \\ &= (1-\eta_x LD + \eta_x^2 L^2 D^2) x_k. \end{aligned}$$

Note that since $0 \leq \eta_x LD \leq 1/2$, we know

$$0 \leq 1 - \eta_x LD + \eta_x^2 L^2 D^2 \leq 1,$$

which implies $0 \leq x_{k+1} \leq 1$. Regarding y, note that

$$\begin{aligned} y_{k+1} &= \mathcal{P}_{[-D,D]}(y_k + \eta_y h(x_{k+1/2}))), \\ y_{k+3/2} &= \mathcal{P}_{[-D,D]}(y_{k+1} + \eta_y h(x_{k+1}))). \end{aligned}$$

As $h(x_{k+1/2}), h(x_{k+1}) \geq 0$ and $y_k = D$, we have $y_{k+1} = y_{k+3/2} = D$. Till now, we have verified the claim.

Note that

$$x_{k+1} = (1-\eta_x LD + \eta_x^2 L^2 D^2) x_k \geq (1-\eta_x LD) x_k.$$

Hence we can unroll the recursion and lower bound the magnitude of $\nabla\Phi_{1/2\ell}(x_T)$, which is similar to the proof of Theorem 4.10. □

C Proof of Stepsize-Independent Lower Bound Results in Nonconvex-Strongly-Concave Setting

In this section, we prove general lower bounds on the convergence rate of GDA/EG/OGDA for the NC-SC setting. In subsection C.1, proof of theorem 5.1 is established giving the lower bound for GDA in NC-SC, and in subsection C.2, the proof of Theorem 5.2 is established, proving the lower bound of EG/OGDA for NC-SC problems.

C.1 Lower Bound for GDA

Theorem C.1 (Theorem 5.1 restated)**.** *For GDA algorithm, given $\eta_y = \Theta(1/\ell)$, for any η_x, there exists a ℓ-smooth function that is nonconvex in x and μ-strongly-concave in y, such that for $\|\Phi(x_T)\| \leq \epsilon$, we must have:*

$$T = \Omega\left(\frac{\kappa\ell\Delta_\phi}{\epsilon^2}\right)$$

Proof. Combining Proposition C.2 and C.3 will conclude the proof. Proposition C.3 shows that when $\eta_x \in (\frac{1}{\kappa\ell}.\infty)$, GDA diverges, and Proposition C.2 shows the lower bound on the convergence rate when $\eta_x \in (0, \frac{1}{\kappa\ell}]$. ☐

Proposition C.2. *For GDA algorithm, given $\eta_y = \Theta(1/\ell)$, for any $\eta_x \in (0, \frac{1}{\kappa\ell}]$, there exists a ℓ-smooth function that is nonconvex in x and μ-strongly-concave in y, such that for $\|\Phi(x_T)\| \leq \epsilon$, we must have:*

$$T = \Omega\left(\frac{\kappa\ell\Delta_\phi}{\epsilon^2}\right)$$

Proof. Recall that we consider the following quadratic NC-SC function $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$

$$f(x, y) := -\tfrac{1}{2}\ell x^2 + bxy - \tfrac{1}{2}\mu y^2.$$

Recall that f is nonconvex in x (it is actually concave in x) and μ strongly concave in y. Assume $\kappa := \ell/\mu \geq 4$ and choose $b = \sqrt{\mu(\ell + \mu_x)}$ for some $0 < \mu_x \leq \ell/2$ to be chosen later. Then we know $b \leq \ell/2$, and it is easy to verify f is ℓ smooth. Note that the primal function

$$\Phi(x) = \max_y f(x, y) = \tfrac{1}{2}\mu_x x^2$$

is actually strongly convex. This also justifies the symbol for μ_x. We use GDA to find the solution for $\min_x \max_y f(x, y)$. Actually, for this problem, the optimal solution is achieved at the origin. The stepsizes ratio is chosen as $r = \frac{\eta_y}{\eta_x}$ and $\eta_y = \frac{1}{\ell}$ for some numerical constants c. Then the GDA update rule can be written as

$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} = (\mathbf{I} + \eta_x \mathbf{M}) \cdot \begin{pmatrix} x_k \\ y_k \end{pmatrix}, \tag{85}$$

where

$$\mathbf{M} := \begin{pmatrix} \ell & -b \\ rb & -\mu r \end{pmatrix}. \tag{86}$$

Note that (85) is a linear time-invariant system. We need to analyze its eigenvalues. Let λ_1 and λ_2 be the two eigenvalues of $\mathbf{M}$, we have

$$\lambda_{1,2} = -\frac{1}{2}(\mu r - \ell) \pm \frac{1}{2}\sqrt{(\mu r - \ell)^2 - 4r\mu\mu_x}.$$

Note that if we choose $\mu_x < \ell/8$, plugging into $r = c\kappa$, we can bound

$$\begin{aligned} 0 \geq \lambda_1 &= -\frac{(2\kappa - 1)\ell}{4}\left(1 - \sqrt{1 - \frac{4c\kappa\mu_x}{(\mu r - \ell)^2}}\right) \\ &\geq -\frac{2\mu r\mu_x}{\mu r - \ell} \geq -4\mu_x. \end{aligned}$$

Let s_1 be the corresponding eigenvalue of $\mathbf{I} + \eta_x \mathbf{M}$, for small enough $c_1 \leq 1$, it satisfies

$$0 \leq 1 - \frac{\mu_x}{r\ell} = 1 - \frac{1}{r\kappa_x} \leq s_1 = 1 + \eta_x \lambda_1 \leq 1.$$

We adversarially choose the initial point (x_0, y_0) such that it is parallel to the eigenvector of $\mathbf{I} + \eta_x \mathbf{M}$ corresponding to s_1. We can always choose $x_0 \geq 0$ for simplicity. Then we have

$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} = (\mathbf{I} + \eta_x \mathbf{M})^T \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = s_1^T \begin{pmatrix} x_0 \\ y_0 \end{pmatrix},$$

so we can compute the magnitude of x_T as $x_T = s_1^T x_0$. Choose $\mu_x = \frac{\kappa\ell}{2T}$, and thus we have:

$$\|\nabla\Phi(x_T)\| = \|\mu_x x_0\| = \mu_x \left(1 - \frac{1}{r\kappa_x}\right)^T |x_0| \geq \mu_x \left(1 - \frac{1}{\kappa\kappa_x}\right)^T |x_0| \geq \mu_x \exp\left(\frac{2T}{\kappa\kappa_x}\right) |x_0| \geq \frac{1}{2}\mu_x |x_0|$$

where we use the inequality that $1 - \frac{z}{2} \geq \exp(z \ln \frac{1}{2})$ and $\exp(z \ln \frac{1}{2}) \geq \frac{1}{2}$ for $z \in [0, 1]$. Recall that we choose $x_0 = \sqrt{\frac{2\Delta_\Phi}{\mu_x}}$, we have:

$$\|\nabla\Phi(x_T)\| \geq \frac{1}{2}\sqrt{2\mu_x \Delta} = \Omega\left(\sqrt{\frac{\kappa\ell\Delta}{T}}\right),$$

which means to guarantee that $\|\nabla\Phi(x_T)\| \leq \epsilon$, we must have $T \geq \Omega\left(\frac{\kappa\ell\Delta_\Phi}{\epsilon^2}\right)$.

□

Proposition C.3. *For GDA algorithm, given $\eta_y = \Theta(1/\ell)$, for any $\eta_x \in (\frac{1}{\kappa\ell}, \infty)$, there exists a ℓ-smooth function that is nonconvex in x and μ-strongly-concave in y, such that:*

$$\|\nabla\Phi(x_T)\| \geq c$$

where c is some constant that does not vanish as T increases.

