A. Preliminary

Notation. Bold upper case letters without subscripts (e.g., $\mathbf{X}, \mathbf{Y}$) denote matrices and bold lower case letters without subscripts (e.g., $\mathbf{x}, \mathbf{y}$) represent vectors. We use $\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})$ to denote the partial gradient with respect to variables of block $\mathbf{x}$.

We provide the proofs of some preliminary lemmas (Lemma 4–Lemma 6) used in the proof of Section B.

First, Lemma 4 and Lemma 5 give the property that quantify the size of the difference of the second-order derivatives of the objective values between two points.

Lemma 4. *If function $f(\cdot)$ is ρ-Hessian Lipschitz, we have*

$$\left\|\int_0^1 \nabla^2 f(\alpha\boldsymbol{\theta})d\alpha - \nabla^2 f(\boldsymbol{\theta}')\right\| \le \rho\left(\|\boldsymbol{\theta}\| + \|\boldsymbol{\theta}'\|\right), \quad \forall\boldsymbol{\theta}, \boldsymbol{\theta}'. \tag{20}$$

Proof. If function $f(\cdot)$ is ρ-Hessian Lipschitz continuous, then we have

$$\begin{aligned}\left\|\int_0^1 (\nabla^2 f(\alpha\boldsymbol{\theta}) - \nabla^2 f(\boldsymbol{\theta}'))d\alpha\right\| &\le \int_0^1 \left\|\nabla^2 f(\alpha\boldsymbol{\theta}) - \nabla^2 f(\boldsymbol{\theta}')\right\| d\alpha \\ \overset{(a)}{\le}\rho\int_0^1 \left\|\alpha\boldsymbol{\theta} - \boldsymbol{\theta}'\right\| d\alpha &\overset{(b)}{\le} \rho\int_0^1 \alpha\|\boldsymbol{\theta}\|d\alpha + \rho\|\boldsymbol{\theta}'\| \le \rho\left(\|\boldsymbol{\theta}\| + \|\boldsymbol{\theta}'\|\right)\end{aligned}$$

where (a) is true because of Hessian Lipschitz (7), in (b) we use the triangle inequality. □

Lemma 5. *Under Assumption 1, we have block-wise Lipschitz continuity as follows:*

$$\left\|\begin{bmatrix} \nabla^2_{\mathbf{x}'\mathbf{x}'} f(\boldsymbol{\theta}') & \nabla^2_{\mathbf{x}'\mathbf{y}'} f(\boldsymbol{\theta}') \\ \mathbf{0} & \nabla^2_{\mathbf{y}''\mathbf{y}''} f(\boldsymbol{\theta}'') \end{bmatrix} - \begin{bmatrix} \nabla^2_{\mathbf{xx}} f(\boldsymbol{\theta}) & \nabla^2_{\mathbf{xy}} f(\boldsymbol{\theta}) \\ \mathbf{0} & \nabla^2_{\mathbf{yy}} f(\boldsymbol{\theta}) \end{bmatrix}\right\| \le \rho\left(\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| + \|\boldsymbol{\theta}'' - \boldsymbol{\theta}\|\right), \forall\boldsymbol{\theta}, \boldsymbol{\theta}', \boldsymbol{\theta}'', \tag{21}$$

and

$$\left\|\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \nabla^2_{\mathbf{y}'\mathbf{x}'} f(\boldsymbol{\theta}') & \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \nabla^2_{\mathbf{yx}} f(\boldsymbol{\theta}) & \mathbf{0} \end{bmatrix}\right\| \le \rho\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|, \forall\boldsymbol{\theta}, \boldsymbol{\theta}', \tag{22}$$

where $\boldsymbol{\theta} = [\mathbf{x}\ \mathbf{y}]^\intercal$, $\boldsymbol{\theta}' = [\mathbf{x}'\ \mathbf{y}']^\intercal$, $\boldsymbol{\theta} = [\mathbf{x}''\ \mathbf{y}'']^\intercal$ and each of them is partitioned as two blocks, $\mathbf{x}, \mathbf{x}', \mathbf{x}'' \in \mathbb{R}^{d_\mathbf{x}}$ and $\mathbf{y}, \mathbf{y}', \mathbf{y}'' \in \mathbb{R}^{d_\mathbf{y}}$.

Proof. There proof involves two parts:

Upper Triangular Matrix Case: Consider three different vectors $\boldsymbol{\theta}$, $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}''$, where each of them is partitioned as two blocks in a same way, i.e., $\boldsymbol{\theta} = [\mathbf{x}\ \mathbf{y}]^\intercal$, $\boldsymbol{\theta}' = [\mathbf{x}'\ \mathbf{y}']^\intercal$, $\boldsymbol{\theta} = [\mathbf{x}''\ \mathbf{y}'']^\intercal$. We can have

$$\begin{aligned}&\left\|\begin{bmatrix} \nabla^2_{\mathbf{x}'\mathbf{x}'} f(\boldsymbol{\theta}') & \nabla^2_{\mathbf{x}'\mathbf{y}'} f(\boldsymbol{\theta}') \\ 0 & \nabla^2_{\mathbf{y}''\mathbf{y}''} f(\boldsymbol{\theta}'') \end{bmatrix} - \begin{bmatrix} \nabla^2_{\mathbf{xx}} f(\boldsymbol{\theta}) & \nabla^2_{\mathbf{xy}} f(\boldsymbol{\theta}) \\ 0 & \nabla^2_{\mathbf{yy}} f(\boldsymbol{\theta}) \end{bmatrix}\right\| \\ \le&\left\|\mathbf{I}_1\left(\begin{bmatrix} \nabla^2_{\mathbf{x}'\mathbf{x}'} f(\boldsymbol{\theta}') & \nabla^2_{\mathbf{x}'\mathbf{y}'} f(\boldsymbol{\theta}') \\ \nabla^2_{\mathbf{y}'\mathbf{x}'} f(\boldsymbol{\theta}') & \nabla^2_{\mathbf{y}'\mathbf{y}'} f(\boldsymbol{\theta}') \end{bmatrix} - \begin{bmatrix} \nabla^2_{\mathbf{xx}} f(\boldsymbol{\theta}) & \nabla^2_{\mathbf{xy}} f(\boldsymbol{\theta}) \\ \nabla^2_{\mathbf{yx}} f(\boldsymbol{\theta}) & \nabla^2_{\mathbf{yy}} f(\boldsymbol{\theta}) \end{bmatrix}\right)\right\| \\ &+\left\|\mathbf{I}_2\left(\begin{bmatrix} \nabla^2_{\mathbf{x}''\mathbf{x}''} f(\boldsymbol{\theta}'') & \nabla^2_{\mathbf{x}''\mathbf{y}''} f(\boldsymbol{\theta}'') \\ \nabla^2_{\mathbf{y}''\mathbf{x}''} f(\boldsymbol{\theta}'') & \nabla^2_{\mathbf{y}''\mathbf{y}''} f(\boldsymbol{\theta}'') \end{bmatrix} - \begin{bmatrix} \nabla^2_{\mathbf{xx}} f(\boldsymbol{\theta}) & \nabla^2_{\mathbf{xy}} f(\boldsymbol{\theta}) \\ \nabla^2_{\mathbf{yx}} f(\boldsymbol{\theta}) & \nabla^2_{\mathbf{yy}} f(\boldsymbol{\theta}) \end{bmatrix}\right)\mathbf{I}_2\right\| \\ \overset{(a)}{\le}&\left\|\begin{bmatrix} \nabla^2_{\mathbf{x}'\mathbf{x}'} f(\boldsymbol{\theta}') & \nabla^2_{\mathbf{x}'\mathbf{y}'} f(\boldsymbol{\theta}') \\ \nabla^2_{\mathbf{y}'\mathbf{x}'} f(\boldsymbol{\theta}') & \nabla^2_{\mathbf{y}'\mathbf{y}'} f(\boldsymbol{\theta}') \end{bmatrix} - \begin{bmatrix} \nabla^2_{\mathbf{xx}} f(\boldsymbol{\theta}) & \nabla^2_{\mathbf{xy}} f(\boldsymbol{\theta}) \\ \nabla^2_{\mathbf{yx}} f(\boldsymbol{\theta}) & \nabla^2_{\mathbf{yy}} f(\boldsymbol{\theta}) \end{bmatrix}\right\| \\ &+\left\|\begin{bmatrix} \nabla^2_{\mathbf{x}''\mathbf{x}''} f(\boldsymbol{\theta}'') & \nabla^2_{\mathbf{x}''\mathbf{y}''} f(\boldsymbol{\theta}'') \\ \nabla^2_{\mathbf{y}''\mathbf{x}''} f(\boldsymbol{\theta}'') & \nabla^2_{\mathbf{y}''\mathbf{y}''} f(\boldsymbol{\theta}'') \end{bmatrix} - \begin{bmatrix} \nabla^2_{\mathbf{xx}} f(\boldsymbol{\theta}) & \nabla^2_{\mathbf{xy}} f(\boldsymbol{\theta}) \\ \nabla^2_{\mathbf{yx}} f(\boldsymbol{\theta}) & \nabla^2_{\mathbf{yy}} f(\boldsymbol{\theta}) \end{bmatrix}\right\| \\ \overset{(7)}{\le}&\rho\left(\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| + \|\boldsymbol{\theta}'' - \boldsymbol{\theta}\|\right)\end{aligned}$$

where in (a) we use

$$\mathbf{I}_1 = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \qquad \mathbf{I}_2 = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \tag{23}$$

and $\|\mathbf{I}_1\| = \|\mathbf{I}_2\| = 1$.

Lower Triangular Matrix Case:

$$\begin{aligned}
&\left\|\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \nabla^2_{\mathbf{y}'\mathbf{x}'}f(\boldsymbol{\theta}') & \mathbf{0}\end{bmatrix} - \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \nabla^2_{\mathbf{y}\mathbf{x}}f(\boldsymbol{\theta}) & \mathbf{0}\end{bmatrix}\right\|\\
=&\left\|\mathbf{I}_2\left(\begin{bmatrix}\nabla^2_{\mathbf{x}'\mathbf{x}'}f(\boldsymbol{\theta}') & \nabla^2_{\mathbf{x}'\mathbf{y}'}f(\boldsymbol{\theta}')\\ \nabla^2_{\mathbf{y}'\mathbf{x}'}f(\boldsymbol{\theta}') & \nabla^2_{\mathbf{y}'\mathbf{y}'}f(\boldsymbol{\theta}')\end{bmatrix} - \begin{bmatrix}\nabla^2_{\mathbf{x}\mathbf{x}}f(\boldsymbol{\theta}) & \nabla^2_{\mathbf{x}\mathbf{y}}f(\boldsymbol{\theta})\\ \nabla^2_{\mathbf{y}\mathbf{x}}f(\boldsymbol{\theta}) & \nabla^2_{\mathbf{y}\mathbf{y}}f(\boldsymbol{\theta})\end{bmatrix}\right)\mathbf{I}_1\right\|\\
\overset{(a)}{\le}&\rho\|\boldsymbol{\theta}'-\boldsymbol{\theta}\|
\end{aligned}$$

where (a) is true because we know $\|\mathbf{I}_1\| = \|\mathbf{I}_2\| = 1$. □

Then, we illustrate that the size of the partial gradient with one round update by the A-GD algorithm has the following relation with the full size of the gradient.

Lemma 6. *If function $f(\cdot)$ is L-smooth with Lipschitz constant, then we have*

$$\|\nabla f(\boldsymbol{\theta}^{(t)})\|^2 \le \left(1+2\left(\frac{L}{L_{\max}}\right)^2\right)\left(\|\nabla_{\mathbf{x}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\|^2 + \|\nabla_{\mathbf{y}}f(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})\|^2\right) \tag{24}$$

where sequence $\boldsymbol{\theta}^{(t)}$ is generated by the A-GD algorithm.

Proof. First, we have

$$\|\nabla_{\mathbf{y}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\|^2 \le 2\|\nabla_{\mathbf{y}}f(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)}) - \nabla_{\mathbf{y}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\|^2 + 2\|\nabla_{\mathbf{y}}f(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})\|^2. \tag{25}$$

Using the block Lipschitz continuity of the objective function, we have

$$\begin{aligned}
\|\nabla_{\mathbf{y}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\|^2 &\le 2L^2\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\|^2 + 2\|\nabla_{\mathbf{y}}f(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})\|^2\\
&\overset{(a)}{=} 2L^2\|\eta\nabla_{\mathbf{x}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\|^2 + 2\|\nabla_{\mathbf{y}}f(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})\|^2\\
&\overset{(b)}{\le} 2\left(\left(\frac{L}{L_{\max}}\right)^2\|\nabla_{\mathbf{x}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\|^2 + \|\nabla_{\mathbf{y}}f(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})\|^2\right)
\end{aligned} \tag{26}$$

where (a) is because we use the update rule of A-GD, (b) is true due to $\eta \le 1/L_{\max}$.

