2020

On the Sublinear Convergence of Randomly Perturbed Alternating Gradient Descent to Second Order Stationary Solutions

Songtao Lu
University of Minnesota - Twin Cities

Mingyi Hong
University of Minnesota Law School

Zhengdao Wang
Iowa State University, zhengdao@iastate.edu

On the Sublinear Convergence of Randomly Perturbed Alternating Gradient Descent to Second Order Stationary Solutions

Abstract

The alternating gradient descent (AGD) is a simple but popular algorithm which has been applied to problems in optimization, machine learning, data ming, and signal processing, etc. The algorithm updates two blocks of variables in an alternating manner, in which a gradient step is taken on one block, while keeping the remaining block fixed. When the objective function is nonconvex, it is well-known the AGD converges to the first-order stationary solution with a global sublinear rate.

In this paper, we show that a variant of AGD-type algorithms will not be trapped by "bad" stationary solutions such as saddle points and local maximum points. In particular, we consider a smooth unconstrained optimization problem, and propose a perturbed AGD (PA-GD) which converges (with high probability) to the set of second-order stationary solutions (SS2) with a global sublinear rate. To the best of our knowledge, this is the first alternating type algorithm which takes O(polylog(d)/ϵ7/3) iterations to achieve SS2 with high probability [where polylog(d) is polynomial of the logarithm of dimension d of the problem].

Disciplines

Electrical and Computer Engineering | Mathematics

Comments

This is a pre-print of the article Lu, Songtao, Mingyi Hong, and Zhengdao Wang. "On the sublinear convergence of randomly perturbed alternating gradient descent to second order stationary solutions." *arXiv preprint arXiv:1802.10418* (2018). Posted with permission.

On the Sublinear Convergence of Randomly Perturbed Alternating Gradient Descent to Second Order Stationary Solutions

Songtao Lu *† lus@umn.edu
Mingyi Hong * mhong@umn.edu
Zhengdao Wang † zhengdao@iastate.edu

Abstract

The alternating gradient descent (AGD) is a simple but popular algorithm which has been applied to problems in optimization, machine learning, data ming, and signal processing, etc. The algorithm updates two blocks of variables in an alternating manner, in which a gradient step is taken on one block, while keeping the remaining block fixed. When the objective function is nonconvex, it is well-known the AGD converges to the first-order stationary solution with a global sublinear rate.

In this paper, we show that a variant of AGD-type algorithms will not be trapped by "bad" stationary solutions such as saddle points and local maximum points. In particular, we consider a smooth unconstrained optimization problem, and propose a perturbed AGD (PA-GD) which converges (with high probability) to the set of second-order stationary solutions (SS2) with a global sublinear rate. To the best of our knowledge, this is the first alternating type algorithm which takes $\mathcal{O}(\text{polylog}(d)/\epsilon^{7/3})$ iterations to achieve SS2 with high probability [where $\text{polylog}(d)$ is polynomial of the logarithm of dimension d of the problem].

1 Introduction

In this paper, we consider a smooth and unconstrained nonconvex optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^{d\times 1}} f(\mathbf{x}) \tag{1}$$

where $f:\mathbb{R}^d\to\mathbb{R}$ is twice differentiable.

There are many ways of solving problem (1), such as gradient descent (GD), accelerated gradient descent (AGD), etc. When the problem dimension is large, it is natural to split the variables into multiple blocks and solve the subproblems with smaller size individually. The block coordinate descent (BCD) algorithm, and many of its variants such as block coordinate gradient descent (BCGD) and alternating gradient descent (AGD) Bertsekas [1999]; Li and Liang [2017], are among the most powerful tools for solving large scale convex/nonconvex optimization problems Nesterov [2012]; Beck and Tetruashvili [2013]; Razaviyayn et al. [2013]; Hong et al. [2017]. The BCD-type algorithms partition the optimization variables into multiple small blocks, and optimize each block one by one following certain block selection rule, such as cyclic rule Tseng [2001], Gauss-Southwell rule Tseng and Yun [2009], etc.

In recent years, there are many applications of BCD-type algorithms in the areas of machine learning and data mining, such as matrix factorization Zhao et al. [2015]; Lu et al. [2017a,b], tensor decomposition, matrix completion/decomposition Xu and Yin [2013]; Jain et al. [2013], and training deep neural networks (DNNs) Zhang and Brand [2017]. Under relatively mild conditions, the convergence of BCD-type algorithms to first-order stationary solutions (SS1) have been broadly investigated for nonconvex and non-differentiable optimization Tseng [2001]; Grippo and Sciandrone [2000]. In particular, it is known that under mild conditions,

*Department of Electrical and Computer Engineering, University of Minnesota – Twin Cities
†Department of Electrical and Computer Engineering, Iowa State University

these algorithms also achieve global sublinear rates Razaviyayn et al. [2014]. However, despite its popularity and significant recent progress in understanding its behavior, it remains unclear whether BCD-type algorithms can converge to the set of second-order stationary solutions (SS2) with a provable global rate, even for the simplest problem with two blocks of variables.

1.1 Motivation

Algorithms that can escape from strict saddle points – those stationary points that have negative eigenvalues – have wide applications. Many recent works have analyzed the saddle points in machine learning problems Kawaguchi [2016]. Such as learning in shallow networks, the stationary points are either global minimum points or strict saddle points. In two-layer porcupine neural networks (PNNs), it has been shown that most local optima of PNN optimizations are also global optimizers Feizi et al. [2017]. Previous work in Ge et al. [2015] has shown that the saddle points in tensor decomposition are indeed strict saddle points. Also, it has been shown that any saddle points are strict in dictionary learning and phase retrieval problems theoretically and numerically in Sun et al. [2015, 2017]; Wang et al. [2017b,a]. More recently, Ge et al. [2017] proposed a unified analysis of saddle points for a board class of low rank matrix factorization problems, and they proved that these saddle points are strict.

1.2 Related Work

Many recent works have been focused on the performance analysis and/or design of algorithms with convergence guarantees to local minimum points/SS2 for nonconvex optimization problems. These include the trust region method Conn et al. [2000], cubic regularized Newton's method Nesterov and Polyak [2006]; Carmon and Duchi [2016], and a mixed approach of the first-order and seconde-order methods Reddi et al. [2017], etc. However, these algorithms typically require second-order information, therefore they incur high computational complexity when problem dimension becomes large.

There has been a line of work on stochastic gradient descent algorithms, where properly scaled Gaussian noise is added to the iterates of the gradient at each time [also known as stochastic gradient Langevin dynamics, (SGLD)]. Some theoretical works have pointed out that SGLD not only converges to the local minimum points asymptotically but also may escape from local minima Zhang et al. [2017]; Raginsky et al. [2017]. Unfortunately, these algorithms require a large number of iterations with $\mathcal{O}(1/\epsilon^4)$ steps to achieve the optimal point. There are fruitful results that show some carefully designed algorithms can escape from strict saddle point efficiently, such as negative-curvature-originated-from noise (Neon) Xu and Yang [2017], Neon2 Allen-Zhu and Li [2017], Neon$^+$Xu et al. [2017] and gradient descent with one-step escaping (GOSE) Yu et al. [2017]. The Neon-type of algorithms utilizes the stochastic first-order updates to find the negative curvature direction, and GOSE just needs one negative curvature descent step with calculation of eigenvectors when the iterates of the algorithm are near the saddle point for saving the computational burden.

On the other hand, there is also a line of work analyzing the deterministic GD type method. With random initializations, it has been shown that GD only converges to SS2 for unconstrained smooth problems Lee et al. [2016]. More recently, block coordinate descent, block mirror descent and proximal block coordinate descent have been proven to almost always converge to SS2 with random initializations Lee et al. [2017], but there is no convergence rate reported. Unfortunately, a follow-up study indicated that GD requires exponential time to escape from saddle points for certain pathological problems Du et al. [2017]. Adding some noise occasionally to the iterates of the algorithm is another way of finding the negative curvature. A perturbed version of GD has been proposed with convergence guarantees to SS2 Jin et al. [2017a], which shows a faster provable convergence rate than the ordinary gradient descent algorithm with random initializations. Furthermore, the accelerated version of PGD (PAGD) is also proposed in Jin et al. [2017b], which shows the fastest convergence rate among all Hessian free algorithms.

Table 1: Convergence rates of algorithms to SS2 with the first order information, where $p \geq 4$, and $\widetilde{\mathcal{O}}$ hides factor ploylog(d).

Algorithm	Iterations	(ϵ, γ)-SS2
SGD Ge et al. [2015]	$\mathcal{O}(d^p/\epsilon^4)$	$(\epsilon, \epsilon^{1/4})$
SGLD Zhang et al. [2017]	$\mathcal{O}(d^p/\epsilon^4)$	$(\epsilon, \epsilon^{1/2})$
Neon+SGD Xu and Yang [2017]	$\widetilde{\mathcal{O}}(1/\epsilon^4)$	$(\epsilon, \epsilon^{1/2})$
Neon+Natasha Xu and Yang [2017]	$\widetilde{\mathcal{O}}(1/\epsilon^{13/4})$	$(\epsilon, \epsilon^{1/4})$
Neon2+SGD Allen-Zhu and Li [2017]	$\widetilde{\mathcal{O}}(1/\epsilon^4)$	$(\epsilon, \epsilon^{1/2})$
Neon$^+$ Xu et al. [2017]	$\widetilde{\mathcal{O}}(1/\epsilon^{7/4})$	$(\epsilon, \epsilon^{1/2})$
PGD Jin et al. [2017a]	$\widetilde{\mathcal{O}}(1/\epsilon^2)$	$(\epsilon, \epsilon^{1/2})$
PAGD Jin et al. [2017b]	$\widetilde{\mathcal{O}}(1/\epsilon^{7/4})$	$(\epsilon, \epsilon^{1/2})$
PA-GD/PA-PP (This work)	$\widetilde{\mathcal{O}}(1/\epsilon^{7/3})$	$(\epsilon, \epsilon^{1/3})$

1.3 Scope of This Paper

In this work, we consider a smooth unconstrained optimization problem, and develop a perturbed AGD algorithm (PA-GD) which converges (with high probability) to the set of SS2 with a global sublinear rate. Our work is inspired by the works Jin et al. [2017a]; Ge et al. [2015], which developed novel perturbed GDs that escapes from strict saddle points. Similarly as in Jin et al. [2017a], we also divide the entire iterates of GD into three types of points: those whose gradients are large, those that are local minimum, and those that are strict saddle points. At a given point, when the size of the gradient is large enough, we just implement the ordinary AGD. When the gradient norm is small, which may be either strict saddle or local minimum, a perturbation will be added on the iterates to help to escape from the saddle points.

From the above section, we know that many works have been developed to make use of negative curvature information around the saddle points. Unfortunately, these techniques cannot be directly applied to the BCD/AGD- type of algorithms. The *key challenge* here is that at each iteration only part of the variables are updated, therefore we have access only to partial second order information at the points of interest. For example, consider a quadratic objective function shown in Figure 1. While fixing one block, the problem is strongly convex with respect to the other block, but the entire problem is nonconvex. Even if the iterates converge for each block to the minimum points within the block, the stationary point could still be a saddle point for the overall objective function. Therefore, the analysis of how AGD type of algorithms exploit the negative curvature is one of the main tasks in this paper.

To the best of our knowledge, there is no work on modifying AGD algorithms to escape from strict saddle points with any convergence rate. The main contributions of this work are as follows.

1.4 Contributions of This Work

In this paper, we design and analyze a perturbed AGD algorithm for solving an unconstrained nonconvex problem, namely perturbed AGD. Through the perturbation of AGD, the algorithm is guaranteed to converge to a set of SS2 of a nonconvex problem with high probability. By utilizing the matrix perturbation theory, convergence rate of the proposed algorithm is also established, which shows that the algorithm takes $\mathcal{O}(\text{polylog}(d)/\epsilon^{7/3})$ iterations to achieve an $(\epsilon, \epsilon^{1/3})$-SS2 with high probability. Also, considering the fact that there is a strong relation between GD and proximal point algorithm, we also study a perturbed alternating proximal point (PA-PP) algorithm with some random perturbation. By leveraging the techniques proposed in this paper, we show that PA-PP, which may not need to calculate the gradient at each step, converges as fast as PA-GD in the order of ϵ . The comparison of the algorithms which only use the first order information for escaping from strict saddle points is summarized as shown in Table 1.

The main contributions of the paper are highlighted below:

1. To the best of our knowledge, it is the first time that the convergence analysis shows that some variants of AGD (using first-order information) can converge to SS2 for nonconvex optimization problems.
2. The convergence rate of the perturbed AGD algorithm is analyzed, where the choice of the step size is only dependent on certain maximum Lipschitz constant over blocks rather than all variables. This is one of the major difference between GD and AGD.
3. By further extending the analysis in this paper, we also show that PA-PP can also escape from the strict points efficiently with the speed of $\mathcal{O}(\text{polylog}(d)/\epsilon^{7/3})$.

2 Preliminaries

2.1 Notation

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. Notation $\mathbf{x}_k$ denotes the kth block of vector $\mathbf{x} \in \mathbb{R}^{d\times 1}$. We use $\nabla_k f(\mathbf{x}_{-k}, \mathbf{x}_k)$ to denote the partial gradient with respect to its kth block variable while the remaining one is fixed. Notation $\mathbb{B}_{\mathbf{x}}(r)$ denotes a d-dimensional ball centered at $\mathbf{x}$ with radius r, and $\lambda_{\min}(\mathbf{X})$, $\lambda_{\max}(\mathbf{X})$ denote the smallest and largest eigenvalues of matrix $\mathbf{X}$ respectively.

2.2 Definitions

The objective function has the following properties.

Definition 1. *A differentiable function $f(\cdot)$ is L-smooth with gradient Lipschitz constant L (uniformly Lipschitz continuous), if*

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \le L\|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y}.$$

The function is called block-wise smooth with gradient Lipschitz constants $\{L_k\}$, if

$$\|\nabla_k f(\mathbf{x}_{-k}, \mathbf{x}_k) - \nabla_k f(\mathbf{x}_{-k}, \mathbf{x}'_k)\| \le L_k\|\mathbf{x}_k - \mathbf{x}'_k\|, \ \forall \mathbf{x}, \mathbf{x}'$$

or with gradient Lipschitz constants $\{\widetilde{L}_k\}$, if

$$\|\nabla_k f(\mathbf{x}_{-k}, \mathbf{x}_k) - \nabla_k f(\mathbf{x}'_{-k}, \mathbf{x}_k)\| \le \widetilde{L}_k\|\mathbf{x}_{-k} - \mathbf{x}'_{-k}\|, \ \forall \mathbf{x}, \mathbf{x}'.$$

Further, let $L_{\max} \triangleq \max\{L_k, \widetilde{L}_k, \forall k\} \le L$.

