Appendix

A. Detailed Convergence Analysis

A.1. Table of Parameters

In Table A1, we summarize the problem and algorithmic parameters used in our convergence analysis.

Table A1: Summary of problem and algorithmic parameters and their descriptions.

parameter	description
d	# of optimization variables
b	mini-batch size
q	# of random direction vectors used in ZO gradient estimation
α	learning rate for ZO-PGD
β	learning rate for ZO-PGA
γ	strongly concavity parameter of $f(\mathbf{x}, \mathbf{y})$ with respect to $\mathbf{y}$
η	upper bound on the gradient norm, implying Lipschitz continuity
L_x, L_y	Lipschitz continuous gradient constant of $f(\mathbf{x}, \mathbf{y})$ with respect to $\mathbf{x}$ and $\mathbf{y}$ respectively
R	diameter of the compact convex set $\mathcal{X}$ or $\mathcal{Y}$
f^*	lower bound on the function value, implying feasibility
σ_x^2, σ_y^2	variances of ZO gradient estimator for variables $\mathbf{x}$ and $\mathbf{y}$, bounded by (3)

A.2. Proof of Lemma 1

Before going into the proof, let's review some preliminaries and give some definitions. Define $h_\mu(\mathbf{x}, \boldsymbol{\xi})$ to be the smoothed version of $h(\mathbf{x}, \boldsymbol{\xi})$ and since $\boldsymbol{\xi}$ models a subsampling process over a finite number of candidate functions, we can further have $h_\mu(\mathbf{x}) \triangleq \mathbb{E}_{\boldsymbol{\xi}}[h_\mu(\mathbf{x}, \boldsymbol{\xi})]$ and $\nabla_{\mathbf{x}} h_\mu(\mathbf{x}) = \mathbb{E}_{\boldsymbol{\xi}}[\nabla_{\mathbf{x}} h_\mu(\mathbf{x}, \boldsymbol{\xi})]$

Recall that in the finite sum setting when $\boldsymbol{\xi}_j$ parameterizes the jth function, the gradient estimator is given by

$$\widehat{\nabla}_{\mathbf{x}} h(\mathbf{x}) = \frac{1}{bq} \sum_{j \in \mathcal{I}} \sum_{i=1}^{q} \frac{d[h(\mathbf{x} + \mu \mathbf{u}_i; \boldsymbol{\xi}_j) - h(\mathbf{x}; \boldsymbol{\xi}_j)]}{\mu} \mathbf{u}_i. \tag{15}$$

where $\mathcal{I}$ is a set with b elements, containing the indices of functions selected for gradient evaluation.

From standard result of the zeroth order gradient estimator, we know

$$\mathbb{E}_{\mathcal{I}} \left[\mathbb{E}_{\mathbf{u}_i, i \in [q]} \left[\widehat{\nabla}_{\mathbf{x}} h(\mathbf{x}) \right] | \mathcal{I} \right] = \mathbb{E}_{\mathcal{I}} \left[\frac{1}{b} \sum_{j \in \mathcal{I}} \nabla_{\mathbf{x}} f_\mu(\mathbf{x}, \boldsymbol{\xi}_j) \right] = \nabla_{\mathbf{x}} h_\mu(\mathbf{x}). \tag{16}$$

Now let's go into the proof. First, we have

$$\begin{aligned}
&\mathbb{E} \left[\| \widehat{\nabla}_{\mathbf{x}} h(\mathbf{x}) - \nabla_{\mathbf{x}} h_\mu(\mathbf{x}) \|_2^2 \right] \\
=&\mathbb{E}_{\mathcal{I}} \left[\mathbb{E}_{\mathbf{u}_i, i \in [q]} \left[\left\| \widehat{\nabla}_{\mathbf{x}} h(\mathbf{x}) - \frac{1}{b} \sum_{j \in \mathcal{I}} \nabla_{\mathbf{x}} f_\mu(\mathbf{x}, \boldsymbol{\xi}_j) + \frac{1}{b} \sum_{j \in \mathcal{I}} \nabla_{\mathbf{x}} f_\mu(\mathbf{x}, \boldsymbol{\xi}_j) - \nabla_{\mathbf{x}} h_\mu(\mathbf{x}) \right\|_2^2 \Big| \mathcal{I} \right] \right] \\
\leq& 2\mathbb{E}_{\mathcal{I}} \left[\mathbb{E}_{\mathbf{u}_i, i \in [q]} \left[\left\| \widehat{\nabla}_{\mathbf{x}} h(\mathbf{x}) - \frac{1}{b} \sum_{j \in \mathcal{I}} \nabla_{\mathbf{x}} f_\mu(\mathbf{x}, \boldsymbol{\xi}_j) \right\|_2^2 + \left\| \frac{1}{b} \sum_{j \in \mathcal{I}} \nabla_{\mathbf{x}} f_\mu(\mathbf{x}, \boldsymbol{\xi}_j) - \nabla_{\mathbf{x}} h_\mu(\mathbf{x}) \right\|_2^2 \Big| \mathcal{I} \right] \right].
\end{aligned} \tag{17}$$

Further, by definition, given $\mathcal{I}$, $\widehat{\nabla}_{\mathbf{x}} h(\mathbf{x})$ is the average of ZO gradient estimates under q *i.i.d.* random directions, each of which has the mean $\frac{1}{b} \sum_{j \in \mathcal{I}} \nabla_{\mathbf{x}} f_\mu(\mathbf{x}, \boldsymbol{\xi}_j)$.

Thus for the first term at the right-hand-side (RHS) of the above inequality, we have

$$\mathbb{E}_{\mathbf{u}_i, i\in[q]}\left[\left\|\widehat{\nabla}_{\mathbf{x}} h(\mathbf{x}) - \frac{1}{b}\sum_{j\in\mathcal{I}} \nabla_{\mathbf{x}} f_\mu(\mathbf{x}, \boldsymbol{\xi}_j)\right\|_2^2 \Big| \mathcal{I}\right] \leq \frac{1}{q}\left(2d\left\|\frac{1}{b}\sum_{j\in\mathcal{I}} \nabla_{\mathbf{x}} f(\mathbf{x}, \boldsymbol{\xi}_j)\right\|^2 + \frac{\mu^2 L_h^2 d^2}{2}\right)$$
$$\leq \frac{1}{q}\left(2d\eta^2 + \frac{\mu^2 L_h^2 d^2}{2}\right) \tag{18}$$

where the first inequality is by the standard bound of the variance of zeroth order estimator and the second inequality is by the assumption that $\|\nabla_{\mathbf{x}} h(x;\boldsymbol{\xi})\|^2 \leq \eta^2$ and thus $\|\frac{1}{b}\sum_{j\in\mathcal{I}} \nabla_{\mathbf{x}} f(\mathbf{x}, \boldsymbol{\xi}_j)\|^2 \leq \eta^2$.

In addition, we have

$$\mathbb{E}_{\mathcal{I}}\left[\mathbb{E}_{\mathbf{u}_i, i\in[q]}\left[\left\|\frac{1}{b}\sum_{j\in\mathcal{I}} \nabla_{\mathbf{x}} f_\mu(\mathbf{x}, \boldsymbol{\xi}_j) - \nabla_{\mathbf{x}} h_\mu(\mathbf{x})\right\|_2^2 \Big| \mathcal{I}\right]\right]$$
$$= \mathbb{E}_{\mathcal{I}}\left[\left\|\frac{1}{b}\sum_{j\in\mathcal{I}} \nabla_{\mathbf{x}} f_\mu(\mathbf{x}, \boldsymbol{\xi}_j) - \nabla_{\mathbf{x}} h_\mu(\mathbf{x})\right\|_2^2\right]$$
$$= \frac{1}{b}\mathbb{E}_{\xi}\left[\|\nabla_{\mathbf{x}} f_\mu(\mathbf{x}, \boldsymbol{\xi}) - \nabla_{\mathbf{x}} h_\mu(\mathbf{x})\|_2^2\right] \leq \frac{\eta^2}{b} \tag{19}$$

where the second equality is because ξ_j are i.i.d. draws from the same distribution as ξ and $\mathbb{E}[\nabla_{\mathbf{x}} f_\mu(\mathbf{x}, \boldsymbol{\xi})] = \nabla_{\mathbf{x}} h_\mu(\mathbf{x})$, the last inequality is because $\|\nabla_{\mathbf{x}} f_\mu(\mathbf{x}, \boldsymbol{\xi})\|_2^2 \leq \eta^2$ by assumption. Substituting (18) and (19) into (17) finishes the proof. □

A.3. Convergence Analysis of ZO-Min-Max by Performing PGA

In this section, we will provide the details of the proofs. Before proceeding, we have the following illustration, which will be useful in the proof.

The order of taking expectation: Since iterates $\mathbf{x}^{(t)}, \mathbf{y}^{(t)}, \forall t$ are random variables, we need to define

$$\mathcal{F}^{(t)} = \{\mathbf{x}^{(t)}, \mathbf{y}^{(t)}, \mathbf{x}^{(t-1)}, \mathbf{y}^{(t-1)}, \ldots, \mathbf{x}^{(1)}, \mathbf{y}^{(1)}\} \tag{20}$$

as the history of the iterates. Throughout the theoretical analysis, taking expectation means that we take expectation over random variable at the tth iteration conditioned on $\mathcal{F}^{(t-1)}$ and then take expectation over $\mathcal{F}^{(t-1)}$.

Subproblem: Also, it is worthy noting that performing (4) and (5) are equivalent to the following optimization problem:

$$\mathbf{x}^{(t)} = \min_{\mathbf{x}\in\mathcal{X}} \left\langle \widehat{\nabla}_{\mathbf{x}} f(\mathbf{x}^{(t-1)}, \mathbf{y}^{(t-1)}), \mathbf{x} - \mathbf{x}^{(t-1)} \right\rangle + \frac{1}{2\alpha}\|\mathbf{x} - \mathbf{x}^{(t-1)}\|^2, \tag{21}$$

$$\mathbf{y}^{(t)} = \max_{\mathbf{y}\in\mathcal{Y}} \left\langle \widehat{\nabla}_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{y} - \mathbf{y}^{(t-1)} \right\rangle - \frac{1}{2\beta}\|\mathbf{y} - \mathbf{y}^{(t-1)}\|^2. \tag{22}$$

When $f(\mathbf{x}, \mathbf{y})$ is white-box w.r.t. $\mathbf{y}$, (22) becomes

$$\mathbf{y}^{(t)} = \max_{\mathbf{y}\in\mathcal{Y}} \left\langle \nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{y} - \mathbf{y}^{(t-1)} \right\rangle - \frac{1}{2\beta}\|\mathbf{y} - \mathbf{y}^{(t-1)}\|^2. \tag{23}$$

In the proof of ZO-Min-Max, we will use the optimality condition of these two problems to derive the descent lemmas.