Summing $\|\nabla_{\mathbf{x}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\|^2$ on both sides of the above equation, we have

$$\|\nabla f(\boldsymbol{\theta}^{(t)})\|^2 = \|\nabla_{\mathbf{x}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\|^2 + \|\nabla_{\mathbf{y}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\|^2 \tag{27}$$

$$\le \left(1+2\left(\frac{L}{L_{\max}}\right)^2\right)\left(\|\nabla_{\mathbf{x}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\|^2 + \|\nabla_{\mathbf{y}}f(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})\|^2\right) \tag{28}$$

$$\le 2\Bigg(\underbrace{1+\frac{L}{L_{\max}}}_{=p_1}\Bigg)^2\left(\|\nabla_{\mathbf{x}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\|^2 + \|\nabla_{\mathbf{y}}f(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})\|^2\right). \tag{29}$$

□

B. Proofs of PA-GD

As stated in the main body of the paper, we can use Lemma 2 and Lemma 3 to prove Theorem 1. Lemma 2 is basically well-known. The main task focuses on proving Lemma 3, which consists of a sequence of lemmas (Lemma 7–Lemma 9) that lead to Lemma 3.

Before discussing the details of Lemma 3, we need to introduce some constants defined as follows,

$$\mathcal{F} \triangleq \frac{\gamma^3}{\rho^2 \widehat{c}^5} \log^{-3}\left(\frac{d\kappa}{\delta}\right) p_1^{-3} p_2^{-2}, \tag{30a}$$

$$\mathcal{G} \triangleq \frac{\gamma^2}{\rho} \log^{-2}\left(\frac{d\kappa}{\delta}\right) p_1^{-2} p_2^{-1}, \tag{30b}$$

$$\mathcal{S} \triangleq \frac{\gamma}{\rho \widehat{c}^2} \log^{-1}\left(\frac{d\kappa}{\delta}\right) p_1^{-1} p_2^{-1}, \tag{30c}$$

$$\mathcal{T} \triangleq \frac{\log\left(\frac{d\kappa}{\delta}\right) p_1}{\eta\gamma}. \tag{30d}$$

These quantities refer to different units of the algorithm. Specifically, $\mathcal{F}$ accounts for the objective value, $\mathcal{G}$ for the size of the gradient, $\mathcal{S}$ for the norm of the difference between iterates, and $\mathcal{T}$ for the number of iterations. Also, we define a condition number in terms of γ as $\kappa \triangleq \frac{L_{\max}}{\gamma} \geq 1$. In the process of the proofs, we also use conditions $\log(\frac{d\kappa}{\delta}) \geq 1$ when $\delta \in (0, \frac{d\kappa}{e}]$, $p_1 \geq 2$ repeatedly to simply the expressions of the parameters.

According to Lemma 3, we consider saddle point $\widetilde{\boldsymbol{\theta}}^{(t)}$ that satisfies the following condition.

Condition 1. *A strict saddle point $\widetilde{\boldsymbol{\theta}}^{(t)}$ satisfies the following conditions:*

$$\|\nabla_{\widetilde{\mathbf{x}}} f(\widetilde{\mathbf{x}}^{(t)}, \widetilde{\mathbf{y}}^{(t)})\|^2 + \|\nabla_{\widetilde{\mathbf{y}}} f(\widetilde{\mathbf{x}}^{(t+1)}, \widetilde{\mathbf{y}}^{(t)})\|^2 \leq g_{th}^2 \quad \text{and} \quad \lambda_{\min}(\nabla^2 f(\widetilde{\boldsymbol{\theta}}^{(t)})) \leq -\gamma \tag{31}$$

where $g_{th} \triangleq \mathcal{G}$, $\gamma = \sqrt{\rho\epsilon}$, $\delta = \frac{dL_{\max}}{\sqrt{\rho\epsilon}} e^{-\chi}$ and $\eta \leq \frac{1}{L_{\max}}$ in Algorithm 1.

Remark 3. Condition 1 implies that point $\widetilde{\boldsymbol{\theta}}^{(t)}$ satisfies $\|\nabla f(\widetilde{\boldsymbol{\theta}}^{(t)})\| \leq \epsilon$ (see Lemma 6 and definition of $\mathcal{G}$ or (37) in Section B.1) and $\lambda_{\min}(\nabla^2 f(\widetilde{\boldsymbol{\theta}}^{(t)})) \leq -\sqrt{\rho\epsilon}$.

Sufficient Decrease after Perturbation Consider $\widetilde{\boldsymbol{\theta}}^{(t)}$ satisfy Condition 1 and let $\boldsymbol{\mathcal{H}} \triangleq \nabla^2 f(\widetilde{\boldsymbol{\theta}}^{(t)})$.

With these definitions of parameters, we will study how PA-GD can escape from strict saddle points. The main part of the proof is to show that when two sequences are apart from each other with a certain distance along the $\vec{\mathbf{e}}$ direction at the starting points, where $\vec{\mathbf{e}}$ denotes the eigenvector of $\mathbf{M}^{-1}\mathbf{T}$ whose eigenvalue is maximum (greater than 1). Then, after a number of iterations at least one of them can give a sufficient decrease of the objective value. This property implies the iterates can easily escape from the saddle points as long as there is a large enough perturbation between the initial points of the two sequences along the $\vec{\mathbf{e}}$ direction. We will introduce the following two lemmas formally which are the main contributions of this work.

Lemma 7. *Under Assumption 1, consider $\widetilde{\boldsymbol{\theta}}^{(t)}$ that satisfies Condition 1 and a generic sequence $\mathbf{u}^{(t)}$ generated by A-GD. For any constant $\widehat{c} \geq 1$, $\delta \in (0, \frac{d\kappa}{e}]$, when initial point $\mathbf{u}^{(1)}$ satisfies*

$$\|\mathbf{u}^{(1)} - \widetilde{\boldsymbol{\theta}}^{(t)}\| \leq 2r, \tag{32}$$

then, with the definition of

$$r \triangleq \frac{\mathcal{S}}{\widehat{c}^3 \kappa \log(\frac{d\kappa}{\delta}) p_1}, \quad \text{and} \quad T \triangleq \min\left\{\inf_{t\geq 1}\{t | f(\mathbf{u}^{(t+2)}) - f(\mathbf{u}^{(1)}) \leq -2\mathcal{F}\}, \widehat{c}\mathcal{T}\right\}, \tag{33}$$

for any $\eta \leq 1/L_{\max}$, the iterates generated by A-GD satisfy $\|\mathbf{u}^{(t+1)} - \widetilde{\boldsymbol{\theta}}^{(t)}\| \leq 3\mathcal{S}, \forall t \leq T$.

Lemma 8. *Under Assumption 1, consider $\widetilde{\boldsymbol{\theta}}^{(t)}$ that satisfies Condition 1. for any $\widehat{c} \geq \min\{136, 8(3p_0^2 + 12p_0 + 12)\}$, $\delta \in (0, \frac{d\kappa}{e}]$ and $\eta \leq 1/L_{\max}$, with the definition of $p_0 \triangleq \max\{6, 4(L/L_{\max})^2 + 1\}$ and*

$$T \triangleq \min\left\{\inf_{t\geq 1}\{t | f(\mathbf{w}^{(t+2)}) - f(\mathbf{w}^{(1)}) \leq -2\mathcal{F}\}, \widehat{c}\mathcal{T}\right\} \tag{34}$$

where two iterates $\{\mathbf{u}^{(t)}\}$ and $\{\mathbf{w}^{(t)}\}$ that are generated by A-GD with initial points $\{\mathbf{u}^{(1)}, \mathbf{w}^{(1)}\}$ satisfying

$$\|\mathbf{u}^{(1)} - \widetilde{\boldsymbol{\theta}}^{(t)}\| \leq r, \ \mathbf{w}^{(1)} = \mathbf{u}^{(1)} + \upsilon r \vec{\mathbf{e}}, \ \upsilon \in [\delta/(2\sqrt{d}), 1], \tag{35}$$

where $\vec{\mathbf{e}}$ denotes the eigenvector of $\mathbf{M}^{-1}\mathbf{T}$ whose eigenvalue is maximum, then, if $\|\mathbf{u}^{(t+1)} - \widetilde{\boldsymbol{\theta}}^{(t)}\| \leq 3\mathcal{S}, \forall t \leq T$, we will have $T < \widehat{c}\mathcal{T}$.

Lemma 7 says that if the $\mathbf{u}^{(t)}$-iterate generated by A-GD cannot provide a sufficient decrease of the objective value, then the iterates are constrained within the area which is very close to the saddle point. With this property, Lemma 8 shows if there exists another A-GD iterate $\mathbf{w}^{(t)}$, which is initialized with a certain distance along the $\vec{\mathbf{e}}$ direction from the $\mathbf{u}$-iterate, then $\mathbf{w}^{(t)}$ will provide a sufficient decrease of the objective value. These two lemmas characterize the convergence behavior of the A-GD iterates.

Escaping from Saddle Points Then, we need to quantify the probability that after adding the perturbation the algorithm cannot escape from strict saddle points. In previous work about escaping from saddle points with GD, a characterization of the geometry around saddle points has been given (Jin et al., 2017, Lemma 15). Once we know that PA-GD also decreases the objective value sufficiently in Lemma 7 and Lemma 8, the following lemma can be claimed straightforwardly. To be more specific, we can obtain the probability that iterates will be stuck at the strict points after T iterations as follows.

$$\begin{aligned}
\mathbb{P}(\mathbf{w}^{(1)} \in \mathcal{X}_{\text{stuck}}) &= \int_{\mathbb{B}_{\widetilde{\boldsymbol{\theta}}^{(t)}}(r)} \mathbb{P}(\mathbf{w}^{(1)} \in \mathcal{X}_{\text{stuck}} | \mathbf{u}^{(1)} \in \mathcal{X}_{\text{stuck}}) \mathbb{P}(\mathbf{u}^{(1)} \in \mathcal{X}_{\text{stuck}}) d\mathbf{u}^{(1)} \\
&\leq \int_{\mathbb{B}_{\widetilde{\boldsymbol{\theta}}^{(t)}}(r)} \mathbb{P}(\mathbf{w}^{(1)} \in \mathcal{X}_{\text{stuck}} | \mathbf{u}^{(1)} \in \mathcal{X}_{\text{stuck}}) \mathbb{P}(\mathbf{u}^{(1)}) d\mathbf{u}^{(1)} \\
&\overset{(a)}{\leq} \delta \int_{\mathbb{B}_{\widetilde{\boldsymbol{\theta}}^{(t)}}(r)} \mathbb{P}(\mathbf{u}^{(1)}) d\mathbf{u}^{(1)} = \delta
\end{aligned}$$

where $\mathcal{X}_{\text{stuck}}$ denotes the set where the algorithm starts such that the sequence cannot escape from the strict saddle point after T iterations, (a) is true because probability $\mathbb{P}(\mathbf{w}^{(1)} \in \mathcal{X}_{\text{stuck}} | \mathbf{u}^{(1)} \in \mathcal{X}_{\text{stuck}})$ can be upper bounded by δ, which is proven in the following lemma.

Lemma 9. *Under Assumption 1, there exists a $\widehat{c} \geq \min\{136, 8(3p_0^2 + 12p_0 + 12)\}$ where $p_0 \triangleq \max\{6, 4(L/L_{\max})^2 + 1\}$, for any $\eta \leq 1/L_{\max}$ and $\delta \in (0, d\kappa/e]$: consider a saddle point $\widetilde{\boldsymbol{\theta}}^{(t)}$ which satisfies Condition 1, let $\boldsymbol{\theta}^{(1)} = \widetilde{\boldsymbol{\theta}}^{(t)} + \xi$ where ξ is generated randomly which follows the uniform distribution over a ball with radius r, and let $\boldsymbol{\theta}^{(t)}$ be the iterates of PA-GD starting from $\boldsymbol{\theta}^{(1)}$. Then, with at least probability $1 - \delta$, we have the following for any $T \geq \widehat{c}\mathcal{T} + 3$*

$$f(\boldsymbol{\theta}^{(T)}) - f(\widetilde{\boldsymbol{\theta}}^{(t)}) \leq -\mathcal{F}. \tag{36}$$

Then, applying $\eta \leq \frac{1}{L_{\max}}$,$\gamma = \sqrt{\rho\epsilon}$, and $\delta = \frac{dL_{\max}}{\sqrt{\rho\epsilon}} e^{-\chi}$ into Lemma 9, we can get Lemma 3 immediately.

With these lemmas, we can give the proof of Theorem 1 as the following.

B.1. Proof of Theorem 1

Next, we prove the main theorem.