Definition 2. *For a differentiable function $f(\cdot)$, if $\|\nabla f(\mathbf{x})\| = 0$, then $\mathbf{x}$ is a first-order stationary point. If $\|\nabla f(\mathbf{x})\| \le \epsilon$, then $\mathbf{x}$ is an ϵ-first-order stationary point.*

Definition 3. *For a differentiable function $f(\cdot)$, if $\mathbf{x}$ is a SS1, and there exists $\epsilon > 0$ so that for any $\mathbf{y}$ in the ϵ-neighborhood of $\mathbf{x}$, we have $f(\mathbf{x}) \le f(\mathbf{y})$, then $\mathbf{x}$ is a local minimum. A saddle point $\mathbf{x}$ is a SS1 that is not a local minimum. If $\lambda_{\min}(\nabla^2 f(\mathbf{x})) < 0$, $\mathbf{x}$ is a strict (non-degenerate) saddle point.*

Definition 4. *A twice-differentiable function $f(\cdot)$ is ρ-Hessian Lipschitz if*

$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\| \le \rho\|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y}. \tag{2}$$

Definition 5. *For a ρ-Hessian Lipschitz function $f(\cdot)$, $\mathbf{x}$ is a second-order stationary point if $\|\nabla f(\mathbf{x})\| = 0$ and $\lambda_{\min}(\nabla^2 f(\mathbf{x})) \ge 0$. If the following holds*

$$\|\nabla f(\mathbf{x})\| \le \epsilon, \quad \textit{and} \quad \lambda_{\min}(\nabla^2 f(\mathbf{x})) \ge -\gamma \tag{3}$$

where $\epsilon, \gamma > 0$, then $\mathbf{x}$ is a (ϵ, γ)-SS2.

Assumption 1. *Function $f(\cdot)$ is L-smooth, block-wise smooth with gradient Lipschitz constants $\{L_k, \widetilde{L}_k\}, k = 1, 2$, and ρ-Hessian Lipschitz.*

Algorithm 1 Perturbed Alternating Gradient Descent (PA-GD) $(\mathbf{x}^{(0)}, L_{\max}, L, \rho, \epsilon, \delta, \Delta f)$

Input: $\mathcal{P}_1 = (1 + \frac{L}{L_{\max}})$, $\mathcal{P}_2 = (1 + \frac{L\log(2d)}{L_{\max}})$, $\chi = 6\max\{\log(\frac{\mathcal{P}_1^6\mathcal{P}_2^2 dL_{\max}^{5/3}\Delta_f}{c^5\rho^{1/3}\epsilon^{7/3}\delta}, 4\}$, $\eta = \frac{c}{L_{\max}}$, $r = \frac{c^3}{\chi^3}\frac{\rho\epsilon}{L_{\max}\mathcal{P}_1^3\mathcal{P}_2}$, $g_{\mathsf{th}} = \frac{c^2\epsilon}{(\chi\mathcal{P}_1)^3\mathcal{P}_2}$, $f_{\mathsf{th}} = \frac{c^5\epsilon^2}{L_{\max}(\chi\mathcal{P}_1)^6\mathcal{P}_2^2}$, $t_{\mathsf{th}} = \frac{L_{\max}\chi\mathcal{P}_1}{c^2(L_{\max}\rho\epsilon)^{\frac{1}{3}}}$
for $t = 0, 1, \ldots$ **do**
  if $\sum_{k=1}^{2}\|\nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t)})\|^2 \le g_{\mathsf{th}}^2$ and $t - t_{\mathsf{p}} > t_{\mathsf{th}}$ **then**
    $\widetilde{\mathbf{x}}^{(t)} \leftarrow \mathbf{x}^{(t)}$ and $t_{\mathrm{p}} \leftarrow t$
    $\mathbf{x}^{(t)} = \widetilde{\mathbf{x}}^{(t)} + \xi^{(t)}$, $\xi^{(t)}$ uniformly taken from $\mathbb{B}_0(r)$
  end if
  if $t - t_{\mathsf{p}} = t_{\mathsf{th}}$ and $f(\mathbf{x}^{(t)}) - f(\widetilde{\mathbf{x}}^{(t_{\mathsf{p}})}) > -f_{\mathsf{th}}$ **then**
    return $\widetilde{\mathbf{x}}^{t_{\mathsf{p}}}$
  end if
  for $k = 1, 2$ **do**
    $\mathbf{x}_k^{(t+1)} = \mathbf{x}_k^{(t)} - \eta\nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t)})$
  end for
end for

3 Perturbed Alternating Gradient Descent

3.1 Algorithm Description

AGD is a classical algorithm that optimizes the variables of an optimization problem in an alternating manner Bertsekas [1999], meaning that when one block of variables is updated, the remaining block is fixed to be the same as its previous solution. Mathematically, the iterates of AGD are updated by the following rule

$$\mathbf{x}_k^{(t+1)} = \mathbf{x}_k^{(t)} - \eta\nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t)}), \quad k = 1, 2 \tag{4}$$

where superscript (t) denotes the iteration counter; $\mathbf{h}_{-1}^{(t)} \triangleq \mathbf{x}_2^{(t)}$ and $\mathbf{h}_{-2}^{(t)} \triangleq \mathbf{x}_1^{(t+1)}$; $\eta > 0$ is the step size. AGD can be considered as a special case of block coordinate gradient descent Nesterov [2012]; Beck and Tetruashvili [2013].

Our proposed algorithm is based on AGD, but modified in a way similar to the recent work [Jin et al., 2017a], which adds some noise in PGD. The details of the implementation of PA-GD are shown in Algorithm 1, where c is a constant so that $\eta = c/L_{\max}$, Δ_f denotes the difference of the objective value at the initial point and global optimal solution, ϵ represents the predefined target error.

In each update of variables, we implement one step of the block gradient descent, and then proceed to the next block. Once the algorithm has sufficient decrease of the objective value, it implies that the algorithm converges to some good solution. Otherwise, some perturbation may be needed to help the iterates escape from the saddle points. If after the perturbation the objective value does not decrease sufficiently after a number of further iterations, the algorithm terminates and returns the iterate before the last perturbation.

To illustrate the practical behavior of the algorithm, we provide an example that shows the trajectory of AGD after a small perturbation at a stationary point. In Figure 1, it is clear that $\mathbf{x} = [0; 0]$ is a SS1 and also a strict saddle point since the eigenvalues of $\mathbf{A}$ are -1 and 3 respectively. When $\mathbf{x}_1$ is fixed, function $f(\mathbf{x})$ is convex with respect to $\mathbf{x}_2$ and vice versa, however, the objective function is nonconvex. It can be observed that PA-GD can escape from the strict saddle point efficiently.

3.2 Convergence Rate Analysis

Despite the fact that PA-GD exploits a different way of updating variables, we will show that it can still escape from strict saddle points with high probability with suitable perturbation. The main theorem is presented as follows.

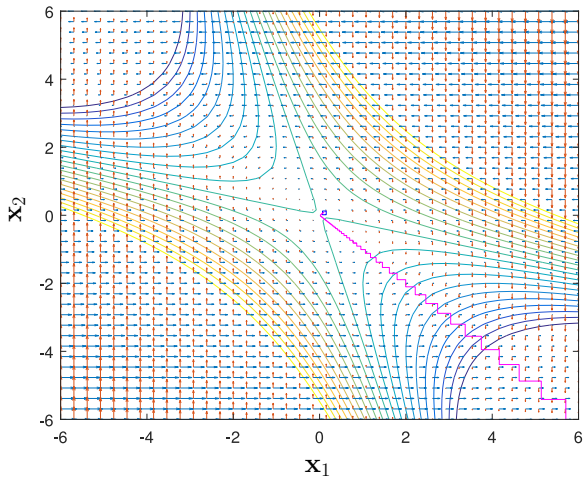


Figure 1: Contour of the objective values and the trajectory (pink color) of PA-GD started near strict saddle point $[0, 0]$. The objective function is $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}, \mathbf{x} = [\mathbf{x}_1; \mathbf{x}_2] \in \mathbb{R}^{2\times 1}$ where $\mathbf{A} \triangleq [1 \quad 2; 2 \quad 1] \in \mathbb{R}^{2\times 2}$, and the length of the arrows indicate the strength of $-\nabla f(\mathbf{x})$ projected onto directions $\mathbf{x}_1, \mathbf{x}_2$.

Theorem 1. *Under Assumption 1, there exists a constant $c_{\max}$ such that: for any $\delta \in (0, 1]$, $\epsilon \leq \frac{L_{\max}^2}{\rho}$, $\Delta_f \triangleq f(\mathbf{h}_{-1}^{(0)}, \mathbf{x}_1^{(0)}) - f^*$, and constant $c \leq c_{\max}$, with probability $1 - \delta$, the iterates generated by PA-GD converge to an ϵ-SS2 $\mathbf{x}$ satisfying*

$$\|\nabla f(\mathbf{x})\| \leq \epsilon, \quad \textit{and} \quad \lambda_{\min}(\nabla^2 f(\mathbf{x})) \geq -(L_{\max}\rho\epsilon)^{1/3}$$

in the following number of iterations:

$$\mathcal{O}\left(\frac{L_{\max}^{5/3}\mathcal{P}_1^7\mathcal{P}_2^2\Delta f}{\rho^{1/3}\epsilon^{7/3}}\log^7\left(\frac{\mathcal{P}_1^6\mathcal{P}_2^2 d L_{\max}^{5/3}\Delta_f}{c^5\rho^{1/3}\epsilon^{7/3}\delta}\right)\right) \tag{5}$$

where f^* denotes the global minimum value of the objective function, and $\mathcal{P}_1 = (1 + L/L_{\max})$ and $\mathcal{P}_2 = (1 + L\log(2d)/L_{\max})$.

Remark 1. When $\eta = c_{\max}/L$ is used, the convergence rate of PA-GD is

$$\mathcal{O}\left(\frac{L_{\max}^{5/3}\log^2(2d)\Delta f}{\rho^{1/3}\epsilon^{7/3}}\log^7\left(\frac{\mathcal{P}_1^6\mathcal{P}_2^2 d L_{\max}^{5/3}\Delta_f}{c^5\rho^{1/3}\epsilon^{7/3}\delta}\right)\right). \tag{6}$$

It shows that if a smaller step size is used, the convergence rate of PA-GD is faster (with smaller constants) since the linear dependency of $\mathcal{P}_1^7$ and $\mathcal{P}_2^2$ in (5) both disappear. This property is consistent with the known result when BCD is used in convex optimization problems, i.e., when a smaller step size is used, the rate could become better; e.g., see [Sun and Hong, 2015, Theorem 2.1].

4 Perturbed Alternating Proximal Point

In many applications, AGD may not be efficient in the sense that the convergence rate of gradient in each block may be very slow. For example, consider matrix factorization problem $\min_{\mathbf{X},\mathbf{Y}} \|\mathbf{Z} - \mathbf{X}\mathbf{Y}\|_F^2$ where $\mathbf{Z} \in \mathbb{R}^{m\times d}$ is the given data, $d \gg m$, and $\mathbf{X} \in \mathbb{R}^{m\times r}, \mathbf{Y} \in \mathbb{R}^{r\times d}$ are two block variables. For this problem,

Algorithm 2 Perturbed Alternating Proximal Point (PA-PP) $(\mathbf{x}^{(0)}, L_{\max}, L, \rho, \epsilon, \delta, \Delta f)$

Input: $\mathcal{P} = (1 + \frac{L\log(2d)}{L_{\max}})$, $\chi = 6\max\{\log(\frac{\mathcal{P}^2 d L_{\max}^{5/3}\Delta_f}{c^5\rho^{1/3}\epsilon^{7/3}\delta}, 4\}$, $\nu = \frac{L_{\max}}{c}$, $r = \frac{c^3}{\chi^3}\frac{\rho\epsilon}{L_{\max}\mathcal{P}}$, $g_{\mathsf{th}} = \frac{c^2\epsilon}{\chi^3\mathcal{P}}$, $f_{\mathsf{th}} = \frac{c^5\epsilon^2}{L_{\max}\chi^6\mathcal{P}^2}$, $t_{\mathsf{th}} = \frac{L_{\max}\chi}{c^2(L_{\max}\rho\epsilon)^{\frac{1}{3}}}$
for $t = 0, 1, \ldots$ **do**
  if $\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\| \le g_{\mathsf{th}}/\nu$ and $t - t_{\mathsf{p}} > t_{\mathsf{th}}$ **then**
    $\widetilde{\mathbf{x}}^{(t)} \leftarrow \mathbf{x}^{(t)}$ and $t_{\mathrm{p}} \leftarrow t$
    $\mathbf{x}^{(t)} = \widetilde{\mathbf{x}}^{(t)} + \xi^{(t)}$, $\xi^{(t)}$ uniformly taken from $\mathbb{B}_0(r)$
  end if
  if $t - t_{\mathsf{p}} = t_{\mathsf{th}}$ and $f(\mathbf{x}^{(t)}) - f(\widetilde{\mathbf{x}}^{(t_{\mathsf{p}})}) > -f_{\mathsf{th}}$ **then**
    return $\widetilde{\mathbf{x}}^{t_{\mathsf{p}}}$
  end if
  for $k = 1, 2$ **do**
    $\mathbf{x}_k^{(t+1)} = \arg\min_{\mathbf{x}_k} f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k) + \frac{\nu}{2}\|\mathbf{x}_k - \mathbf{x}_k^{(t)}\|^2$
  end for
end for

the alternating least squares algorithm (which exactly minimizes each block) would be a faster algorithm compared with the AGD which only uses gradient steps.

In this section, we consider the classical proximal point algorithm Parikh et al. [2014] in which each block of variables is exactly minimized with respect to certain quadratic surrogate. To be specific, we can replace (4) in Algorithm 1 by

$$\mathbf{x}_k^{(t+1)} = \arg\min_{\mathbf{x}_k} f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k) + \frac{\nu}{2}\|\mathbf{x}_k - \mathbf{x}_k^{(t)}\|^2,\ k = 1, 2 \tag{7}$$

where $\nu > 0$ is penalty parameter. The iteration can be explicitly written as

$$\mathbf{x}_k^{(t+1)} = \mathbf{x}_k^{(t)} - \frac{1}{\nu}\nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t+1)}), \quad k = 1, 2, \tag{8}$$

which has the similar form as the PA-GD algorithm, but with the step size being $\eta \triangleq 1/\nu$, and with gradient evaluated at the new iterate. The resulting algorithm, detailed in the table above, is referred to as the perturbed alternating proximal point (PA-PP). It is worth noting that when the subproblem is convex, such as $\min_{\mathbf{X},\mathbf{Y}} \|\mathbf{Z} - \mathbf{X}\mathbf{Y}\|_F^2$, ν only needs to be a small number to make the corresponding subproblem strongly convex. This property is useful in practice.

Next, we can also give the convergence rate of PA-PP.

Corollary 1. *Under Assumption 1, there exists a constant $c_{\max}$ such that: for any $\delta \in (0, 1]$, $\epsilon \le \frac{L_{\max}^2}{\rho}$, $\Delta_f \triangleq f(\mathbf{h}_{-1}^{(0)}, \mathbf{x}_1^{(0)}) - f^*$, and constant $c \le c_{\max}$, with probability $1 - \delta$, the iterates generated by PA-PP converges to an ϵ-SS2 $\mathbf{x}$ satisfying*

$$\|\nabla f(\mathbf{x})\| \le \epsilon, \quad \textit{and} \quad \lambda_{\min}(\nabla^2 f(\mathbf{x})) \ge -(L_{\max}\rho\epsilon)^{1/3}$$

in the following number of iterations:

$$\mathcal{O}\left(\frac{L_{\max}^{5/3}\mathcal{P}^2\Delta f}{\rho^{1/3}\epsilon^{7/3}}\log^7\left(\frac{\mathcal{P}^2 d L_{\max}^{5/3}\Delta_f}{c^5\rho^{1/3}\epsilon^{7/3}\delta}\right)\right)$$

where f^* denotes the global minimum value of the objective function, and $\mathcal{P} = (1 + L\log(2d)/L_{\max})$.