Relationship with smoothing function We denote by $f_{\mu,x}(\mathbf{x}, \mathbf{y})$ the smoothing version of f w.r.t. $\mathbf{x}$ with parameter $\mu > 0$. The similar definition holds for $f_{\mu,y}(\mathbf{x}, \mathbf{y})$. By taking $f_{\mu,x}(\mathbf{x}, \mathbf{y})$ as an example, under **A2** f and $f_{\mu,x}$ has the

following relationship (Gao et al., 2014, Lemma 4.1):

$$|f_{\mu,\mathbf{x}}(\mathbf{x},\mathbf{y}) - f(\mathbf{x},\mathbf{y}))| \leq \frac{L_x\mu^2}{2} \quad \text{and} \quad \|\nabla_{\mathbf{x}} f_{\mu,\mathbf{x}}(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})\|_2^2 \leq \frac{\mu^2 d^2 L_x^2}{4}, \tag{24}$$

$$|f_{\mu,\mathbf{y}}(\mathbf{x},\mathbf{y}) - f(\mathbf{x},\mathbf{y}))| \leq \frac{L_y\mu^2}{2} \quad \text{and} \quad \|\nabla_{\mathbf{y}} f_{\mu,\mathbf{y}}(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y})\|_2^2 \leq \frac{\mu^2 d^2 L_y^2}{4}. \tag{25}$$

First, we will show the descent lemma in minimization as follows.

A.3.1. Proof of Lemma 2

Proof: Since $f(\mathbf{x},\mathbf{y})$ has L_x Lipschtiz continuous gradients with respect to $\mathbf{x}$, we have

$$\begin{aligned} f_\mu(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)}) \leq & f_\mu(\mathbf{x}^{(t)},\mathbf{y}^{(t)}) + \langle \nabla_{\mathbf{x}} f_\mu(\mathbf{x}^{(t)},\mathbf{y}^{(t)}), \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \rangle + \frac{L_x}{2} \| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \|^2 \\ = & f_\mu(\mathbf{x}^{(t)},\mathbf{y}^{(t)}) + \langle \widehat{\nabla}_{\mathbf{x}} f(\mathbf{x}^{(t)},\mathbf{y}^{(t)}), \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \rangle + \frac{L_x}{2} \| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \|^2 \\ & + \langle \nabla_{\mathbf{x}} f_\mu(\mathbf{x}^{(t)},\mathbf{y}^{(t)}) - \widehat{\nabla}_{\mathbf{x}} f(\mathbf{x}^{(t)},\mathbf{y}^{(t)}), \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \rangle. \end{aligned} \tag{26}$$

Recall that

$$\mathbf{x}^{(t+1)} = \mathrm{proj}_{\mathcal{X}}(\mathbf{x}^{(t)} - \alpha \widehat{\nabla}_{\mathbf{x}} f(\mathbf{x}^{(t)},\mathbf{y}^{(t)})), \tag{27}$$

From the optimality condition of $\mathbf{x}$-subproblem (21), we have

$$\langle \widehat{\nabla}_{\mathbf{x}} f(\mathbf{x}^{(t)},\mathbf{y}^{(t)}), \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \rangle \leq -\frac{1}{\alpha} \| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \|^2. \tag{28}$$

Here we use the fact that the optimality condition of problem (21) at the solution $\mathbf{x}^{(t+1)}$ yields $\langle \widehat{\nabla}_{\mathbf{x}} f(\mathbf{x}^{(t)},\mathbf{y}^{(t)}) + (\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)})/\alpha, \mathbf{x}^{(t+1)} - \mathbf{x} \rangle \leq 0$ for any $\mathbf{x} \in \mathcal{X}$. By setting $\mathbf{x} = \mathbf{x}^{(t)}$, we obtain (28).

In addition, we define another iterate generated by $\nabla_{\mathbf{x}} f_\mu(\mathbf{x}^{(t)},\mathbf{y}^{(t)})$

$$\widehat{\mathbf{x}}^{(t+1)} = \mathrm{proj}_{\mathcal{X}}(\mathbf{x}^{(t)} - \alpha \nabla_{\mathbf{x}} f_\mu(\mathbf{x}^{(t)},\mathbf{y}^{(t)})). \tag{29}$$

Then, we can have

$$\begin{aligned} & \langle \nabla_{\mathbf{x}} f_\mu(\mathbf{x}^{(t)},\mathbf{y}^{(t)}) - \widehat{\nabla}_{\mathbf{x}} f(\mathbf{x}^{(t)},\mathbf{y}^{(t)}), \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \rangle \\ = & \langle \nabla_{\mathbf{x}} f_\mu(\mathbf{x}^{(t)},\mathbf{y}^{(t)}) - \widehat{\nabla}_{\mathbf{x}} f(\mathbf{x}^{(t)},\mathbf{y}^{(t)}), \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} - (\widehat{\mathbf{x}}^{(t+1)} - \mathbf{x}^{(t)}) \rangle \\ & + \langle \nabla_{\mathbf{x}} f_\mu(\mathbf{x}^{(t)},\mathbf{y}^{(t)}) - \widehat{\nabla}_{\mathbf{x}} f(\mathbf{x}^{(t)},\mathbf{y}^{(t)}), \widehat{\mathbf{x}}^{(t+1)} - \mathbf{x}^{(t)} \rangle. \end{aligned} \tag{30}$$

Due to the fact that $\mathbb{E}_{\mathbf{u}}[\widehat{\nabla}_{\mathbf{x}} f(\mathbf{x}^{(t)},\mathbf{y}^{(t)})] = \nabla_{\mathbf{x}} f_\mu(\mathbf{x}^{(t)},\mathbf{y}^{(t)})$, we further have

$$\mathbb{E}_{\mathbf{u}}[\langle \nabla_{\mathbf{x}} f_\mu(\mathbf{x}^{(t)},\mathbf{y}^{(t)}) - \widehat{\nabla}_{\mathbf{x}} f(\mathbf{x}^{(t)},\mathbf{y}^{(t)}), \widehat{\mathbf{x}}^{(t+1)} - \mathbf{x}^{(t)} \rangle] = 0. \tag{31}$$

Finally, we also have

$$\begin{aligned} & \langle \nabla_{\mathbf{x}} f_\mu(\mathbf{x}^{(t)},\mathbf{y}^{(t)}) - \widehat{\nabla}_{\mathbf{x}} f(\mathbf{x}^{(t)},\mathbf{y}^{(t)}), \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} - (\widehat{\mathbf{x}}^{(t+1)} - \mathbf{x}^{(t)}) \rangle \\ \leq & \frac{\alpha}{2} \| \nabla_{\mathbf{x}} f_\mu(\mathbf{x}^{(t)},\mathbf{y}^{(t)}) - \widehat{\nabla}_{\mathbf{x}} f(\mathbf{x}^{(t)},\mathbf{y}^{(t)}) \|^2 + \frac{1}{2\alpha} \| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} - (\widehat{\mathbf{x}}^{(t+1)} - \mathbf{x}^{(t)}) \|^2 \\ \leq & \alpha \| \nabla_{\mathbf{x}} f_\mu(\mathbf{x}^{(t)},\mathbf{y}^{(t)}) - \widehat{\nabla}_{\mathbf{x}} f(\mathbf{x}^{(t)},\mathbf{y}^{(t)}) \|^2 \end{aligned} \tag{32}$$

where the first inequality is due to Young's inequality, the second inequality is due to non-expansiveness of the projection operator. Thus

$$\begin{aligned} & \mathbb{E}_{\mathbf{u}}[\langle \nabla_{\mathbf{x}} f_\mu(\mathbf{x}^{(t)},\mathbf{y}^{(t)}) - \widehat{\nabla}_{\mathbf{x}} f(\mathbf{x}^{(t)},\mathbf{y}^{(t)}), \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} - (\widehat{\mathbf{x}}^{(t+1)} - \mathbf{x}^{(t)}) \rangle] \\ \leq & \mathbb{E}_{\mathbf{u}}[\alpha \| \nabla_{\mathbf{x}} f_\mu(\mathbf{x}^{(t)},\mathbf{y}^{(t)}) - \widehat{\nabla}_{\mathbf{x}} f(\mathbf{x}^{(t)},\mathbf{y}^{(t)}) \|^2] \leq \alpha \sigma_x^2 \end{aligned} \tag{33}$$

where $\sigma_x^2 := \sigma^2(L_x, b, q, d)$ which was defined in (3).

Combining all above, we have

$$\mathbb{E}[f_\mu(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)})] \leq \mathbb{E}[f_\mu(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})] - \left(\frac{1}{\alpha} - \frac{L_x}{2}\right) \| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \|^2 + \alpha\sigma^2, \tag{34}$$

and we request $\alpha \leq 1/L_x$, which completes the proof.

Using $|f_{\mu,x}(\mathbf{x}, \mathbf{y}) - f(\mathbf{x}, \mathbf{y}))| \leq \frac{L_x\mu^2}{2}$, we can get

$$\mathbb{E}[f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)})] - \frac{L_x\mu^2}{2} \leq \mathbb{E}[f_\mu(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)})] \leq \mathbb{E}[f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)})] + \frac{L_x\mu^2}{2}, \tag{35}$$

so we are able to obtain from (3)

$$\mathbb{E}[f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)})] \leq \mathbb{E}[f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})] - \left(\frac{1}{\alpha} - \frac{L_x}{2}\right) \| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \|^2 + \alpha\sigma_x^2 + L_x\mu^2. \tag{36}$$

□

Corollary 1.

$$\mathbb{E}\left\langle \widehat{\nabla} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}) - \nabla f_\mu(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \right\rangle \leq \beta\sigma_y^2 \tag{37}$$

$\sigma_y^2 := \sigma^2(L_y, b, q, d)$ *which was defined in* (3).