Proof. Recall the transition matrix in (86). We notice that

$$\mathsf{trace}(\mathbf{M}) = \lambda_1 + \lambda_2 = L - \mu r.$$

Since $r \leq \kappa$, then $\lambda_1 + \lambda_2 \geq 0$, which means that $\max\{Re[\lambda_1], Re[\lambda_2]\} \geq 0$, so:

$$\|(\mathbf{I} + \eta_x \mathbf{M})^T\| \geq \max\{|1 + \eta_x\lambda_1|, |1 + \eta_x\lambda_2|\}^T \geq \alpha^T$$

where α is some constant larger than 1. If we choose the initialization to be $[x_0, 0]$, the gradient $\|\nabla\Phi(x_T)\| = \mu_x \|(\mathbf{I} + \eta_x \mathbf{M})^T\| x_0$ diverges. □

C.2 Lower bound for EG/OGDA

Theorem C.4 (Theorem 5.2 restated)**.** *For deterministic EG/OGDA algorithm, given $\eta_y = \Theta(1/\ell)$, for any η_x, there exists a ℓ-smooth function that is nonconvex in x and μ-strongly-concave in y, such that for $\|\Phi(x_T)\| \leq \epsilon$, we must have:*

$$T = \Omega\left(\frac{\kappa\ell\Delta_\phi}{\epsilon^2}\right)$$

Proof of Theorem C.4 for EG. We consider the same quadratic hard example f and notation used in the proof of Theorem 5.1. For simplicity, denote $\boldsymbol{w} = (x, y)$. Then the updating rule for EG can be written as:

$$\begin{aligned} \boldsymbol{w}_{k+1/2} =& (\mathbf{I} + \eta_x \mathbf{M})\boldsymbol{w}_k, \\ \boldsymbol{w}_{k+1} =& \boldsymbol{w}_k + \eta_x \mathbf{M}\boldsymbol{w}_{k+1/2} \\ =& (\mathbf{I} + \eta_x \mathbf{M} + \eta_x^2 \mathbf{M}^2)\boldsymbol{w}_k. \end{aligned}$$

Therefore, similar to GDA, EG is also a linear time-invariant system with the difference that the transition matrix now becomes as $\mathbf{M}' = (\mathbf{I} + \eta_x \mathbf{M} + \eta_x^2 \mathbf{M}^2)$.

The rest of the analysis is the same as that of GDA in Proposition C.2. Then, we are going to show that when $\eta_x \in (\frac{1}{c_x \kappa \ell}, +\infty)$ for some c_x, the EG method diverges. Consider

$$f(x, y) := -\tfrac{1}{2}\ell x^2 + bxy - \tfrac{1}{2}\mu y^2.$$

Then according to Proposition C.2, we have:

$$\begin{aligned} \mathsf{trace}(\mathbf{M}') &= \mathsf{trace}(\mathbf{I} + \eta_x \mathbf{M} + \eta_x^2 \mathbf{M}^2) \\ &= 1 + \eta_x(\ell - \mu r) + \eta_x^2(\ell^2 + \mu^2 r^2 - 2rb^2) \\ &= 1 + \eta_x(\ell - \mu r) + \eta_x^2\left((\ell - \mu r)^2 - 2r\mu\mu_x\right) \end{aligned} \tag{87}$$

Now note that since $r \leq \kappa$, to show $\mathsf{trace}(\mathbf{M}') \geq 1$, it is enough to have $\mu_x \leq \frac{(\ell - \mu r)^2}{2r\mu}$. However, by choosing $\mu_x = \Theta(\epsilon^2)$, and by choosing the small enough ϵ, we can satisfy the condition that $\mu_x \leq \frac{(\ell - \mu r)^2}{2r\mu}$, thus we can conclude that under this situation $\mathsf{trace}(\mathbf{M}') \geq 1$, which means that same step as the Proposition C.3 can be taken to prove the divergence of $\|\nabla \Phi(x_T)\|^2$.

□

Proof of Theorem C.4 for OGDA. Assuming the same setup as the proof of EG, the update rule can be written as follows: The dynamics of OGDA is

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + 2\eta_x \mathbf{M}\boldsymbol{w}_k - \eta_x \mathbf{M}\boldsymbol{w}_{k-1}.$$

If we initialize $\boldsymbol{w}_0$ parallel to the eigenvector of $\mathbf{M}$ corresponding to λ_1 and let $\boldsymbol{w}_1 = \boldsymbol{w}_0$, we know every $\boldsymbol{w}_k$ is parallel to it, i.e., $\boldsymbol{w}_k = z_k \boldsymbol{w}_0$ for some scalar z_k which satisfies

$$z_{k+1} = z_k + 2\eta_x \lambda_1 z_k - \eta_x \lambda_1 z_{k-1}.$$

The general solution of the above recurrence relation is

$$z_k = a\alpha^k + b\beta^k$$

for some constant a, b and

$$\begin{aligned} \alpha =& \frac{1}{2}\left(1 + 2\eta_x\lambda_1 + \sqrt{1 + 4\eta_x^2\lambda_1^2}\right), \\ \beta =& \frac{1}{2}\left(1 + 2\eta_x\lambda_1 - \sqrt{1 + 4\eta_x^2\lambda_1^2}\right). \end{aligned}$$

We have

$$1 + \eta_x\lambda_1 \leq \alpha \leq 1, \quad \eta_x\lambda_1 \leq \beta \leq 0.$$

Using the initial condition $z_{-1} = z_0 = 1$, we can get the constants

$$\begin{aligned} a =& \frac{\alpha(1-\beta)}{\alpha - \beta} = \frac{1}{2} + \frac{1}{2\sqrt{1 + 4\eta_x^2\lambda_1^2}} \geq 1/2, \\ b =& -\frac{\beta(1-\alpha)}{\alpha - \beta} = \frac{\sqrt{1 + 4\eta_x^2\lambda_1^2} - 1}{2\sqrt{1 + 4\eta_x^2\lambda_1^2}} \leq \eta_x^2\lambda_1^2. \end{aligned}$$

We can bound

$$\begin{aligned} |z_T| &\geq \frac{1}{2}(1 + \eta_x\lambda_1)^T - |\eta_x\lambda_1|^{k+2} \\ &\geq \frac{1}{2}\left(1 - \frac{4c_1\mu_x}{\kappa}\right)^T - \frac{1}{4}, \end{aligned}$$

where we use the fact $|\eta_x \lambda_1| \leq 1/2$. Similar to the analysis for GDA, choosing $\mu_x = 50\epsilon^2/\Delta_\Phi$, we have

$$\begin{aligned}|\nabla\Phi(\bar{x})| = \mu_x \bar{x} \geq & \mu_x x_T \geq \mu_x x_0 \left[\frac{1}{2}\left(1 - \frac{4c_1\mu_x}{\kappa}\right)^T - \frac{1}{4}\right] \\ = & 10\epsilon \left[\frac{1}{2}\left(1 - \frac{4c_1\mu_x}{\kappa}\right)^T - \frac{1}{4}\right].\end{aligned}$$

Therefore, if $|\nabla\Phi(\bar{x})| \leq \epsilon$, we must have

$$T = \Omega\left(\frac{\kappa}{\mu_x}\right) = \Omega\left(\frac{\kappa\Delta_\Phi}{\epsilon^2}\right).$$

Now, we will show that Proposition C.3 also holds for OGDA. Consider the following 4×4 matrix $\mathbf{M}'$:

$$\mathbf{M}' = \left[\begin{array}{c:c} (\mathbf{I} + 2\eta_x\mathbf{M})^2 & -\eta_x(\mathbf{I} + 2\eta_x\mathbf{M})\mathbf{M} \\ \hdashline \mathbf{I} + 2\eta_x\mathbf{M} & -\eta_x\mathbf{M} \end{array}\right] \tag{88}$$