Proof. Submitting $\gamma = \sqrt{\rho\epsilon}$, and $\delta = \frac{dL_{\max}}{\sqrt{\rho\epsilon}} e^{-\chi}$ into the definitions of $\mathcal{F}, \mathcal{G}, \mathcal{T}$, we will have the following definitions.

$$\begin{aligned}
f_{\text{th}} &\triangleq \mathcal{F} = \frac{\sqrt{\frac{\epsilon^3}{\rho}}}{\widehat{c}^5 (\chi p_1)^3 p_2^2}, \\
g_{\text{th}} &\triangleq \mathcal{G} = \frac{\epsilon}{(\chi p_1)^2 p_2}, \\
t_{\text{th}} &\triangleq \widehat{c}\mathcal{T} + 3 = \frac{\widehat{c} L_{\max} \chi p_1}{\sqrt{\rho\epsilon}} + 3.
\end{aligned}$$

After applying Lemma 6, we know that

$$\|\nabla f(\widetilde{\boldsymbol{\theta}}^{(t)})\| \le \frac{\epsilon}{\chi^2 p_2} \le \epsilon \tag{37}$$

where $\chi, p_2 \ge 1$.

With a set of necessary lemmas and leveraging the proof of PGD (Jin et al., 2017, Theorem 3), we have the following convergence analysis of PA-GD. Specifically, at any iteration, we need to consider two cases (we use the first iteration as an example):

1. In this case the gradient is large such that $\|\nabla_{\mathbf{x}} f(\mathbf{x}^{(1)}, \mathbf{y}^{(1)})\|^2 + \|\nabla_{\mathbf{y}} f(\mathbf{x}^{(2)}, \mathbf{y}^{(1)})\|^2 > g_{\mathsf{th}}^2$: According to Lemma 2, we have

$$\begin{aligned} f(\boldsymbol{\theta}^{(2)}) - f(\boldsymbol{\theta}^{(1)}) &\le -\frac{\eta}{2}\left(\|\nabla_{\mathbf{x}} f(\mathbf{x}^{(1)}, \mathbf{y}^{(1)})\|^2 + \|\nabla_{\mathbf{y}} f(\mathbf{x}^{(2)}, \mathbf{y}^{(1)})\|^2\right) \le -\frac{\eta}{2} g_{\mathsf{th}}^2 \\ &\overset{(a)}{\le} -\frac{1}{2(\chi p_1)^4 p_2^2}\frac{\epsilon^2}{L_{\max}} \end{aligned} \tag{38}$$

where in (a) use the definition of g_{th}^2 and $\eta \le 1/L_{\max}$.

2. The gradient is small in all block directions, namely $\|\nabla_{\mathbf{x}} f(\mathbf{x}^{(1)}, \mathbf{y}^{(1)})\|^2 + \|\nabla_{\mathbf{y}} f(\mathbf{x}^{(2)}, \mathbf{y}^{(1)})\|^2 \le g_{\mathsf{th}}^2$: in this case, we will add the perturbation to the iterates, and implement A-GD for the next t_{th} steps and then check the termination condition. If the termination condition is not satisfied, we must have

$$f(\boldsymbol{\theta}^{(t_{\mathsf{th}})}) - f(\boldsymbol{\theta}^{(1)}) \le -f_{\mathsf{th}} = -\frac{\sqrt{\frac{\epsilon^3}{\rho}}}{\widehat{c}^5 (\chi p_1)^3 p_2^2}, \tag{39}$$

which implies that the objective value in each step on average is decreased by

$$\frac{f(\boldsymbol{\theta}^{(t_{\mathsf{th}})}) - f(\boldsymbol{\theta}^{(1)})}{t_{\mathsf{th}}} \le -\frac{1}{2\widehat{c}^6 (\chi p_1)^4 p_2^2}\frac{\epsilon^2}{L_{\max}} \tag{40}$$

where we use $t_{\mathsf{th}} < 2\frac{\widehat{c} L_{\max} \chi p_1}{\sqrt{\rho\epsilon}}$ since $p_1 \ge 1, \chi > 1, \widehat{c} \ge 136$.

With the results of these two cases, we can know that if there is a large size of the gradient, we can know the decrease of the objective function value by the result of case 1, and if not, we use the result of case 2. In summary, PA-GD can have a sufficient decrease of the objective function value by $\frac{1}{2\widehat{c}^6(\chi p_1)^4 p_2^2}\frac{\epsilon^2}{L_{\max}}$ per iteration on average. This means that Algorithm 1 must stop within a finite number of iterations, which is

$$\frac{f(\boldsymbol{\theta}^{(1)}) - f^\star}{\frac{1}{2\widehat{c}^6(\chi p_1)^4 p_2^2}\frac{\epsilon^2}{L_{\max}}} = 2\widehat{c}^6 (\chi p_1)^4 p_2^2 \frac{L_{\max}\Delta_f}{\epsilon^2} = \mathcal{O}\left(\frac{\widehat{c}^6 \Delta_f (\chi p_1)^4 p_2^2 L_{\max}}{\epsilon^2}\right) \tag{41}$$

where $\Delta_f \triangleq f(\boldsymbol{\theta}^{(1)}) - f^\star$.

According to Lemma 3, we know that with probability $1 - \frac{dL_{\max}}{\sqrt{\rho\epsilon}} e^{-\chi}$ the algorithm can give a sufficient descent with the perturbation when $\|\nabla_{\mathbf{x}} f(\mathbf{x}^{(1)}, \mathbf{y}^{(1)})\|^2 + \|\nabla_{\mathbf{y}} f(\mathbf{x}^{(2)}, \mathbf{y}^{(1)})\|^2 \le g_{\mathsf{th}}^2$. Since the total number of perturbation we can add is at most

$$n = \frac{1}{t_{\mathsf{th}}} 2\widehat{c}^6 (\chi p_1)^4 p_2^2 \frac{L_{\max}\Delta_f}{\epsilon^2} \le \frac{1}{\widehat{c}\mathcal{T}} 2\widehat{c}^6 (\chi p_1)^4 p_2^2 \frac{L_{\max}\Delta_f}{\epsilon^2} = 2(p_1\chi)^3 p_2^2 \widehat{c}^5 \frac{\sqrt{\rho\epsilon}\Delta_f}{\epsilon^2} \tag{42}$$

where we use $t_{\mathsf{th}} \ge \widehat{c}\mathcal{T}$.

Using the union bound, the probability of Lemma 3 being satisfied for all perturbations is

$$1 - n\frac{dL_{\max}}{\sqrt{\rho\epsilon}} e^{-\chi} = 1 - \frac{dL_{\max}}{\sqrt{\rho\epsilon}} e^{-\chi} 2\widehat{c}^5 (p_1\chi)^3 p_2^2 \frac{\sqrt{\rho\epsilon}\Delta_f}{\epsilon^2} = 1 - \underbrace{2\widehat{c}^5 d L_{\max} p_1^3 p_2^2 \frac{\Delta_f}{\epsilon^2}}_{\triangleq \mathcal{C}} \chi^3 e^{-\chi}. \tag{43}$$

With chosen $\chi = 3\max\{\ln(\mathcal{C}/\delta), 4\}$, we have $\chi^3 e^{-\chi} \le e^{-\chi/3}$, which implies $\chi^3 e^{-\chi}\mathcal{C} \le e^{-\chi/3}\mathcal{C} \le \delta$.

The proof is complete. □

B.2. Proof of Lemma 1

Proof. Recall the definitions:

$$\boldsymbol{\mathcal{H}}_u \triangleq \begin{bmatrix} \nabla^2_{\mathbf{xx}} f(\widetilde{\boldsymbol{\theta}}^{(t)}) & \nabla^2_{\mathbf{xy}} f(\widetilde{\boldsymbol{\theta}}^{(t)}) \\ 0 & \nabla^2_{\mathbf{yy}} f(\widetilde{\boldsymbol{\theta}}^{(t)}) \end{bmatrix} \quad \boldsymbol{\mathcal{H}}_l \triangleq \begin{bmatrix} 0 & 0 \\ \nabla^2_{\mathbf{yx}} f(\widetilde{\boldsymbol{\theta}}^{(t)}) & 0 \end{bmatrix}, \tag{44}$$

where $\widetilde{\boldsymbol{\theta}}^{(t)}$ is an ϵ-SOSP, and

$$\mathbf{M} \triangleq \mathbf{I} + \eta\boldsymbol{\mathcal{H}}_l, \quad \mathbf{T} \triangleq \mathbf{I} - \eta\boldsymbol{\mathcal{H}}_u. \tag{45}$$

Our goal of this lemma is to show that the maximum eigenvalue of $\mathbf{M}^{-1}\mathbf{T}$ is strictly greater than 1 so that we can project iterates $\mathbf{v}^{(t)}$ onto the two subspaces, where the first subspace is spanned by the eigenvector of $\mathbf{M}^{-1}\mathbf{T}$ whose eigenvalue is the largest (greater than 1) and the other one is spanned by the remaining eigenvectors.

Note that $\mathbf{M}$ is a lower triangular matrix and $\det(\mathbf{M}) = 1$, which implies that $\det(\mathbf{M}^{-1}\mathbf{T} - \lambda\mathbf{I}) = \det(\mathbf{T} - \lambda\mathbf{M})$, where λ denotes the eigenvalue. We can analyze the determinant of $\mathbf{T} - \lambda\mathbf{M}$, i.e.,

$$\begin{aligned} \det[\mathbf{T} - \lambda\mathbf{M}] =& \det[\mathbf{I} - \eta\boldsymbol{\mathcal{H}}_u - \lambda(\mathbf{I} + \eta\boldsymbol{\mathcal{H}}_l)] \\ =& \det\left[\underbrace{\begin{matrix} (1-\lambda)\mathbf{I} - \eta\nabla^2_{\mathbf{xx}} f(\widetilde{\boldsymbol{\theta}}^{(t)}) & -\eta\nabla^2_{\mathbf{xy}} f(\widetilde{\boldsymbol{\theta}}^{(t)}) \\ -\lambda\eta\nabla^2_{\mathbf{yx}} f(\widetilde{\boldsymbol{\theta}}^{(t)}) & (1-\lambda)\mathbf{I} - \eta\nabla^2_{\mathbf{yy}} f(\widetilde{\boldsymbol{\theta}}^{(t)}) \end{matrix}}_{\triangleq\mathbf{Q}(\lambda)}\right]. \end{aligned}$$

Then, we use the following two steps to show $\lambda_{\max}(\mathbf{M}^{-1}\mathbf{T}) > 1$: 1) we can show that all eigenvalues of $\mathbf{Q}(1+\delta)$ are real where $\delta > 0$; 2) there exists a $\lambda > 1$ such that $\det(\mathbf{Q}(\lambda)) = 0$.

All the eigenvalues of $\mathbf{Q}(1+\delta)$ are real:

Consider a $\delta > 0$. We have

$$\mathbf{Q}(1+\delta) = -\left(\underbrace{\eta\boldsymbol{\mathcal{H}} + \delta(\mathbf{I} + \eta\boldsymbol{\mathcal{H}}_l)}_{\triangleq\mathbf{F}(\delta)}\right) \tag{46}$$

where

$$\begin{aligned} \mathbf{F}(\delta) =& \delta\mathbf{I} + \eta\begin{bmatrix} \nabla^2_{\mathbf{xx}} f(\widetilde{\boldsymbol{\theta}}^{(t)}) & \nabla^2_{\mathbf{xy}} f(\widetilde{\boldsymbol{\theta}}^{(t)}) \\ (1+\delta)\nabla^2_{\mathbf{yx}} f(\widetilde{\boldsymbol{\theta}}^{(t)}) & \nabla^2_{\mathbf{yy}} f(\widetilde{\boldsymbol{\theta}}^{(t)}) \end{bmatrix} \\ =& \begin{bmatrix} \mathbf{I} & \\ & \sqrt{1+\delta} \end{bmatrix} \underbrace{\begin{bmatrix} \delta\mathbf{I} + \eta\nabla^2_{\mathbf{xx}} f(\widetilde{\boldsymbol{\theta}}^{(t)}) & \eta\sqrt{1+\delta}\nabla^2_{\mathbf{xy}} f(\widetilde{\boldsymbol{\theta}}^{(t)}) \\ \eta\sqrt{1+\delta}\nabla^2_{\mathbf{yx}} f(\widetilde{\boldsymbol{\theta}}^{(t)}) & \delta\mathbf{I} + \eta\nabla^2_{\mathbf{yy}} f(\widetilde{\boldsymbol{\theta}}^{(t)}) \end{bmatrix}}_{\mathbf{G}(\delta)} \begin{bmatrix} \mathbf{I} & \\ & \frac{1}{\sqrt{1+\delta}} \end{bmatrix}, \end{aligned}$$

meaning that $\mathbf{F}(\delta)$ is similar to $\mathbf{G}(\delta)$. Hence, $\mathbf{F}(\delta)$ has the same eigenvalues of $\mathbf{G}(\delta)$. Since $\mathbf{G}(\delta)$ is a symmetric matrix, all the eigenvalues of $\mathbf{G}(\delta)$ are real. From (46), we can conclude that all the the eigenvalues of $\mathbf{Q}(1+\delta)$ are real.