Comparing with Theorem 1, we can find that term $\mathcal{P}_1^7, \mathcal{P}_1 > 2$ is removed so the convergence rate of PA-PP is slightly faster than PA-GD.

5 Convergence Analysis

In this section, we will present the main proof steps of convergence analysis of PA-GD.

5.1 The Main Difficulty of the Proof

Gradient Descent: GD searches the descent direction of the objective function in the entire space $\mathbb{R}^d$. Without loss of generality, we assume $\mathbf{x}^{(0)} = 0$. According to the mean value theorem, the GD update can be expressed as

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta \nabla f(\mathbf{x}^{(t)}) = \mathbf{x}^{(t)} - \eta \nabla f(0) - \eta \left(\int_0^1 \nabla^2 f(\theta \mathbf{x}^{(t)}) d\theta \right) \mathbf{x}^{(t)}. \tag{9}$$

It can be observed that the update rule of GD contains the information of the Hessian matrix at point $\mathbf{x}^{(t)}$, i.e., $\nabla^2 f(\theta \mathbf{x}^{(t)})$. To be more specific, letting $\boldsymbol{\mathcal{H}} \triangleq \nabla^2 f(\mathbf{x}^*)$ where $\mathbf{x}^*$ denotes an ϵ-SS2 satisfying (3), we can rewrite (9) as

$$\mathbf{x}^{(t+1)} = (\mathbf{I} - \eta \boldsymbol{\mathcal{H}}) \mathbf{x}^{(t)} - \eta \Delta^{(t)} \mathbf{x}^{(t)} - \eta \nabla f(0) \tag{10}$$

where $\Delta^{(t)} \triangleq \int_0^1 (\nabla^2 f(\theta \mathbf{x}^{(t)}) - \boldsymbol{\mathcal{H}}) d\theta$.

Based on the ρ-Hessian Lipschitz property, we can quantify $\|\Delta^{(t)}\|$ that is upper bounded by the difference of iterates. By exploiting the negative curvature of the Hessian matrix at saddle point $\mathbf{x}^*$, we can project the iterate onto the direction $\vec{d}$ where the eigenvalue of $\mathbf{I} - \eta \boldsymbol{\mathcal{H}}$ is greater than 1. This leads to the fact that the norm of the iterates projected along direction $\vec{d}$ will be increasing exponentially as the algorithm proceeds around point $\mathbf{x}^*$, implying the sequence generated by GD is escaping from the saddle point. The details of characterizing the convergence rate have been analyzed previously in Jin et al. [2017a].

Alternating Gradient Descent: However, the AGD algorithm only updates partial variables of vector $\mathbf{x}$, which belong to a subspace of the feasible set. Similarly, from the mean value theorem we can express the AGD rule of updating variables with assuming $\mathbf{x}^{(0)} = 0$ as follows:

$$\begin{aligned} \mathbf{x}^{(t+1)} &= \mathbf{x}^{(t)} - \eta \begin{bmatrix} \nabla_1 f(\mathbf{x}_1^{(t)}, \mathbf{x}_2^{(t)}) \\ \nabla_2 f(\mathbf{x}_1^{(t+1)}, \mathbf{x}_2^{(t)}) \end{bmatrix} \\ &= \mathbf{x}^{(t)} - \eta \nabla f(0) - \eta \int_0^1 \boldsymbol{\mathcal{H}}_l^{(t)} d\theta \mathbf{x}^{(t+1)} - \eta \int_0^1 \boldsymbol{\mathcal{H}}_u^{(t)} d\theta \mathbf{x}^{(t)} \end{aligned} \tag{11}$$

where

$$\boldsymbol{\mathcal{H}}_l^{(t)} \triangleq \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \nabla_{21}^2 f(\theta \mathbf{x}_1^{(t+1)}, \theta \mathbf{x}_2^{(t)}) & \mathbf{0} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\mathcal{H}}_u^{(t)} \triangleq \begin{bmatrix} \nabla_{11}^2 f(\theta \mathbf{x}_1^{(t)}, \theta \mathbf{x}_2^{(t)}) & \nabla_{12}^2 f(\theta \mathbf{x}_1^{(t)}, \theta \mathbf{x}_2^{(t)}) \\ \mathbf{0} & \nabla_{22}^2 f(\theta \mathbf{x}_1^{(t+1)}, \theta \mathbf{x}_2^{(t)}) \end{bmatrix}.$$

From the above expression, it can be seen clearly that the update rule of AGD does not include a full Hessian matrix at any point but only partial ones. Furthermore, the right hand side of (11) not only contains the second order information of the previous point, i.e., $[\mathbf{x}_1^{(t)}, \mathbf{x}_2^{(t)}]$ but also the one of the most recently updated point, i.e., $[\mathbf{x}_1^{(t+1)}, \mathbf{x}_2^{(t)}]$. These represent the main challenges in understanding the behavior of the sequence generated by the AGD algorithm.

5.2 The Main Idea of the Proof

Although the second order information is divided into two parts, we can still characterize the recursion of the iterates around strict saddle points. We can also split $\boldsymbol{\mathcal{H}}$ as two parts, which are

$$\boldsymbol{\mathcal{H}}_u = \begin{bmatrix} \nabla_{11}^2 f(\mathbf{x}^*) & \nabla_{12}^2 f(\mathbf{x}^*) \\ \mathbf{0} & \nabla_{22}^2 f(\mathbf{x}^*) \end{bmatrix}, \quad \boldsymbol{\mathcal{H}}_l = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \nabla_{21}^2 f(\mathbf{x}^*) & \mathbf{0} \end{bmatrix}, \tag{12}$$

and obviously we have $\boldsymbol{\mathcal{H}} = \boldsymbol{\mathcal{H}}_l + \boldsymbol{\mathcal{H}}_u$.

Then, recursion (11) can be written as

$$\mathbf{x}^{(t+1)} + \eta\boldsymbol{\mathcal{H}}_l\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta\boldsymbol{\mathcal{H}}_u\mathbf{x}^{(t)} - \eta\Delta_u^{(t)}\mathbf{x}^{(t)} - \eta\Delta_l^{(t)}\mathbf{x}^{(t+1)} \tag{13}$$

where $\Delta_u^{(t)} \triangleq \int_0^1(\boldsymbol{\mathcal{H}}_u^{(t)}(\theta) - \boldsymbol{\mathcal{H}}_u)d\theta$, $\Delta_l^{(t)} \triangleq \int_0^1(\boldsymbol{\mathcal{H}}_l^{(t)}(\theta) - \boldsymbol{\mathcal{H}}_l)d\theta$. However, it is still unclear from (13) how the iteration evolves around the strict saddle point.

To highlight ideas, let us define

$$\mathbf{M} \triangleq \mathbf{I} + \eta\boldsymbol{\mathcal{H}}_l, \quad \mathbf{T} \triangleq \mathbf{I} - \eta\boldsymbol{\mathcal{H}}_u. \tag{14}$$

It can be observed that $\mathbf{M}$ is a lower triangular matrix where the diagonal entries are all 1s; therefore it is invertible. After taking the inverse of matrix $\mathbf{M}$ on both sides of (13), we can obtain

$$\mathbf{x}^{(t+1)} = \mathbf{M}^{-1}\mathbf{T}\mathbf{x}^{(t)} - \eta\mathbf{M}^{-1}\Delta_u^{(t)}\mathbf{x}^{(t)} - \eta\mathbf{M}^{-1}\Delta_l^{(t)}\mathbf{x}^{(t+1)}.$$

Our goal of analyzing the recursion of $\mathbf{x}^{(t)}$ becomes to find the maximum eigenvalue of $\mathbf{M}^{-1}\mathbf{T}$. With the help of the matrix perturbation theory, we can quantify the difference between the eigenvalues of matrix $\boldsymbol{\mathcal{H}}$ that contains the negative curvature and matrix $\mathbf{M}^{-1}\mathbf{T}$ that we are interested in analyzing. To be more precise, we give the following lemma.

Lemma 1. *Under Assumption 1, let $\boldsymbol{\mathcal{H}} \triangleq \nabla^2 f(\mathbf{x})$ denote the Hessian matrix at an ϵ-SS2 $\mathbf{x}$ where $\lambda_{\min}(\boldsymbol{\mathcal{H}}) \leq -\gamma$ and $\gamma > 0$. We have*

$$\lambda_{\max}(\mathbf{M}^{-1}\mathbf{T}) > 1 + \frac{\eta\gamma}{1 + L/L_{\max}} \tag{15}$$

where $\mathbf{M}, \mathbf{T}$ are defined in (12) *and* (14).

Lemma 1 illustrates that there exits a subspace spanned by the eigenvector of $\mathbf{M}^{-1}\mathbf{T}$ whose eigenvalue is greater than 1, indicating that the sequence generated by AGD can still potentially escape from the strict saddle point by leveraging such negative curvature information. Next, we can give a sketch of the proof of Theorem 1.

5.3 The Sketch of the Proof

The structure of the proof for quantifying the sufficient decrease of the objective function after the perturbation is borrowed from the proof of PGD Jin et al. [2017a], but PA-GD updates the variables block by block, so we have to provide the new proofs to show that PA-GD can still escape from saddle points with the perturbation technique.

First, if the size of the gradient is large enough, Algorithm 1 just implements the ordinary AGD. We give the descent lemma of AGD as follows.

Lemma 2. *Under Assumption 1, for the AGD algorithm with step size $\eta < 1/L_{\max}$, we have*

$$f(\mathbf{x}^{(t+1)}) \leq f(\mathbf{x}^{(t)}) - \sum_{k=1}^{2}\frac{\eta}{2}\|\nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t)})\|^2.$$

Second, if the iterates are near a strict saddle point, we can show that the AGD algorithm after a perturbation can give a sufficient decrease with high probability in terms of the objective value. To be more precise, the statement is given as follows.

Lemma 3. *Under Assumption 1, there exists a absolute constant $c_{\max}$. Let $c \leq c_{\max}$, $\chi \geq 1$, and η, r, g_{th}, t_{th} calculated as Algorithm 1 describes. Let $\widetilde{\mathbf{x}}^{(t)}$ be a strict saddle point, which satisfies*

$$\|\nabla f(\widetilde{\mathbf{x}}^{(t)})\|^2 \leq 4\sum_{k=1}^{2}\|\nabla_k f(\widetilde{\mathbf{h}}_{-k}^{(t)}, \widetilde{\mathbf{x}}_k^{(t)})\|^2 \leq 4g_{th}^2 \tag{16}$$

and $\lambda_{\min}(\nabla^2 f(\widetilde{\mathbf{x}}^{(t)})) \leq -\gamma$, *where* $\widetilde{\mathbf{h}}_{-1}^{(t)} \triangleq \widetilde{\mathbf{x}}_2^{(t)}$ *and* $\widetilde{\mathbf{h}}_{-2}^{(t)} \triangleq \mathbf{x}_1^{(t+1)}$.

Let $\mathbf{x}^{(t)} = \widetilde{\mathbf{x}}^{(t)} + \xi^{(t)}$ *where* $\xi^{(t)}$ *is generated randomly which follows the uniform distribution over* $\mathbb{B}_0(r)$, *and let* $\mathbf{x}^{(t+t_{th})}$ *be the iterates of PA-GD. With at least probability* $1 - \frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{1/3}}e^{-\chi}$, *we have* $f(\mathbf{x}^{(t+t_{th})}) - f(\widetilde{\mathbf{x}}^{(t)}) \leq -f_{th}$.

We remark that Lemma 2 is well-known and Lemma 3 is the core technique. In the following, we outline the main idea used in proving the latter. The formal statements of these steps are shown in the appendix; see Lemma 8–Lemma 10 therein.

We emphasize that the main contributions of this paper lies in the analysis of the first two steps, where the special update rule of PA-GD is analyzed so that the negative curvature of $\boldsymbol{\mathcal{H}}$ around the saddle points can be utilized.

Step 1 (Lemma 8) Consider a generic sequence $\mathbf{u}^{(t)}$ generated by PA-GD. As long as the initial point of $\mathbf{u}^{(t)}$ is close to saddle point $\widetilde{\mathbf{x}}^{(t)}$, the distance between $\mathbf{u}^{(t)}$ and $\widetilde{\mathbf{x}}^{(t)}$ can be upper bounded by using the ρ-Hessian Lipschitz continuity property.

Step 2 (Lemma 9) Leveraging the negative curvature around the strict saddle point, we know that there exits a direction, i.e., $\vec{\mathbf{e}}$, which is spanned by the eigenvector of $\mathbf{M}^{-1}\mathbf{T}$ whose corresponding eigenvalue is largest (greater than 1). Consider two sequences generated by PA-GD, $\mathbf{u}^{(t)}, \mathbf{w}^{(t)}$ initialized around the saddle point. When the initial points of these two iterates are separated apart away from each other along direction $\vec{\mathbf{e}}$ with a small distance, meaning that $\mathbf{w}^{(0)} = \mathbf{u}^{(0)} + \upsilon r\vec{\mathbf{e}},\ \upsilon \in [\delta/(2\sqrt{d}), 1]$ where r denotes the radius of the perturbation ball defined in Algorithm 1, we can show that if iterate $\mathbf{u}^{(t)}$ is still near the saddle point after T steps, the other sequence $\mathbf{w}^{(t)}$ will give a sufficient decrease of the objective value with less than T steps, implying that iterates $\mathbf{w}^{(t)}$ can escape from the saddle point with less than T steps.

Step 3 (Lemma 10) Consider $\mathbf{u}^{(0)}, \mathbf{w}^{(0)}$ as the points after the perturbation from the saddle point. We can quantify the probability that the AGD sequence will give a sufficient decrease of the objective value within T iterations after the perturbation [Jin et al., 2017a, Lemma 14,15].

5.4 Extension to PA-PP

By leveraging the convergence analysis of PA-GD and relation between PA-GD and PA-PP shown in (8), we can also write the recursion of the PA-PP iteration as

$$\mathbf{x}^{(t+1)} + \eta\boldsymbol{\mathcal{H}}_l'\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta\boldsymbol{\mathcal{H}}_u'\mathbf{v}^{(t)} - \eta\Delta_u'^{(t)}\mathbf{x}^{(t)} - \eta\Delta_l'^{(t)}\mathbf{x}^{(t+1)} \tag{17}$$

where $\eta = 1/\nu$, $\Delta_u'^{(t)} \triangleq \int_0^1(\boldsymbol{\mathcal{H}}_u'^{(t)}(\theta) - \boldsymbol{\mathcal{H}}_u')d\theta$, $\Delta_l'^{(t)} \triangleq \int_0^1(\boldsymbol{\mathcal{H}}_l'^{(t)}(\theta) - \boldsymbol{\mathcal{H}}_l')d\theta$,

$$\boldsymbol{\mathcal{H}}_u' = \begin{bmatrix} \mathbf{0} & \nabla_{12}^2 f(\widetilde{\mathbf{x}}^{(t)}) \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \boldsymbol{\mathcal{H}}_l' = \begin{bmatrix} \nabla_{11}^2 f(\widetilde{\mathbf{x}}^{(t)}) & \mathbf{0} \\ \nabla_{21}^2 f(\widetilde{\mathbf{x}}^{(t)}) & \nabla_{22}^2 f(\widetilde{\mathbf{x}}^{(t)}) \end{bmatrix}, \tag{18}$$

and

$$\boldsymbol{\mathcal{H}}_l'^{(t)} \triangleq \begin{bmatrix} \nabla_{11}^2 f(\theta\mathbf{x}_1^{(t+1)}, \theta\mathbf{x}_2^{(t)}) & \mathbf{0} \\ \nabla_{21}^2 f(\theta\mathbf{x}_1^{(t+1)}, \theta\mathbf{x}_2^{(t+1)}) & \nabla_{22}^2 f(\theta\mathbf{x}_1^{(t+1)}, \theta\mathbf{x}_2^{(t+1)}) \end{bmatrix}, \quad \boldsymbol{\mathcal{H}}_u'^{(t)} \triangleq \begin{bmatrix} \mathbf{0} & \nabla_{12}^2 f(\theta\mathbf{x}_1^{(t+1)}, \theta\mathbf{x}_2^{(t)}) \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Let

$$\mathbf{M}' \triangleq \mathbf{I} + \eta\boldsymbol{\mathcal{H}}_l' \quad \mathbf{T}' \triangleq \mathbf{I} - \eta\boldsymbol{\mathcal{H}}_u'. \tag{19}$$

We know that $\mathbf{T}'$ is an upper triangular matrix where the diagonal entries are all 1s, so it is invertible. Different from the case of PA-GD, we take the inverse of matrix $\mathbf{T}'$ on both sides of (17) and obtain

$$\mathbf{T}'^{-1}\mathbf{M}'\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta\mathbf{T}'^{-1}\Delta_u'^{(t)}\mathbf{x}^{(t)} - \eta\mathbf{T}'^{-1}\Delta_l'^{(t)}\mathbf{x}^{(t+1)}.$$

Then, we can give the following result that characterizes the recursion of $\mathbf{x}^{(t)}$ generated by PA-PP.