It can be easily shown that, the OGDA dynamic can be written as follows:

$$\begin{bmatrix} \boldsymbol{w}_{k+1} \\ \boldsymbol{w}_k \end{bmatrix} = \left[\begin{array}{c:c} (\mathbf{I} + 2\eta_x\mathbf{M})^2 & -\eta_x(\mathbf{I} + 2\eta_x\mathbf{M})\mathbf{M} \\ \hdashline \mathbf{I} + 2\eta_x\mathbf{M} & -\eta_x\mathbf{M} \end{array}\right] \begin{bmatrix} \boldsymbol{w}_{k-1} \\ \boldsymbol{w}_{k-2} \end{bmatrix} \tag{89}$$

Now similar to proof of Proposition C.3 for GDA, it suffices to show that the $\mathsf{trace}(\mathbf{M}') \geq 1$ given the conditions on the learning rate. To this end, note that we can write:

$$\begin{aligned}\mathsf{trace}(\mathbf{M}') &= \mathsf{trace}(-\eta_x\mathbf{M}) + \mathsf{trace}(\mathbf{I} + 4\eta_x\mathbf{M} + 4\eta_x^2\mathbf{M}^2) \\ &= 1 - \eta_x(\ell - \mu r) + 4\eta_x(\ell - \mu r) + 4\eta_x^2(\ell^2 + \mu^2 r^2 - 2rb^2) \\ &= 1 + 3\eta_x(\ell - \mu r) + 4\eta_x^2\left((\ell - \mu r)^2 - 2r\mu\mu_x\right)\end{aligned} \tag{90}$$

Now note that since $r \leq \kappa$, to show $\mathsf{trace}(\mathbf{M}') \geq 1$, it is enough to have $\mu_x \leq \frac{(\ell-\mu r)^2}{2r\mu}$. However, note that we let $\mu_x = \frac{50\epsilon^2}{\Delta_\Phi}$, thus by choosing the small enough ϵ, we can satisfy the condition that $\mu_x \leq \frac{(\ell-\mu r)^2}{2r\mu}$, thus we can conclude that $\mathsf{trace}(\mathbf{M}') \geq 1$ holds. Consequently, similar argument as the Proposition C.3 can be made to prove the divergence of $\|\nabla\Phi(x_T)\|^2$. $\square$

D Extension to Generalized OGDA

In this section, we analyze the convergence of generalized OGDA (Algorithm 4) where we utilize different learning rates for descent/ascent gradients and correction terms. Specifically, we propose to use different learning rates for $\nabla_x f(\boldsymbol{x}_t, \boldsymbol{y}_t)$, and $\nabla_x f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})$ terms, and also $\nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t)$, and $\nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})$, in order to make the algorithm more stable. We believe this algorithm is more convenient in practice due to the more flexibility it provides in deciding the learning rates. We demonstrated this stabilizing effect of generalized OGDA in our empirical results in Section 6. Also, note that if we let $\eta_{x,1} = \eta_{x,2}$, and $\eta_{y,1} = \eta_{y,2}$ in Algorithm 4, it reduces to stochastic OGDA. Theorem D.1 establishes the convergence rate of generalized OGDA in NC-SC. However, it still remains open to analyze this algorithm in C-C/SC-SC and NC-C settings.

We remark that the analysis of generalized OGDA was only known for the restricted bilinear functions, which is established in [39], and convergence analysis beyond these simple functions previously was unknown that we provide here.

Algorithm 4 Generalized Stochastic OGDA

```
Input: $(\boldsymbol{x}_0, \boldsymbol{y}_0)$, stepsizes $(\eta_{x,1}, \eta_{x,2}, \eta_{y,1}, \eta_{y,2})$
for $t = 1, 2, \ldots, T$ do
    $\boldsymbol{x}_t \leftarrow \boldsymbol{x}_{t-1} - \eta_{x,1}\boldsymbol{g}_{x,t-1} - \eta_{x,2}(\boldsymbol{g}_{x,t-1} - \boldsymbol{g}_{x,t-2})$
    $\boldsymbol{y}_t \leftarrow \boldsymbol{y}_{t-1} + \eta_{y,1}\boldsymbol{g}_{y,t-1} + \eta_{y,2}(\boldsymbol{g}_{y,t-1} - \boldsymbol{g}_{y,t-2})$
end for
Randomly choose $\bar{\boldsymbol{x}}$ from $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$
Output:$\bar{\boldsymbol{x}}$
```

Theorem D.1. *Let* $\eta_{x,1} = \frac{1}{50\kappa^2\ell}$, $\eta_{y,2} = \frac{1}{6\ell}$. *Also, let* $\alpha = \frac{\eta_{x,2}}{\eta_{x,1}}$, *and* $\beta = \frac{\eta_{y,1}}{\eta_{y,2}}$. *Then assuming* $\beta \leq 1$, *and* $\alpha \leq 2\kappa^2\sqrt{\beta}$, *under Assumptions 4.1, and 4.3 for Algorithm 4 we have:*

$$\mathbb{E}[\|\nabla\Phi(\bar{\boldsymbol{x}})\|^2] \leq O\Big(\frac{\kappa^2\ell\Delta}{T} + \frac{(\kappa+\alpha^2)\ell^2 D}{\beta T} + \frac{\kappa\sigma^2}{M_y} + \frac{(1+\alpha^2)\sigma^2}{M_x}\Big), \tag{91}$$

where $D = \max(\|\boldsymbol{y}_1 - \boldsymbol{y}_1^*\|^2, \|\boldsymbol{y}_1 - \boldsymbol{y}_0\|^2, \|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2)$, *and* $\Delta = \phi(\boldsymbol{x}_1) - \min_{\boldsymbol{x}} \Phi(\boldsymbol{x})$.

A few observations about the obtained rate are in place.

Corollary D.2. *Let* $\sigma = 0$, *and pick an* $\alpha \leq \sqrt{\kappa}$. *Then deterministic generalized OGDA converges to* ϵ*-stationary point of* $\Phi(\boldsymbol{x})$ *with gradient complexity of* $O(\frac{\kappa^2}{\epsilon^2})$.

Corollary D.3. *For any* $\alpha = O(\sqrt{k})$, *and any* $\mu \leq \beta \leq 1$, *if we choose* $M_x = O(\kappa\frac{\sigma^2}{\epsilon^2})$, *and* $M_y = O(\frac{\kappa}{\epsilon^2})$, *then stochastic generalized OGDA converges to* ϵ*-stationary point of* $\Phi(\boldsymbol{x})$ *with gradient complexity of* $O(\frac{\kappa^3}{\epsilon^4})$.

Remark D.4. Theorem D.1 establishes the convergence rate under broad range of primal learning rates ratio ($0 \leq \alpha \leq O(\kappa^2)$), and it shows that as long as $\alpha \leq \sqrt{\kappa}$, we can achieve the same convergence rate as OGDA if we assume $\mu \leq \beta \leq 1$.

D.1 Nonconvex-strongly-concave setting

We follow exact same steps as Lemma A.4, to derive the following lemmas.