Quantify $\|\eta\boldsymbol{\mathcal{H}} - \mathbf{G}(\delta)\|$:

Since we know that $\boldsymbol{\mathcal{H}}$ and $\mathbf{G}(\delta)$ are diagonalizable (normal matrices), then we have the following result (Weyl, 1912) (or (Holbrook, 1992)) of quantifying the difference of the eigenvalues of the two normal matrices

$$\max_{1\le i\le d} |\lambda_i(\eta\boldsymbol{\mathcal{H}}) - \lambda_i(\mathbf{G}(\delta))| \le \|\eta\boldsymbol{\mathcal{H}} - \mathbf{G}(\delta)\| \tag{47}$$

where $\lambda_i(\boldsymbol{\mathcal{H}})$ and $\lambda_i(\mathbf{G}(\delta))$ denote the ith eigenvalue of $\boldsymbol{\mathcal{H}}$ and $\mathbf{G}(\delta)$, which are listed in a decreasing order.

With the help of (47), we can check

$$\begin{aligned}
&\|\eta\boldsymbol{\mathcal{H}}-\mathbf{G}(\delta)\|\\
=&\left\|\delta\mathbf{I}+\begin{bmatrix} 0 & (\sqrt{1+\delta}-1)\eta\nabla^2_{\mathbf{xy}}f(\widetilde{\boldsymbol{\theta}}^{(t)}) \\ (\sqrt{1+\delta}-1)\eta\nabla^2_{\mathbf{yx}}f(\widetilde{\boldsymbol{\theta}}^{(t)}) & 0\end{bmatrix}\right\|\\
\le&\delta+\left(\sqrt{1+\delta}-1\right)\eta\|\boldsymbol{\mathcal{H}}\|+\left(\sqrt{1+\delta}-1\right)\eta\left\|\begin{matrix}\nabla^2_{\mathbf{xx}}f(\widetilde{\boldsymbol{\theta}}^{(t)}) & 0\\ 0 & \nabla^2_{\mathbf{yy}}f(\widetilde{\boldsymbol{\theta}}^{(t)})\end{matrix}\right\|\\
\overset{(a)}{\le}&\delta+\left(\sqrt{1+\delta}-1\right)\left(\frac{L}{L_{\max}}+1\right)
\end{aligned}\tag{48}$$

where (a) is true since we used $\eta\le 1/L_{\max}$ and the fact that $\|\boldsymbol{\mathcal{H}}\|\le L$ and $\|\boldsymbol{\mathcal{H}}_d\|\le L_{\max}$. Also, it can be observed that when $\delta=0$, matrix $\mathbf{G}(\delta)$ is reduced to $\eta\boldsymbol{\mathcal{H}}$. Note that if $\eta=1/L$ is used, then we have $\|\eta\boldsymbol{\mathcal{H}}-\mathbf{G}(\delta)\|\le\delta+2(\sqrt{1+\delta}-1)$.

Quantify δ:

From Condition 1, we know that the minimum eigenvalue of $\eta\boldsymbol{\mathcal{H}}$ is less than $-\eta\gamma$ and the maximum difference of the eigenvalues between $\eta\boldsymbol{\mathcal{H}}$ and $\mathbf{G}(\delta)$ is upper bounded by (48). Then, we can choose a sufficient small δ such that $\mathbf{G}(\delta)$ also has a negative eigenvalue, meaning that we need to find a δ such that

$$\delta+\left(\sqrt{1+\delta}-1\right)\left(\frac{L}{L_{\max}}+1\right)<\eta\gamma.\tag{49}$$

In other words, if we choose

$$\delta^*=\frac{\eta\gamma}{1+\frac{L}{L_{\max}}}$$

then we can conclude that $\mathbf{G}(\delta^*)$ has a negative eigenvalue which is less than $-\eta\gamma+\delta^*=-\frac{\eta\gamma}{1+\frac{L_{\max}}{L}}$.

In the following, we will check that δ^* is a valid choice, meaning that equation (49) holds when $\delta^*=\frac{\eta\gamma}{1+\frac{L}{L_{\max}}}$.

First step: since $L/L_{\max}\ge 1$, we have $\eta\gamma/(1+L/L_{\max})\le\eta\gamma/2$.

Second step: we only need to check

$$\left(\sqrt{1+\delta}-1\right)\left(\frac{L}{L_{\max}}+1\right)<\frac{\eta\gamma}{2},$$

meaning that it is sufficient to check

$$\left(\frac{L}{L_{\max}}+1\right)^2(1+\delta)<\left(\frac{L}{L_{\max}}+1+\frac{\eta\gamma}{2}\right)^2.\tag{50}$$

It can be easily check that the left-hand side (LHS) of (50) with chosen δ^* is

$$\begin{aligned}
\left(\frac{L}{L_{\max}}+1\right)^2\left(1+\frac{\eta\gamma}{\frac{L}{L_{\max}}+1}\right)\le&\left(\frac{L}{L_{\max}}+1\right)^2+\left(\frac{L}{L_{\max}}+1\right)^2\eta\gamma\\
<&\left(\frac{L}{L_{\max}}+1\right)^2+\left(\frac{L}{L_{\max}}+1\right)^2\eta\gamma+\frac{\eta^2\gamma^2}{4},
\end{aligned}$$

which is RHS of (50). Therefore, we can conclude that $\mathbf{G}(\delta^*)$ has a negative eigenvalue.

There exists a $\lambda>1$ such that $\det(\mathbf{Q}(\lambda))=0$:

When δ is large, it is easy to check $\mathbf{G}(\delta)$ has a positive eigenvalue, since term $\delta^2\mathbf{I}$ dominates the spectrum of matrix $\mathbf{G}(\delta)$ in (46). Since the eigenvalue is continuous with respect to δ, we can conclude there exists a largest δ, i.e., $\widehat{\delta}$, such that $\mathbf{G}(\widehat{\delta})$ has a zero eigenvalue, i.e., $\det(\mathbf{Q}(\widehat{\delta}))=0$ where $\widehat{\delta}$ is at least $\frac{\eta\gamma}{1+\frac{L}{L_{\max}}}$. From (46), we know that there exits a $\lambda=1+\widehat{\delta}$

such that $\det(\mathbf{G}(1+\widehat{\delta}))=0$ since $\det(\mathbf{G}(\widehat{\delta}))=\det(\mathbf{Q}(1+\widehat{\delta}))=0$. Equivalently, there exits a eigenvalue of $\mathbf{M}^{-1}\mathbf{T}$, which is

$$1+\widehat{\delta}>1+\delta^*=1+\frac{\eta\gamma}{1+\frac{L}{L_{\max}}}. \tag{51}$$

Therefore, we can conclude that the largest real eigenvalue of $\mathbf{M}^{-1}\mathbf{T}$ is at least $1+\delta^*>1$. □

B.3. Proof of Lemma 2

Proof. Under Assumption 1, we have (descent lemma)

$$\begin{aligned}
f(\boldsymbol{\theta}^{(t+1)})\leq& f(\boldsymbol{\theta}^{(t)})+\nabla_{\mathbf{x}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t)})^{\mathsf{T}}(\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)})+\frac{L_{\mathbf{x}}}{2}\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\|^2\\
&+\nabla_{\mathbf{y}}f(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})^{\mathsf{T}}(\mathbf{y}^{(t+1)}-\mathbf{y}^{(t)})+\frac{L_{\mathbf{y}}}{2}\|\mathbf{y}^{(t+1)}-\mathbf{y}^{(t)}\|^2\\
\overset{(a)}{\leq}& f(\boldsymbol{\theta}^{(t)})-\eta\left(\|\nabla_{\mathbf{x}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\|^2+\|\nabla_{\mathbf{y}}f(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})\|^2\right)\\
&+\frac{\eta^2L_{\mathbf{x}}}{2}\|\nabla_{\mathbf{x}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\|^2+\frac{\eta^2L_{\mathbf{y}}}{2}\|\nabla_{\mathbf{y}}f(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})\|^2\\
\overset{(b)}{\leq}& f(\boldsymbol{\theta}^{(t)})-\frac{\eta}{2}\left(\|\nabla_{\mathbf{x}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\|^2+\|\nabla_{\mathbf{y}}f(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})\|^2\right)
\end{aligned} \tag{52}$$

where (a) is true because of the update rule of gradient descent in each block and Assumption 1, in (b) we used $\eta\leq 1/L_{\max}$. □

B.4. Proof of Lemma 7

Proof. Consider a generic sequence $\mathbf{u}^{(t)}$ and partition it into two blocks, i.e., $\mathbf{u}^{(t)}=[\mathbf{x}^{(t)}\ \mathbf{y}^{(t)}]^{\mathsf{T}}$. Without loss of generality, let $\mathbf{u}^{(1)}$ be the origin, i.e., $\mathbf{u}^{(1)}=0$. According to the A-GD update rules, we have

$$\mathbf{u}^{(t+1)}=\mathbf{u}^{(t)}-\eta\begin{bmatrix}\nabla_{\mathbf{x}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\\ \nabla_{\mathbf{y}}f(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})\end{bmatrix}. \tag{53}$$

From (52), we also know that

$$\begin{aligned}
f(\mathbf{u}^{(t+1)})\leq& f(\mathbf{u}^{(t)})-\frac{\eta}{2}\left(\|\nabla_{\mathbf{x}}f(\mathbf{x}^{(t)},\mathbf{y}^{(t)})\|^2+\|\nabla_{\mathbf{y}}f(\mathbf{x}^{(t+1)},\mathbf{y}^{(t)})\|^2\right),\quad\forall t\leq T.\\
\overset{(53)}{=}& f(\mathbf{u}^{(t)})-\frac{1}{2\eta}\|\mathbf{u}^{(t+1)}-\mathbf{u}^{(t)}\|^2
\end{aligned} \tag{54}$$

By applying telescoping sum of (54), we have

$$f(\mathbf{u}^{(t+1)})\leq f(\mathbf{u}^{(1)})-\frac{1}{2\eta}\sum_{\tau=1}^{t}\left(\|\mathbf{u}^{(\tau+1)}-\mathbf{u}^{(\tau)}\|^2\right),\quad\forall t\leq T. \tag{55}$$

According to the definition of T, we know that

$$f(\mathbf{u}^{(1)})-f(\mathbf{u}^{(t+1)})<2\mathcal{F}\quad\forall t\leq T. \tag{56}$$

Combining (56) and (55), we know that

$$\sum_{\tau=1}^{t}\|\mathbf{u}^{(\tau+1)}-\mathbf{u}^{(\tau)}\|^2<4\eta\mathcal{F}. \tag{57}$$

Next, we will get the upper bound of $\|\mathbf{u}^{(t+1)}-\mathbf{u}^{(1)}\|,\forall t\leq T$ as the following, first, by the triangle inequality, we know

$$\|\mathbf{u}^{(t+1)}-\mathbf{u}^{(1)}\|\leq\sum_{\tau=1}^{t}\|\mathbf{u}^{(\tau+1)}-\mathbf{u}^{(\tau)}\|, \tag{58}$$

so we have

$$\|\mathbf{u}^{(t+1)}-\mathbf{u}^{(1)}\|^2 \le t\sum_{\tau=1}^{t}\|\mathbf{u}^{(\tau+1)}-\mathbf{u}^{(\tau)}\|^2 \tag{59}$$

$$\le T\sum_{\tau=1}^{t}\|\mathbf{u}^{(\tau+1)}-\mathbf{u}^{(\tau)}\|^2 \tag{60}$$

$$\overset{(57)}{\le} T4\eta\mathcal{F} \overset{(33)}{\le} \widehat{c}\mathcal{T}4\eta\mathcal{F} \overset{(a)}{\le} 4\mathcal{S}^2, \tag{61}$$

where in (a) we use the relation $\widehat{c}\eta\mathcal{T}\mathcal{F}=\mathcal{S}^2$ by applying (30a)(30c)(30d).