Proof:

Define

$$\widetilde{\mathbf{y}}^{(t)} = \mathrm{proj}_{\mathcal{Y}}(\mathbf{y}^{(t)} - \beta\nabla_{\mathbf{y}} f_\mu(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)})), \tag{38}$$

we have

$$\begin{aligned}
&\langle \nabla_{\mathbf{y}} f_\mu(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}) - \widehat{\nabla}_{\mathbf{x}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \rangle \\
=&\langle \nabla_{\mathbf{y}} f_\mu(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}) - \widehat{\nabla}_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} - (\widetilde{\mathbf{y}}^{(t)} - \mathbf{y}^{(t-1)}) \rangle \\
&+ \langle \nabla_{\mathbf{y}} f_\mu(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}) - \widehat{\nabla}_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \widetilde{\mathbf{y}}^{(t)} - \mathbf{y}^{(t-1)} \rangle.
\end{aligned} \tag{39}$$

Due to the fact that $\mathbb{E}_{\mathbf{u}}[\widehat{\nabla}_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)})] = \nabla_{\mathbf{y}} f_\mu(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)})$, we further have

$$\mathbb{E}_{\mathbf{u}}[\langle \nabla_{\mathbf{y}} f_\mu(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}) - \widehat{\nabla}_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \widetilde{\mathbf{y}}^{(t)} - \mathbf{y}^{(t-1)} \rangle] = 0. \tag{40}$$

Finally, we also have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{u}}[\langle \nabla_{\mathbf{y}} f_\mu(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}) - \widehat{\nabla}_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} - (\widetilde{\mathbf{y}}^{(t)} - \mathbf{y}^{(t-1)}) \rangle] \\
\leq&\mathbb{E}_{\mathbf{u}}[\frac{\beta}{2} \| \langle \nabla_{\mathbf{y}} f_\mu(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}) - \widehat{\nabla}_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}) \|^2 + \frac{1}{2\beta} \| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} - (\widetilde{\mathbf{y}}^{(t)} - \mathbf{y}^{(t-1)}) \|^2] \\
\leq&\mathbb{E}_{\mathbf{u}}[\beta \| \nabla_{\mathbf{y}} f_\mu(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}) - \widehat{\nabla}_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}) \|^2] \leq \beta\sigma_y^2
\end{aligned} \tag{41}$$

where $\sigma_y^2 := \sigma^2(L_y, b, q, d)$ which was defined in (3).

□

Next, before showing the proof of Lemma 3, we need the following lemma to show the recurrence of the size of the successive difference between two iterations.

Lemma 5. *Under assumption 1, assume iterates* $\mathbf{x}^{(t)}, \mathbf{y}^{(t)}$ *generated by algorithm 1. When* $f(\mathbf{x}^{(t)}, \mathbf{y})$ *is white-box, we have*

$$\begin{aligned}
\frac{2}{\beta^2\gamma}\mathbb{E}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 - \frac{2}{\beta^2\gamma}\mathbb{E}\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 \leq& \frac{2L_x^2}{\beta\gamma^2}\mathbb{E}\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \|^2 \\
&+ \frac{2}{\beta}\mathbb{E}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 - \left(\frac{4}{\beta} - \frac{2L_y^2}{\gamma}\right) \mathbb{E}\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2.
\end{aligned} \tag{42}$$

Proof: from the optimality condition of $\mathbf{y}$-subproblem (23) at iteration t and $t-1$, we have the following two inequalities:

$$-\langle \nabla_{\mathbf{y}} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) - \frac{1}{\beta}(\mathbf{y}^{(t+1)} - \mathbf{y}^{(t)}), \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \rangle \leq 0, \tag{43}$$

$$\langle \nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}) - \frac{1}{\beta}(\mathbf{y}^{(t)} - \mathbf{y}^{(t-1)}), \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \rangle \leq 0. \tag{44}$$

Adding the above inequalities, we can get

$$\begin{aligned}\frac{1}{\beta}\langle \mathbf{v}^{(t+1)}, \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \rangle \leq & \left\langle \nabla_{\mathbf{y}} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) - \nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}), \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \right\rangle \\ & + \left\langle \nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \right\rangle \end{aligned} \tag{45}$$

where $\mathbf{v}^{(t+1)} = \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} - (\mathbf{y}^{(t)} - \mathbf{y}^{(t-1)})$.

According to the quadrilateral indentity, we know

$$\left\langle \mathbf{v}^{(t+1)}, \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \right\rangle = \frac{1}{2}\left(\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 + \| \mathbf{v}^{(t+1)} \|^2 - \| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 \right). \tag{46}$$

Based on the definition of $\mathbf{v}^{(t+1)}$, we substituting (46) into (45), which gives

$$\begin{aligned}\frac{1}{2\beta}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 \leq & \frac{1}{2\beta}\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 - \frac{1}{2\beta}\| \mathbf{v}^{(t+1)} \|^2 \\ & + \left\langle \nabla_{\mathbf{y}} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) - \nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}), \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \right\rangle \\ & + \left\langle \nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \right\rangle \end{aligned} \tag{47}$$

$$\begin{aligned}\overset{(a)}{\leq} & \frac{1}{2\beta}\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 + \left\langle \nabla_{\mathbf{y}} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) - \nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}), \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \right\rangle \\ & + \frac{\beta L_y^2}{2}\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 - \gamma \| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 \\ \overset{(b)}{\leq} & \frac{1}{2\beta}\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 + \frac{\gamma}{2}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 \\ & + \frac{L_x^2}{2\gamma}\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \|^2 - (\gamma - \frac{\beta L_y^2}{2})\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 \end{aligned} \tag{48}$$

where in (a) we use the strong concavity of function $f(\mathbf{x}, \mathbf{y})$ in $\mathbf{y}$ (with parameter $\gamma > 0$) and Young's inequality, i.e.,

$$\begin{aligned} & \langle \nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \rangle \\ = & \langle \nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{v}^{(t+1)} + \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \rangle \\ \leq & \frac{\beta L_y^2}{2}\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 + \frac{1}{2\beta}\| \mathbf{v}^{(t+1)} \|^2 - \gamma \| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 \end{aligned} \tag{49}$$

and in (b) we apply the Young's inequality, i.e.,

$$\left\langle \nabla_{\mathbf{y}} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) - \nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}), \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \right\rangle \leq \frac{L_x^2}{2\gamma}\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \|^2 + \frac{\gamma}{2}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2. \tag{50}$$

Therefore, we have

$$\begin{aligned}\frac{1}{2\beta}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 \leq & \frac{1}{2\beta}\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 + \frac{L_x^2}{2\gamma}\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \|^2 \\ & + \frac{\gamma}{2}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 - \left(\gamma - \frac{\beta L_y^2}{2} \right) \| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2, \end{aligned} \tag{51}$$

which implies

$$\frac{2}{\beta^2\gamma}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 \leq \frac{2}{\beta^2\gamma}\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 + \frac{2L_x^2}{\beta\gamma^2}\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \|^2 + \frac{2}{\beta}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 - \left(\frac{4}{\beta} - \frac{2L_y^2}{\gamma}\right)\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2. \tag{52}$$

By taking the expectation on both sides of (52), we can get the results of Lemma 5. □

Lemma 5 basically gives the recursion of $\|\Delta_y^{(t)}\|^2$. It can be observed that term $(4/\beta - 2L_y^2/\gamma)\|\Delta_{\mathbf{y}}^{(t)}\|$ provides the descent of the recursion when β is small enough, which will take an important role in the proof of Lemma 3 when we quantify the descent in maximization.

Then, we can quantify the descent of the objective value by the following descent lemma.

A.3.2. Proof of Lemma 3

Proof: let $f'(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)}) = f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)}) - 1(\mathbf{y}^{(t+1)})$ and $1(\mathbf{y})$ denote the indicator function with respect to the constraint of $\mathbf{y}$. From the optimality condition of sub-problem $\mathbf{y}$ in (22), we have

$$\nabla_{\mathbf{y}} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) - \frac{1}{\beta}(\mathbf{y}^{(t+1)} - \mathbf{y}^{(t)}) - \xi^{(t+1)} = 0 \tag{53}$$

where $\xi^{(t)}$ denote the subgradient of $1(\mathbf{y}^{(t)})$. Since function $f'(\mathbf{x}, \mathbf{y})$ is concave with respect to $\mathbf{y}$, we have

$$\begin{aligned} & f'(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)}) - f'(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) \leq \langle \nabla_{\mathbf{y}} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}), \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \rangle - \langle \xi^{(t)}, \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \rangle \\ \overset{(a)}{=} & \frac{1}{\beta}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 - \langle \xi^{(t)} - \xi^{(t+1)}, \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \rangle \\ = & \frac{1}{\beta}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 + \left\langle \nabla_{\mathbf{y}} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) - \nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \right\rangle \\ & - \frac{1}{\beta}\left\langle \mathbf{v}^{(t+1)}, \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \right\rangle \end{aligned} \tag{54}$$

where in (a) we use $\xi^{(t+1)} = \nabla_{\mathbf{y}} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) - \frac{1}{\beta}(\mathbf{y}^{(t+1)} - \mathbf{y}^{(t)})$.

The last two terms of (54) is the same as the RHS of (45). We can apply the similar steps from (47) to (48). To be more specific, the derivations are shown as follows: First, we know

$$f'(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)}) - f'(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) \leq \frac{1}{\beta}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 + \left\langle \nabla_{\mathbf{y}} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) - \nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \right\rangle - \frac{1}{\beta}\left\langle \mathbf{v}^{(t+1)}, \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \right\rangle. \tag{55}$$

Then, we move term $1/\beta\langle \mathbf{v}^{(t+1)}, \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)}\rangle$ to RHS of (54) and have

$$\begin{aligned}
& f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)}) - f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) \\
\leq & \frac{1}{2\beta}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 + \frac{1}{2\beta}\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 - \frac{1}{2\beta}\| \mathbf{v}^{(t+1)} \|^2 \\
& + \left\langle \nabla_{\mathbf{y}} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) - \nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}), \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \right\rangle \\
& + \left\langle \nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \right\rangle \\
\leq & \frac{1}{2\beta}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 + \left\langle \nabla_{\mathbf{y}} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) - \nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}), \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \right\rangle \\
& + \frac{\beta L_y^2}{2}\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 - \gamma \| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 \\
\overset{(a)}{\leq} & \frac{1}{\beta}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 + \frac{1}{2\beta}\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 \\
& + \frac{\beta L_x^2}{2}\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \|^2 - (\gamma - \frac{\beta L_y^2}{2})\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2
\end{aligned} \tag{56}$$

where in (a) we use

$$\langle \nabla_{\mathbf{y}} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) - \nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) \rangle \leq \frac{\beta L_x^2}{2}\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \|^2 + \frac{1}{2\beta}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 \tag{57}$$

which is different from (50); also $\mathbf{y}^{(t)}, \mathbf{y}^{(t+1)} \in \mathcal{Y}$ so have $f'(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)}) = f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)})$ and $f'(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) = f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)})$.