Corollary 2. *Under Assumption 1, let* $\boldsymbol{\mathcal{H}} \triangleq \nabla^2 f(\mathbf{x})$ *denote the Hessian matrix at an* ϵ*-SS2* $\mathbf{x}$ *where* $\lambda_{\min}(\boldsymbol{\mathcal{H}}) \leq -\gamma$ *and* $\gamma > 0$*. Let* $\lambda^{+}_{\min}(\cdot)$ *denote the minimum positive eigenvalue of a matrix. Then we have*

$$\lambda^{+}_{\min}(\mathbf{T}'^{-1}\mathbf{M}') \leq 1 - \eta\gamma/2 \tag{20}$$

where $\mathbf{M}', \mathbf{T}'$ *are defined in* (18) *and* (19)*;* $\eta \leq 1/L_{\max}$ *and* $\gamma \leq L_{\max}$*.*

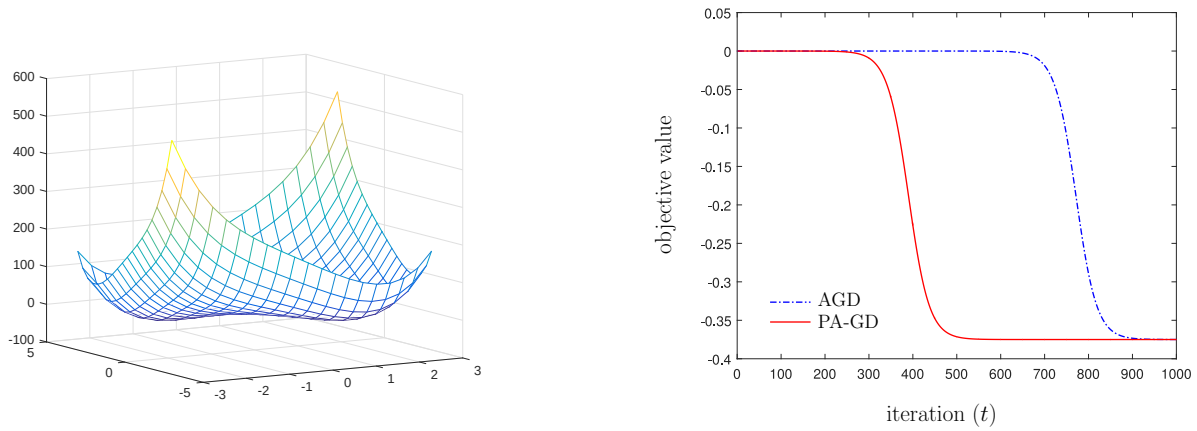
We remark that Corollary 2 is useful since it can be leveraged to show that the norm of the iterates around saddle points can increase exponentially. Then, we can apply the similar analysis steps as the case of proving the convergence rate of PA-GD and obtain the results shown in Corollary 1.

6 Connection with Existing Works

Remark 2. In Theorem 1 we characterized the convergence rate to an $(\epsilon, \epsilon^{1/3})$-SS2. We can also translate this bound to the one for achieving an $(\epsilon, \sqrt{\epsilon})$-SS2, and in this case PA-GD needs $\widetilde{\mathcal{O}}(1/\epsilon^{3.5})$ iterations. Compared with the existing recent works Jin et al. [2017a], the convergence rate of PA-GD/PA-PP is slower than GD. The main reason is the fact that different from GD-type algorithms, PA-GD and PA-PP cannot fully utilize the Hessian information because they never see a full iteration. Similar situation happens for SGD-type of algorithms which also cannot get the exact negative curvature around strict saddle points.

From Table 1, it can be seen that the convergence rate of PA-GD/PA-PP is still faster than SGD Ge et al. [2015], SGLD Zhang et al. [2017], Neon+SGD Xu and Yang [2017], and Neon2+SGD Allen-Zhu and Li [2017] to achieve an $(\epsilon, \sqrt{\epsilon})$-SS2, but slower than the rest. We emphasize that PA-GD and PA-PP represent the first BCD-type algorithms with the convergence rate guarantee to escape from the strict saddle points efficiently. At this point, it is unclear whether our rate is the best that is achievable, and the question of whether the resulting rate can be improved will be left to future work.

7 Numerical Results



(a) Objective function in 2D.

(b) Objective value versus the number of iterations

Figure 2: Convergence comparison between AGD and PA-GD, where $\epsilon = 10^{-4}$, $g_{\mathsf{th}} = \epsilon/10$, $\eta = 0.02$, $t_{\mathsf{th}} = 10/\epsilon^{1/3}$, $r = \epsilon/10$.

In this section, we present a simple example that shows the convergence behavior of PA-GD. Consider a nonconvex objective function, i.e.,

$$f(\mathbf{x}) \triangleq \mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} + \frac{1}{4}\|\mathbf{x}\|_4^4. \tag{21}$$

First, we have the following properties of function $f(\mathbf{x})$ such that $f(\mathbf{x})$ satisfies the assumptions of the analysis.

Lemma 4. *For any* $\tau \geq \lambda_{\max}(\mathbf{A})$ *and* $\mathbf{x} \in \{\mathbf{x} | \|\mathbf{x}\|^2 \leq \tau\}$, $f(\mathbf{x})$ *defined in* (21) *is* 5τ*-smooth and* $6\sqrt{\tau}$*-Hessian Lipschitz.*

Here, we can easily show the shape of objective function (21) in the two dimensional (2D) case in Figure 2(a), where $\mathbf{A} = [1\ 2; 2\ 1] \in \mathbb{R}^{2\times 2}$. It can be observed clearly that there exits a strict saddle point at $[0, 0]$ and two other local optimal points. We randomly initialize the algorithms around strict saddle point $[0, 0]$. The convergence comparison between AGD and PA-GD is shown in Figure 2(b). It can be observed that PA-GD converges faster than AGD to a local optimal point.

8 Conclusion

In this paper, the perturbed variants of AGD and alternating proximal point (APP) algorithms are proposed, with the objective of finding the second order stationary solutions of nonconvex smooth problems. Leveraging the recently developed idea of random perturbation for the first-order methods, the proposed algorithms add suitable perturbation to the AGD or APP iterates. The main contribution of this work is a new analysis that takes into consideration the block structure of the updates for the perturbed AGD and APP algorithms. By exploiting the negative curvature, it is established that with high probability the algorithms can converge to an $(\epsilon, \epsilon^{1/3})$-SS2 with $\mathcal{O}(\text{polylog}(d)/\epsilon^{7/3})$ iterations.

9 Acknowledgment

The authors would like to thank Chi Jin for discussion on the perturbed gradient descent algorithm.

References

Zeyuan Allen-Zhu and Yuanzhi Li. Neon2: Finding local minima via first-order oracles. *arXiv preprint arXiv:1711.06673*, 2017.

James R. Angelos, Carl C. Cowen, and Sivaram K. Narayan. Triangular truncation and finding the norm of a Hadamard multiplier. *Linear Algebra and its Applications*, 170:117–135, 1992.

Amir Beck and Luba Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23(4):2037–2060, 2013.

Dimitri P. Bertsekas. *Nonlinear Programming, 2nd ed.* Athena Scientific, Belmont, MA, 1999.

Yair Carmon and John C Duchi. Gradient descent efficiently finds the cubic-regularized non-convex Newton step. *arXiv preprint arXiv:1612.00547*, 2016.

Andrew R. Conn, Nicholas I. M. Gould, and Philippe L. Toint. *Trust region methods*. SIAM, 2000.

Simon S. Du, Chi Jin, Jason D. Lee, Michael I. Jordan, Barnabás Póczos, and Aarti Singh. Gradient descent can take exponential time to escape saddle points. In *Proceedings of Neural Information Processing Systems (NIPS)*, 2017.

Soheil Feizi, Hamid Javadi, Jesse Zhang, and David Tse. Porcupine neural networks: (almost) all local optima are global. *arXiv:1710.02196 [stat.ML]*, 2017.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points — online stochastic gradient for tensor decomposition. In *Proceedings of Annual Conference on Learning Theory (COLT)*, pages 797–842, 2015.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 1233–1242, 2017.

L. Grippo and M. Sciandrone. On the convergence of the block nonlinear Gauss-Seidel method under convex constraints. *Operations Research Letters*, 26:127–136, 2000.

John A Holbrook. Spectral variation of normal matrices. *Linear Algebra and its Applications*, 174:131–144, 1992.

Mingyi Hong, Xiangfeng Wang, Meisam Razaviyayn, and Zhi-Quan Luo. Iteration complexity analysis of block coordinate descent methods. *Mathematical Programming Series A*, 163(1):85–114, May 2017.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of Annual ACM Symposium on Theory of Computing*, pages 665–674, 2013.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 1724–1732, 2017a.

Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. *arXiv preprint arXiv:1711.10456*, 2017b.

Kenji Kawaguchi. Deep learning without poor local minima. In *Proceedings of Neural Information Processing Systems (NIPS)*, pages 586–594, 2016.

Jason D. Lee, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Proceedings of Annual Conference on Learning Theory (COLT)*, pages 1246–1257, 2016.

Jason D. Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. First-order methods almost always avoid saddle points. *arXiv:1710.07406v1 [stat.ML]*, 2017.

Yuanzhi Li and Yingyu Liang. Provable alternating gradient descent for non-negative matrix factorization with strong correlations. In *Proceedings of International Conference on Machine Learning (ICML)*, volume 70, pages 2062–2070, 2017.

Songtao Lu, Mingyi Hong, and Zhengdao Wang. A stochastic nonconvex splitting method for symmetric nonnegative matrix factorization. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54, pages 812–821, 2017a.

Songtao Lu, Mingyi Hong, and Zhengdao Wang. A nonconvex splitting method for symmetric nonnegative matrix factorization: Convergence analysis and optimality. *IEEE Transactions on Signal Processing*, 65 (12):3120–3135, June 2017b.

Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Yurii Nesterov and Boris T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3): 127–239, 2014.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Proceedings of Annual Conference on Learning Theory (COLT)*, pages 1674–1703, 2017.

Meisam Razaviyayn, Mingyi Hong, and Zhi-Quan Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013.

Meisam Razaviyayn, Mingyi Hong, Zhi-Quan Luo, and Jong-Shi Pang. Parallel successive convex approximation for nonsmooth nonconvex optimization. In *Proceedings of Neural Information Processing Systems (NIPS)*, 2014.

Sashank J. Reddi, Manzil Zaheer, Suvrit Sra, Barnabás Póczos, Francis Bach, Ruslan Salakhutdinov, and Alexander J Smola. A generic approach for escaping saddle points. *arXiv:1709.01434 [cs.LG]*, 2017.

Ju Sun, Qing Qu, and John Wright. When are nonconvex problems not scary? In *Proceedings of NIPS Workshop on Non-convex Optimization for Machine Learning: Theory and Practice*, 2015.

Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *arXiv:1602.06664 [cs.IT]*, 2017.

Ruoyu Sun and Mingyi Hong. Improved iteration complexity bounds of cyclic block coordinate descent for convex problems. In *Proceedings of Neural Information Processing Systems (NIPS)*, pages 1306–1314, 2015.

Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001.

Paul Tseng and Sangwoon Yun. Block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization. *Journal of Optimization Theory and Applications*, 140(3):513, 2009.

Gang Wang, Georgios B. Giannakis, and Yonina C. Eldar. Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Transactions on Information Theory*, 2017a.

Gang Wang, Georgios B. Giannakis, Yousef Saad, and Jie Chen. Solving almost all systems of random quadratic equations. In *Proceedings of Neural Information Processing Systems (NIPS)*, 2017b.

Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912.

Yangyang Xu and Wotao Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on Imaging Sciences*, 6(3):1758–1789, 2013.

Yi Xu and Tianbao Yang. First-order stochastic algorithms for escaping from saddle points in almost linear time. *arXiv preprint arXiv:1711.01944*, 2017.

Yi Xu, Rong Jin, and Tianbao Yang. Neon+: Accelerated gradient methods for extracting negative curvature for non-convex optimization. *arXiv preprint arXiv:1712.01033*, 2017.

Yaodong Yu, Difan Zou, and Quanquan Gu. Saving gradient and negative curvature computations: Finding local minima more efficiently. *arXiv preprint arXiv:1712.03950*, 2017.

Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient langevin dynamics. In *Proceedings of Annual Conference on Learning Theory (COLT)*, pages 1980–2022, 2017.

Ziming Zhang and Matthew Brand. On the convergence of block coordinate descent in training DNNs with Tikhonov regularization. In *Proceedings of Neural Information Processing Systems (NIPS)*, 2017.

Tuo Zhao, Zhaoran Wang, and Han Liu. A nonconvex optimization framework for low rank matrix estimation. In *Proceedings of Neural Information Processing Systems (NIPS)*, pages 559–567, 2015.

Appendix

A Preliminary

We provide the proofs of some preliminary lemmas (Lemma 5–Lemma 7) used in the proof of Section B.

First, Lemma 5 and Lemma 6 give the property that quantify the size of the difference of the second-order information of the objective values between two points.