Due to the following fact

$$\|\mathbf{u}^{(t+1)}-\widetilde{\boldsymbol{\theta}}^{(t)}\| = \|\mathbf{u}^{(t+1)}-\mathbf{u}^{(1)}+\mathbf{u}^{(1)}-\widetilde{\boldsymbol{\theta}}^{(t)}\| \le \underbrace{\|\mathbf{u}^{(t+1)}-\mathbf{u}^{(1)}\|}_{\le 2\mathcal{S}} + \underbrace{\|\mathbf{u}^{(1)}-\widetilde{\boldsymbol{\theta}}^{(t)}\|}_{\le \frac{\mathcal{S}}{\widehat{c}\log(\frac{d\kappa}{\delta})}} \le 3\mathcal{S} \tag{62}$$

where the last inequality is true when $\widehat{c}\ge 1$. Therefore, we know that $\|\mathbf{u}^{(t+1)}-\widetilde{\boldsymbol{\theta}}^{(t)}\|\le 3\mathcal{S}, \forall t\le T$ where $\eta\le 1/L_{\max}$, which completes the proof. □

B.5. Proof of Lemma 8

Proof. Let $\mathbf{u}^{(1)}=0$ and define $\mathbf{v}^{(t)}\triangleq\mathbf{w}^{(t)}-\mathbf{u}^{(t)}$ where $\mathbf{v}^{(t)}$, $\mathbf{w}^{(t)}$ and $\mathbf{u}^{(t)}$ are partitioned as $\mathbf{u}^{(t)}=[\mathbf{x}_u^{(t)}\ \mathbf{y}_u^{(t)}]^\intercal$, $\mathbf{v}=[\mathbf{x}_v^{(t)}\ \mathbf{y}_v^{(t)}]^\intercal$ and $\mathbf{w}=[\mathbf{x}_w^{(t)}\ \mathbf{y}_w^{(t)}]^\intercal$. According to the assumption of Lemma 8, we know that

$$\mathbf{v}^{(1)}=\upsilon\frac{\mathcal{S}}{\widehat{c}^3\kappa\log(\frac{d\kappa}{\delta})p_1}\vec{\mathbf{e}} \tag{63}$$

where $\upsilon\in[\delta/(2\sqrt{d}),1]$. First, defining an auxiliary function

$$h(\alpha)\triangleq\begin{bmatrix}\nabla_{\mathbf{x}}f(\mathbf{x}_u^{(t)}+\alpha\mathbf{x}_v^{(t)},\mathbf{y}_u^{(t)}+\alpha\mathbf{y}_v^{(t)})\\ \nabla_{\mathbf{y}}f(\mathbf{x}_u^{(t+1)}+\alpha\mathbf{x}_v^{(t+1)},\mathbf{y}_u^{(t)}+\alpha\mathbf{y}_v^{(t)})\end{bmatrix},$$

we can have

$$h(0)=\begin{bmatrix}\nabla_{\mathbf{x}}f(\mathbf{x}_u^{(t)},\mathbf{y}_u^{(t)})\\ \nabla_{\mathbf{y}}f(\mathbf{x}_u^{(t+1)},\mathbf{y}_u^{(t)})\end{bmatrix},\quad h(1)=\begin{bmatrix}\nabla_{\mathbf{x}}f(\mathbf{x}_u^{(t)}+\mathbf{y}_v^{(t)},\mathbf{x}_u^{(t)}+\mathbf{y}_v^{(t)})\\ \nabla_{\mathbf{y}}f(\mathbf{x}_u^{(t+1)}+\mathbf{y}_v^{(t+1)},\mathbf{x}_u^{(t)}+\mathbf{y}_v^{(t)})\end{bmatrix},$$

$$g(\alpha)=\frac{dh(\alpha)}{d\alpha}=\underbrace{\begin{bmatrix}\nabla^2_{\mathbf{xx}}f(\mathbf{x}_u^{(t)}+\alpha\mathbf{y}_v^{(t)},\mathbf{x}_u^{(t)}+\alpha\mathbf{y}_v^{(t)}) & \nabla^2_{\mathbf{xy}}f(\mathbf{x}_u^{(t)}+\alpha\mathbf{x}_v^{(t)},\mathbf{y}_u^{(t)}+\alpha\mathbf{y}_v^{(t)})\\ \mathbf{0} & \nabla^2_{\mathbf{yy}}f(\mathbf{x}_u^{(t+1)}+\alpha\mathbf{x}_u^{(t+1)},\mathbf{y}_u^{(t)}+\alpha\mathbf{y}_v^{(t)})\end{bmatrix}}_{\triangleq\widetilde{\boldsymbol{\mathcal{H}}}_u^{(t)}(\alpha)}\mathbf{v}^{(t)}$$

$$+\underbrace{\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \nabla^2_{\mathbf{yx}}f(\mathbf{x}_u^{(t+1)}+\alpha\mathbf{x}_v^{(t+1)},\mathbf{y}_u^{(t)}+\alpha\mathbf{y}_v^{(t)}) & \mathbf{0}\end{bmatrix}}_{\triangleq\widetilde{\boldsymbol{\mathcal{H}}}_l^{(t)}(\alpha)}\mathbf{v}^{(t+1)},$$

which also give

$$\begin{bmatrix}\nabla_{\mathbf{x}}f(\mathbf{x}_w^{(t)},\mathbf{y}_w^{(t)})\\ \nabla_{\mathbf{y}}f(\mathbf{x}_w^{(t+1)},\mathbf{y}_w^{(t)})\end{bmatrix}=\int_0^1 g(\alpha)d\alpha+\begin{bmatrix}\nabla_{\mathbf{x}}f(\mathbf{x}_u^{(t)},\mathbf{y}_u^{(t)})\\ \nabla_{\mathbf{y}}f(\mathbf{x}_u^{(t+1)},\mathbf{y}_u^{(t)})\end{bmatrix}. \tag{64}$$

Second, we consider sequence $\mathbf{w}^{(t)}$, i.e.,

$$\begin{aligned}
&\mathbf{u}^{(t+1)}+\mathbf{v}^{(t+1)}\\
=&\mathbf{w}^{(t+1)}=\mathbf{w}^{(t)}-\eta\left[\begin{array}{c}\nabla_{\mathbf{x}}f(\mathbf{x}_w^{(t)},\mathbf{y}_w^{(t)})\\ \nabla_{\mathbf{y}}f(\mathbf{x}_w^{(t+1)},\mathbf{y}_w^{(t)})\end{array}\right] &(65)\\
\overset{(64)}{=}&\mathbf{u}^{(t)}+\mathbf{v}^{(t)}-\eta\left[\begin{array}{c}\nabla_{\mathbf{x}}f(\mathbf{x}_u^{(t)},\mathbf{y}_u^{(t)})\\ \nabla_{\mathbf{y}}f(\mathbf{x}_u^{(t+1)},\mathbf{y}_u^{(t)})\end{array}\right]-\eta\int_0^1 g(\alpha)d\alpha &(66)\\
\overset{(a)}{=}&\mathbf{u}^{(t)}+\mathbf{v}^{(t)}-\eta\left[\begin{array}{c}\nabla_{\mathbf{x}}f(\mathbf{x}_u^{(t)},\mathbf{y}_u^{(t)})\\ \nabla_{\mathbf{y}}f(\mathbf{x}_u^{(t+1)},\mathbf{y}_u^{(t)})\end{array}\right]-\eta\widetilde{\Delta}_u^{(t)}\mathbf{v}^{(t)}-\eta\boldsymbol{\mathcal{H}}_u\mathbf{v}^{(t)}-\eta\widetilde{\Delta}_l^{(t)}\mathbf{v}^{(t+1)}-\eta\boldsymbol{\mathcal{H}}_l\mathbf{v}^{(t+1)} &(67)
\end{aligned}$$

where in (a) we use the following definitions:

$$\widetilde{\Delta}_u^{(t)}\triangleq\int_0^1\widetilde{\boldsymbol{\mathcal{H}}}_u^{(t)}(\alpha)d\alpha-\boldsymbol{\mathcal{H}}_u, \tag{68}$$

$$\widetilde{\Delta}_l^{(t)}\triangleq\int_0^1\widetilde{\boldsymbol{\mathcal{H}}}_l^{(t)}(\alpha)d\alpha-\boldsymbol{\mathcal{H}}_l, \tag{69}$$

and

$$\boldsymbol{\mathcal{H}}_u\triangleq\left[\begin{array}{cc}\nabla^2_{\mathbf{xx}}f(\widetilde{\boldsymbol{\theta}}^{(t)}) & \nabla^2_{\mathbf{xy}}f(\widetilde{\boldsymbol{\theta}}^{(t)})\\ \mathbf{0} & \nabla^2_{\mathbf{yy}}f(\widetilde{\boldsymbol{\theta}}^{(t)})\end{array}\right],\quad \boldsymbol{\mathcal{H}}_l\triangleq\left[\begin{array}{cc}\mathbf{0} & \mathbf{0}\\ \nabla^2_{\mathbf{yx}}f(\widetilde{\boldsymbol{\theta}}^{(t)}) & \mathbf{0}\end{array}\right]. \tag{70}$$

Obviously, $\boldsymbol{\mathcal{H}}=\boldsymbol{\mathcal{H}}_l+\boldsymbol{\mathcal{H}}_u$.

Third, since the first two terms at RHS of (67) combined with $\mathbf{u}^{(t)}$ at LHS of (67) are exactly the same as (53). It can be observed that equation (67) gives the dynamic of $\mathbf{v}^{(t)}$, i.e.,

$$\mathbf{v}^{(t+1)}=\mathbf{v}^{(t)}-\eta\widetilde{\Delta}_u^{(t)}\mathbf{v}^{(t)}-\eta\boldsymbol{\mathcal{H}}_u\mathbf{v}^{(t)}-\eta\widetilde{\Delta}_l^{(t)}\mathbf{v}^{(t+1)}-\eta\boldsymbol{\mathcal{H}}_l\mathbf{v}^{(t+1)}. \tag{71}$$

Then, we can rewrite (71) in a matrix form as the following.

$$\underbrace{(\mathbf{I}+\eta\boldsymbol{\mathcal{H}}_l)}_{\triangleq\mathbf{M}}\mathbf{v}^{(t+1)}+\eta\widetilde{\Delta}_l^{(t)}\mathbf{v}^{(t+1)}\overset{(67)}{=}\underbrace{(\mathbf{I}-\eta\boldsymbol{\mathcal{H}}_u)}_{\triangleq\mathbf{T}}\mathbf{v}^{(t)}-\eta\widetilde{\Delta}_u^{(t)}\mathbf{v}^{(t)}. \tag{72}$$

It is worth noting that matrix $\mathbf{M}$ is a lower triangular matrix where the diagonal entries are all 1s, so it is invertible.

Taking the inverse of $\mathbf{M}$ on both sides of (72), we can obtain

$$\mathbf{v}^{(t+1)}+\mathbf{M}^{-1}\eta\widetilde{\Delta}_l^{(t)}\mathbf{v}^{(t+1)}=\mathbf{M}^{-1}\mathbf{T}\mathbf{v}^{(t)}-\mathbf{M}^{-1}\eta\widetilde{\Delta}_u^{(t)}\mathbf{v}^{(t)}. \tag{73}$$

Projecting $\mathbf{v}^{(t)}$ to the Direction with the Negative Curvature:
Let $\mathbf{P}$ denote the projection operator that projects the vector onto the space spanned by the eigenvector of $\mathbf{M}^{-1}\mathbf{T}$ whose eigenvalue is maximum. Taking the projection on both sides of (73), we have

$$\mathbf{P}\mathbf{v}^{(t+1)}+\mathbf{P}\mathbf{M}^{-1}\eta\widetilde{\Delta}_l^{(t)}\mathbf{v}^{(t+1)}=\mathbf{P}(\mathbf{M}^{-1}\mathbf{T})\mathbf{v}^{(t)}-\mathbf{P}\mathbf{M}^{-1}\eta\widetilde{\Delta}_u^{(t)}\mathbf{v}^{(t)}. \tag{74}$$

From Lemma 1, we know that the maximum eigenvalue of $\mathbf{M}^{-1}\mathbf{T}$ is greater than 1. Next, we will leverage this property to show how A-GD can escape from the saddle point.