Combing (52), we have

$$\begin{aligned}
& f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)}) + \left(\frac{2}{\beta^2\gamma} + \frac{1}{2\beta}\right)\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 - 4\left(\frac{1}{\beta} - \frac{L_y^2}{2\gamma}\right)\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 \\
\leq & f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) + \left(\frac{2}{\beta^2\gamma} + \frac{1}{2\beta}\right)\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 - 4\left(\frac{1}{\beta} - \frac{L_y^2}{2\gamma}\right)\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 \\
& - \left(\frac{1}{2\beta} - \frac{2L_y^2}{\gamma}\right)\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 + \left(\frac{2L_x^2}{\gamma^2\beta} + \frac{\beta L_x^2}{2}\right)\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \|^2.
\end{aligned} \tag{58}$$

By taking the expectation on both sides of (52), we can get the results of Lemma 3. □

Next, we use the following lemma to show the descent of the objective value after solving $\mathbf{x}$-subproblem by (4).

A.3.3. Proof of Theorem 1

Proof:

From Lemma 3, we know

$$\begin{aligned}
& \mathbb{E}[f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)})] + \left(\frac{2}{\beta^2\gamma} + \frac{1}{2\beta}\right)\mathbb{E}[\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2] \\
& - 4\left(\frac{1}{\beta} - \frac{L_y^2}{2\gamma}\right)\mathbb{E}[\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2] \leq \mathbb{E}[f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)})] \\
& + \left(\frac{2}{\beta^2\gamma} + \frac{1}{2\beta}\right)\mathbb{E}[\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2] - 4\left(\frac{1}{\beta} - \frac{L_y^2}{2\gamma}\right)\mathbb{E}[\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2] \\
& - \left(\frac{1}{2\beta} - \frac{2L_y^2}{\gamma}\right)\mathbb{E}[\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2] + \left(\frac{2L_x^2}{\gamma^2\beta} + \frac{\beta L_x^2}{2}\right)\mathbb{E}[\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \|^2].
\end{aligned} \tag{59}$$

Combining Lemma 2, we have

$$\begin{aligned}
&\mathbb{E}[f(\mathbf{x}^{(t+1)},\mathbf{y}^{(t+1)})]+\left(\frac{2}{\beta^2\gamma}+\frac{1}{2\beta}\right)\mathbb{E}\left[\|\mathbf{y}^{(t+1)}-\mathbf{y}^{(t)}\|^2\right]\\
&-4\left(\frac{1}{\beta}-\frac{L_y^2}{2\gamma}\right)\mathbb{E}\left[\|\mathbf{y}^{(t+1)}-\mathbf{y}^{(t)}\|^2\right]\leq\mathbb{E}[f(\mathbf{x}^{(t)},\mathbf{y}^{(t)})]+\left(\frac{2}{\beta^2\gamma}+\frac{1}{2\beta}\right)\mathbb{E}\left[\|\mathbf{y}^{(t)}-\mathbf{y}^{(t-1)}\|^2\right]\\
&-4\left(\frac{1}{\beta}-\frac{L_y^2}{2\gamma}\right)\mathbb{E}\left[\|\mathbf{y}^{(t)}-\mathbf{y}^{(t-1)}\|^2\right]-\underbrace{\left(\frac{1}{2\beta}-\frac{2L_y^2}{\gamma}\right)}_{c_1}\mathbb{E}\left[\|\mathbf{y}^{(t+1)}-\mathbf{y}^{(t)}\|^2\right]\\
&-\underbrace{\left(\frac{1}{\alpha}-\left(\frac{L_x}{2}+\frac{2L_x^2}{\gamma^2\beta}+\frac{\beta L_x^2}{2}\right)\right)}_{c_2}\mathbb{E}\left[\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\|^2\right]+\alpha\sigma_x^2+L_x\mu^2. \qquad (60)
\end{aligned}$$

If

$$\beta<\frac{\gamma}{4L_y^2}\quad\text{and}\quad\alpha<\frac{1}{\frac{L_x}{2}+\frac{2L_x^2}{\gamma^2\beta}+\frac{\beta L_x^2}{2}}, \qquad (61)$$

then we have that there exist positive constants c_1 and c_2 such that

$$\begin{aligned}
&\mathcal{P}(\mathbf{x}^{(t+1)},\mathbf{y}^{(t+1)},\Delta_{\mathbf{y}}^{(t+1)})-\mathcal{P}(\mathbf{x}^{(t)},\mathbf{y}^{(t)},\Delta_{\mathbf{y}}^{(t)})\\
\leq&-c_1\mathbb{E}\left[\|\mathbf{y}^{(t+1)}-\mathbf{y}^{(t)}\|^2\right]-c_2\mathbb{E}\left[\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\|^2\right]+\alpha\sigma_x^2+L_x\mu^2\\
\leq&-\zeta\left(\mathbb{E}\left[\|\mathbf{y}^{(t+1)}-\mathbf{y}^{(t)}\|^2\right]+\mathbb{E}\left[\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\|^2\right]\right)+\alpha\sigma_x^2+L_x\mu^2 \qquad (62)
\end{aligned}$$

where $\zeta=\min\{c_1,c_2\}$.

From (6), we can have

$$\begin{aligned}
&\|\mathcal{G}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\|\\
\leq&\frac{1}{\alpha}\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\|+\frac{1}{\alpha}\|\mathbf{x}^{(t+1)}-\text{proj}_{\mathcal{X}}(\mathbf{x}^{(t)}-\alpha\nabla_{\mathbf{x}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t)}))\|+\frac{1}{\beta}\|\mathbf{y}^{(t+1)}-\mathbf{y}^{(t)}\|\\
&+\frac{1}{\beta}\|\mathbf{y}^{(t+1)}-\text{proj}_{\mathcal{Y}}(\mathbf{y}^{(t)}+\beta\nabla_{\mathbf{y}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\|\\
\overset{(a)}{\leq}&\frac{1}{\alpha}\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\|\\
&+\frac{1}{\alpha}\|\text{proj}_{\mathcal{X}}(\mathbf{x}^{(t+1)}-\alpha(\nabla_{\mathbf{x}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t)})+\frac{1}{\alpha}(\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}))-\text{proj}_{\mathcal{X}}(\mathbf{x}^{(t)}-\alpha\nabla_{\mathbf{x}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t)}))\|\\
&+\frac{1}{\beta}\|\mathbf{y}^{(t+1)}-\mathbf{y}^{(t)}\|\\
&+\frac{1}{\beta}\|\text{proj}_{\mathcal{Y}}(\mathbf{y}^{(t+1)}+\beta(\nabla_{\mathbf{y}}f(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})-\frac{1}{\beta}(\mathbf{y}^{(t+1)}-\mathbf{y}^{(t)}))-\text{proj}_{\mathcal{Y}}(\mathbf{y}^{(t)}+\beta\nabla_{\mathbf{y}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t)}))\|\\
\overset{(b)}{\leq}&\frac{3}{\alpha}\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\|+\|\nabla_{\mathbf{y}}f(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)}))-\nabla_{\mathbf{y}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t)}))\|+\frac{3}{\beta}\|\mathbf{y}^{(t+1)}-\mathbf{y}^{(t)}\|\\
\overset{(c)}{\leq}&\left(\frac{3}{\alpha}+L_x\right)\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\|+\frac{3}{\beta}\|\mathbf{y}^{(t+1)}-\mathbf{y}^{(t)}\|
\end{aligned}$$

where in (a) we use $\mathbf{x}^{(t+1)}=\text{proj}_{\mathcal{X}}(\mathbf{x}^{(t+1)}-\alpha\nabla f(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})-(\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}))$; in (b) we use nonexpansiveness of the projection operator; in (c) we apply the Lipschitz continuous of function $f(\mathbf{x},\mathbf{y})$ with respect to $\mathbf{x}$ and $\mathbf{y}$ under assumption **A2**.

Therefore, we can know that there exist a constant $c=\max\{L_x+\frac{3}{\alpha},\frac{3}{\beta}\}$ such that

$$\|\mathcal{G}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\|^2\leq c\left(\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\|^2+\|\mathbf{y}^{(t+1)}-\mathbf{y}^{(t)}\|^2\right). \qquad (63)$$

After applying the telescope sum on (62) and taking expectation over (63), we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\mathcal{G}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\|^2 \leq \frac{c}{\zeta}\left(\frac{\mathcal{P}_1-\mathcal{P}_{T+1}}{T}+\alpha\sigma_x^2+L_x\mu^2\right). \tag{64}$$

Recall from **A1** that $f \geq f^*$ and $\mathcal{Y}$ is bounded with diameter R, therefore, $\mathcal{P}_t$ given by (8) yields

$$\mathcal{P}_t \geq f^* + \left(\frac{\min\{4+4\beta^2L_y^2-7\beta\gamma,0\}}{2\beta^2\gamma}\right)R^2, \quad \forall t. \tag{65}$$

And let $(\mathbf{x}^{(r)},\mathbf{y}^{(r)})$ be uniformly and randomly picked from $\{(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\}_{t=1}^T$, based on (64) and (65), we obtain

$$\mathbb{E}_r[\mathbb{E}\|\mathcal{G}(\mathbf{x}^{(r)},\mathbf{y}^{(r)})\|^2] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\mathcal{G}(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\|^2 \leq \frac{c}{\zeta}\left(\frac{\mathcal{P}_1-f^*-\nu R^2}{T}+\alpha\sigma_x^2+L_x\mu^2\right), \tag{66}$$

where recall that $\zeta=\min\{c_1,c_2\}$, $c=\max\{L_x+\frac{3}{\alpha},\frac{3}{\beta}\}$ and $\nu=\frac{\min\{4+4\beta^2L_y^2-7\beta\gamma,0\}}{2\beta^2\gamma}$.

The proof is now complete. □

A.4. Convergence Analysis of ZO-Min-Max by Performing ZO-PGA

Before showing the proof of Lemma 4, we first give the following lemma regarding to recursion of the difference between two successive iterates of variable $\mathbf{y}$.