Lemma D.5. *Let* $\Phi(\boldsymbol{x}) = \max_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})$, *and* $\boldsymbol{y}^*(\boldsymbol{x}) = \arg\max_{\boldsymbol{y}} f(\boldsymbol{x}, \boldsymbol{y})$. *Also, let* $\boldsymbol{g}_i = \boldsymbol{g}_{x,i} + \alpha(\boldsymbol{g}_{x,i} - \boldsymbol{g}_{x,i-1})$, *where* $\alpha = \frac{\eta_{x,2}}{\eta_{x,1}}$. *Therefore, we have* $\boldsymbol{x}_i = \boldsymbol{x}_{i-1} - \eta_{x,1}\boldsymbol{g}_i$. *Then for Algorithm 4, we have:*

$$\begin{aligned}\mathbb{E}[\Phi(\boldsymbol{x}_t)] \leq\ & \mathbb{E}[\Phi(\boldsymbol{x}_{t-1})] - \frac{\eta_{x,1}}{2}\mathbb{E}[\|\nabla\Phi(\boldsymbol{x}_{t-1})\|^2] - \frac{\eta_{x,1}}{2}(1 - 2\kappa\ell\eta_{x,1})\mathbb{E}[\|\boldsymbol{g}_{t-1}\|^2] \\ & + \frac{3}{2}\eta_{x,1}^3\alpha^2\ell^2\mathbb{E}[\|\boldsymbol{g}_{t-2}\|^2] + \frac{3}{2}\eta_{x,1}\ell^2\mathbb{E}[\|\boldsymbol{y}_{t-1}^* - \boldsymbol{y}_{t-1}\|^2] + \frac{3}{2}\eta_{x,1}\alpha^2\ell^2\mathbb{E}[\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2] \\ & + 3((1+\alpha)^2 + 1)\eta_{x,1}\frac{\sigma^2}{M_x}\end{aligned} \tag{92}$$

Proof of Lemma D.5. Proof is pretty much similar to proof of Lemma A.4, and we only include this proof for sake of completeness. First, let $\boldsymbol{\delta}_i^x = \boldsymbol{g}_{x,i} - \nabla_x f(\boldsymbol{x}_i, \boldsymbol{y}_i)$. By definition of $\boldsymbol{g}_{x,i}$, we have $\mathbb{E}[\boldsymbol{\delta}_i^x] = 0$, for all $i \in [T]$.

Using the fact that $\Phi(\boldsymbol{x})$ is $2\kappa\ell$ smooth, we have:

$$\begin{aligned}
\Phi(\boldsymbol{x}_t) &\le \Phi(\boldsymbol{x}_{t-1}) + \langle \nabla\Phi(\boldsymbol{x}_{t-1}), \boldsymbol{x}_t - \boldsymbol{x}_{t-1}\rangle + \kappa\ell\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|^2 \\
&= \Phi(\boldsymbol{x}_{t-1}) - \eta_{x,1}\langle \nabla\Phi(\boldsymbol{x}_{t-1}), \boldsymbol{g}_{t-1}\rangle + \kappa\ell\eta_{x,1}^2\|\boldsymbol{g}_{t-1}\|^2 \\
&= \Phi(\boldsymbol{x}_{t-1}) - \frac{\eta_{x,1}}{2}\|\nabla\Phi(\boldsymbol{x}_{t-1})\|^2 - \frac{\eta_{x,1}}{2}\|\boldsymbol{g}_{t-1}\|^2 + \frac{\eta_{x,1}}{2}\|\nabla\Phi(\boldsymbol{x}_{t-1}) - \boldsymbol{g}_{t-1}\|^2 \\
&\quad + \kappa\ell\eta_{x,1}^2\|\boldsymbol{g}_{t-1}\|^2 \\
&= \Phi(\boldsymbol{x}_{t-1}) - \frac{\eta_{x,1}}{2}\|\nabla\Phi(\boldsymbol{x}_{t-1})\|^2 - \frac{\eta_{x,1}}{2}(1 - 2\kappa\ell\eta_{x,1})\|\boldsymbol{g}_{t-1}\|^2 \\
&\quad + \frac{\eta_{x,1}}{2}\|\nabla\Phi(\boldsymbol{x}_{t-1}) - \boldsymbol{g}_{t-1}\|^2
\end{aligned} \tag{93}$$

Now using ℓ-smoothness of f, and κ-Lipschitzness of $\boldsymbol{y}^*(\boldsymbol{x})$ (Lemma A.1) we have:

$$\begin{aligned}
\|\nabla\Phi(\boldsymbol{x}_{t-1}) - \boldsymbol{g}_{t-1}\|^2 &= \|\nabla\Phi(\boldsymbol{x}_{t-1}) - \nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) \\
&\quad - \alpha\left(\nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_x f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2})\right) - ((\alpha+1)\boldsymbol{\delta}_{t-1}^x - \boldsymbol{\delta}_{t-2}^x)\|^2 \\
&\le 3\|\nabla\Phi(\boldsymbol{x}_{t-1}) - \nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\|^2 + 3\alpha^2\|\nabla_x f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_x f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2})\|^2 \\
&\quad + 3\|(\alpha+1)\boldsymbol{\delta}_{t-1}^x - \boldsymbol{\delta}_{t-2}^x\|^2 \\
&\le 3\ell^2\|\boldsymbol{y}^*(\boldsymbol{x}_{t-1}) - \boldsymbol{y}_{t-1}\|^2 + 3\alpha^2\ell^2\|\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-2}\|^2 + 3\alpha^2\ell^2\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2 \\
&\quad + 6(\alpha+1)^2\|\boldsymbol{\delta}_{t-1}^x\|^2 + 6\|\boldsymbol{\delta}_{t-2}^x\|^2
\end{aligned} \tag{94}$$

where in the first and second inequalities we used Young's inequality.

By combining Equations 93 and 94 we have:

$$\begin{aligned}
\Phi(\boldsymbol{x}_t) &\le \Phi(\boldsymbol{x}_{t-1}) - \frac{\eta_{x,1}}{2}\|\nabla\Phi(\boldsymbol{x}_{t-1})\|^2 - \frac{\eta_{x,1}}{2}(1 - 2\kappa\ell\eta_{x,1})\|\boldsymbol{g}_{t-1}\|^2 \\
&\quad + \frac{3}{2}\eta_{x,1}\ell^2\|\boldsymbol{y}_{t-1}^* - \boldsymbol{y}_{t-1}\|^2 + \frac{3}{2}\eta_{x,1}\alpha^2\ell^2\|\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-2}\|^2 + \frac{3}{2}\eta_{x,1}\alpha^2\ell^2\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2 \\
&\quad + 3\eta_{x,1}(\alpha+1)^2\|\boldsymbol{\delta}_{t-1}^x\|^2 + 3\eta_{x,1}\|\boldsymbol{\delta}_{t-2}^x\|^2 \\
&\le \Phi(\boldsymbol{x}_{t-1}) - \frac{\eta_{x,1}}{2}\|\nabla\Phi(\boldsymbol{x}_{t-1})\|^2 - \frac{\eta_{x,1}}{2}(1 - 2\kappa\ell\eta_{x,1})\|\boldsymbol{g}_{t-1}\|^2 + \frac{3}{2}\eta_{x,1}^3\alpha^2\ell^2\|\boldsymbol{g}_{t-2}\|^2 \\
&\quad + \frac{3}{2}\eta_{x,1}\ell^2\|\boldsymbol{y}_{t-1}^* - \boldsymbol{y}_{t-1}\|^2 + \frac{3}{2}\eta_{x,1}\ell^2\alpha^2\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2 + 3\eta_{x,1}(\alpha+1)^2\|\boldsymbol{\delta}_{t-1}^x\|^2 \\
&\quad + 3\eta_{x,1}\|\boldsymbol{\delta}_{t-2}^x\|^2
\end{aligned} \tag{95}$$

We proceed by taking expectation on both side of Equation 95, to get:

$$\begin{aligned}
\mathbb{E}[\Phi(\boldsymbol{x}_t)] &\le \mathbb{E}[\Phi(\boldsymbol{x}_{t-1})] - \frac{\eta_{x,1}}{2}\mathbb{E}[\|\nabla\Phi(\boldsymbol{x}_{t-1})\|^2] - \frac{\eta_{x,1}}{2}(1 - 2\kappa\ell\eta_{x,1})\mathbb{E}[\|\boldsymbol{g}_{t-1}\|^2] \\
&\quad + \frac{3}{2}\eta_{x,1}^3\alpha^2\ell^2\mathbb{E}[\|\boldsymbol{g}_{t-2}\|^2] + \frac{3}{2}\eta_{x,1}\ell^2\mathbb{E}[\|\boldsymbol{y}_{t-1}^* - \boldsymbol{y}_{t-1}\|^2] \\
&\quad + \frac{3}{2}\eta_{x,1}\alpha^2\ell^2\mathbb{E}[\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2] + 3((1+\alpha)^2 + 1)\eta_{x,1}\frac{\sigma^2}{M_x}
\end{aligned} \tag{96}$$

where we used the fact that $\mathbb{E}[\boldsymbol{\delta}_i^x] \le \frac{\sigma^2}{M_x}$ for all $i \in [T]$.