Lemma 5. *If function $f(\cdot)$ is ρ-Hessian Lipschitz, we have*

$$\left\| \int_0^1 \nabla^2 f(\theta \mathbf{x}) d\theta - \nabla^2 f(\mathbf{y}) \right\| \leq \rho \left(\|\mathbf{x}\| + \|\mathbf{y}\| \right), \quad \forall \mathbf{x}, \mathbf{y}. \tag{22}$$

Lemma 6. *Under Assumption 1, we have block-wise Lipschitz continuity as follows:*

$$\left\| \begin{bmatrix} \nabla_{11}^2 f(\mathbf{x}) & \nabla_{12}^2 f(\mathbf{x}) \\ \mathbf{0} & \nabla_{22}^2 f(\mathbf{y}) \end{bmatrix} - \begin{bmatrix} \nabla_{11}^2 f(\mathbf{z}) & \nabla_{12}^2 f(\mathbf{z}) \\ \mathbf{0} & \nabla_{22}^2 f(\mathbf{z}) \end{bmatrix} \right\| \leq \rho \left(\|\mathbf{x} - \mathbf{z}\| + \|\mathbf{y} - \mathbf{z}\| \right), \forall \mathbf{x}, \mathbf{y}, \mathbf{z}, \tag{23}$$

and

$$\left\| \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \nabla_{21}^2 f(\mathbf{x}) & \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \nabla_{21}^2 f(\mathbf{y}) & \mathbf{0} \end{bmatrix} \right\| \leq \rho \|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y}. \tag{24}$$

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

Lemma 7. *If function $f(\cdot)$ is L-smooth with Lipschitz constant, then we have*

$$\|\nabla f(\mathbf{x}^{(t)})\|^2 \leq 4 \sum_{k=1}^{2} \|\nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t)})\|^2 \tag{25}$$

where sequence $\mathbf{x}_k^{(t)}, k = 1, 2$ is generated by the AGD algorithm.

A.1 Proof of Lemma 5

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

$$\begin{aligned} \left\| \int_0^1 (\nabla^2 f(\theta \mathbf{x}) - \nabla^2 f(\mathbf{y})) d\theta \right\| &\leq \int_0^1 \left\| \nabla^2 f(\theta \mathbf{x}) - \nabla^2 f(\mathbf{y}) \right\| d\theta \\ \overset{(a)}{\leq} \rho \int_0^1 \|\theta \mathbf{x} - \mathbf{y}\| \, d\theta &\overset{(b)}{\leq} \rho \int_0^1 \theta \|\mathbf{x}\| d\theta + \rho \|\mathbf{y}\| \leq \rho \left(\|\mathbf{x}\| + \|\mathbf{y}\| \right) \end{aligned}$$

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

A.2 Proof of Lemma 6

There proof involves two parts:

Upper Triangular Matrix: Consider three different vectors $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$. We can have

$$
\begin{aligned}
&\left\|\begin{bmatrix} \nabla^2_{11} f(\mathbf{x}) & \nabla^2_{12} f(\mathbf{x}) \\ 0 & \nabla^2_{22} f(\mathbf{y}) \end{bmatrix} - \begin{bmatrix} \nabla^2_{11} f(\mathbf{z}) & \nabla^2_{12} f(\mathbf{z}) \\ 0 & \nabla^2_{22} f(\mathbf{z}) \end{bmatrix}\right\| \\
\leq &\left\|\mathbf{I}_1 \left(\begin{bmatrix} \nabla^2_{11} f(\mathbf{x}) & \nabla^2_{12} f(\mathbf{x}) \\ \nabla^2_{21} f(\mathbf{x}) & \nabla^2_{22} f(\mathbf{x}) \end{bmatrix} - \begin{bmatrix} \nabla^2_{11} f(\mathbf{z}) & \nabla^2_{12} f(\mathbf{z}) \\ \nabla^2_{21} f(\mathbf{z}) & \nabla^2_{22} f(\mathbf{z}) \end{bmatrix} \right)\right\| \\
&+ \left\|\mathbf{I}_2 \left(\begin{bmatrix} \nabla^2_{11} f(\mathbf{y}) & \nabla^2_{12} f(\mathbf{y}) \\ \nabla^2_{21} f(\mathbf{y}) & \nabla^2_{22} f(\mathbf{y}) \end{bmatrix} - \begin{bmatrix} \nabla^2_{11} f(\mathbf{z}) & \nabla^2_{12} f(\mathbf{z}) \\ \nabla^2_{21} f(\mathbf{z}) & \nabla^2_{22} f(\mathbf{z}) \end{bmatrix} \right) \mathbf{I}_2\right\| \\
\overset{(a)}{\leq} &\left\|\begin{bmatrix} \nabla^2_{11} f(\mathbf{x}) & \nabla^2_{12} f(\mathbf{x}) \\ \nabla^2_{21} f(\mathbf{x}) & \nabla^2_{22} f(\mathbf{x}) \end{bmatrix} - \begin{bmatrix} \nabla^2_{11} f(\mathbf{z}) & \nabla^2_{12} f(\mathbf{z}) \\ \nabla^2_{21} f(\mathbf{z}) & \nabla^2_{22} f(\mathbf{z}) \end{bmatrix}\right\| + \left\|\begin{bmatrix} \nabla^2_{11} f(\mathbf{y}) & \nabla^2_{12} f(\mathbf{y}) \\ \nabla^2_{21} f(\mathbf{y}) & \nabla^2_{22} f(\mathbf{y}) \end{bmatrix} - \begin{bmatrix} \nabla^2_{11} f(\mathbf{z}) & \nabla^2_{12} f(\mathbf{z}) \\ \nabla^2_{21} f(\mathbf{z}) & \nabla^2_{22} f(\mathbf{z}) \end{bmatrix}\right\| \\
\leq &\rho \left(\|\mathbf{x} - \mathbf{z}\| + \|\mathbf{y} - \mathbf{z}\| \right)
\end{aligned}
$$

where in (a) we used

$$
\mathbf{I}_1 = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & 0 \end{bmatrix} \qquad \mathbf{I}_2 = \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{I} \end{bmatrix} \tag{26}
$$

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

Lower Triangular Matrix:

$$
\begin{aligned}
&\left\|\begin{bmatrix} 0 & 0 \\ \nabla^2_{21} f(\mathbf{x}) & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \nabla^2_{21} f(\mathbf{y}) & 0 \end{bmatrix}\right\| \\
= &\left\|\mathbf{I}_2 \left(\begin{bmatrix} \nabla^2_{11} f(\mathbf{x}) & \nabla^2_{12} f(\mathbf{x}) \\ \nabla^2_{21} f(\mathbf{x}) & \nabla^2_{22} f(\mathbf{x}) \end{bmatrix} - \begin{bmatrix} \nabla^2_{11} f(\mathbf{y}) & \nabla^2_{12} f(\mathbf{y}) \\ \nabla^2_{21} f(\mathbf{y}) & \nabla^2_{22} f(\mathbf{y}) \end{bmatrix} \right) \mathbf{I}_1\right\| \\
\overset{(a)}{\leq} &\rho \|\mathbf{x} - \mathbf{y}\|
\end{aligned}
$$

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

A.3 Proof of Lemma 7

Proof. Recall the definition

$$
\mathbf{h}^{(t)}_{-1} \triangleq \mathbf{x}^{(t)}_2 \quad \text{and} \quad \mathbf{h}^{(t)}_{-2} \triangleq \mathbf{x}^{(t+1)}_1.
$$

First, we have

$$
\|\nabla_2 f(\mathbf{x}^{(t)}_1, \mathbf{x}^{(t)}_2)\|^2 \leq 2\|\nabla_2 f(\mathbf{x}^{(t+1)}_1, \mathbf{x}^{(t)}_2) - \nabla_2 f(\mathbf{x}^{(t)}_1, \mathbf{x}^{(t)}_2)\|^2 + 2\|\nabla_2 f(\mathbf{x}^{(t+1)}_1, \mathbf{x}^{(t)}_2)\|^2. \tag{27}
$$

Using block-wise Lipschitz continuity, we have

$$
\begin{aligned}
\|\nabla_2 f(\mathbf{x}^{(t)}_1, \mathbf{x}^{(t)}_2)\|^2 &\leq 2L^2_{\max} \|\mathbf{x}^{(t+1)}_1 - \mathbf{x}^{(t)}_1\|^2 + 2\|\nabla_2 f(\mathbf{x}^{(t+1)}_1, \mathbf{x}^{(t)}_2)\|^2 \\
&\overset{(a)}{=} 2L^2_{\max} \|\eta \nabla_1 f(\mathbf{x}^{(t)}_1, \mathbf{x}^{(t)}_2)\|^2 + 2\|\nabla_2 f(\mathbf{x}^{(t+1)}_1, \mathbf{x}^{(t)}_2)\|^2 \\
&\overset{(b)}{\leq} 2 \sum_{k=1}^{2} \|\nabla_k f(\mathbf{h}^{(t)}_{-k}, \mathbf{x}^{(t)}_k)\|^2
\end{aligned} \tag{28}
$$

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

Summing $\|\nabla_1 f(\mathbf{x}^{(t)}_1, \mathbf{x}^{(t)}_2)\|^2$ on both sides of the above equation, we have

$$
\|\nabla f(\mathbf{x}^{(t)})\|^2 \leq \sum_{k=1}^{2} \|\nabla_k f(\mathbf{x}^{(t)}_k)\|^2 \leq 4 \sum_{k=1}^{2} \|\nabla_k f(\mathbf{h}^{(t)}_{-k}, \mathbf{x}^{(t)}_k)\|^2. \tag{29}
$$

□

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 8–Lemma 10) that lead to Lemma 3.

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

$$\begin{aligned}
\mathcal{F} \triangleq& \eta^5 L_{\max}^5 \frac{\gamma^3}{\kappa^3\rho^2} \log^{-6}\left(\frac{d\kappa}{\delta}\right) \mathcal{P}_1^{-6}\mathcal{P}_2^{-2},\\
\mathcal{G} \triangleq& \eta^2 L_{\max}^2 \frac{\gamma^2}{\rho} \log^{-3}\left(\frac{d\kappa}{\delta}\right) \mathcal{P}_1^{-3}\mathcal{P}_2^{-1},\\
\mathcal{S} \triangleq& \eta^2 L_{\max}^2 \frac{\gamma}{\kappa\rho} \log^{-2}\left(\frac{d\kappa}{\delta}\right) \mathcal{P}_1^{-2}\mathcal{P}_2^{-1},\\
\mathcal{T} \triangleq& \frac{\log\left(\frac{d\kappa}{\delta}\right)\mathcal{P}_1}{\eta\gamma}.
\end{aligned}$$

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 $\kappa \triangleq \frac{L_{\max}}{\gamma} \geq 1$.

These quantities, $\mathcal{F}$, $\mathcal{G}$, $\mathcal{S}$ and $\mathcal{T}$ have some certain relations as follows, which are useful of simplifying the expressions in the proofs.

$$\sqrt{\mathcal{F}} = \frac{\sqrt{\eta}\mathcal{G}}{\kappa}, \tag{30a}$$

$$\frac{\eta\mathcal{G}\mathcal{T}}{\kappa} = \mathcal{S}, \tag{30b}$$

$$\rho\mathcal{S}^3 = \frac{\eta L_{\max}\mathcal{F}}{\mathcal{P}_2}, \tag{30c}$$

$$\eta\rho\mathcal{S}\mathcal{T} = \frac{\eta^2 L_{\max}^2}{\kappa\log(\frac{d\kappa}{\delta})\mathcal{P}_1\mathcal{P}_2}. \tag{30d}$$

In the process of the proofs, we used conditions $\log(\frac{d\kappa}{\delta}) \geq 1$, $\mathcal{P}_1 \geq 2$ repeatedly to simply the expressions of the parameters. We also consider saddle point $\widetilde{\mathbf{x}}^{(t)}$ that satisfies the following condition.

Condition 1. *An ϵ-second order stationary point $\widetilde{\mathbf{x}}^{(t)}$ satisfies the following conditions:*

$$\sum_{k=1}^{2} \|\nabla_k f(\widetilde{\mathbf{h}}_{-k}^{(t)}, \widetilde{\mathbf{x}}_k^{(t)})\|^2 \leq g_{th}^2 \quad \textit{and} \quad \lambda_{\min}(\nabla^2 f(\widetilde{\mathbf{x}}^{(t)})) \leq -\gamma \tag{31}$$

where $g_{th} \triangleq \frac{\mathcal{G}}{2\kappa}$.

Condition 1 implies that point $\widetilde{\mathbf{x}}^{(t)}$ satisfies $\|\nabla f(\widetilde{\mathbf{x}}^{(t)})\| \leq \mathcal{G}/\kappa$ (see Lemma 7) and $\lambda_{\min}(\nabla^2 f(\widetilde{\mathbf{x}}^{(t)})) \leq -\gamma$.

Sufficient Decrease after Perturbation Consider $\widetilde{\mathbf{x}}^{(t)}$ satisfy Condition 1 and let $\boldsymbol{\mathcal{H}} \triangleq \nabla^2 f(\widetilde{\mathbf{x}}^{(t)})$. We consider a second order approximation as the following

$$\widehat{f}_{\mathbf{y}}(\mathbf{x}) \triangleq f(\mathbf{y}) + \nabla f(\mathbf{y})^{\mathsf{T}}(\mathbf{x}-\mathbf{y}) + \frac{1}{2}(\mathbf{x}-\mathbf{y})^{\mathsf{T}}\boldsymbol{\mathcal{H}}(\mathbf{x}-\mathbf{y}). \tag{32}$$

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 8. *Under Assumption 1, consider $\widetilde{\mathbf{x}}^{(t)}$ that satisfies Condition 1 and a generic sequence $\mathbf{u}^{(t)}$ generated by AGD. For any constant $\widehat{c} \geq 2$, $\delta \in (0, \frac{d\kappa}{e}]$, when initial point $\mathbf{u}^{(0)}$ satisfies*

$$\|\mathbf{u}^{(0)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 2r, \tag{33}$$

then, with the definition of

$$r \triangleq \frac{\eta L_{\max}\mathcal{S}}{\kappa \log(\frac{d\kappa}{\delta})\mathcal{P}_1}, \quad \textit{and} \quad T \triangleq \min\{\inf_t\{t|\widehat{f}_{\mathbf{u}^{(0)}}(\mathbf{u}^{(t)}) - f(\mathbf{u}^{(0)}) \leq -3\mathcal{F}\}, \widehat{c}\mathcal{T}\}, \tag{34}$$

there exits constants $c_{\max}^{(1)}, \widehat{c}$ such that for any $\eta \leq c_{\max}^{(1)}/L_{\max}$, the iterates generated by PA-GD satisfy $\|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 5\widehat{c}\mathcal{S}, \forall t < T$.

Lemma 9. *Under Assumption 1, consider $\widetilde{\mathbf{x}}^{(t)}$ that satisfies Condition 1. There exist constants $c_{\max}^{(2)}$, $\widehat{c}$ such that: for any $\delta \in (0, \frac{d\kappa}{e}]$ and $\eta \leq c_{\max}^{(2)}/L_{\max}$, with the definition of*

$$T \triangleq \min\left\{\inf_t\{t|\widehat{f}_{\mathbf{w}_0}(\mathbf{w}^{(t)}) - f(\mathbf{w}^{(0)}) \leq -3\mathcal{F}\}, \widehat{c}\mathcal{T}\right\}$$

where two iterates $\{\mathbf{u}^{(t)}\}$ and $\{\mathbf{w}^{(t)}\}$ that are generated by PA-GD with initial points $\{\mathbf{u}^{(0)}, \mathbf{w}^{(0)}\}$ satisfying

$$\|\mathbf{u}^{(0)} - \widetilde{\mathbf{x}}^{(t)}\| \leq r, \ \mathbf{w}^{(0)} = \mathbf{u}^{(0)} + \upsilon r\vec{\mathbf{e}}, \ \upsilon \in [\delta/(2\sqrt{d}), 1], \tag{35}$$

where $\vec{\mathbf{e}}$ denotes the eigenvector of $\mathbf{M}^{-1}\mathbf{T}$ whose eigenvalue is maximum, then, if $\|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 5\widehat{c}\mathcal{S}, \forall t < T$, we will have $T < \widehat{c}\mathcal{T}$.