Lemma 6. *Under assumption 1, assume iterates $\mathbf{x}^{(t)},\mathbf{y}^{(t)}$ generated by algorithm 1. When function $f(\mathbf{x}^{(t)},\mathbf{y})$ is black-box, we have*

$$\begin{aligned}\frac{2}{\beta^2\gamma}\mathbb{E}\|\,\mathbf{y}^{(t+1)}-\mathbf{y}^{(t)}\,\|^2 \leq& \frac{2}{\beta^2\gamma}\mathbb{E}\|\,\mathbf{y}^{(t)}-\mathbf{y}^{(t-1)}\,\|^2+\frac{2}{\beta}\mathbb{E}\|\,\mathbf{y}^{(t+1)}-\mathbf{y}^{(t)}\,\|^2\\ &+\frac{6L_y^2}{\beta\gamma^2}\mathbb{E}\|\,\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\,\|^2-\left(\frac{4}{\beta}-\frac{6L_y^2+4}{\gamma}\right)\mathbb{E}\|\,\mathbf{y}^{(t)}-\mathbf{y}^{(t-1)}\,\|^2\\ &+\frac{4\sigma_y^2}{\beta\gamma}\left(\frac{3}{\gamma}+4\beta\right)+\frac{\mu^2d^2L_y^2}{\beta^2\gamma}.\end{aligned} \tag{67}$$

From the optimality condition of $\mathbf{y}$-subproblem in (22) at iteration t and $t-1$, we have

$$-\left\langle\widehat{\nabla}_{\mathbf{y}}f(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})-\frac{1}{\beta}(\mathbf{y}^{(t+1)}-\mathbf{y}^{(t)}),\mathbf{y}^{(t+1)}-\mathbf{y}^{(t)}\right\rangle \leq 0, \tag{68}$$

$$\left\langle\widehat{\nabla}_{\mathbf{y}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t-1)})-\frac{1}{\beta}(\mathbf{y}^{(t)}-\mathbf{y}^{(t-1)}),\mathbf{y}^{(t+1)}-\mathbf{y}^{(t)}\right\rangle \leq 0. \tag{69}$$

Adding the above inequalities and applying the definition of $\mathbf{v}^{(t+1)}$, we can get

$$\begin{aligned}\frac{1}{\beta}\langle\mathbf{v}^{(t+1)},\mathbf{y}^{(t+1)}-\mathbf{y}^{(t)}\rangle \leq& \underbrace{\left\langle\widehat{\nabla}_{\mathbf{y}}f(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})-\widehat{\nabla}_{\mathbf{y}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t)}),\mathbf{y}^{(t+1)}-\mathbf{y}^{(t)}\right\rangle}_{\mathbf{I}}\\ &+\underbrace{\left\langle\widehat{\nabla}_{\mathbf{y}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t)})-\widehat{\nabla}_{\mathbf{y}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t-1)}),\mathbf{y}^{(t+1)}-\mathbf{y}^{(t)}\right\rangle}_{\mathbf{II}}.\end{aligned} \tag{70}$$

Next, we will bound $\mathbb{E}[\mathbf{I}]$ and $\mathbb{E}[\mathbf{II}]$ separably as follows.

First, we give an upper bound of $\mathbb{E}[\mathbf{I}]$ as the following,

$$\begin{aligned}
&\mathbb{E}\left\langle \widehat{\nabla}_{\mathbf{y}} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) - \widehat{\nabla}_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}), \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \right\rangle \\
\leq & \frac{3}{2\gamma}\mathbb{E}\|\widehat{\nabla}_{\mathbf{y}} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) - \nabla_{\mathbf{y}} f_{\mu,\mathbf{y}}(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)})\|^2 + \frac{\gamma}{6}\mathbb{E}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 \\
&+ \frac{3}{2\gamma}\mathbb{E}\|\nabla_{\mathbf{y}} f_{\mu,\mathbf{y}}(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) - \nabla_{\mathbf{y}} f_{\mu,\mathbf{y}}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\|^2 + \frac{\gamma}{6}\mathbb{E}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 \\
&+ \frac{3}{2\gamma}\mathbb{E}\|\nabla_{\mathbf{y}} f_{\mu,\mathbf{y}}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \widehat{\nabla} f_{\mathbf{y}}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\|^2 + \frac{\gamma}{6}\mathbb{E}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 \\
\leq & \frac{3\sigma_y^2}{\gamma} + \frac{3L_x^2}{2\gamma}\mathbb{E}\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \|^2 + \frac{\gamma}{2}\mathbb{E}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2
\end{aligned} \tag{71}$$

where Lemma 1 is used.

Second, we need to give an upper bound of $\mathbb{E}[\mathbf{II}]$ as follows:

$$\begin{aligned}
&\langle \widehat{\nabla} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \widehat{\nabla} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \rangle \\
= & \langle \widehat{\nabla} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \widehat{\nabla} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{v}^{(t+1)} + \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \rangle \\
= & \left\langle \nabla f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \nabla f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \right\rangle \\
&+ \left\langle \nabla f_{\mu,\mathbf{y}}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \nabla f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}), \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \right\rangle \\
&+ \left\langle \widehat{\nabla} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \nabla f_{\mu,\mathbf{y}}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}), \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \right\rangle \\
&- \left\langle \nabla f_{\mu,\mathbf{y}}(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}) - \nabla f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \right\rangle \\
&- \left\langle \widehat{\nabla} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}) - \nabla f_{\mu,\mathbf{y}}(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \right\rangle \\
&+ \langle \widehat{\nabla} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \widehat{\nabla} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{v}^{(t+1)} \rangle.
\end{aligned}$$

Next, we take expectation on both sides of the above equality and obtain

$$\begin{aligned}
&\mathbb{E}\langle \widehat{\nabla} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \widehat{\nabla} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \rangle \\
\overset{(a)}{\leq} & \left(\frac{3\beta L_y^2}{2} + \beta \right) \| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 + \frac{1}{2\beta} \| \mathbf{v}^{(t+1)} \|^2 - \gamma \| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 \\
&+ \frac{\mu^2 d^2 L_y^2}{4\beta} + 4\beta\sigma_y^2
\end{aligned} \tag{72}$$

where in (a) we use the fact that 1) γ-strong concavity of f with respect to $\mathbf{y}$:

$$\left\langle \nabla f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \nabla f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \right\rangle \leq -\gamma \| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2; \tag{73}$$

and the facts that 2) smoothing property (25) and Young's inequality

$$\mathbb{E}\left\langle \nabla f_{\mu,\mathbf{y}}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \nabla f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}), \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \right\rangle \leq \frac{\mu^2 d^2 L_y^2}{8\beta} + \frac{\beta}{2} \| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2; \tag{74}$$

and the fact that 3) the ZO estimator is unbiased according to Lemma 1

$$\mathbb{E}\left\langle \widehat{\nabla} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \nabla f_{\mu,\mathbf{y}}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}), \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \right\rangle = 0; \tag{75}$$

and

$$\mathbb{E}\left\langle \nabla f_{\mu,\mathbf{y}}(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}) - \nabla f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \right\rangle \leq \frac{\mu^2 d^2 L_y^2}{8\beta} + \frac{\beta}{2} \| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2; \tag{76}$$

and from Corollary 1 we have

$$\mathbb{E}\left\langle \widehat{\nabla} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}) - \nabla f_{\mu,\mathbf{y}}(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \right\rangle \leq \beta\sigma_y^2; \tag{77}$$

and

$$\begin{aligned}
&\mathbb{E}\langle \widehat{\nabla} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \widehat{\nabla} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{v}^{(t+1)} \rangle \\
\leq & \frac{3\beta}{2}\mathbb{E}\|\nabla f_{\mu,\mathbf{y}}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \widehat{\nabla} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\|^2 + \frac{1}{6\beta}\| \mathbf{v}^{(t+1)} \|^2 \\
& + \frac{3\beta}{2}\mathbb{E}\|\nabla f_{\mu,\mathbf{y}}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \nabla f_{\mu,\mathbf{y}}(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)})\|^2 + \frac{1}{6\beta}\| \mathbf{v}^{(t+1)} \|^2 \\
& + \frac{3\beta}{2}\mathbb{E}\|\nabla f_{\mu,\mathbf{y}}(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}) - \widehat{\nabla} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)})\|^2 + \frac{1}{6\beta}\| \mathbf{v}^{(t+1)} \|^2 \\
\leq & 3\beta\sigma_y^2 + \frac{1}{2\beta}\| \mathbf{v}^{(t+1)} \|^2 + \frac{3\beta L_y^2}{2}\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2.
\end{aligned} \tag{78}$$

Then, from (70), we can have

$$\begin{aligned}
\frac{1}{2\beta}\mathbb{E}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 \leq & \frac{1}{2\beta}\mathbb{E}\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 - \frac{1}{2\beta}\mathbb{E}\| \mathbf{v}^{(t+1)} \|^2 \\
& + \frac{3\sigma_y^2}{\gamma} + \frac{3L_x^2}{2\gamma}\mathbb{E}\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \|^2 + \frac{\gamma}{2}\mathbb{E}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 \\
& + \left\langle \widehat{\nabla} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \widehat{\nabla} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \right\rangle \\
\leq & \frac{1}{2\beta}\mathbb{E}\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 + \frac{\gamma}{2}\mathbb{E}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 \\
& + \frac{3L_y^2}{2\gamma}\mathbb{E}\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \|^2 - \left(\gamma - \left(\frac{3\beta L_y^2}{2} + \beta\right)\right)\mathbb{E}\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 \\
& + \frac{3\sigma_y^2}{\gamma} + 4\beta\sigma_y^2 + \frac{\mu^2 d^2 L_y^2}{4\beta},
\end{aligned} \tag{79}$$

which implies

$$\begin{aligned}
\frac{2}{\beta^2\gamma}\mathbb{E}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 \leq & \frac{2}{\beta^2\gamma}\mathbb{E}\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 + \frac{2}{\beta}\mathbb{E}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 \\
& + \frac{6L_y^2}{\beta\gamma^2}\mathbb{E}\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \|^2 - \left(\frac{4}{\beta} - \frac{6L_y^2 + 4}{\gamma}\right)\mathbb{E}\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 \\
& + \frac{4\sigma_y^2}{\beta\gamma}\left(\frac{3}{\gamma} + 4\beta\right) + \frac{\mu^2 d^2 L_y^2}{\beta^2\gamma}.
\end{aligned} \tag{80}$$