□

Lemma D.6. *Let $\eta_{y,2} = \frac{1}{6\ell}$, then the following inequality holds true for generalized OGDA iterates:*

$$\begin{aligned}
\sum_{i=1}^{t+1}\mathbb{E}[\|\boldsymbol{y}_i - \boldsymbol{y}_i^*\|^2] &\le \frac{9}{7}\mathbb{E}[\|\boldsymbol{y}_1 - \boldsymbol{y}_1^*\|^2] + \frac{36}{7}\sum_{i=2}^{t+1}\mathbb{E}[\|\boldsymbol{z}_i - \boldsymbol{y}_i^*\|^2] + \frac{18}{7}\eta_{x,1}^2\kappa^2\sum_{i=1}^{t}\mathbb{E}[\|\boldsymbol{g}_i\|^2] \\
&\quad + \frac{2T\sigma^2}{7\ell^2 M_y}
\end{aligned} \tag{97}$$

Proof of Lemma D.6. Using Young's inequality, and κ-Lipschitzness of $\boldsymbol{y}^*(\boldsymbol{x})$ we have:

$$\begin{aligned}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}^*_{t+1}\|^2 &\le 2\|\boldsymbol{y}_{t+1} - \boldsymbol{y}^*_t\|^2 + 2\|\boldsymbol{y}^*_{t+1} - \boldsymbol{y}^*_t\|^2 \\ &\le 2\|\boldsymbol{y}_{t+1} - \boldsymbol{y}^*_t\|^2 + 2\kappa^2\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2\end{aligned} \tag{98}$$

Similar to Lemma A.5, we try to find an upper bound for $\|\boldsymbol{y}_{t+1} - \boldsymbol{y}^*_t\|^2$. Let $\boldsymbol{z}_{t+1} = \boldsymbol{y}_t + \eta_{y,1}\boldsymbol{g}_{y,t} - \eta_{y,2}\boldsymbol{g}_{y,t-1}$, and $\boldsymbol{\delta}^y_i = \boldsymbol{g}_{y,i} - \nabla_y f(\boldsymbol{x}_i, \boldsymbol{y}_i)$. Then we have:

$$\begin{aligned}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}^*_t\|^2 &= \|\boldsymbol{z}_{t+1} - \boldsymbol{y}^*_t + \eta_{y,2}\boldsymbol{g}_{y,t}\|^2 \\ &\le 2\|\boldsymbol{z}_{t+1} - \boldsymbol{y}^*_t\|^2 + 2\eta_{y,2}^2\|\boldsymbol{g}_{y,t}\|^2 \\ &\le 2\|\boldsymbol{z}_{t+1} - \boldsymbol{y}^*_t\|^2 + 4\eta_{y,2}^2\|\nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t)\|^2 + 4\eta_{y,2}^2\|\boldsymbol{\delta}^y_t\|^2 \\ &\le 2\|\boldsymbol{z}_{t+1} - \boldsymbol{y}^*_t\|^2 + 4\eta_{y,2}^2\ell^2\|\boldsymbol{y}_t - \boldsymbol{y}^*_t\|^2 + 4\eta_{y,2}^2\|\boldsymbol{\delta}^y_t\|^2\end{aligned} \tag{99}$$

The rest of the proof is exactly same as proof of Lemma A.5. □

Similar to Lemma A.6, we have:

Lemma D.7. *Let* $\boldsymbol{z}_{t+1} = \boldsymbol{y}_t + \eta_{y,1}\boldsymbol{g}_{y,t} - \eta_{y,2}\boldsymbol{g}_{y,t-1}$, $\boldsymbol{r}_t = \|\boldsymbol{z}_{t+1} - \boldsymbol{y}^*_t\|^2 + \frac{\beta}{4}\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2$ *and* $\eta_{y,2} = \frac{1}{6\ell}$. *Also let* $\frac{\eta_{y,1}}{\eta_{y,2}} = \beta$, *and assume* $\beta \le 1$. *Then OGDA iterates satisfy the following inequalities:*

$$\mathbb{E}[\boldsymbol{r}_t] \le (1 - \frac{\beta}{12\kappa})\mathbb{E}[\boldsymbol{r}_{t-1}] + 12\eta_{x,1}^2\kappa^3\mathbb{E}[\|\boldsymbol{g}_{t-1}\|^2] + \frac{\beta\eta_{x,1}^2}{18}\mathbb{E}[\|\boldsymbol{g}_{t-2}\|^2] + \frac{\beta\sigma^2}{3\ell^2 M_y} \tag{100}$$

$$\sum_{i=1}^{t}\mathbb{E}[\boldsymbol{r}_i] \le \frac{12\kappa}{\beta}\mathbb{E}[\boldsymbol{r}_1] + \frac{2}{3}\kappa\mathbb{E}[\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2] + 145\frac{\eta_{x,1}^2\kappa^4}{\beta}\sum_{i=1}^{t-1}\mathbb{E}[\|\boldsymbol{g}_i\|^2] + \frac{4\kappa\sigma^2(t-1)}{\ell^2 M_y} \tag{101}$$

Proof of Lemma D.7. Let $\boldsymbol{\delta}^y_i = \boldsymbol{g}_{y,i} - \nabla_y f(\boldsymbol{x}_i, \boldsymbol{y}_i)$, and note that we have $\boldsymbol{z}_{t+1} - \boldsymbol{z}_t = \eta_{y,1}\boldsymbol{g}_{y,t}$. We have:

$$\begin{aligned}\|\boldsymbol{z}_{t+1} - \boldsymbol{y}^*_t\|^2 &= \|\boldsymbol{z}_t - \boldsymbol{y}^*_t + \eta_{y,1}\boldsymbol{g}_{y,t}\|^2 \\ &= \|\boldsymbol{z}_t - \boldsymbol{y}^*_t\|^2 + 2\eta_{y,1}\langle\boldsymbol{g}_{y,t}, \boldsymbol{z}_t - \boldsymbol{y}^*_t\rangle + \eta_{y,1}^2\|\boldsymbol{g}_{y,t}\|^2 \\ &= \|\boldsymbol{z}_t - \boldsymbol{y}^*_t\|^2 - 2\eta_{y,1}\eta_{y,2}\langle\boldsymbol{g}_{y,t}, \boldsymbol{g}_{y,t-1}\rangle + 2\eta_{y,1}\langle\boldsymbol{g}_{y,t}, \boldsymbol{y}_t - \boldsymbol{y}^*_t\rangle + \eta_{y,1}^2\|\boldsymbol{g}_{y,t}\|^2 \\ &= \|\boldsymbol{z}_t - \boldsymbol{y}^*_t\|^2 + \eta_{y,1}\eta_{y,2}\|\boldsymbol{g}_{y,t} - \boldsymbol{g}_{y,t-1}\|^2 + 2\eta_{y,1}\langle\boldsymbol{g}_{y,t}, \boldsymbol{y}_t - \boldsymbol{y}^*_t\rangle \\ &\quad - \eta_{y,1}\eta_{y,2}\|\boldsymbol{g}_{y,t-1}\|^2 - \eta_{y,1}(\eta_{y,2} - \eta_{y,1})\|\boldsymbol{g}_{y,t}\|^2 \\ &\le \|\boldsymbol{z}_t - \boldsymbol{y}^*_t\|^2 + 3\eta_{y,1}\eta_{y,2}\|\nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t) - \nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\|^2 \\ &\quad + 2\eta_{y,1}\langle\nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t), \boldsymbol{y}_t - \boldsymbol{y}^*_t\rangle - \eta_{y,1}\eta_{y,2}\|\boldsymbol{g}_{y,t-1}\|^2 - \eta_{y,1}(\eta_{y,2} - \eta_{y,1})\|\boldsymbol{g}_{y,t}\|^2 \\ &\quad + 3\eta_{y,1}\eta_{y,2}\|\boldsymbol{\delta}^y_t\|^2 + 3\eta_{y,1}\eta_{y,2}\|\boldsymbol{\delta}^y_{t-1}\|^2 + 2\eta_{y,1}\langle\boldsymbol{\delta}^y_t, \boldsymbol{y}_t - \boldsymbol{y}^*_t\rangle \\ &\le \|\boldsymbol{z}_t - \boldsymbol{y}^*_t\|^2 + 3\eta_{y,1}\eta_{y,2}\ell^2\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|^2 + 3\eta_{y,1}\eta_{y,2}\ell^2\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 \\ &\quad - 2\eta_{y,1}\mu\|\boldsymbol{y}_t - \boldsymbol{y}^*_t\|^2 - \eta_{y,1}\eta_{y,2}\|\boldsymbol{g}_{y,t-1}\|^2 - \eta_{y,1}(\eta_{y,2} - \eta_{y,1})\|\boldsymbol{g}_{y,t}\|^2 \\ &\quad + 3\eta_{y,1}\eta_{y,2}\|\boldsymbol{\delta}^y_t\|^2 + 3\eta_{y,1}\eta_{y,2}\|\boldsymbol{\delta}^y_{t-1}\|^2 + 2\eta_{y,1}\langle\boldsymbol{\delta}^y_t, \boldsymbol{y}_t - \boldsymbol{y}^*_t\rangle\end{aligned} \tag{102}$$