Relationship of the Norm of $\mathbf{v}^{(t)}$ Projected in the Two Subspaces:
Let $\phi^{(t)}$ denote the norm of $\mathbf{v}^{(t)}$ projected onto the space spanned by the eigenvector of $\mathbf{M}^{-1}\mathbf{T}$ whose maximum eigenvalue is $1+\widehat{\delta}$ where $\widehat{\delta}\geq\eta\gamma/(1+\frac{L}{L_{\max}})$ due to Lemma 1, and $\psi^{(t)}$ denote the norm of $\mathbf{v}^{(t)}$ projected onto the remaining space. From (74), we can have

$$\phi^{(t+1)}\geq(1+\widehat{\delta})\phi^{(t)}-\eta\|\mathbf{M}^{-1}\|\|\widetilde{\Delta}_l^{(t)}\|\|\mathbf{v}^{(t+1)}\|-\eta\|\mathbf{M}^{-1}\|\|\widetilde{\Delta}_u^{(t)}\|\|\mathbf{v}^{(t)}\|, \tag{75}$$

$$\psi^{(t+1)}\leq(1+\widehat{\delta})\psi^{(t)}+\eta\|\mathbf{M}^{-1}\|\|\widetilde{\Delta}_l^{(t)}\|\|\mathbf{v}^{(t+1)}\|+\eta\|\mathbf{M}^{-1}\|\|\widetilde{\Delta}_u^{(t)}\|\|\mathbf{v}^{(t)}\| \tag{76}$$

where we apply the triangle inequality. Also, $\mathbf{M}$ is a 2×2 block matrix, so we have $\mathbf{M}^{-1} = \mathbf{I} - \eta \boldsymbol{\mathcal{H}}_l$ and the size of $\mathbf{M}^{-1}$ can be bounded as follows.

$$\begin{aligned}
\|\mathbf{M}^{-1}\| \leq & 1 + \eta \|\boldsymbol{\mathcal{H}}_l\| \\
\overset{(a)}{=} & 1 + \|\eta \boldsymbol{\mathcal{H}} \odot \mathbf{D} - \eta \boldsymbol{\mathcal{H}}_d\| \\
\leq & 1 + \eta \|\boldsymbol{\mathcal{H}} \odot \mathbf{D}\| + \eta \|\boldsymbol{\mathcal{H}}_d\| \\
\overset{(b)}{\leq} & 1 + \eta \left(1 + \frac{1}{\pi} + \frac{\log(d)}{\pi}\right) \|\boldsymbol{\mathcal{H}}\| + \eta \|\boldsymbol{\mathcal{H}}_d\| \\
\overset{(c)}{\leq} & 1 + \eta \log(4d) \|\boldsymbol{\mathcal{H}}\| + \eta \|\boldsymbol{\mathcal{H}}_d\| \\
\overset{(d)}{\leq} & 1 + \eta L \log(4d) + \eta L_{\max} \\
\leq & 1 + \frac{L}{L_{\max}} \log(4d) + 1 < 2 \Bigg(\underbrace{1 + \frac{L \log(4d)}{2 L_{\max}}}_{p_2} \Bigg)
\end{aligned} \tag{77}$$

where in (a) operator $\odot$ denotes the Hadamard product and

$$\boldsymbol{\mathcal{H}}_d \triangleq \begin{bmatrix} \nabla^2_{\mathbf{xx}} f(\widetilde{\boldsymbol{\theta}}^{(t)}) & \mathbf{0} \\ \mathbf{0} & \nabla^2_{\mathbf{yy}} f(\widetilde{\boldsymbol{\theta}}^{(t)}) \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \cdots & 1 & 1 \end{bmatrix} \in \mathbb{R}^{d \times d}$$

and inequality (b) comes from the result on the spectral norm of the triangular truncation operator (please see [Theorem 1](Angelos et al., 1992)). In particular, by defining

$$\zeta(\mathbf{D}) \triangleq \max \left\{ \frac{\|\boldsymbol{\mathcal{H}} \odot \mathbf{D}\|}{\|\boldsymbol{\mathcal{H}}\|}, \boldsymbol{\mathcal{H}} \neq 0 \right\},$$

we have

$$\left| \frac{\zeta(\mathbf{D})}{\log(d)} - \frac{1}{\pi} \right| \leq \left(1 + \frac{1}{\pi}\right) \frac{1}{\log(d)}, \tag{78}$$

(c) is true because $1 + \frac{1}{\pi} + \frac{\log(d)}{\pi} \leq \log(4d)$ for $d \geq 1$, in (d) we use the fact that $\|\boldsymbol{\mathcal{H}}\| \leq L$ and $\|\boldsymbol{\mathcal{H}}_d\| \leq L_{\max}$.

Since $\|\mathbf{w}^{(1)} - \widetilde{\boldsymbol{\theta}}^{(t)}\| \leq \underbrace{\|\mathbf{u}^{(1)} - \widetilde{\boldsymbol{\theta}}^{(t)}\|}_{\leq r} + \underbrace{\|\mathbf{v}^{(1)}\|}_{\leq r} \leq 2r$, we can apply Lemma 7. Then, we know $\|\mathbf{w}^{(t+1)} - \widetilde{\boldsymbol{\theta}}^{(t)}\| \leq 3\mathcal{S}, \forall t \leq T$. According to the assumptions of Lemma 8, we have $\|\mathbf{u}^{(t+1)} - \widetilde{\boldsymbol{\theta}}^{(t)}\| \leq 3\mathcal{S}, \forall t \leq T$, and

$$\|\mathbf{v}^{(t+1)}\| = \|\mathbf{w}^{(t+1)} - \mathbf{u}^{(t+1)}\| \leq \|\mathbf{u}^{(t+1)} - \widetilde{\boldsymbol{\theta}}^{(t)}\| + \|\mathbf{w}^{(t+1)} - \widetilde{\boldsymbol{\theta}}^{(t)}\| \leq 6\mathcal{S}. \tag{79}$$

According to Lipsichiz continuity, we have the relation betweenn $\|\mathbf{v}^{(t)}\|$ and $\|\mathbf{v}^{(t+1)}\|$, and following upper bounds of $\|\widetilde{\Delta}_u^{(t)}\|$ and $\|\widetilde{\Delta}_l^{(t)}\|$.

1. Relation between $\|\mathbf{v}^{(t)}\|$ and $\|\mathbf{v}^{(t+1)}\|$:

According to the definition of $\mathbf{v}^{(t)}$, we know that

$$
\begin{aligned}
\|\mathbf{x}_v^{(t+1)}\|^2 + \|\mathbf{y}_v^{(t+1)}\|^2 =& \|\mathbf{v}^{(t+1)}\|^2 = \|\mathbf{w}^{(t+1)} - \mathbf{u}^{(t+1)}\|^2 \\
=& \left\| \mathbf{w}^{(t)} - \eta \begin{bmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}_w^{(t)}, \mathbf{y}_w^{(t)}) \\ \nabla_{\mathbf{y}} f(\mathbf{x}_w^{(t+1)}, \mathbf{y}_w^{(t)}) \end{bmatrix} - \left(\mathbf{u}^{(t)} - \eta \begin{bmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}_u^{(t)}, \mathbf{y}_u^{(t)}) \\ \nabla_{\mathbf{y}} f(\mathbf{x}_u^{(t+1)}, \mathbf{y}_u^{(t)}) \end{bmatrix} \right) \right\|^2 \\
\leq& 2\|\mathbf{v}^{(t)}\|^2 + 4\eta^2 \left\| \begin{bmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}_w^{(t)}, \mathbf{y}_w^{(t)}) \\ \nabla_{\mathbf{y}} f(\mathbf{x}_w^{(t+1)}, \mathbf{y}_w^{(t)}) \end{bmatrix} - \begin{bmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}_u^{(t)}, \mathbf{y}_w^{(t)}) \\ \nabla_{\mathbf{y}} f(\mathbf{x}_u^{(t+1)}, \mathbf{y}_w^{(t)}) \end{bmatrix} \right\|^2 \\
&+ 4\eta^2 \left\| \begin{bmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}_u^{(t)}, \mathbf{y}_w^{(t)}) \\ \nabla_{\mathbf{y}} f(\mathbf{x}_u^{(t+1)}, \mathbf{y}_w^{(t)}) \end{bmatrix} - \begin{bmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}_u^{(t)}, \mathbf{y}_u^{(t)}) \\ \nabla_{\mathbf{y}} f(\mathbf{x}_u^{(t+1)}, \mathbf{y}_u^{(t)}) \end{bmatrix} \right\|^2 \\
\overset{(a)}{\leq}& 2\|\mathbf{v}^{(t)}\|^2 + 4\eta^2 (L^2 \|\mathbf{x}_v^{(t+1)}\|^2 + L_{\max}^2 \|\mathbf{x}_v^{(t)}\|^2) + 4\eta^2 (L^2 + L_{\max}^2) \|\mathbf{y}_v^{(t)}\|^2 \qquad (80) \\
\overset{(b)}{\leq}& 6\|\mathbf{v}^{(t)}\|^2 + 4\left(\frac{L}{L_{\max}}\right)^2 \left(\|\mathbf{x}_v^{(t+1)}\|^2 + \|\mathbf{y}_v^{(t)}\|^2 \right) \qquad (81)
\end{aligned}
$$

where (a) is true due to gradient Lipschitz continuity, and in (b) we use $\eta \leq 1/L_{\max}$.

Also, we need the relation between $\|\mathbf{x}_v^{(t+1)}\|$ and $\|\mathbf{x}_v^{(t+1)}\|$, which is

$$
\begin{aligned}
\|\mathbf{x}_v^{(t+1)}\|^2 =& \|\mathbf{x}_w^{(t+1)} - \mathbf{x}_u^{(t+1)}\|^2 \leq 2\|\mathbf{x}_w^{(t)} - \mathbf{x}_u^{(t)}\|^2 + 2\|\nabla_{\mathbf{x}} f(\mathbf{x}_w^{(t)}, \mathbf{y}_w^{(t)}) - \nabla_{\mathbf{x}} f(\mathbf{x}_u^{(t)}, \mathbf{y}_u^{(t)})\|^2 \\
\leq& 2\|\mathbf{x}_v^{(t)}\|^2 + 4\left(\|\nabla_{\mathbf{x}} f(\mathbf{x}_w^{(t)}, \mathbf{y}_w^{(t)}) - \nabla_{\mathbf{x}} f(\mathbf{x}_w^{(t)}, \mathbf{y}_u^{(t)})\|^2 + \|\nabla_{\mathbf{x}} f(\mathbf{x}_w^{(t)}, \mathbf{y}_u^{(t)}) - \nabla_{\mathbf{x}} f(\mathbf{x}_u^{(t)}, \mathbf{y}_u^{(t)})\|^2 \right) \\
\leq& 6\|\mathbf{x}_v^{(t)}\|^2 + 4\left(\frac{L}{L_{\max}}\right)^2 \|\mathbf{y}_v^{(t)}\|^2. \qquad (82)
\end{aligned}
$$

Combing (81) and (82), we have

$$
\begin{aligned}
\|\mathbf{v}^{(t+1)}\|^2 \leq& 6\|\mathbf{v}^{(t)}\|^2 + 4\left(\frac{L}{L_{\max}}\right)^2 \left(6\|\mathbf{x}_v^{(t)}\|^2 + \left(4\left(\frac{L}{L_{\max}}\right)^2 + 1 \right) \|\mathbf{y}_v^{(t)}\|^2 \right) \\
&\leq 6\|\mathbf{v}^{(t)}\|^2 + 4\left(\frac{L}{L_{\max}}\right)^2 p_0 \|\mathbf{v}^{(t)}\|^2 \leq p_0^2 \|\mathbf{v}^{(t)}\|^2 \qquad (83)
\end{aligned}
$$

where $p_0 \triangleq \max\left\{6, 4\left(\frac{L}{L_{\max}}\right)^2 + 1\right\}$.

2. Upper bounds of $\|\widetilde{\Delta}_u^{(t)}\|$ and $\|\widetilde{\Delta}_l^{(t)}\|$:

According to ρ-Hessian Lipschitz continuity and Lemma 5, we have the size of $\widetilde{\Delta}_u^{(t)}$ as the following.

$$
\begin{aligned}
\|(\widetilde{\Delta}_u^{(t)})\| \leq& \int_0^1 \|\widetilde{\boldsymbol{\mathcal{H}}}_u^{(t)}(\alpha) - \boldsymbol{\mathcal{H}}_u\| d\alpha \\
\overset{(21)}{\leq}& \int_0^1 \rho \left(\|\mathbf{u}^{(t)} + \alpha\mathbf{v}^{(t)} - \widetilde{\boldsymbol{\theta}}^{(t)}\| + \left\| \begin{bmatrix} \mathbf{x}_u^{(t+1)} + \alpha\mathbf{x}_v^{(t+1)} \\ \mathbf{y}_u^{(t)} + \alpha\mathbf{y}_v^{(t)} \end{bmatrix} - \widetilde{\boldsymbol{\theta}}^{(t)} \right\| \right) d\alpha \qquad (84) \\
\overset{(a)}{\leq}& \int_0^1 \rho \left(2\|\mathbf{u}^{(t)} + \alpha\mathbf{v}^{(t)} - \widetilde{\boldsymbol{\theta}}^{(t)}\| + \|\mathbf{u}^{(t+1)} + \alpha\mathbf{v}^{(t+1)} - \widetilde{\boldsymbol{\theta}}^{(t)}\| \right) d\alpha \\
\leq& \rho(\|\mathbf{u}^{(t+1)} - \widetilde{\boldsymbol{\theta}}^{(t)}\| + 2\|\mathbf{u}^{(t)} - \widetilde{\boldsymbol{\theta}}^{(t)}\|) + \rho \int_0^1 \alpha(\|\mathbf{v}^{(t+1)}\| + \|\mathbf{v}^{(t)}\|) d\alpha \\
\leq& \rho \left(\|\mathbf{u}^{(t+1)} - \mathbf{u}^{(t)}\| + \|\mathbf{u}^{(t)} - \widetilde{\boldsymbol{\theta}}^{(t)}\| + 2\|\mathbf{u}^{(t)} - \widetilde{\boldsymbol{\theta}}^{(t)}\|) + 0.5\|\mathbf{v}^{(t+1)}\| + 0.5\|\mathbf{v}^{(t)}\| \right) \\
\overset{(83)}{\leq}& \rho \left(3\mathcal{S} + 6\mathcal{S} + 0.5(p_0 + 1)\|\mathbf{v}^{(t)}\| \right) \\
\leq& (3p_0 + 12)\rho\mathcal{S} \qquad (85)
\end{aligned}
$$

where (a) is true because

$$\left\|\begin{bmatrix} \mathbf{x}_u^{(t+1)}+\alpha\mathbf{x}_v^{(t+1)} \\ \mathbf{y}_u^{(t)}+\alpha\mathbf{y}_v^{(t)}\end{bmatrix} - \widetilde{\boldsymbol{\theta}}^{(t)}\right\| \le \left\|\mathbf{I}_1\left(\mathbf{u}^{(t+1)}+\alpha\mathbf{v}^{(t+1)}-\widetilde{\boldsymbol{\theta}}^{(t)}\right)\right\| + \left\|\mathbf{I}_2\left(\mathbf{u}^{(t)}+\alpha\mathbf{v}^{(t)}-\widetilde{\boldsymbol{\theta}}^{(t)}\right)\right\|$$
$$\overset{(23)}{\le}\|\mathbf{u}^{(t+1)}+\alpha\mathbf{v}^{(t+1)}-\widetilde{\boldsymbol{\theta}}^{(t)}\| + \|\mathbf{u}^{(t)}+\alpha\mathbf{v}^{(t)}-\widetilde{\boldsymbol{\theta}}^{(t)}\|. \tag{86}$$