A.4.1. Proof of Lemma 4

Proof: Similarly as A.3.2, let $f'(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)}) = f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)}) - 1(\mathbf{y}^{(t+1)})$, $1(\cdot)$ denotes the indicator function and $\xi^{(t)}$ denote the subgradient of $1(\mathbf{y}^{(t)})$. Since function $f'(\mathbf{x}, \mathbf{y})$ is concave with respect to $\mathbf{y}$, we have

$$\begin{aligned}
& f'(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)}) - f'(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) \leq \langle \nabla f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}), \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \rangle - \langle \xi^{(t)}, \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \rangle \\
\overset{(a)}{=} & \frac{1}{\beta}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 - \langle \xi^{(t)} - \xi^{(t+1)}, \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \rangle \\
= & \frac{1}{\beta}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 + \left\langle \widehat{\nabla} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) - \widehat{\nabla} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \right\rangle \\
& - \frac{1}{\beta}\left\langle \mathbf{v}^{(t+1)}, \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \right\rangle
\end{aligned} \tag{81}$$

where in (a) we use $\xi^{(t+1)} = \widehat{\nabla} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) - \frac{1}{\beta}(\mathbf{y}^{(t+1)} - \mathbf{y}^{(t)})$. Then, we have

$$\begin{aligned}
&\mathbb{E} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)}) - \mathbb{E} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) + \frac{1}{\beta}\left\langle \mathbf{v}^{(t+1)}, \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \right\rangle \\
\leq &\frac{1}{\beta}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 + \left\langle \widehat{\nabla} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) - \widehat{\nabla} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t-1)}), \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \right\rangle.
\end{aligned}$$

Applying the steps from (72) to (79), we can have

$$\begin{aligned}
&\mathbb{E} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)}) - \mathbb{E} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) \\
\leq &\frac{1}{\beta}\mathbb{E}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 + \frac{1}{2\beta}\mathbb{E}\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 - \left(\gamma - \left(\frac{3\beta L_y^2}{2} + \beta\right)\right)\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 \\
&+ \frac{3\beta L_x^2}{2}\mathbb{E}\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \|^2 + 7\beta\sigma_y^2 + \frac{\mu^2 d^2 L_y^2}{4\beta}
\end{aligned} \tag{82}$$

where we use

$$\begin{aligned}
&\mathbb{E}\left\langle \widehat{\nabla}_{\mathbf{y}} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) - \widehat{\nabla}_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}), \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \right\rangle \\
\leq &3\beta\sigma_y^2 + \frac{3\beta L_x^2}{2}\mathbb{E}\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \|^2 + \frac{1}{2\beta}\mathbb{E}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2.
\end{aligned} \tag{83}$$

Combing (80), we have

$$\begin{aligned}
&\mathbb{E} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)}) + \left(\frac{2}{\beta^2\gamma} + \frac{1}{2\beta}\right)\mathbb{E}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 - \left(\frac{4}{\beta} - \frac{6L_y^2 + 4}{\gamma}\right)\mathbb{E}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 \\
\leq &\mathbb{E} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) + \left(\frac{2}{\beta^2\gamma} + \frac{1}{2\beta}\right)\mathbb{E}\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 - \left(\frac{4}{\beta} - \frac{6L_y^2 + 4}{\gamma}\right)\mathbb{E}\| \mathbf{y}^{(t)} - \mathbf{y}^{(t-1)} \|^2 \\
&- \left(\frac{1}{2\beta} - \frac{6L_y^2 + 4}{\gamma}\right)\mathbb{E}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 + \left(\frac{6L_x^2}{\gamma^2\beta} + \frac{3\beta L_x^2}{2}\right)\mathbb{E}\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \|^2. \\
&+ \frac{\mu^2 d^2 L_y^2}{\beta}\left(\frac{1}{4} + \frac{1}{\beta\gamma}\right) + \left(7\beta + \frac{4}{\beta\gamma}\left(\frac{3}{\gamma} + 7\beta\right)\right)\sigma_y^2.
\end{aligned} \tag{84}$$

□

A.4.2. Proof of Theorem 2

Proof: From (36), we know the "descent" of the minimization step, i.e., the changes from $\mathcal{P}'(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}, \Delta_{\mathbf{y}}^{(t)})$ to $\mathcal{P}'(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}, \Delta_{\mathbf{y}}^{(t)})$.

Combining the "descent" of the maximization step by Lemma 4 shown in (84), we can obtain the following:

$$\begin{aligned}
&\mathcal{P}'(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)}, \Delta_{\mathbf{y}}^{(t+1)}) \\
\leq &\mathcal{P}'(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}, \Delta_{\mathbf{y}}^{(t)}) - \underbrace{\left(\frac{1}{2\beta} - \frac{6L_y^2 + 4}{\gamma}\right)}_{a_1}\mathbb{E}\left[\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2\right] \\
&- \underbrace{\left(\frac{1}{\alpha} - \left(\frac{L_x}{2} + \frac{6L_x^2}{\gamma^2\beta} + \frac{3\beta L_x^2}{2}\right)\right)}_{a_2}\mathbb{E}\left[\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \|^2\right] \\
&+ \mu^2\underbrace{\left(L_x + \frac{d^2 L_y^2}{\beta}\left(\frac{1}{4} + \frac{1}{\beta\gamma}\right)\right)}_{b_1} + \alpha\sigma_x^2 + \underbrace{\left(7\beta + \frac{4}{\beta\gamma}\left(\frac{3}{\gamma} + 4\beta\right)\right)}_{b_2}\sigma_y^2.
\end{aligned} \tag{85}$$

When β, α satisfy the following conditions:

$$\beta < \frac{\gamma}{4(3L_y^2+2)}, \quad \text{and} \quad \alpha < \frac{1}{\frac{L_x}{2} + \frac{6L_x^2}{\gamma^2\beta} + \frac{3\beta L_x^2}{2}}, \tag{86}$$

we can conclude that there exist $b_1, b_2 > 0$ such that

$$\begin{aligned}
&\mathcal{P}'(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)}, \Delta_{\mathbf{y}}^{(t+1)}) \\
\leq &\mathcal{P}'(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}, \Delta_{\mathbf{y}}^{(t)}) - a_1\mathbb{E}\left[\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2\right] \\
&- a_2\left[\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \|^2\right] + b_1\mu^2 + \alpha\sigma_x^2 + b_2\sigma_y^2 \\
\leq &-\zeta'\mathbb{E}\left[\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2 + \| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \|^2\right] + b_1\mu^2 + \alpha\sigma_x^2 + b_2\sigma^2
\end{aligned} \tag{87}$$

where $\zeta' = \min\{a_1, a_2\}$.

From (6), we can have

$$\begin{aligned}
&\mathbb{E}\|\mathcal{G}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\| \\
\leq &\frac{1}{\alpha}\mathbb{E}\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \| + \frac{1}{\alpha}\mathbb{E}\| \mathbf{x}^{(t+1)} - \text{proj}_{\mathcal{X}}(\mathbf{x}^{(t)} - \alpha\nabla_{\mathbf{x}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}))\| \\
&+ \frac{1}{\beta}\mathbb{E}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \| + \frac{1}{\beta}\mathbb{E}\| \mathbf{y}^{(t+1)} - \text{proj}_{\mathcal{Y}}(\mathbf{y}^{(t)} + \beta\nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\| \\
\overset{(a)}{\leq} &\frac{1}{\alpha}\mathbb{E}\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \| + \frac{1}{\beta}\mathbb{E}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \| \\
&+ \frac{1}{\alpha}\mathbb{E}\|\text{proj}_{\mathcal{X}}(\mathbf{x}^{(t+1)} - \alpha(\widehat{\nabla}_{\mathbf{x}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) + \frac{1}{\alpha}(\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)})) - \text{proj}_{\mathcal{X}}(\mathbf{x}^{(t)} - \alpha\nabla_{\mathbf{x}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}))\| \\
&+ \frac{1}{\beta}\mathbb{E}\|\text{proj}_{\mathcal{Y}}(\mathbf{y}^{(t+1)} + \beta(\widehat{\nabla}_{\mathbf{y}} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) - \frac{1}{\beta}(\mathbf{y}^{(t+1)} - \mathbf{y}^{(t)})) - \text{proj}_{\mathcal{Y}}(\mathbf{y}^{(t)} + \beta\nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}))\| \\
\overset{(b)}{\leq} &\frac{3}{\alpha}\mathbb{E}\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \| + \mathbb{E}\|\widehat{\nabla}_{\mathbf{x}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})) - \nabla_{\mathbf{x}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}))\| \\
&+ \frac{3}{\beta}\mathbb{E}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \| + \mathbb{E}\|\widehat{\nabla}_{\mathbf{y}} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) - \nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\| \\
\leq &\frac{3}{\alpha}\mathbb{E}\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \| + \mathbb{E}\|\widehat{\nabla}_{\mathbf{x}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})) - \nabla_{\mathbf{x}} f_{\mu,\mathbf{y}}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}))\| \\
&+ \mathbb{E}\|\nabla_{\mathbf{x}} f_{\mu,\mathbf{y}}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})) - \nabla_{\mathbf{x}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}))\| \\
&+ \frac{3}{\beta}\mathbb{E}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \| + \mathbb{E}\|\widehat{\nabla}_{\mathbf{y}} f(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) - \nabla_{\mathbf{y}} f_{\mu,\mathbf{y}}(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)})\| \\
&+ \mathbb{E}\|\nabla_{\mathbf{y}} f_{\mu,\mathbf{y}}(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t)}) - \nabla_{\mathbf{y}} f_{\mu,\mathbf{y}}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\| \\
&+ \mathbb{E}\|\nabla_{\mathbf{y}} f_{\mu,\mathbf{y}}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) - \nabla_{\mathbf{y}} f(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\| \\
\overset{(c)}{\leq} &\left(\frac{3}{\alpha} + L_x\right)\mathbb{E}\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \| + \frac{3}{\beta}\mathbb{E}\| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \| + 2\sigma_y^2 + \mu^2 d^2 L_y^2
\end{aligned}$$

where in (a) we use the optimality condition of $\mathbf{x}^{(t)}$-subproblem; in (b) we use nonexpansiveness of the projection operator; in (c) we apply the Lipschitz continuous of function $f(\mathbf{x}, \mathbf{y})$ under assumption **A2**.