Lemma 8 says that if the $\mathbf{u}^{(t)}$-iterate generated by PA-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 9 shows if there exists another PA-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 PA-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., 2017a, Lemma 15]. Once we know that PA-GD also decreases the objective value sufficiently in Lemma 8 and Lemma 9, 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.

$$\begin{aligned}
\mathbb{P}(\mathbf{w}^{(0)} \in \mathcal{X}_{\text{stuck}}) &= \int_{\mathbb{B}_{\widetilde{\mathbf{x}}^{(t)}}(r)} \mathbb{P}(\mathbf{w}^{(0)} \in \mathcal{X}_{\text{stuck}}|\mathbf{u}^{(0)} \in \mathcal{X}_{\text{stuck}})\mathbb{P}(\mathbf{u}^{(0)} \in \mathcal{X}_{\text{stuck}})d\mathbf{u}^{(0)} \\
&\leq \int_{\mathbb{B}_{\widetilde{\mathbf{x}}^{(t)}}(r)} \mathbb{P}(\mathbf{w}^{(0)} \in \mathcal{X}_{\text{stuck}}|\mathbf{u}^{(0)} \in \mathcal{X}_{\text{stuck}})\mathbb{P}(\mathbf{u}^{(0)})d\mathbf{u}^{(0)} \\
&\overset{(a)}{\leq} \delta \int_{\mathbb{B}_{\widetilde{\mathbf{x}}^{(t)}}(r)} \mathbb{P}(\mathbf{u}^{(0)})d\mathbf{u}^{(0)} = \delta
\end{aligned}$$

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 $\mathbb{P}(\mathbf{w}^{(0)} \in \mathcal{X}_{\text{stuck}}|\mathbf{u}^{(0)} \in \mathcal{X}_{\text{stuck}})$ can be upper bounded by δ, which is proven in the following lemma.

Lemma 10. *Under Assumption 1, there exists a universal constant $c_{\max}$, for any $\delta \in (0, d\kappa/e]$: consider a saddle point $\widetilde{\mathbf{x}}^{(t)}$ which satisfies Condition 1, let $\mathbf{x}^{(0)} = \widetilde{\mathbf{x}}^{(t)} + \xi$ where ξ is generated randomly which follows the uniform distribution over a ball with radius r, and let $\mathbf{x}^{(t)}$ be the iterates of PA-GD starting from $\mathbf{x}^{(0)}$. Then, when step size $\eta \leq c_{\max}/L_{\max}$, with at least probability $1-\delta$, we have the following for any $T \geq \mathcal{T}/c_{\max}$*

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

Then, applying $\eta = \frac{c}{L_{\max}}$,$\gamma = (L_{\max}\rho\epsilon)^{1/3}$, and $\delta = \frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{1/3}}e^{-\chi}$ into Lemma 10, 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 $\eta = \frac{c}{L_{\max}}$,$\gamma = (L_{\max}\rho\epsilon)^{1/3}$, and $\delta = \frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{1/3}}e^{-\chi}$ into the definitions of $\mathcal{F}, \mathcal{G}, \mathcal{T}$, we will have the following definitions.

$$\begin{aligned}
f_{\mathsf{th}} &\triangleq \mathcal{F} = \frac{c^5\epsilon^2}{L_{\max}(\chi\mathcal{P}_1)^6\mathcal{P}_2^2}, \\
g_{\mathsf{th}} &\triangleq \frac{\mathcal{G}}{2\kappa} = \frac{c^2\epsilon}{2(\chi\mathcal{P}_1)^3\mathcal{P}_2}, \\
t_{\mathsf{th}} &\triangleq \frac{\mathcal{T}}{c} = \frac{L_{\max}\chi\mathcal{P}_1}{c^2(L_{\max}\rho\epsilon)^{\frac{1}{3}}}.
\end{aligned}$$

After applying Lemma 7, we know that

$$\|\nabla f(\mathbf{x})\| \leq \frac{c}{\chi^3\mathcal{P}_1^3\mathcal{P}_2}\epsilon$$

where $c \leq 1, \chi, \mathcal{P}_1, \mathcal{P}_2 \geq 1$.

With a set of necessary lemmas and leveraging the proof of PGD [Jin et al., 2017a, 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 $\sum_{k=1}^{2}\|\nabla_k f(\mathbf{h}_{-k}^{(0)}, \mathbf{x}_k^{(0)})\|^2 > g_{\mathsf{th}}^2$: According to Lemma 2, we have

$$\begin{aligned}
f(\mathbf{x}^{(1)}) - f(\mathbf{x}^{(0)}) &\leq -\sum_{k=1}^{2}\frac{\eta}{2}\|\nabla_k f(\mathbf{h}_{-k}^{(0)}, \mathbf{x}_k^{(0)})\|^2 \leq -\frac{\eta}{2}g_{\mathsf{th}}^2 \\
&\overset{(a)}{=} -\frac{c^5}{8(\chi\mathcal{P}_1)^6\mathcal{P}_2^2}\frac{\epsilon^2}{L_{\max}}
\end{aligned} \tag{37}$$

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

2. The gradient is small in all block directions, namely $\sum_{k=1}^{2}\|\nabla_k f(\mathbf{h}_{-k}^{(0)}, \mathbf{x}_k^{(0)})\|^2 \leq g_{\mathsf{th}}^2$: in this case, we will add the perturbation to the iterates, and implement AGD for the next t_{th} steps and then check the termination condition. If the termination condition is not satisfied, we must have

$$f(\mathbf{x}^{(t_{\mathsf{th}})}) - f(\mathbf{x}^{(0)}) \leq -f_{\mathsf{th}} = -\frac{c^5\epsilon^2}{L_{\max}(\chi\mathcal{P}_1)^6\mathcal{P}_2^2}, \tag{38}$$

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

$$\frac{f(\mathbf{x}^{(t_{\mathsf{th}})}) - f(\mathbf{x}^{(0)})}{t_{\mathsf{th}}} \leq -\frac{c^7}{(\chi\mathcal{P}_1)^7\mathcal{P}_2^2}\frac{\epsilon^2}{L_{\max}}\frac{(L_{\max}\rho\epsilon)^{\frac{1}{3}}}{L_{\max}}. \tag{39}$$

Since $\kappa = L_{\max}/(L_{\max}\rho\epsilon)^{1/3} \geq 1$, we know that the right-hand side (RHS) of (39) is greater than RHS of (37).

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 $\frac{c^7}{(\chi\mathcal{P}_1)^7\mathcal{P}_2^2}\frac{\epsilon^2}{L_{\max}}\frac{(L_{\max}\rho\epsilon)^{1/3}}{L_{\max}}$ per iteration on average. This means that Algorithm 1 must stop within a finite number of iterations, which is

$$\frac{f(\mathbf{h}_{-1}^{(0)}, \mathbf{x}_1^{(0)}) - f^*}{\frac{c^7}{(\chi\mathcal{P}_1)^7\mathcal{P}_2^2}\frac{\epsilon^2}{L_{\max}}\frac{(L_{\max}\rho\epsilon)^{1/3}}{L_{\max}}} = \frac{(\chi\mathcal{P}_1)^7\mathcal{P}_2^2}{c^7}\frac{L_{\max}^2\Delta f}{\epsilon^2(L_{\max}\rho\epsilon)^{1/3}} = \mathcal{O}\left(\frac{\Delta f(\chi\mathcal{P}_1)^7\mathcal{P}_2^2 L_{\max}^{5/3}}{\rho^{1/3}\epsilon^{7/3}}\right) \tag{40}$$

where $\Delta f \triangleq f(\mathbf{h}_{-1}^{(0)}, \mathbf{x}_1^{(0)}) - f^*$.

According to Lemma 3, we know that with probability $1 - \frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{1/3}}e^{-\chi}$ the algorithm can give a sufficient descent with the perturbation when $\sum_{k=1}^{2}\|\nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t)})\|^2 \leq g_{\mathsf{th}}^2$. Since the total number of perturbation we can add is at most

$$n = \frac{1}{t_{\mathsf{th}}}\frac{(\chi\mathcal{P}_1)^7\mathcal{P}_2^2}{c^7}\frac{L_{\max}^2\Delta f}{\epsilon^2(L_{\max}\rho\epsilon)^{1/3}} = \frac{(\mathcal{P}_1\chi)^6\mathcal{P}_2^2}{c^5}\frac{L_{\max}\Delta_f}{\epsilon^2}. \tag{41}$$

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

$$1 - n\frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{\frac{1}{3}}}e^{-\chi} = 1 - \frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{\frac{1}{3}}}e^{-\chi}\frac{(\mathcal{P}_1\chi)^6\mathcal{P}_2^2}{c^5}\frac{L_{\max}\Delta_f}{\epsilon^2} = 1 - \underbrace{\frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{\frac{1}{3}}}\frac{\mathcal{P}_1^6\mathcal{P}_2^2}{c^5}\frac{\Delta_f}{\epsilon^2}}_{\triangleq\mathcal{C}}\chi^6 e^{-\chi}. \tag{42}$$

With chosen $\chi = 6\max\{\ln(\mathcal{C}/\delta), 4\}$, we have $\chi^6 e^{-\chi} \leq e^{-\chi/6}$, which implies $\chi^6 e^{-\chi}\mathcal{C} \leq e^{-\chi/6}\mathcal{C} \leq \delta$.

The proof is complete. □

B.2 Proof of Lemma 1

Proof. Recall the definitions:

$$\boldsymbol{\mathcal{H}}_u \triangleq \begin{bmatrix} \nabla_{11}^2 f(\widetilde{\mathbf{x}}^{(t)}) & \nabla_{12}^2 f(\widetilde{\mathbf{x}}^{(t)}) \\ 0 & \nabla_{22}^2 f(\widetilde{\mathbf{x}}^{(t)}) \end{bmatrix} \quad \boldsymbol{\mathcal{H}}_l \triangleq \begin{bmatrix} 0 & 0 \\ \nabla_{21}^2 f(\widetilde{\mathbf{x}}^{(t)}) & 0 \end{bmatrix}, \tag{43}$$

where $\widetilde{\mathbf{x}}^{(t)}$ is an ϵ-second order stationary point, and

$$\mathbf{M} \triangleq \mathbf{I} + \eta\boldsymbol{\mathcal{H}}_l, \quad \mathbf{T} \triangleq \mathbf{I} - \eta\boldsymbol{\mathcal{H}}_u. \tag{44}$$

Our goal of this lemma is to show that the maximum eigenvalue of $\mathbf{M}^{-1}\mathbf{T}$ is 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 $\det(\mathbf{M}) = 1$, 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.,

$$\begin{aligned}\det[\mathbf{T} - \lambda\mathbf{M}] =& \det[\mathbf{I} - \eta\boldsymbol{\mathcal{H}}_u - \lambda(\mathbf{I} + \eta\boldsymbol{\mathcal{H}}_l)] \\ =& \det\left[\underbrace{\begin{array}{cc}(1-\lambda)\mathbf{I} - \eta\nabla_{11}^2 f(\widetilde{\mathbf{x}}^{(t)}) & -\eta\nabla_{12}^2 f(\widetilde{\mathbf{x}}^{(t)}) \\ -\lambda\eta\nabla_{21}^2 f(\widetilde{\mathbf{x}}^{(t)}) & (1-\lambda)\mathbf{I} - \eta\nabla_{22}^2 f(\widetilde{\mathbf{x}}^{(t)})\end{array}}_{\triangleq\mathbf{Q}(\lambda)}\right].\end{aligned}$$

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

Consider a $\delta > 0$. We have

$$\mathbf{Q}(1+\delta) = -\left(\underbrace{\eta\boldsymbol{\mathcal{H}} + \delta(\mathbf{I} + \eta\boldsymbol{\mathcal{H}}_l)}_{\triangleq\mathbf{F}(\delta)}\right) \tag{45}$$

where

$$\begin{aligned}\mathbf{F}(\delta) =& \delta\mathbf{I} + \eta\begin{bmatrix}\nabla_{11}^2 f(\widetilde{\mathbf{x}}^{(t)}) & \nabla_{12}^2 f(\widetilde{\mathbf{x}}^{(t)}) \\ (1+\delta)\nabla_{21}^2 f(\widetilde{\mathbf{x}}^{(t)}) & \nabla_{22}^2 f(\widetilde{\mathbf{x}}^{(t)})\end{bmatrix} \\ =& \begin{bmatrix}\mathbf{I} & \\ & \sqrt{1+\delta}\end{bmatrix}\underbrace{\begin{bmatrix}\delta\mathbf{I} + \eta\nabla_{11}^2 f(\widetilde{\mathbf{x}}^{(t)}) & \eta\sqrt{1+\delta}\nabla_{12}^2 f(\widetilde{\mathbf{x}}^{(t)}) \\ \eta\sqrt{1+\delta}\nabla_{21}^2 f(\widetilde{\mathbf{x}}^{(t)}) & \delta\mathbf{I} + \eta\nabla_{22}^2 f(\widetilde{\mathbf{x}}^{(t)})\end{bmatrix}}_{\mathbf{G}(\delta)}\begin{bmatrix}\mathbf{I} & \\ & \frac{1}{\sqrt{1+\delta}}\end{bmatrix},\end{aligned}$$

meaning that $\mathbf{F}(\delta)$ is similar to $\mathbf{G}(\delta)$. Consequently, we can conclude that $\mathbf{F}(\delta)$ has the same eigenvalues of $\mathbf{G}(\delta)$. 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

$$\max_{1\le i\le d}|\lambda_i(\eta\boldsymbol{\mathcal{H}}) - \lambda_i(\mathbf{G}(\delta))| \le \|\eta\boldsymbol{\mathcal{H}} - \mathbf{G}(\delta)\| \tag{46}$$

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 (46), we can check

$$\begin{aligned}&\|\eta\boldsymbol{\mathcal{H}} - \mathbf{G}(\delta)\| \\ =& \left\|\delta\mathbf{I} + \begin{bmatrix}0 & (\sqrt{1+\delta}-1)\eta\nabla_{12}^2 f(\widetilde{\mathbf{x}}^{(t)}) \\ (\sqrt{1+\delta}-1)\eta\nabla_{21}^2 f(\widetilde{\mathbf{x}}^{(t)}) & 0\end{bmatrix}\right\| \\ \le& \delta + (\sqrt{1+\delta}-1)\eta\|\boldsymbol{\mathcal{H}}\| + (\sqrt{1+\delta}-1)\eta\left\|\begin{array}{cc}\nabla_{11}^2 f(\widetilde{\mathbf{x}}^{(t)}) & 0 \\ 0 & \nabla_{22}^2 f(\widetilde{\mathbf{x}}^{(t)})\end{array}\right\| \\ \overset{(a)}{\le}& \delta + (\sqrt{1+\delta}-1)(\frac{L}{L_{\max}}+1).\end{aligned} \tag{47}$$

where (a) is true since we used $\eta \le c_{\max}/L_{\max}$ 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 $\|\eta\boldsymbol{\mathcal{H}} - \mathbf{G}(\delta)\| \le \delta + 2(\sqrt{1+\delta}-1)$.