where the last inequality follows from smoothness of f, and strong concavity of $f(\boldsymbol{x}_t, .)$. Now note that using Young's inequality we can write:

$$\|\boldsymbol{y}_t - \boldsymbol{y}^*_t\|^2 \ge \frac{1}{2}\|\boldsymbol{z}_t - \boldsymbol{y}^*_t\|^2 - \eta_{y,2}^2\|\boldsymbol{g}_{y,t-1}\|^2 \tag{103}$$

Now plugging Equation 103 back to Equation 102, and letting $\eta_{y,1} = \beta\eta_{y,2}$, we have:

$$\begin{aligned}\|\boldsymbol{z}_{t+1} - \boldsymbol{y}^*_t\|^2 &\le (1 - \beta\eta_{y,2}\mu)\|\boldsymbol{z}_t - \boldsymbol{y}^*_t\|^2 + 3\beta\eta_{y,2}^2\ell^2\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|^2 + 3\beta\eta_{y,2}^2\ell^2\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 \\ &\quad - \beta\eta_{y,2}^2(1 - 2\eta_{y,2}\mu)\|\boldsymbol{g}_{y,t-1}\|^2 - \beta\eta_{y,2}^2(1 - \beta)\|\boldsymbol{g}_{y,t}\|^2 \\ &\quad + 3\beta\eta_{y,2}^2\|\boldsymbol{\delta}^y_t\|^2 + 3\beta\eta_{y,2}^2\|\boldsymbol{\delta}^y_{t-1}\|^2 + 2\beta\eta_{y,2}\langle\boldsymbol{\delta}^y_t, \boldsymbol{y}_t - \boldsymbol{y}^*_t\rangle\end{aligned} \tag{104}$$

We can also write:

$$\begin{aligned}
\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 &= \|\eta_{y,1}\boldsymbol{g}_{y,t-1} + \eta_{y,2}(\boldsymbol{g}_{y,t-1} - \boldsymbol{g}_{y,t-2})\|^2 \\
&\leq 2\eta_{y,1}^2\|\boldsymbol{g}_{y,t-1}\|^2 + 2\eta_{y,2}^2\|\boldsymbol{g}_{y,t-1} - \boldsymbol{g}_{y,t-2}\|^2 \\
&\leq 2\eta_{y,1}^2\|\boldsymbol{g}_{y,t-1}\|^2 + 6\eta_{y,2}^2\|\nabla_y f(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}) - \nabla_y f(\boldsymbol{x}_{t-2}, \boldsymbol{y}_{t-2})\|^2 \\
&\quad + 6\eta_{y,2}^2\|\boldsymbol{\delta}_{t-1}^y\|^2 + 6\eta_{y,2}^2\|\boldsymbol{\delta}_{t-2}^y\|^2 \\
&\leq 2\eta_{y,1}^2\|\boldsymbol{g}_{y,t-1}\|^2 + 6\eta_{y,2}^2\ell^2\|\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-2}\|^2 + 6\eta_{y,2}^2\ell^2\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2 \\
&\quad + 6\eta_{y,2}^2\|\boldsymbol{\delta}_{t-1}^y\|^2 + 6\eta_{y,2}^2\|\boldsymbol{\delta}_{t-2}^y\|^2 \\
&= 2\beta^2\eta_{y,2}^2\|\boldsymbol{g}_{y,t-1}\|^2 + 6\eta_{y,2}^2\ell^2\|\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-2}\|^2 + 6\eta_{y,2}^2\ell^2\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2 \\
&\quad + 6\eta_{y,2}^2\|\boldsymbol{\delta}_{t-1}^y\|^2 + 6\eta_{y,2}^2\|\boldsymbol{\delta}_{t-2}^y\|^2
\end{aligned} \tag{105}$$

Now adding $9\beta\eta_{y,2}^2\ell^2\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2$ to both side of Equation 104, and using Equation 105 we have:

$$\begin{aligned}
&\|\boldsymbol{z}_{t+1} - \boldsymbol{y}_t^*\|^2 + 9\beta\eta_y^2\ell^2\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 \leq (1 - \beta\eta_{y,2}\mu)\|\boldsymbol{z}_t - \boldsymbol{y}_t^*\|^2 + 3\beta\eta_{y,1}^2\ell^2\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|^2 \\
&\quad - \beta\eta_{y,2}^2(1 - 2\eta_{y,2}\mu - 24\beta^2\eta_{y,2}^2\ell^2)\|\boldsymbol{g}_{y,t-1}\|^2 - \beta\eta_{y,2}^2(1-\beta)\|\boldsymbol{g}_{y,t}\|^2 \\
&\quad + 72\beta\eta_{y,2}^4\ell^4\|\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-2}\|^2 + 72\beta\eta_{y,2}^4\ell^4\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2 \\
&\quad + 3\beta\eta_{y,2}^2(1 + 24\eta_{y,2}^2\ell^2)\|\boldsymbol{\delta}_t^y\|^2 + 3\beta\eta_{y,2}^2(1 + 24\eta_{y,2}^2\ell^2)\|\boldsymbol{\delta}_{t-1}^y\|^2 \\
&\quad + 2\beta\eta_{y,2}\langle\boldsymbol{\delta}_t^y, \boldsymbol{y}_t - \boldsymbol{y}_t^*\rangle
\end{aligned} \tag{106}$$

Now plugging $\eta_{y,2} = \frac{1}{6\ell}$ into Equation 106, and assuming $\beta \leq 1$ we have:

$$\begin{aligned}
\|\boldsymbol{z}_{t+1} - \boldsymbol{y}_t^*\|^2 + \frac{\beta}{4}\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2 &\leq (1 - \frac{\beta}{6\kappa})\left(\|\boldsymbol{z}_t - \boldsymbol{y}_t^*\|^2\right) + \frac{\beta}{18}\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2 \\
&\quad + \frac{\beta}{12}\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|^2 + \frac{\beta}{18}\|\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-2}\|^2 \\
&\quad + \frac{\beta}{6\ell^2}\|\boldsymbol{\delta}_t^y\|^2 + \frac{\beta}{6\ell^2}\|\boldsymbol{\delta}_{t-1}^y\|^2 + \frac{2\beta}{6\ell}\langle\boldsymbol{\delta}_t^y, \boldsymbol{y}_t - \boldsymbol{y}_t^*\rangle
\end{aligned} \tag{107}$$