Therefore, we can know that

$$\mathbb{E}\left[\|\mathcal{G}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\|^2\right] \leq c\left(\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \|^2 + \| \mathbf{y}^{(t+1)} - \mathbf{y}^{(t)} \|^2\right) + 2\sigma_y^2 + \mu^2 d^2 L_y^2. \tag{88}$$

After applying the telescope sum on (87) and taking expectation over (88), we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\mathcal{G}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\|^2\right] \leq \frac{c}{\zeta'}\frac{\mathcal{P}_1 - \mathcal{P}_{T+1}}{T} + \frac{cb_1}{\zeta'}\mu^2 + \frac{c\alpha\sigma_x^2}{\zeta'} + \frac{cb_2}{\zeta'}\sigma_y^2 + 2\sigma_y^2 + \mu^2 d^2 L_y^2. \tag{89}$$

Recall from **A1** that $f \geq f^*$ and $\mathcal{Y}$ is bounded with diameter R, therefore, $\mathcal{P}_t$ given by (11) yields

$$\mathcal{P}_t \geq f^* + \left(\frac{\min\{4 + 4(3L_y^2 + 2)\beta^2 - 7\beta\gamma, 0\}}{\beta^2 \gamma} \right) R^2, \quad \forall t. \tag{90}$$

And let $(\mathbf{x}^{(r)}, \mathbf{y}^{(r)})$ be uniformly and randomly picked from $\{(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\}_{t=1}^T$, based on (90) and (89), we obtain

$$\begin{aligned} &\mathbb{E}_r \left[\mathbb{E} \left[\|\mathcal{G}(\mathbf{x}^{(r)}, \mathbf{y}^{(r)})\|^2 \right] \right] = \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\|\mathcal{G}(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\|^2 \right] \\ \leq &\frac{c}{\zeta'} \frac{\mathcal{P}_1 - f^* - \nu' R^2}{T} + \frac{cb_1}{\zeta'} \mu^2 + \frac{c\alpha\sigma_x^2}{\zeta'} + \frac{cb_2}{\zeta'} \sigma_y^2 + 2\sigma_y^2 + \mu^2 d^2 L_y^2 \end{aligned} \tag{91}$$

where recall that $\zeta' = \min\{a_1, a_2\}$, $c = \max\{L_x + \frac{3}{\alpha}, \frac{3}{\beta}\}$, and $\nu' = \frac{\min\{4+4(3L_y^2+2)\beta^2-7\beta\gamma,0\}}{\beta^2\gamma}$.

The proof is now complete. □

B. Toy Example in (Bogunovic et al., 2018): ZO-Min-Max versus BO

We review the example in (Bogunovic et al., 2018) as below,

$$\begin{aligned}\underset{\mathbf{x}\in\mathcal{C}}{\text{maximize}}\ \underset{\|\boldsymbol{\delta}\|_2\le 0.5}{\text{minimize}}\quad & f(\mathbf{x}-\boldsymbol{\delta}) := -2(x_1-\delta_1)^6+12.2(x_1-\delta_1)^5-21.2(x_1-\delta_1)^4\\ & -6.2(x_1-\delta_1)+6.4(x_1-\delta_1)^3+4.7(x_1-\delta_1)^2-(x_2-\delta_2)^6\\ & +11(x_2-\delta_2)^5-43.3(x_2-\delta_2)^4+10(x_2-\delta_2)+74.8(x_2-\delta_2)^3\\ & -56.9(x_2-\delta_2)^2+4.1(x_1-\delta_1)(x_2-\delta_2)+0.1(x_1-\delta_1)^2(x_2-\delta_2)^2\\ & -0.4(x_2-\delta_2)^2(x_1-\delta_1)-0.4(x_1-\delta_1)^2(x_2-\delta_2),\end{aligned} \tag{92}$$

where $\mathbf{x}\in\mathbb{R}^2$, and $\mathcal{C}=\{x_1\in(-0.95,3.2),x_2\in(-0.45,4.4)\}$.

Problem (92) can be equivalently transformed to the min-max setting consistent with ours

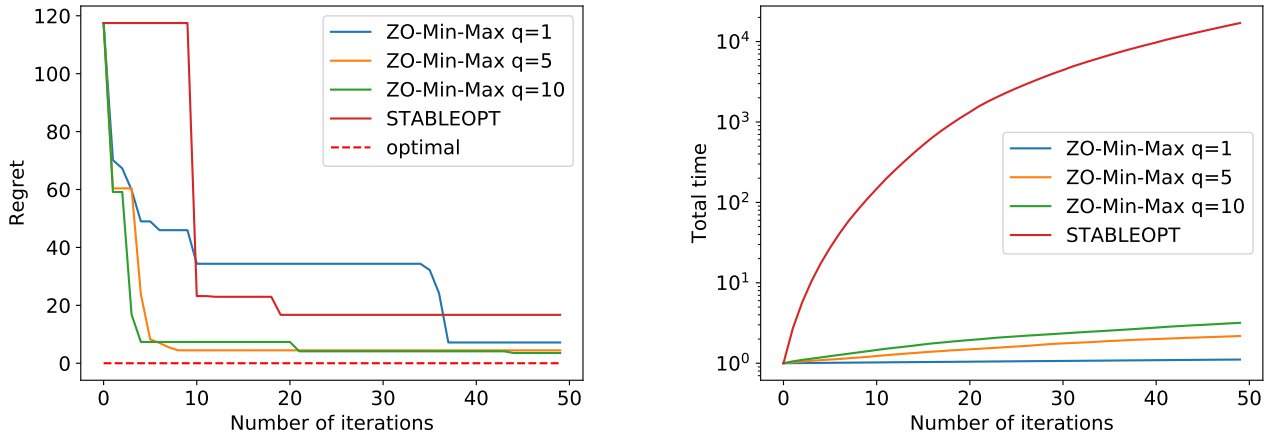
$$\underset{\mathbf{x}\in\mathcal{C}}{\text{minimize}}\ \underset{\|\boldsymbol{\delta}\|_2\le 0.5}{\text{maximize}}\quad -f(\mathbf{x}-\boldsymbol{\delta}). \tag{93}$$

The optimality of solving problem (92) is measured by regret versus iteration t,

$$\text{Regret}(t)=\underset{\|\boldsymbol{\delta}\|_2\le 0.5}{\text{minimize}}\, f(\mathbf{x}^*-\boldsymbol{\delta})-\underset{\|\boldsymbol{\delta}\|_2\le 0.5}{\text{minimize}}\, f(\mathbf{x}^{(t)}-\boldsymbol{\delta}), \tag{94}$$

where $\text{minimize}_{\|\boldsymbol{\delta}\|_2\le 0.5}\, f(\mathbf{x}^*-\boldsymbol{\delta})=-4.33$ and $\mathbf{x}^*=[-0.195,0.284]^T$ (Bogunovic et al., 2018).

In Figure A1, we compare the convergence performance and computation time of ZO-Min-Max with the BO based approach STABLEOPT proposed in (Bogunovic et al., 2018). Here we choose the same initial point for both ZO-Min-Max and STABLEOPT. And we set the same number of function queries per iteration for ZO-Min-Max (with $q=1$) and STABLEOPT. We recall from (2) that the larger q is, the more queries ZO-Min-Max takes. In our experiments, we present the best achieved regret up to time t and report the average performance of each method over 5 random trials. As we can see, ZO-Min-Max is more stable, with lower regret and less running time. Besides, as q becomes larger, ZO-Min-Max has a faster convergence rate. We remark that BO is slow since learning the accurate GP model and solving the acquisition problem takes intensive computation cost.



(a) (b)

Figure A1: Comparison of ZO-Min-Max against STABLEOPT (Bogunovic et al., 2018): a) Convergence performance; b) Computation time (seconds).

C. Additional Details on Ensemble Evasion Attack

Experiment setup. We specify the attack loss F_{ij} in (13) as the C&W untargeted attack loss (Carlini & Wagner, 2017),

$$F_{ij}\left(\mathbf{x};\Omega_i\right)=(1/|\Omega_i|)\sum_{\mathbf{z}\in\Omega_i}\max\{g_j(\mathbf{z}+\mathbf{x})_i-\max_{k\neq i}g_j(\mathbf{z}+\mathbf{x})_k,0\},\tag{95}$$

where $|\Omega_i|$ is the cardinality of the set Ω_i, $g_j(\mathbf{z}+\mathbf{x})_k$ denotes the prediction score of class k given the input $\mathbf{z}+\mathbf{x}$ using model j. In (13), the regularization parameter λ trikes a balance between the worse-case attack loss and the average loss (Wang et al., 2019b). The rationale behind that is from two manifolds. First, as $\gamma=0$, then $\text{maximize}_{\mathbf{w}\in\mathcal{W}}\sum_{j=1}^{J}\sum_{i=1}^{I}\left[w_{ij}F_{ij}\left(\mathbf{x};\Omega_i\right)\right]=F_{i^*j^*}\left(\mathbf{x};\Omega_i\right)$, where $w_{i^*j^*}=1$ and 0s for $(i,j)\neq(i^*,j^*)$, and $(i^*,j^*)=\arg\max_{i,j}F_{ij}\left(\mathbf{x};\Omega_i\right)$ given $\mathbf{x}$. On the other hand, as $\gamma\to\infty$, then $\mathbf{w}\to\mathbf{1}/(IJ)$.