Applying Lemma 5, we can also get the upper bound of $\|\widetilde{\Delta}_l^{(t)}\|$, i.e.,

$$\begin{aligned}
\|\widetilde{\Delta}_l^{(t)}\| \le & \int_0^1 \|\widetilde{\boldsymbol{\mathcal{H}}}_l^{(t)}(\alpha) - \boldsymbol{\mathcal{H}}_l\| d\alpha \\
\overset{(22)}{\le} & \int_0^1 \rho \left\|\begin{bmatrix} \mathbf{x}_u^{(t+1)}+\alpha\mathbf{x}_v^{(t+1)} \\ \mathbf{y}_u^{(t)}+\alpha\mathbf{y}_v^{(t)}\end{bmatrix} - \widetilde{\boldsymbol{\theta}}^{(t)}\right\| d\alpha & (87) \\
\le & \int_0^1 \rho(\|\mathbf{u}^{(t)}+\alpha\mathbf{v}^{(t)}-\widetilde{\boldsymbol{\theta}}^{(t)}\| + \|\mathbf{u}^{(t+1)}+\alpha\mathbf{v}^{(t+1)}-\widetilde{\boldsymbol{\theta}}^{(t)})\| d\alpha \\
\le & \rho(\|\mathbf{u}^{(t+1)}-\widetilde{\boldsymbol{\theta}}^{(t)}\| + \|\mathbf{u}^{(t)}-\widetilde{\boldsymbol{\theta}}^{(t)}\|) + \rho\int_0^1 \alpha(\|\mathbf{v}^{(t+1)}\| + \|\mathbf{v}^{(t)}\|)d\alpha \\
\overset{(79)}{\le} & \rho\left(3\mathcal{S}+3\mathcal{S}+0.5(p_0+1)\|\mathbf{v}^{(t)}\|\right) \\
\le & (3p_0+9)\rho\mathcal{S}. & (88)
\end{aligned}$$

With the upper bounds of $\|\widetilde{\Delta}_u^{(t)}\|$, $\|\widetilde{\Delta}_l^{(t)}\|$ and relation between $\|\mathbf{v}^{(t+1)}\|$ and $\|\mathbf{v}^{(t)}\|$ as shown in (83), we can further simply (75) and (76) as follows,

$$\phi^{(t+1)} \overset{(75)}{\ge} (1+\widehat{\delta})\phi^{(t)} - \eta(p_0\|\widetilde{\Delta}_l^{(t)}\| + \|\widetilde{\Delta}_u^{(t)}\|)\|\mathbf{M}^{-1}\|\|\mathbf{v}^{(t)}\|,$$
$$\psi^{(t+1)} \overset{(76)}{\le} (1+\widehat{\delta})\psi^{(t)} + \eta(p_0\|\widetilde{\Delta}_l^{(t)}\| + \|\widetilde{\Delta}_u^{(t)}\|)\|\mathbf{M}^{-1}\|\|\mathbf{v}^{(t)}\|,$$

and further we have

$$\phi^{(t+1)} \ge (1+\widehat{\delta})\phi^{(t)} - \eta(p_0\|\widetilde{\Delta}_l^{(t)}\| + \|\widetilde{\Delta}_u^{(t)}\|)\|\mathbf{M}^{-1}\|\sqrt{(\phi^{(t)})^2+(\theta^{(t)})^2},$$
$$\psi^{(t+1)} \le (1+\widehat{\delta})\psi^{(t)} + \eta(p_0\|\widetilde{\Delta}_l^{(t)}\| + \|\widetilde{\Delta}_u^{(t)}\|)\|\mathbf{M}^{-1}\|\sqrt{(\phi^{(t)})^2+(\theta^{(t)})^2},$$

since $\|\mathbf{v}^{(t)}\| = \sqrt{(\phi^{(t)})^2+(\psi^{(t)})^2}$.

Consequently, we can arrive at

$$\phi^{(t+1)} \ge (1+\widehat{\delta})\phi^{(t)} - \mu\sqrt{(\phi^{(t)})^2+(\psi^{(t)})^2}, \tag{89}$$
$$\psi^{(t+1)} \le (1+\widehat{\delta})\psi^{(t)} + \mu\sqrt{(\phi^{(t)})^2+(\psi^{(t)})^2}, \tag{90}$$

where μ is the upper bound of $\eta(p_0\|\widetilde{\Delta}_l^{(t)}\| + \|\widetilde{\Delta}_u^{(t)}\|)\|\mathbf{M}^{-1}\|$ and can be obtained as

$$\mu \triangleq 2(3p_0^2+12p_0+12)\eta\rho p_2\mathcal{S} \tag{91}$$

by applying (77)(85)(88).

Quantifying the Norm of $\mathbf{v}^{(t)}$ Projected at Different Subspaces:
Then, we will use mathematical induction to prove

$$\psi^{(t)} \le 4\mu t\phi^{(t)}. \tag{92}$$

It is true when $t = 1$ since $\|\psi^{(1)}\| \overset{(35)}{=} 0$.

Assuming that equation (92) is true at the tth iteration, we need to prove

$$\psi^{(t+1)} \le 4\mu(t+1)\phi^{(t+1)}. \tag{93}$$

We use (89) and (90) separately into RHS and LHS of (93) to get the lower and upper bound of $4\mu(t+1)\phi^{(t+1)}$ and $\psi^{(t+1)}$. Applying (89) into RHS of (93), we have the lower bound of $4\mu(t+1)\phi^{(t+1)}$ as the following:

$$4\mu(t+1)\phi^{(t+1)} \ge 4\mu(t+1)\left((1+\widehat{\delta})\phi^{(t)} - \mu\sqrt{(\phi^{(t)})^2 + (\psi^{(t)})^2}\right) \tag{94}$$

and substituting (90) into LHS of (93), we have the upper bound of $\psi^{(t+1)}$, i.e.,

$$\psi^{(t+1)} \le (1+\widehat{\delta})(4\mu t\phi^{(t)}) + \mu\sqrt{(\phi^{(t)})^2 + (\psi^{(t)})^2}. \tag{95}$$

If RHS of (94) is greater than RHS of (95), it is equivalent to prove (93). After some manipulations, it is sufficient to show

$$(1+4\mu(t+1))\left(\sqrt{(\phi^{(t)})^2 + (\psi^{(t)})^2}\right) \le 4\phi^{(t)}. \tag{96}$$

In the following, we will show that the above relation (96) is true, i.e., RHS of (94) is greater than RHS of (95).

First step: We know that

$$4\mu(t+1) \le 4\mu T \overset{(91)}{\le} 8(3p_0^2 + 12p_0 + 12)\eta\rho\mathcal{S}p_2\widehat{c}\mathcal{T} \overset{(a)}{\le} \frac{8(3p_0^2 + 12p_0 + 12)}{\widehat{c}} \overset{(b)}{\le} 1 \tag{97}$$

where in (a) we use the relation $\eta\rho\mathcal{S}p_2\widehat{c}\mathcal{T} = \frac{1}{\widehat{c}}$ by applying (30c)(30d); (b) is true when $\widehat{c} \ge 8(3p_0^2 + 12p_0 + 12)$.

Second step: Also, we know that

$$4\phi^{(t)} \ge 2\sqrt{2(\phi^{(t)})^2} \overset{(92),(97)}{\ge} (1+4\mu(t+1))\sqrt{(\phi^{(t)})^2 + (\psi^{(t)})^2}, \tag{98}$$

which gives (96). Therefore, we can conclude that $\psi^{(t+1)} \le 4\mu(t+1)\phi^{(t+1)}$ is true, which completes the induction.

Recursion of $\phi^{(t)}$:

Using (92), we have $\psi^{(t)} \overset{(92)}{\le} 4\mu t\phi^{(t)} \overset{(97)}{\le} \phi^{(t)}$, which implies

$$\begin{aligned}
\phi^{(t+1)} \overset{(89)}{\ge} & (1+\widehat{\delta})\phi^{(t)} - \mu\sqrt{(\phi^{(t)})^2 + (\psi^{(t)})^2} \\
\overset{(a)}{\ge} & (1 + \frac{\gamma\eta}{1+L/L_{\max}})\phi^{(t)} - \mu\sqrt{(\phi^{(t)})^2 + (\psi^{(t)})^2} \\
\overset{(b)}{\ge} & (1 + \frac{1}{1+L/L_{\max}}\frac{\gamma\eta}{2})\phi^{(t)}
\end{aligned} \tag{99}$$

where in (a) we used Lemma 1, and (b) is true because $\psi^{(t)} \le \phi^{(t)}$ and

$$\begin{aligned}
\mu = & \eta\rho\mathcal{S}p_2 8(3p_0^2 + 12p_0 + 12) \\
\overset{(30c)}{=} & \frac{\gamma\eta}{1+L/L_{\max}}\frac{8(3p_0^2 + 12p_0 + 12)}{\widehat{c}^2\log(\frac{d\kappa}{\delta})} \\
\overset{(a)}{\le} & \frac{1}{1+L/L_{\max}}\frac{\gamma\eta}{2\sqrt{2}}
\end{aligned}$$

where (a) is true when $\widehat{c} \ge 4\sqrt{\sqrt{2}(3p_0^2 + 12p_0 + 12)}$.

Quantifying Escaping Time:
From (79), we have

$$
\begin{aligned}
6\mathcal{S} \geq & \|\mathbf{v}^{(t)}\| \geq \phi^{(t)} \\
\overset{(99)}{\geq} & (1+\frac{\gamma\eta}{2(1+L/L_{\max})})^t \phi^{(1)} \\
\overset{(a)}{\geq} & (1+\frac{\gamma\eta}{2(1+L/L_{\max})})^t \frac{\delta}{2\sqrt{d}} \frac{\mathcal{S}}{\widehat{c}^3\kappa} \log^{-1}(\frac{d\kappa}{\delta}) p_1^{-1} \quad \forall t \leq T
\end{aligned} \tag{100}
$$

where in (a) we use condition $\upsilon \in [\delta/(2\sqrt{d}), 1]$.

Since (100) is true for all $t \leq T$, we can have

$$
\begin{aligned}
T \leq & \frac{\log(12\widehat{c}^3(\frac{\kappa\sqrt{d}}{\delta})\log(\frac{d\kappa}{\delta})p_1)}{\log(1+\frac{\eta\gamma}{2(1+L/L_{\max})})} \\
\overset{(a)}{<} & \frac{4(1+L/L_{\max})\log(12\widehat{c}^3(\frac{\sqrt{d}\kappa}{\delta})\log(\frac{d\kappa}{\delta})p_1)}{\eta\gamma} \\
\overset{(b)}{<} & \frac{4(1+L/L_{\max})\log(12\widehat{c}^3(\frac{d\kappa}{\delta})^2 p_1)}{\eta\gamma} \\
\overset{(c)}{<} & \frac{4(\log(12\widehat{c}^4(\frac{d\kappa}{\delta})^2))p_1}{\eta\gamma} \\
\overset{(d),(30d)}{\leq} & 4(2+\log(12\widehat{c}^4))\mathcal{T}
\end{aligned} \tag{101}
$$

where (a) comes from inequality $\log(1+x) > x/2$ when $x < 1$, in (b) we used relation $\log(x) < x, x > 0$, (c) is true because $\widehat{c} > p_0 > p_1$, and (d) is true because $\delta \in (0, \frac{d\kappa}{e}]$ and $\log(d\kappa/\delta) > 1$ so that $\log(12\widehat{c}^4) + 2\log(\frac{d\kappa}{\delta}) \leq (\log(12\widehat{c}^4)+2)\log(\frac{d\kappa}{\delta})$.