We know that the minimum eigenvalue of $\eta\boldsymbol{\mathcal{H}}$ which is $-\eta\gamma$ and the maximum difference of the eigenvalues between $\eta\boldsymbol{\mathcal{H}}$ and $\mathbf{G}(\delta)$ is upper bounded by (47). 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

$$\delta + (\sqrt{1+\delta}-1)(\frac{L}{L_{\max}}+1) < \eta\gamma. \tag{48}$$

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 (48) holds when $\delta^* = \frac{\eta\gamma}{1+\frac{L}{L_{\max}}}$.

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

Second step : we only need to check

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

meaning that it is sufficient to check

$$(\frac{L}{L_{\max}}+1)^2(1+\delta) \leq \left(\frac{L}{L_{\max}}+1+\frac{\eta\gamma}{2}\right)^2. \tag{49}$$

It can be easily check that the left-hand side (LHS) of (49) with chosen δ^* is

$$(\frac{L}{L_{\max}}+1)^2(1+\frac{\eta\gamma}{\frac{L}{L_{\max}}+1}) \leq (\frac{L}{L_{\max}}+1)^2+(\frac{L}{L_{\max}}+1)^2\eta\gamma < (\frac{L}{L_{\max}}+1)^2+(\frac{L}{L_{\max}}+1)^2\eta\gamma+\frac{\eta^2\gamma^2}{4},$$

which is RHS of (49).

Therefore, we can conclude that $\mathbf{Q}(1+\delta^*)$ has a negative eigenvalue.

When δ is large, it is easy to check $\mathbf{Q}(1+\delta)$ has a positive eigenvalue, since term $\delta^2\mathbf{I}$ dominates the spectrum of matrix $\mathbf{Q}(1+\delta)$ in (45). Since the eigenvalue is continuous with respect to δ, we can conclude there exists a largest δ, i.e., $\widehat{\delta}$, such that $\mathbf{Q}(1+\widehat{\delta})$ has a zero eigenvalue, i.e., $\det(\mathbf{Q}(1+\widehat{\delta}))=0$ where $1+\widehat{\delta}$ is at least

$$1+\delta^* = 1+\frac{\eta\gamma}{L/L_{\max}+1}. \tag{50}$$

Therefore, we can conclude that there exits a largest real eigenvalue of $\mathbf{M}^{-1}\mathbf{T}$ which is $1+\widehat{\delta} > 1+\delta^* > 1$. □

B.3 Proof of Lemma 2

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

$$\begin{aligned}
f(\mathbf{x}^{(t+1)}) \leq & f(\mathbf{x}^{(t)}) + \sum_{k=1}^{2} \nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t)})^{\mathsf{T}}(\mathbf{x}_k^{(t+1)} - \mathbf{x}_k^{(t)}) + \sum_{k=1}^{2} \frac{L_k}{2}\|\mathbf{x}_k^{(t+1)} - \mathbf{x}_k^{(t)}\|^2 \\
\overset{(a)}{\leq} & f(\mathbf{x}^{(t)}) - \sum_{k=1}^{2} \eta\|\nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t)})\|^2 + \sum_{k=1}^{2} \frac{\eta^2 L_k}{2}\|\nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t)})\|^2 \\
\overset{(b)}{\leq} & f(\mathbf{x}^{(t)}) - \sum_{k=1}^{2} \frac{\eta}{2}\|\nabla_k f(\mathbf{h}_{-k}^{(t)}, \mathbf{x}_k^{(t)})\|^2
\end{aligned} \tag{51}$$

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 8

Proof. Without loss of generality, let $\mathbf{u}^{(0)}$ be the origin, i.e., $\mathbf{u}^{(0)} = 0$. According to the AGD update rules, we have

$$\mathbf{u}^{(t+1)} = \mathbf{u}^{(t)} - \eta \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t)}, \mathbf{u}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t)}) \end{bmatrix}. \tag{52}$$

Then, we use the mathematical induction to prove that

$$\|\mathbf{u}^{(t)}\| \leq 5\widehat{c}\mathcal{S}, \forall t < T. \tag{53}$$

When $t = 0$, we have $\mathbf{u}^{(0)} = 0$, so (53) is true.

Suppose (53) is true for the case where $\tau \leq t$. We will show that (53) is also true for the case where $\tau = t + 1$.

First, we need to show the upper bound of $\|\mathbf{u}^{(t+1)} - \mathbf{u}^{(t)}\|$. According to the Taylor expansion and ρ-Hessian Lipschitz continuity, we have

$$f(\mathbf{u}^{(t)}) \leq f(\mathbf{u}^{(0)}) + \nabla f(\mathbf{u}^{(0)})^\mathsf{T}(\mathbf{u}^{(t)} - \mathbf{u}^{(0)}) + \frac{1}{2}(\mathbf{u}^{(0)} - \mathbf{u}^{(t)})^\mathsf{T}\nabla^2 f(\mathbf{u}^{(0)})(\mathbf{u}^{(0)} - \mathbf{u}^{(t)}) + \frac{\rho}{6}\|\mathbf{u}^{(t)} - \mathbf{u}^{(0)}\|^3.$$

Comparing with the definition of $\widehat{f}_{\mathbf{u}^{(0)}}(\mathbf{u}^{(t)})$, we have

$$\begin{aligned} |f(\mathbf{u}^{(t)}) - \widehat{f}_{\mathbf{u}^{(0)}}(\mathbf{u}^{(t)})| &\overset{(32)}{\leq} \frac{1}{2}(\mathbf{u}^{(0)} - \mathbf{u}^{(t)})^\mathsf{T}\left(\nabla^2 f(\mathbf{u}^{(0)}) - \boldsymbol{\mathcal{H}}\right)(\mathbf{u}^{(0)} - \mathbf{u}^{(t)}) + \frac{\rho}{6}\|\mathbf{u}^{(t)} - \mathbf{u}^{(0)}\|^3 \\ &\overset{(a)}{\leq} \frac{\rho}{2}\|\mathbf{u}^{(0)} - \widetilde{\mathbf{x}}^{(t)}\|\|\mathbf{u}^{(t)} - \mathbf{u}^{(0)}\|^2 + \frac{\rho}{6}\|\mathbf{u}^{(t)} - \mathbf{u}^{(0)}\|^3 \end{aligned}$$

where in (a) we also used ρ-Hessian Lipschitz continuity.

According to the definition of T, we know that $f(\mathbf{u}^{(0)}) - \widehat{f}_{\mathbf{u}^{(0)}}(\mathbf{u}^{(t)}) \leq 3\mathcal{F}$ for all $t < T$, which implies that

$$\begin{aligned} f(\mathbf{u}^{(0)}) - f(\mathbf{u}^{(t)}) \leq& |f(\mathbf{u}^{(0)}) - \widehat{f}_{\mathbf{u}^{(0)}}(\mathbf{u}^{(t)})| + |\widehat{f}_{\mathbf{u}^{(0)}}(\mathbf{u}^{(t)}) - f(\mathbf{u}^{(t)})| \\ \overset{(34)}{\leq}& 3\mathcal{F} + \frac{\rho}{2}\|\widetilde{\mathbf{x}}^{(t)} - \mathbf{u}^{(0)}\|\|\mathbf{u}^{(t)} - \mathbf{u}^{(0)}\|^2 + \frac{\rho}{6}\|\mathbf{u}^{(t)} - \mathbf{u}^{(0)}\|^3 \\ \leq& 3\mathcal{F} + \frac{\rho}{2}\frac{\eta L_{\max}\mathcal{S}}{\kappa \log(\frac{d\kappa}{\delta})\mathcal{P}_1}(5\widehat{c}\mathcal{S})^2 + \frac{\rho}{6}(5\widehat{c}\mathcal{S})^3 \qquad (54) \\ \leq& 3\mathcal{F} + ((5\widehat{c})^2/4 + (5\widehat{c})^3/6)\rho\mathcal{S}^3 \\ \overset{(30c)}{\leq}& 3\mathcal{F} + \eta L_{\max}(5\widehat{c})^3\mathcal{F}\mathcal{P}_2^{-1} \qquad (55) \\ \leq& 4\mathcal{F} \qquad (56) \end{aligned}$$

where in (56) we used $c_{\max} = \mathcal{P}_2/(5\widehat{c})^3$ and $\eta \leq c_{\max}/L_{\max}$.

From (51), we also know that

$$f(\mathbf{u}^{(t+1)}) \leq f(\mathbf{u}^{(t)}) - \frac{\eta}{2}\left(\|\nabla_1 f(\mathbf{u}_1^{(t)}, \mathbf{u}_2^{(t)})\|^2 + \|\nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t)})\|^2\right), \quad \forall t < T. \tag{57}$$

For simplification of expression, we define

$$\mathbf{z}_{-1}^{(t)} \triangleq \mathbf{u}_2^{(t)} \quad \text{and} \quad \mathbf{z}_{-2}^{(t)} \triangleq \mathbf{u}_1^{(t+1)}, \quad \forall t < T. \tag{58}$$

Summing up (57) for $\tau = 0, \ldots, t$, we have

$$f(\mathbf{u}^{(t)}) \leq f(\mathbf{u}^{(0)}) - \sum_{\tau=0}^{t-1}\sum_{k=1}^{2}\frac{\eta}{2}\|\nabla_k f(\mathbf{z}_{-k}^{(\tau)}, \mathbf{u}_k^{(\tau)})\|^2, \quad \forall t < T. \tag{59}$$

Combining (56) and (59), we know that

$$\sum_{\tau=0}^{t-1}\sum_{k=1}^{2}\frac{\eta}{2}\|\nabla_k f(\mathbf{z}_{-k}^{(\tau)},\mathbf{u}_k^{(\tau)})\|^2\leq 4\mathcal{F}, \tag{60}$$

which implies

$$\max_{\tau}\sum_{k=1}^{2}\frac{\eta}{2}\|\nabla_k f(\mathbf{z}_{-k}^{(\tau)},\mathbf{u}_k^{(\tau)})\|^2\leq 4\mathcal{F}, \tau\leq t-1. \tag{61}$$

According to (52), we know

$$\begin{aligned}
&\|\mathbf{u}^{(t+1)}-\mathbf{u}^{(t)}\|^2\\
=&\eta^2\sum_{k=1}^{2}\|\nabla_k f(\mathbf{z}_{-k}^{(t)},\mathbf{u}_k^{(t)})\|^2\\
=&2\eta^2\sum_{k=1}^{2}\|\nabla_k f(\mathbf{z}_{-k}^{(t)},\mathbf{u}_k^{(t)})-\nabla_k f(\mathbf{z}_{-k}^{(t-1)},\mathbf{u}_k^{(t-1)})\|^2+2\eta^2\sum_{k=1}^{2}\|\nabla_k f(\mathbf{z}_{-k}^{(t-1)},\mathbf{u}_k^{(t-1)})\|^2\\
=&2\eta^2\left(2\sum_{k=1}^{2}\|\nabla_k f(\mathbf{z}_{-k}^{(t)},\mathbf{u}_k^{(t)})-\nabla_k f(\mathbf{z}_{-k}^{(t-1)},\mathbf{u}_k^{(t)})\|^2+2\sum_{k=1}^{2}\|\nabla_k f(\mathbf{z}_{-k}^{(t-1)},\mathbf{u}_k^{(t)})-\nabla_k f(\mathbf{z}_{-k}^{(t-1)},\mathbf{u}_k^{(t-1)})\|^2\right)\\
&+2\eta^2\sum_{k=1}^{2}\|\nabla_k f(\mathbf{z}_{-k}^{(t-1)},\mathbf{u}_k^{(t-1)})\|^2\\
\overset{(a)}{\leq}&8\eta^2L_{\max}^2\|\mathbf{u}^{(t+1)}-\mathbf{u}^{(t)}\|^2+4\eta^2L_{\max}^2\|\mathbf{u}^{(t)}-\mathbf{u}^{(t-1)}\|^2+16\eta\mathcal{F}.
\end{aligned}$$

where in (a) we used Lipschitz continuity, i.e., $\sum_{k=1}^{2}\|\nabla_k f(\mathbf{z}_{-k}^{(t)},\mathbf{u}_k^{(t)})-\nabla_k f(\mathbf{z}_{-k}^{(t-1)},\mathbf{u}_k^{(t)})\|^2\leq L_{\max}^2\|\mathbf{u}_1^{(t+1)}-\mathbf{u}_1^{(t)}\|^2+L_{\max}^2\|\mathbf{u}_2^{(t)}-\mathbf{u}_2^{(t-1)}\|^2$, and $\sum_{k=1}^{2}\|\nabla_k f(\mathbf{z}_{-k}^{(t-1)},\mathbf{u}_k^{(t)})-\nabla_k f(\mathbf{z}_{-k}^{(t-1)},\mathbf{u}_k^{(t-1)})\|^2\leq L_{\max}^2\|\mathbf{u}_1^{(t+1)}-\mathbf{u}_1^{(t)}\|$.

Then, we have

$$\begin{aligned}
\|\mathbf{u}^{(t+1)}-\mathbf{u}^{(t)}\|^2\leq&\underbrace{\frac{4\eta^2L_{\max}^2}{(1-8\eta^2L_{\max}^2)}}_{\triangleq\omega}\|\mathbf{u}^{(t)}-\mathbf{u}^{(t-1)}\|^2+\frac{16\eta\mathcal{F}}{(1-8\eta^2L_{\max}^2)}\\
=&\omega^t\|\mathbf{u}^{(1)}-\mathbf{u}^{(0)}\|^2+\sum_{\tau=0}^{t-1}\omega^\tau\frac{16\eta\mathcal{F}}{(1-8\eta^2L_{\max}^2)}\\
\overset{(a)}{\leq}&\frac{1-\omega^t}{1-\omega}\frac{16\eta\mathcal{F}}{(1-8\eta^2L_{\max}^2)}\leq\frac{1}{1-\omega}\frac{16\eta\mathcal{F}}{(1-8\eta^2L_{\max}^2)}<1.14*16\eta\mathcal{F}<18.2\eta\mathcal{F}
\end{aligned}$$

where (a) is true because we have $\|\mathbf{u}^{(1)}-\mathbf{u}^{(0)}\|^2\leq 16\eta\mathcal{F}$ since $t<T$ and (61), and we used $\eta\leq c'_{\max}/L_{\max}$ where $c'_{\max}=1/10$ such that $\omega\approx 0.0435<1$.

Then, we can obtain

$$\|\mathbf{u}^{(t+1)}-\mathbf{u}^{(t)}\|\leq 4.3\sqrt{\eta\mathcal{F}}\overset{(30a)}{\leq}\frac{4.3\eta\mathcal{G}}{\kappa}. \tag{62}$$

Based on (62), we can get the upper bound of the sum of $\|\mathbf{u}^{(t+1)}-\mathbf{u}^{(t)}\|, \forall t<T$ as the following,

$$\sum_{\tau=1}^{t+1}\|\mathbf{u}^{(\tau)}-\mathbf{u}^{(\tau-1)}\|\leq\sqrt{t\sum_{\tau=1}^{t+1}\|\mathbf{u}^{(\tau)}-\mathbf{u}^{(\tau-1)}\|^2}\overset{(62)}{\leq}T\cdot\frac{4.3\eta\mathcal{G}}{\kappa}\leq\widehat{c}\mathcal{T}\frac{4.3\eta\mathcal{G}}{\kappa}\overset{(30b)}{\leq}4.3\widehat{c}\mathcal{S}, \tag{63}$$

which implies

$$\|\mathbf{u}^{(t+1)}\| \overset{(a)}{\leq} \sum_{\tau=1}^{t+1} \|\mathbf{u}^{(\tau)} - \mathbf{u}^{(\tau-1)}\| + \|\mathbf{u}^{(0)}\| \leq 4.3\widehat{c}\mathcal{S} \tag{64}$$

where in (a) we used the triangle inequality and $\mathbf{u}^{(0)} = 0$.