Implementation of ZO-PGD for solving problem (13). To solve problem (13), the baseline method ZO-PGD performs single-objective ZO minimization under the equivalent form of (13), $\min_{\mathbf{x}\in\mathcal{X}}h(\mathbf{x})$, where $h(\mathbf{x})=\max_{\mathbf{w}\in\mathcal{W}}f(\mathbf{x},\mathbf{w})$. It is worth noting that we report the best convergence performance of ZO-PGD by searching its learning rate over 5 grid points in $[0.01,0.05]$. Also, when querying the function value of h (at a given point $\mathbf{x}$), we need the solution to the inner maximization problem

$$\underset{\mathbf{w}\in\mathcal{W}}{\text{maximize}}\quad\sum_{j=1}^{J}\sum_{i=1}^{I}\left[w_{ij}F_{ij}\left(\mathbf{x};\Omega_i\right)\right]-\lambda\|\mathbf{w}-\mathbf{1}/(IJ)\|_2^2.\tag{96}$$

Problem (96) is equivalent to

$$\begin{array}{ll}\underset{\mathbf{w}}{\text{minimize}} & \lambda\|\mathbf{w}-\mathbf{1}/(IJ)-(1/(2\lambda))\mathbf{f}(\mathbf{x})\|_2^2\\ \text{subject to} & \mathbf{1}^T\mathbf{w}=1,\mathbf{w}\geq 0\end{array}\tag{97}$$

where $\mathbf{f}(\mathbf{x}):=[F_{11}(\mathbf{x}),\ldots,f_{IJ}(\mathbf{x})]^T$. The solution is given by the projection of the point $\mathbf{1}/(IJ)+(1/(2\lambda))\mathbf{f}(\mathbf{x})$ on the the probabilistic simplex (Parikh et al., 2014)

$$\mathbf{w}^*=\left[\mathbf{1}/(IJ)+(1/(2\lambda))\mathbf{f}(\mathbf{x})-\mu\mathbf{1}\right]_+,\tag{98}$$

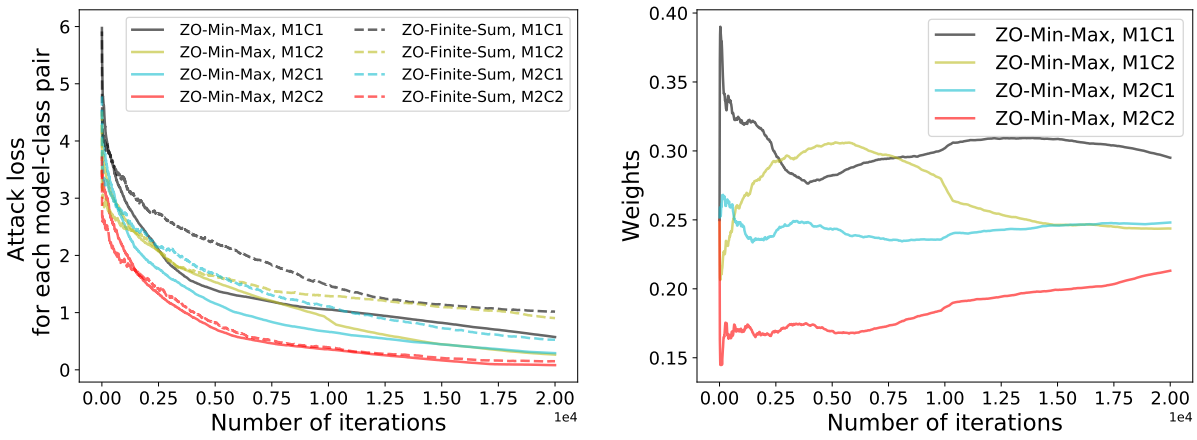
where $[\cdot]_+$ is element-wise non-negative operator, and μ is the root of the equation

$$\mathbf{1}^T\left[\mathbf{1}/(IJ)+(1/(2\lambda))\mathbf{f}(\mathbf{x})-\mu\mathbf{1}\right]_+=\sum_i\max\{0,1/(IJ)+f_i(\mathbf{x})/(2\lambda)-\mu\}=1.\tag{99}$$

The above equation in μ can be solved using the bisection method at a given $\mathbf{x}$ (Boyd & Vandenberghe, 2004).

Additional results. In Figure A2-(a), We compare ZO-Min-Max with ZO-Finite-Sum, where the latter minimizes the average loss over all model-class combinations. As we can see, our approach significantly improves the worst-case attack performance (corresponding to M1C1). Here the worst case represents the most robust model-class pair against the attack. This suggests that ZO-Min-Max takes into account different robustness levels of model-class pairs through the design of importance weights $\mathbf{w}$. This can also be evidenced from Figure A2-(b): M1C1 has the largest weight while M2C2 corresponds to the smallest weight.

In Figure A3, we contrast the success or failure (marked by blue or red in the plot) of attacking each image using the obtained universal perturbation $\mathbf{x}$ with the attacking difficulty (in terms of required iterations for successful adversarial example) of using per-image non-universal PGD attack (Madry et al., 2018). We observe that the success rate of the ensemble universal attack is around 80% at each model-class pair, where the failed cases (red cross markers) also need a large amount of iterations to succeed at the case of per-image PGD attack. And images that are difficult to attack keep consistent across models; see dash lines to associate the same images between two models in Figure A3.



(a) (b)

Figure A2: Convergence performance of ZO-Min-Max in design of black-box ensemble attack. a) Attack loss of using ZO-Min-Max vs. ZO-Finite-Sum, and b) importance weights learnt from ZO-Min-Max.

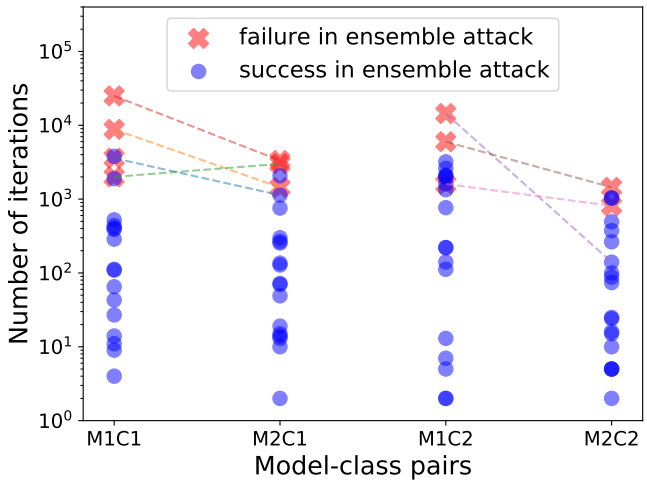


Figure A3: Success or failure of our ensemble attack versus successful per-image PGD attack.

D. Additional Details on Poisoning Attack

Experiment setup. In our experiment, we generate a synthetic dataset that contains $n = 1000$ samples $(\mathbf{z}_i, t_i)$. We randomly draw the feature vector $\mathbf{z}_i \in \mathbb{R}^{100}$ from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and determine $t_i = 1$ if $1/(1 + e^{-(\mathbf{z}_i^T \boldsymbol{\theta}^* + \nu_i)}) > 0.5$. Here we choose $\boldsymbol{\theta}^* = \mathbf{1}$ as the ground-truth model parameters, and $\nu_i \in \mathcal{N}(0, 10^{-3})$ as random noise. We randomly split the generated dataset into the training dataset $\mathcal{D}_{\rm tr}$ (70%) and the testing dataset $\mathcal{D}_{\rm te}$ (30%). We specify our learning model as the logistic regression model for binary classification. Thus, the loss function in problem (14) is chosen as $F_{\rm tr}(\mathbf{x}, \boldsymbol{\theta}; \mathcal{D}_{\rm tr}) := h(\mathbf{x}, \boldsymbol{\theta}; \mathcal{D}_{{\rm tr},1}) + h(\mathbf{0}, \boldsymbol{\theta}; D_{{\rm tr},2})$, where $\mathcal{D}_{\rm tr} = \mathcal{D}_{{\rm tr},1} \cup \mathcal{D}_{{\rm tr},2}$, $\mathcal{D}_{{\rm tr},1}$ represents the subset of the training dataset that will be poisoned, $|\mathcal{D}_{{\rm tr},1}|/|\mathcal{D}_{\rm tr}|$ denotes the poisoning ratio, $h(\mathbf{x}, \boldsymbol{\theta}; \mathcal{D}) = -(1/|\mathcal{D}|) \sum_{(\mathbf{z}_i, t_i) \in \mathcal{D}} [t_i \log(h(\mathbf{x}, \boldsymbol{\theta}; \mathbf{z}_i)) + (1 - t_i) \log(1 - h(\mathbf{x}, \boldsymbol{\theta}; \mathbf{z}_i))]$, and $h(\mathbf{x}, \boldsymbol{\theta}; \mathbf{z}_i) = 1/(1 + e^{-(\mathbf{z}_i + \mathbf{x})^T \boldsymbol{\theta}})$. In problem (14), we also set $\epsilon = 2$ and $\lambda = 10^{-3}$.

In Algorithm 1, unless specified otherwise we choose the the mini-batch size $b = 100$, the number of random direction vectors $q = 5$, the learning rate $\alpha = 0.02$ and $\beta = 0.05$, and the total number of iterations $T = 50000$. We report the empirical results over 10 independent trials with random initialization.

Addition results. In Figure A4, we show the testing accuracy of the poisoned model as the regularization parameter λ varies. We observe that the poisoned model accuracy could be improved as λ increases, e.g., $\lambda = 1$. However, this leads to a decrease in clean model accuracy (below 90% at $\lambda = 1$). This implies a robustness-accuracy tradeoff. If λ continues to increase, both the clean and poisoned accuracy will decrease dramatically as the training loss in (14) is less optimized.

In Figure A5, we present the testing accuracy of the learnt model under different data poisoning ratios. As we can see, only 5% poisoned training data can significantly break the testing accuracy of a well-trained model.

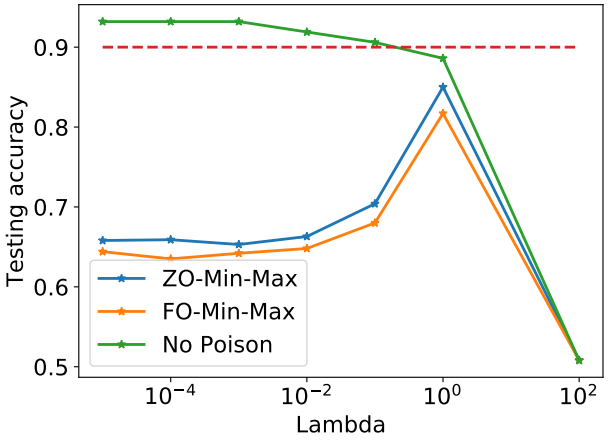


Figure A4: Empirical performance of ZO-Min-Max in design of poisoning attack: Testing accuracy versus regularization parameter λ.

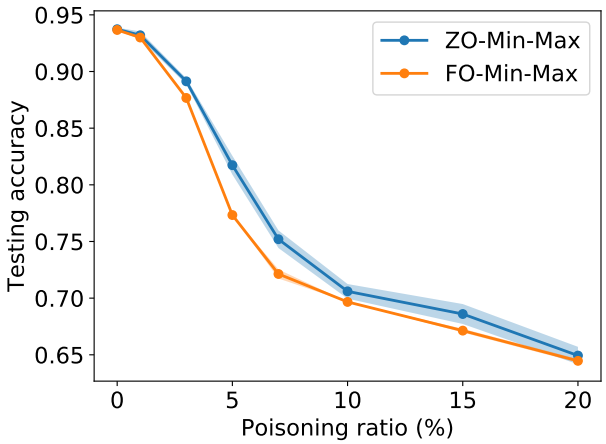


Figure A5: Testing accuracy of data poisoning attacks generated by ZO-Min-Max (vs. FO-Min-Max) for different data poisoning ratios.