Taking expectation from both side of Equation 107, we have:

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{z}_{t+1} - \boldsymbol{y}_t^*\|^2 + \frac{\beta}{4}\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2\right] &\leq (1 - \frac{\beta}{6\kappa})\mathbb{E}\left[\|\boldsymbol{z}_t - \boldsymbol{y}_t^*\|^2\right] + \frac{\beta}{18}\mathbb{E}[\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2] \\
&\quad + \frac{\beta}{12}\mathbb{E}[\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|^2] + \frac{\beta}{18}\mathbb{E}[\|\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-2}\|^2] \\
&\quad + \frac{\beta\sigma^2}{3\ell^2 M_y}
\end{aligned} \tag{108}$$

Also using Young's inequality we have:

$$\|\boldsymbol{z}_t - \boldsymbol{y}_t^*\|^2 \leq (1 + \frac{\beta}{12\kappa})\|\boldsymbol{z}_t - \boldsymbol{y}_{t-1}^*\|^2 + (1 + 12\frac{\kappa}{\beta})\kappa^2\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|^2 \tag{109}$$

where we used the fact that for any $\alpha > 0$, $\|\boldsymbol{x} + \boldsymbol{y}\|^2 \leq (1+\alpha)\|\boldsymbol{x}\|^2 + (1 + \frac{1}{\alpha})\|\boldsymbol{y}\|^2$, and κ-lipschitzness of $\boldsymbol{y}^*(\boldsymbol{x})$. Plugging Equation 109 back to Equation 108, we have:

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{z}_{t+1} - \boldsymbol{y}_t^*\|^2 + \frac{\beta}{4}\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2\right] &\leq (1 - \frac{\beta}{12\kappa})\mathbb{E}\left[\|\boldsymbol{z}_t - \boldsymbol{y}_{t-1}^*\|^2 + \frac{\beta}{4}\mathbb{E}[\|\boldsymbol{y}_{t-1} - \boldsymbol{y}_{t-2}\|^2]\right] \\
&\quad + 12\kappa^3\mathbb{E}[\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|^2] + \frac{\beta}{18}\mathbb{E}[\|\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-2}\|^2] \\
&\quad + \frac{\beta\sigma^2}{3\ell^2 M_y}
\end{aligned} \tag{110}$$

Therefore, if we let $\boldsymbol{r}_t = \|\boldsymbol{z}_{t+1} - \boldsymbol{y}_t^*\|^2 + \frac{\beta}{4}\|\boldsymbol{y}_t - \boldsymbol{y}_{t-1}\|^2$, then we have:

$$\mathbb{E}[\boldsymbol{r}_t] \leq (1 - \frac{\beta}{12\kappa})\mathbb{E}[\boldsymbol{r}_{t-1}] + 12\eta_{x,1}^2\kappa^3\mathbb{E}[\|\boldsymbol{g}_{t-1}\|^2] + \frac{\beta\eta_{x,1}^2}{18}\mathbb{E}[\|\boldsymbol{g}_{t-2}\|^2] + \frac{\beta\sigma^2}{3\ell^2 M_y} \tag{111}$$

We can derive the following equation, by applying Lemma A.2.

$$\begin{aligned}\sum_{i=1}^{t}\mathbb{E}[\boldsymbol{r}_i] \leq \frac{12\kappa}{\beta}\mathbb{E}[\boldsymbol{r}_1] + 144\frac{\eta_{x,1}^2\kappa^4}{\beta}\sum_{i=1}^{t-1}\mathbb{E}[\|\boldsymbol{g}_i\|^2] + \frac{2}{3}\eta_{x,1}^2\kappa\sum_{i=1}^{t-2}\mathbb{E}[\|\boldsymbol{g}_i\|^2] + \frac{2}{3}\kappa\mathbb{E}[\|\boldsymbol{x}_1 - \boldsymbol{x}_0]\|^2 \\ + \frac{4\kappa\sigma^2(t-1)}{\ell^2 M_y}\end{aligned} \tag{112}$$

Or equivalently we have:

$$\sum_{i=1}^{t}\mathbb{E}[\boldsymbol{r}_i] \leq \frac{12\kappa}{\beta}\mathbb{E}[\boldsymbol{r}_1] + \frac{2}{3}\kappa\mathbb{E}[\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2] + 145\frac{\eta_{x,1}^2\kappa^4}{\beta}\sum_{i=1}^{t-1}\mathbb{E}[\|\boldsymbol{g}_i\|^2] + \frac{4\kappa\sigma^2(t-1)}{\ell^2 M_y} \tag{113}$$

□

***Proof of Theorem** D.1.* We begin by taking summation of Equation 92 (Lemma D.5) from $t = 2$ to $t = T$ which yields:

$$\begin{aligned}\frac{\eta_{x,1}}{2}\sum_{i=1}^{T-1}\mathbb{E}[\|\nabla\Phi(\boldsymbol{x}_i)\|^2] \leq \Phi(\boldsymbol{x}_1) - \mathbb{E}[\Phi(\boldsymbol{x}_T)] + \frac{3}{2}\eta_{x,1}\alpha^2\ell^2\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2 \\ - \frac{\eta_{x,1}}{2}(1 - 2\kappa\ell\eta_{x,1})\sum_{i=1}^{T-1}\mathbb{E}[\|\boldsymbol{g}_i\|^2] + \frac{3}{2}\eta_{x,1}^3\alpha^2\ell^2\sum_{i=1}^{T-2}\mathbb{E}[\|\boldsymbol{g}_i\|^2] \\ + \frac{3}{2}\eta_{x,1}\ell^2\sum_{i=1}^{T-1}\|\boldsymbol{y}_i - \boldsymbol{y}_i^*\|^2 + \frac{3}{2}\eta_{x,1}\alpha^2\ell^2\sum_{i=1}^{T-1}\mathbb{E}[\|\boldsymbol{y}_i - \boldsymbol{y}_{i-1}\|^2] \\ + 3((1+\alpha)^2+1)\eta_{x,1}\frac{(T-1)\sigma^2}{M_x}\end{aligned} \tag{114}$$

Now note that if $\eta_x \leq \frac{1}{2\kappa\ell}$ then we can drop $\|\boldsymbol{g}_{T-1}\|^2$ term in above equation. By considering this, and multiplying both sides by $\frac{2}{\eta_{x,1}}$ we get (also let $\Delta = \Phi(\boldsymbol{x}_1) - \min_{\boldsymbol{x}}\Phi(\boldsymbol{x})$) :

$$\begin{aligned}\sum_{i=1}^{T-1}\mathbb{E}[\|\nabla\Phi(\boldsymbol{x}_i)\|^2] \leq \frac{2\Delta}{\eta_{x,1}} + 3\alpha^2\ell^2\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2 \\ - (1 - 2\kappa\ell\eta_{x,1} - 3\eta_{x,1}^2\alpha^2\ell^2)\sum_{i=1}^{T-2}\mathbb{E}[\|\boldsymbol{g}_i\|^2] \\ + 3\ell^2\sum_{i=1}^{T-1}\mathbb{E}[\|\boldsymbol{y}_i^* - \boldsymbol{y}_i\|^2] + 3\alpha^2\ell^2\sum_{i=1}^{T-1}\mathbb{E}[\|\boldsymbol{y}_i - \boldsymbol{y}_{i-1}\|^2] \\ + 6((1+\alpha)^2+1)\frac{(T-1)\sigma^2}{M_x}\end{aligned} \tag{115}$$