From (101), we know that when

$$
4(2+\log(12\widehat{c}^4)) < \widehat{c}, \tag{102}
$$

we will have $T < \widehat{c}\mathcal{T}$.

It can be observed that LHS of (102) is a logarithmic with respect to $\widehat{c}$ and RHS of (102) is a linear function in terms of $\widehat{c}$, implying that when $\widehat{c}$ is large enough inequality (102) holds. It is can be numerically checked that when $\widehat{c} \geq 136$ inequality (102) holds. The proof is complete. □

B.6. Proof of Lemma 9

Proof. The proof of Lemma 9 is similar as the one of proving convergence of PGD shown in (Jin et al., 2017, Lemma 14,15). Considering the completeness of the whole proof in this paper, here we give the following proof of this lemma in details.

Step 1: Let $\mathbf{u}^{(1)}$ be a vector that follows uniform distribution within the ball $\mathbb{B}^{(d)}_{\widetilde{\boldsymbol{\theta}}^{(t)}}(r)$, where $\mathbb{B}^{(d)}_{\widetilde{\boldsymbol{\theta}}^{(t)}}$ denotes the d-dimensional ball centered at $\widetilde{\boldsymbol{\theta}}^{(t)} = [\widetilde{\mathbf{x}}^{(t)} \ \widetilde{\mathbf{y}}^{(t)}]^{\mathsf{T}}$ with radius r. Let $\xi \triangleq [\xi_{\mathbf{x}} \ \xi_{\mathbf{y}}]^{\mathsf{T}}$ represent the difference between the random generated vector $\mathbf{u}^{(1)}$ and saddle point $\widetilde{\boldsymbol{\theta}}^{(t)}$. We can show that the objective function value in the worst case is increased at most by

$$
\begin{aligned}
& f(\mathbf{u}^{(1)}) - f(\widetilde{\boldsymbol{\theta}}^{(t)}) \\
= & f(\mathbf{u}^{(1)}) - f(\widetilde{\mathbf{x}}^{(t+1)}, \mathbf{y}^{(t)}) + f(\widetilde{\mathbf{x}}^{(t+1)}, \mathbf{y}^{(t)}) - f(\widetilde{\boldsymbol{\theta}}^{(t)}) && (103) \\
\le & \nabla_{\mathbf{x}} f(\widetilde{\mathbf{x}}^{(t)}, \widetilde{\mathbf{y}}^{(t)})^{\mathsf{T}} \xi_{\mathbf{x}} + \frac{L_{\mathbf{x}}}{2}\|\xi_{\mathbf{x}}\|^2 + \nabla_{\mathbf{y}} f(\widetilde{\mathbf{x}}^{(t+1)}, \widetilde{\mathbf{y}}^{(t)})^{\mathsf{T}} \xi_{\mathbf{y}} + \frac{L_{\mathbf{y}}}{2}\|\xi_{\mathbf{y}}\|^2 && (104) \\
\le & \|\nabla_{\mathbf{x}} f(\widetilde{\mathbf{x}}^{(t)}, \widetilde{\mathbf{y}}^{(t)})\|\|\xi_{\mathbf{x}}\| + \|\nabla_{\mathbf{y}} f(\widetilde{\mathbf{x}}^{(t+1)}, \widetilde{\mathbf{y}}^{(t)})\|\|\xi_{\mathbf{y}}\| + \frac{L_{\max}}{2}\|\xi\|^2 \\
\overset{(a)}{\le} & \|\xi\|\sqrt{\|\nabla_{\mathbf{x}} f(\widetilde{\mathbf{x}}^{(t)}, \widetilde{\mathbf{y}}^{(t)})\|^2 + \|\nabla_{\mathbf{y}} f(\widetilde{\mathbf{x}}^{(t+1)}, \widetilde{\mathbf{y}}^{(t)})\|^2} + \frac{L_{\max}}{2}\|\xi\|^2 \\
\overset{(b)}{\le} & \underbrace{\mathcal{G}\frac{\mathcal{S}}{\widehat{c}^3\kappa\log(\frac{d\kappa}{\delta})p_1}}_{\le\frac{1}{2}\mathcal{F}} + \underbrace{\frac{L_{\max}}{2}\left(\frac{\mathcal{S}}{\widehat{c}^3\kappa\log(\frac{d\kappa}{\delta})p_1}\right)^2}_{\le\frac{1}{2}\mathcal{F}} \overset{(c)}{\le} \mathcal{F} && (105)
\end{aligned}
$$

where (a) is true because $\max\{\|\xi_{\mathbf{x}}\|, \|\xi_{\mathbf{y}}\|\} \le \|\xi\|$, and in (b) we used $\kappa > 1$, $\log(d\kappa/\delta) > 1$ and Condition 1, and in (c) we just use the definitions of $\mathcal{S}, \mathcal{G}, \mathcal{F}$, i.e., $\mathcal{G}\mathcal{S}/(\widehat{c}^3 p_1) < \frac{1}{2}\mathcal{F}$ and $L_{\max}\mathcal{S}^2/(\widehat{c}\kappa\log(d\kappa/\delta)p_1) \le \mathcal{F}$.

Step 2: we will quantify the decrease of the objective value after T number of iterations. Under Assumption 1, let $\widetilde{\boldsymbol{\theta}}^{(t)}$ satisfy conditions Condition 1, and two PA-GD iterates $\{\mathbf{u}^{(t)}\}$ $\{\mathbf{w}^{(t)}\}$ satisfy the conditions as in Lemma 8.

First, let us discuss the following two cases.

1. If $f(\mathbf{u}^{(2)}) - f(\mathbf{u}^{(1)}) \le -2\mathcal{F}$: the objective value is decreased by at least $2\mathcal{F}$, meaning that we can have a sufficient decrease of the objective value. Then, we can proceed to step 3.
2. If $f(\mathbf{u}^{(2)}) - f(\mathbf{u}^{(1)}) > -2\mathcal{F}$: there is no sufficient decrease, implying that we need to more number of iterations so that the objective can be decreased by at least $2\mathcal{F}$. Applying Lemma 7 and Lemma 8, we have the following analysis.

Second, let $T^* \triangleq \widehat{c}\mathcal{T}$ and $T' \triangleq \inf_{t\ge1}\left\{t|f(\mathbf{u}^{(t+2)}) - f(\mathbf{u}^{(1)}) \le -2\mathcal{F}\right\}$. Then, we have the following two cases to analyze the decrease of the objective value.

1. Case $T' \le T^*$: Applying Lemma 7, we know that

$$
f(\mathbf{u}^{(T'+2)}) - f(\mathbf{u}^{(1)}) \le -2\mathcal{F}. \tag{106}
$$

   Based on Lemma 2, we know that A-GD is always decreasing the objective function. For any $T > \widehat{c}\mathcal{T} + 2 = T^* + 2 \ge T' + 2$, we have

$$
f(\mathbf{u}^{(T)}) - f(\mathbf{u}^{(1)}) \le f(\mathbf{u}^{(T^*+2)}) - f(\mathbf{u}^{(1)}) \le f(\mathbf{u}^{(T'+2)}) - f(\mathbf{u}^{(1)}) \le -2\mathcal{F} \tag{107}
$$

   when $\widehat{c} \ge 1$.

2. Case $T' > T^*$: Applying Lemma 7, we know that $\|\mathbf{u}^{(t+1)} - \mathbf{u}^{(1)}\| \le 3\mathcal{S}$ for $t \le T^*$. Define $T'' = \inf_{t\ge1}\left\{t|f(\mathbf{w}^{(t+2)}) - f(\mathbf{w}^{(1)}) \le -2\mathcal{F}\right\}$. Then, after applying Lemma 8, we know $T'' < T^*$. Using the same argument as the above case, for $T \ge \widehat{c}\mathcal{T} + 2 \ge T^* + 2 \ge T'' + 2$, we also have

$$
f(\mathbf{w}^{(T)}) - f(\mathbf{w}^{(1)}) \le f(\mathbf{w}^{(T^*+2)}) - f(\mathbf{w}^{(1)}) \le f(\mathbf{w}^{(T''+2)}) - f(\mathbf{w}^{(1)}) \le -2\mathcal{F}. \tag{108}
$$

Step 3: combining (107) and (108), we have

$$
\min\left\{f(\mathbf{u}^{(T)}) - f(\mathbf{u}^{(1)}), f(\mathbf{w}^{(T)}) - f(\mathbf{w}^{(1)})\right\} \le -2\mathcal{F}, \forall T \ge \widehat{c}\mathcal{T} + 3, \tag{109}
$$

meaning that at least one of the sequences can give a sufficient decrease of the objective function if the initial points of the two sequences are separated apart with each other far enough along direction $\vec{\mathbf{e}}$.

Therefore, we can conclude that if $\mathbf{u}^{(1)} \in \mathcal{X}_{\mathsf{stuck}}$, then $(\mathbf{u}^{(1)} \pm \upsilon r\vec{\mathbf{e}}) \notin \mathcal{X}_{\mathsf{stuck}}$ where $\upsilon \in [\frac{\delta}{2\sqrt{d}}, 1]$.

Step 4: finally, we give the upper bound of the volume of $\mathcal{X}_{\mathsf{stuck}}$,

$$\begin{aligned}\mathsf{Vol}(\mathcal{X}_{\mathsf{stuck}}) &= \int_{\mathbb{B}^{(d)}_{\widetilde{\boldsymbol{\theta}}^{(t)}}} d\mathbf{u} I_{\mathcal{X}_{\mathsf{stuck}}}(\mathbf{u}) = \int_{\mathbb{B}^{(d-1)}_{\widetilde{\boldsymbol{\theta}}^{(t)}}} d\mathbf{u}_{-1} \int_{\widetilde{\boldsymbol{\theta}}^{(t)}_1 - \sqrt{r^2 - \|\widetilde{\boldsymbol{\theta}}^{(t)}_{-1} - \mathbf{u}_{-1}\|^2}}^{\widetilde{\boldsymbol{\theta}}^{(t)}_1 + \sqrt{r^2 - \|\widetilde{\boldsymbol{\theta}}^{(t)}_{-1} - \mathbf{u}_{-1}\|^2}} d\mathbf{u}_1 I_{\mathcal{X}_{\mathsf{stuck}}}(\mathbf{u}) \\ &\le \int_{\mathbb{B}^{(d-1)}_{\widetilde{\boldsymbol{\theta}}^{(t)}}} d\mathbf{u}_{-1} \left(2\frac{\delta}{2\sqrt{d}r}\right) = \mathsf{Vol}\left(\mathbb{B}^{(d-1)}_{\widetilde{\boldsymbol{\theta}}^{(t)}}(r)\right)\frac{r\delta}{\sqrt{d}}\end{aligned}$$

where $I_{\mathcal{X}_{\mathsf{stuck}}}(\mathbf{u})$ is an indicator function showing that $\mathbf{u}$ belongs to set $\mathcal{X}_{\mathsf{stuck}}$, and $\mathbf{u}_1$ represents the component of vector $\mathbf{u}$ along $\vec{\mathbf{e}}$ direction, and $\mathbf{u}_{-1}$ is the remaining $d-1$ dimensional vector.

Then, the ratio of $\mathsf{Vol}(\mathcal{X}_{\mathsf{stuck}})$ over the whole volume of the perturbation ball can be upper bounded by

$$\frac{\mathsf{Vol}(\mathcal{X}_{\mathsf{stuck}})}{\mathsf{Vol}(\mathbb{B}^{(d)}_{\widetilde{\boldsymbol{\theta}}^{(t)}}(r))} \le \frac{\frac{r\delta}{\sqrt{d}}\mathsf{Vol}(\mathbb{B}^{(d-1)}_{\widetilde{\boldsymbol{\theta}}^{(t)}}(r))}{\mathsf{Vol}(\mathbb{B}^{(d)}_{\widetilde{\boldsymbol{\theta}}^{(t)}}(r))} = \frac{\delta}{\sqrt{d\pi}}\frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{d}{2}+1)} \le \frac{\delta}{\sqrt{d\pi}}\sqrt{\frac{d}{2}+\frac{1}{2}} \le \delta$$

where $\Gamma(\cdot)$ denotes the Gamma function, and inequality is true due to the fact that $\Gamma(x+1)/\Gamma(x+1/2) < \sqrt{x+1/2}$ when $x \ge 0$.

Step 5: combining (105) and (109), we can show that

$$f(\boldsymbol{\theta}^{(T)}) - f(\widetilde{\boldsymbol{\theta}}^{(t)}) = \underbrace{f(\boldsymbol{\theta}^{(T)}) - f(\mathbf{u}^{(1)})}_{\le -2\mathcal{F}} + \underbrace{f(\mathbf{u}^{(1)}) - f(\widetilde{\boldsymbol{\theta}}^{(t)})}_{\le \mathcal{F}} \le -\mathcal{F} \tag{110}$$

with at least probability $1-\delta$. □