We can replace $\sum_{i=1}^{T-1} \|\boldsymbol{y}_i^* - \boldsymbol{y}_i\|^2$ with its upper bound obtained in Lemma D.6 to get:

$$\begin{aligned}\sum_{i=1}^{T-1} \|\nabla\Phi(\boldsymbol{x}_i)\|^2 \leq \frac{2\Delta}{\eta_{x,1}} &+ 3\alpha^2\ell^2\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2 + \frac{27}{7}\ell^2\|\boldsymbol{y}_1 - \boldsymbol{y}_1^*\|^2 \\ &- (1 - 2\kappa\ell\eta_{x,1} - 3\eta_{x,1}^2\alpha^2\ell^2 - \frac{54}{7}\eta_{x,1}^2\kappa^2\ell^2)\sum_{i=1}^{T-2}\mathbb{E}[\|\boldsymbol{g}_i\|^2] \\ &+ \frac{108}{7}\ell^2\sum_{i=2}^{T-1}\mathbb{E}[\|\boldsymbol{z}_i - \boldsymbol{y}_{i-1}^*\|^2] + 3\ell^2\sum_{i=1}^{T-1}\mathbb{E}[\|\boldsymbol{y}_i - \boldsymbol{y}_{i-1}\|^2] \\ &+ 6((1+\alpha)^2+1)\frac{(T-1)\sigma^2}{M_x} + \frac{6}{7}\frac{(T-2)\sigma^2}{M_y}\end{aligned} \tag{116}$$

Now note that $\frac{108}{7}\mathbb{E}[\|\boldsymbol{z}_{i+1} - \boldsymbol{y}_i^*\|^2] + 3\beta\sum_{i=2}^{T-1}\mathbb{E}[\|\boldsymbol{y}_i - \boldsymbol{y}_{i-1}\|^2] \leq 15.5\mathbb{E}[\boldsymbol{r}_i]$. Therefore we have:

$$\begin{aligned}\sum_{i=1}^{T-1} \|\nabla\Phi(\boldsymbol{x}_i)\|^2 \leq \frac{2\Delta}{\eta_{x,1}} &+ 3\alpha^2\ell^2\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2 + \frac{27}{7}\ell^2\|\boldsymbol{y}_1 - \boldsymbol{y}_1^*\|^2 \\ &- (1 - 2\kappa\ell\eta_{x,1} - 3\eta_{x,1}^2\alpha^2\ell^2 - \frac{54}{7}\eta_{x,1}^2\kappa^2\ell^2)\sum_{i=1}^{T-2}\mathbb{E}[\|\boldsymbol{g}_i\|^2] \\ &+ 15.5\ell^2\sum_{i=1}^{T-1}\mathbb{E}[\boldsymbol{r}_i] + 6((1+\alpha)^2+1)\frac{(T-1)\sigma^2}{M_x} + \frac{6}{7}\frac{(T-2)\sigma^2}{M_y}\end{aligned} \tag{117}$$

Furthermore, using Lemma D.7, we can find an upper bound on $\sum_{i=1}^{T-1}\mathbb{E}[\boldsymbol{r}_i]$, and replacing it in above equation yields:

$$\begin{aligned}\sum_{i=1}^{T-1} \|\nabla\Phi(\boldsymbol{x}_i)\|^2 \leq \frac{2\Delta}{\eta_{x,1}} &+ 186\frac{\kappa\ell^2}{\beta}\mathbb{E}[\boldsymbol{r}_1] + 11\kappa\ell^2\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2 + 3\alpha^2\ell^2\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2 \\ &+ \frac{27}{7}\ell^2\|\boldsymbol{y}_1 - \boldsymbol{y}_1^*\|^2 - (1 - 2\kappa\ell\eta_{x,1} - 3\eta_{x,1}^2\alpha^2\ell^2 - \frac{54}{7}\eta_{x,1}^2\kappa^2\ell^2 - 2248\eta_{x,1}^2\frac{\kappa^4\ell^2}{\beta})\sum_{i=1}^{T-2}\mathbb{E}[\|\boldsymbol{g}_i\|^2] \\ &+ \frac{62\kappa\sigma^2(T-2)}{M_y} + 6((1+\alpha)^2+1)\frac{(T-1)\sigma^2}{M_x} + \frac{6}{7}\frac{(T-2)\sigma^2}{M_y}\end{aligned} \tag{118}$$

By letting $\eta_{x,1} = \frac{\sqrt{\beta}}{50\kappa^2\ell}$, and $\eta_{x,2} \leq \frac{1}{25\ell}$, it holds that $-(1 - 2\kappa\ell\eta_{x,1} - 3\eta_{x,1}^2\alpha^2\ell^2 - \frac{54}{7}\eta_{x,1}^2\kappa^2\ell^2 - 2248\eta_{x,1}^2\frac{\kappa^4\ell^2}{\beta})\sum_{i=1}^{T-2}\mathbb{E}[\|\boldsymbol{g}_i\|^2] \leq 0$. Therefore, with the choice of letting rate $\eta_{x,1} = \frac{\sqrt{\beta}}{50\kappa^2\ell}$ and simplifying the terms, we have:

$$\begin{aligned}\frac{1}{T-1}\sum_{i=1}^{T-1}\mathbb{E}[\|\nabla\Phi(\boldsymbol{x}_i)\|^2] \leq 100\frac{\kappa^2\ell\Delta}{\sqrt{\beta}(T-1)} &+ 186\frac{\kappa\ell^2}{\beta(T-1)}\|\boldsymbol{y}_1 - \boldsymbol{y}_1^* + \eta_{y,1}\boldsymbol{g}_{y,1} - \eta_{y,2}\boldsymbol{g}_{y,0}\|^2 \\ &+ 47\beta\frac{\kappa\ell^2}{T-1}\|\boldsymbol{y}_1 - \boldsymbol{y}_0\|^2 + \frac{(11\kappa + 3\alpha^2)\ell^2}{T-1}\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2 \\ &+ \frac{27}{7}\frac{\ell^2}{T-1}\|\boldsymbol{y}_1 - \boldsymbol{y}_1^*\|^2 + \frac{63\kappa\sigma^2}{M_y} + 6((1+\alpha)^2+1)\frac{\sigma^2}{M_x}\end{aligned} \tag{119}$$

Using Young's inequality, and ℓ-smoothness of f, we have:

$$\begin{aligned}\|\boldsymbol{y}_1 - \boldsymbol{y}_1^* + \eta_{y,1}\boldsymbol{g}_{y,1} - \eta_{y,2}\boldsymbol{g}_{y,0}\|^2 &\leq 2\|\boldsymbol{y}_1 - \boldsymbol{y}_1^*\|^2 + 2\|\eta_{y,2}(\boldsymbol{g}_{y,1} - \boldsymbol{g}_{y,0}) + \eta_{y,2}(\beta - 1)\boldsymbol{g}_{y,1}\|^2 \\ &\leq 2\|\boldsymbol{y}_1 - \boldsymbol{y}_1^*\|^2 + \frac{1}{9}\|x_1 - x_0\|^2 + \frac{1}{9}\|y_1 - y_0\|^2 + \frac{1-\beta}{9}\|\boldsymbol{y}_1 - \boldsymbol{y}_1^*\|^2\end{aligned} \tag{120}$$

Plugging this into Equation 119, we have:

$$\begin{aligned}\frac{1}{T-1}\sum_{i=1}^{T-1}\mathbb{E}[\|\nabla\Phi(\boldsymbol{x}_i)\|^2] \leq 100\frac{\kappa^2\ell\Delta}{T-1} + 376\frac{\kappa\ell^2}{\beta(T-1)}\|\boldsymbol{y}_1 - \boldsymbol{y}_1^*\|^2 \\ + 68\frac{\kappa\ell^2}{\beta(T-1)}\|\boldsymbol{y}_1 - \boldsymbol{y}_0\|^2 + \frac{(32\kappa + 3\alpha^2)\ell^2}{\beta(T-1)}\|\boldsymbol{x}_1 - \boldsymbol{x}_0\|^2 \\ + \frac{63\kappa\sigma^2}{M_y} + 6((1+\alpha)^2+1)\frac{\sigma^2}{M_x}\end{aligned} \tag{121}$$

which completes the proof as stated. □