Due to the following fact

$$\|\mathbf{u}^{(t+1)} - \widetilde{\mathbf{x}}^{(t)}\| = \|\mathbf{u}^{(t+1)} - \mathbf{u}^{(0)} + \mathbf{u}^{(0)} - \widetilde{\mathbf{x}}^{(t)}\| \leq \|\mathbf{u}^{(t+1)} - \mathbf{u}^{(0)}\| + \|\mathbf{u}^{(0)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 4.3\widehat{c}\mathcal{S} + \mathcal{S}/(2\kappa \log(\frac{d\kappa}{\delta})), \tag{65}$$

we have $\|\mathbf{u}^{(t+1)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 5\widehat{c}\mathcal{S}$ since $\widehat{c} \geq 2$. Therefore, we know that there exits $c_{\max}^{(1)} = \min\{c_{\max}, c'_{\max}\}$ such that $\|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 5\widehat{c}\mathcal{S}, \forall t < T$ when $\eta \leq c_{\max}^{(1)}/L_{\max}$, which completes the proof. □

B.5 Proof of Lemma 9

Proof. Let $\mathbf{u}^{(0)} = 0$ and define $\mathbf{v}^{(t)} \triangleq \mathbf{w}^{(t)} - \mathbf{u}^{(t)}$. According to the assumption of Lemma 9, we know that $\mathbf{v}^{(0)} = \upsilon[\eta L_{\max}\mathcal{S}/(\kappa \log(\frac{d\kappa}{\delta})\mathcal{P}_1)]\vec{\mathbf{e}}$ when $\upsilon \in [\delta/(2\sqrt{d}), 1]$. First, we define an auxiliary function

$$h(\theta) \triangleq \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t)} + \theta\mathbf{v}_1^{(t)}, \mathbf{u}_2^{(t)} + \theta\mathbf{v}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)} + \theta\mathbf{v}_1^{(t+1)}, \mathbf{u}_2^{(t)} + \theta\mathbf{v}_2^{(t)}) \end{bmatrix},$$

then have

$$h(0) = \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t)}, \mathbf{u}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t)}) \end{bmatrix}, \quad h(1) = \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t)} + \mathbf{v}_1^{(t)}, \mathbf{u}_2^{(t)} + \mathbf{v}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)} + \mathbf{v}_1^{(t+1)}, \mathbf{u}_2^{(t)} + \mathbf{v}_2^{(t)}) \end{bmatrix},$$

$$g(\theta) = \frac{dh(\theta)}{d\theta} = \underbrace{\begin{bmatrix} \nabla_{11}^2 f(\mathbf{u}_1^{(t)} + \theta\mathbf{v}_1^{(t)}, \mathbf{u}_2^{(t)} + \theta\mathbf{v}_2^{(t)}) & \nabla_{12}^2 f(\mathbf{u}_1^{(t)} + \theta\mathbf{v}_1^{(t)}, \mathbf{u}_2^{(t)} + \theta\mathbf{v}_2^{(t)}) \\ 0 & \nabla_{22}^2 f(\mathbf{u}_1^{(t+1)} + \theta\mathbf{v}_1^{(t+1)}, \mathbf{u}_2^{(t)} + \theta\mathbf{v}_2^{(t)}) \end{bmatrix}}_{\widetilde{\boldsymbol{\mathcal{H}}}_u^{(t)}(\theta)} \mathbf{v}^{(t)}$$

$$+ \underbrace{\begin{bmatrix} 0 & 0 \\ \nabla_{21}^2 f(\mathbf{u}_1^{(t+1)} + \theta\mathbf{v}_1^{(t+1)}, \mathbf{u}_2^{(t)} + \theta\mathbf{v}_2^{(t)}) & 0 \end{bmatrix}}_{\widetilde{\boldsymbol{\mathcal{H}}}_l^{(t)}(\theta)} \mathbf{v}^{(t+1)},$$

$$\begin{bmatrix} \nabla_1 f(\mathbf{w}_1^{(t)}, \mathbf{w}_2^{(t)}) \\ \nabla_2 f(\mathbf{w}_1^{(t+1)}, \mathbf{w}_2^{(t)}) \end{bmatrix} = \int_0^1 g(\theta)d\theta + \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t)}, \mathbf{u}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t)}) \end{bmatrix}.$$

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

$$\mathbf{u}^{(t+1)} + \mathbf{v}^{(t+1)} = \mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \begin{bmatrix} \nabla_1 f(\mathbf{w}_1^{(t)}, \mathbf{w}_2^{(t)}) \\ \nabla_2 f(\mathbf{w}_1^{(t+1)}, \mathbf{w}_2^{(t)}) \end{bmatrix} \tag{66}$$

$$= \mathbf{u}^{(t)} + \mathbf{v}^{(t)} - \eta \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t)} + \mathbf{v}_1^{(t)}, \mathbf{u}_2^{(t)} + \mathbf{v}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)} + \mathbf{v}_1^{(t+1)}, \mathbf{u}_2^{(t)} + \mathbf{v}_2^{(t)}) \end{bmatrix}$$

$$= \mathbf{u}^{(t)} + \mathbf{v}^{(t)} - \eta \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t)}, \mathbf{u}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t)}) \end{bmatrix} - \int_0^1 g(\theta)d\theta \tag{67}$$

$$\overset{(a)}{=} \mathbf{u}^{(t)} + \mathbf{v}^{(t)} - \eta \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t)}, \mathbf{u}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t)}) \end{bmatrix} - \eta\widetilde{\Delta}_u^{(t)}\mathbf{v}^{(t)} - \eta\boldsymbol{\mathcal{H}}_u\mathbf{v}^{(t)} - \eta\widetilde{\Delta}_l^{(t)}\mathbf{v}^{(t+1)} - \eta\boldsymbol{\mathcal{H}}_l\mathbf{v}^{(t+1)} \tag{68}$$

where in (a) we used the following definitions:

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

$$\widetilde{\Delta}_l^{(t)} \triangleq \int_0^1 \widetilde{\boldsymbol{\mathcal{H}}}_l^{(t)}(\theta)d\theta - \boldsymbol{\mathcal{H}}_l, \tag{70}$$

and

$$\boldsymbol{\mathcal{H}}_u \triangleq \begin{bmatrix} \nabla_{11}^2 f(\widetilde{\mathbf{x}}^{(t)}) & \nabla_{12}^2 f(\widetilde{\mathbf{x}}^{(t)}) \\ 0 & \nabla_{22}^2 f(\widetilde{\mathbf{x}}^{(t)}) \end{bmatrix} \quad \boldsymbol{\mathcal{H}}_l \triangleq \begin{bmatrix} 0 & 0 \\ \nabla_{21}^2 f(\widetilde{\mathbf{x}}^{(t)}) & 0 \end{bmatrix}. \tag{71}$$

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

Dynamics of $\mathbf{v}^{(t)}$: Since the first two terms at RHS of (68) combined with $\mathbf{u}^{(t)}$ at LHS of (68) are exactly the same as (52). It can be observed that equation (68) gives the dynamic of $\mathbf{v}^{(t)}$, i.e.,

$$\mathbf{v}^{(t+1)} = \mathbf{v}^{(t)} - \eta\widetilde{\Delta}_u^{(t)}\mathbf{v}^{(t)} - \eta\boldsymbol{\mathcal{H}}_u\mathbf{v}^{(t)} - \eta\widetilde{\Delta}_l^{(t)}\mathbf{v}^{(t+1)} - \eta\boldsymbol{\mathcal{H}}_l\mathbf{v}^{(t+1)}. \tag{72}$$

Then, we can rewrite (72) in a matrix form as the following.

$$\underbrace{(\mathbf{I} + \eta\boldsymbol{\mathcal{H}}_l)}_{\triangleq\mathbf{M}}\mathbf{v}^{(t+1)} + \eta\widetilde{\Delta}_l^{(t)}\mathbf{v}^{(t+1)} \overset{(68)}{=} \underbrace{(\mathbf{I} - \eta\boldsymbol{\mathcal{H}}_u)}_{\triangleq\mathbf{T}}\mathbf{v}^{(t)} - \eta\widetilde{\Delta}_u^{(t)}\mathbf{v}^{(t)}. \tag{73}$$

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 (73), we can obtain

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

Let $\mathbb{P}_{\texttt{left}}$ 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 (74), we have

$$\mathbb{P}_{\texttt{left}}\widehat{\mathbf{v}}^{(t+1)} + \mathbb{P}_{\texttt{left}}\mathbf{M}^{-1}\eta\widetilde{\Delta}_l^{(t)}\mathbf{v}^{(t+1)} = \mathbb{P}_{\texttt{left}}(\mathbf{M}^{-1}\mathbf{T})\widehat{\mathbf{v}}^{(t)} - \mathbb{P}_{\texttt{left}}\mathbf{M}^{-1}\eta\widetilde{\Delta}_u^{(t)}\mathbf{v}^{(t)}. \tag{75}$$

From Lemma 1, we know that the maximum eigenvalue of $\mathbf{M}^{-1}\mathbf{T}$ is greater than 1.

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 $\widehat{\delta} \geq \eta\gamma/(1+L/L_{\max})$ due to Lemma 1, and $\theta^{(t)}$ denote the norm of $\mathbf{v}^{(t)}$ projected onto the remaining space. From (75), we can have

$$\phi^{(t+1)} \overset{(a)}{\geq} (1+\widehat{\delta})\phi^{(t)} - \eta\|\mathbf{M}^{-1}\|\|\widetilde{\Delta}_l^{(t)}\|\|\widehat{\mathbf{v}}^{(t+1)}\| - \eta\|\mathbf{M}^{-1}\|\|\widetilde{\Delta}_u^{(t)}\|\|\mathbf{v}^{(t)}\|, \tag{76}$$

$$\theta^{(t+1)} \leq (1+\widehat{\delta})\theta^{(t)} + \eta\|\mathbf{M}^{-1}\|\|\widetilde{\Delta}_l^{(t)}\|\|\widehat{\mathbf{v}}^{(t+1)}\| + \eta\|\mathbf{M}^{-1}\|\|\widetilde{\Delta}_u^{(t)}\|\|\mathbf{v}^{(t)}\|. \tag{77}$$

where (a) is true because we applied the triangle inequality since η is sufficiently small. Also, since $\mathbf{M}^{-1} = \mathbf{I} - \eta\boldsymbol{\mathcal{H}}_l$, we have

$$\begin{aligned}
\|\mathbf{M}^{-1}\| \leq & 1 + \eta\|\boldsymbol{\mathcal{H}}_l\| \\
\overset{(a)}{=} & 1 + \|\eta\boldsymbol{\mathcal{H}} \odot \mathbf{D} - \eta\boldsymbol{\mathcal{H}}_d\| \\
\leq & 1 + \eta\|\boldsymbol{\mathcal{H}} \odot \mathbf{D}\| + \eta\|\boldsymbol{\mathcal{H}}_d\| \\
\overset{(b)}{\leq} & 1 + \eta(1 + \frac{1}{\pi} + \frac{\log(d)}{\pi})\|\boldsymbol{\mathcal{H}}\| + \eta\|\boldsymbol{\mathcal{H}}_d\| \\
\overset{(c)}{\leq} & 1 + \eta\log(2d)\|\boldsymbol{\mathcal{H}}\| + \eta\|\boldsymbol{\mathcal{H}}_d\| \\
\overset{(d)}{\leq} & 1 + \eta L\log(2d) + \eta L_{\max} \\
\leq & 1 + \frac{L}{L_{\max}}\log(2d) + 1 < 2(1 + \frac{L\log(2d)}{L_{\max}})
\end{aligned} \tag{78}$$

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

$$\boldsymbol{\mathcal{H}}_d \triangleq \begin{bmatrix} \nabla^2_{11} f(\widetilde{\mathbf{x}}^{(t)}) & 0 \\ 0 & \nabla^2_{22} f(\widetilde{\mathbf{x}}^{(t)}) \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \cdots & 1 & 1 \end{bmatrix} \in \mathbb{R}^{d\times d}$$

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

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

we have

$$\left|\frac{Y(\mathbf{D})}{\log(d)} - \frac{1}{\pi}\right| \leq (1 + \frac{1}{\pi})\frac{1}{\log(d)}, \tag{79}$$

(c) is true for $d \geq 3$, in (d) we used the fact that $\|\boldsymbol{\mathcal{H}}\| \leq L$ and $\|\boldsymbol{\mathcal{H}}_d\| \leq L_{\max}$.

Since $\|\mathbf{w}^{(0)} - \widetilde{\mathbf{x}}^{(t)}\| \leq \|\mathbf{u}^{(0)} - \widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{v}^{(0)}\| \leq 2r$, we can apply Lemma 8. Then, we know $\|\mathbf{w}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 5\widehat{c}\mathcal{S}, \forall t < T$. According to the assumptions of Lemma 9, we have $\|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 5\widehat{c}\mathcal{S}$, and

$$\|\mathbf{v}^{(t)}\| = \|\mathbf{w}^{(t)} - \mathbf{u}^{(t)}\| \leq \|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{w}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 10\widehat{c}\mathcal{S}. \tag{80}$$

From (62), we know that

$$\|\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}\| \leq \frac{4.3\eta\mathcal{G}}{\kappa} = \frac{4.3\eta^3 L^3_{\max}\frac{\gamma}{\rho}}{\kappa^2\log^3\frac{d\kappa}{\delta}\mathcal{P}_1^3\mathcal{P}_2} \leq \mathcal{S},$$

since $\mathcal{P}_1 \geq 2$ and we choose $\eta \leq c_{\max}/L_{\max}$ and $c_{\max} = 1/10$. Similarly, we also have $\|\mathbf{u}^{(t+1)} - \mathbf{u}^{(t)}\| \leq \mathcal{S}$.

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

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

$$\begin{aligned}
\|\mathbf{v}^{(t+1)}\|^2 =&\|\mathbf{w}^{(t+1)} - \mathbf{u}^{(t+1)}\|^2 = \left\|\mathbf{w}^{(t)} - \eta \begin{bmatrix} \nabla_1 f(\mathbf{w}_1^{(t)}, \mathbf{w}_2^{(t)}) \\ \nabla_2 f(\mathbf{w}_1^{(t+1)}, \mathbf{w}_2^{(t)}) \end{bmatrix} - \left(\mathbf{u}^{(t)} - \eta \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t)}, \mathbf{u}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t)}) \end{bmatrix}\right)\right\|^2 \\
\leq& 2\|\mathbf{v}^{(t)}\|^2 + 4\eta^2 \left\| \begin{bmatrix} \nabla_1 f(\mathbf{w}_1^{(t)}, \mathbf{w}_2^{(t)}) \\ \nabla_2 f(\mathbf{w}_1^{(t+1)}, \mathbf{w}_2^{(t)}) \end{bmatrix} - \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t)}, \mathbf{w}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{w}_2^{(t)}) \end{bmatrix} \right\|^2 \\
&+ 4\eta^2 \left\| \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t)}, \mathbf{w}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{w}_2^{(t)}) \end{bmatrix} - \begin{bmatrix} \nabla_1 f(\mathbf{u}_1^{(t)}, \mathbf{u}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t)}) \end{bmatrix} \right\|^2 \\
\overset{(a)}{\leq}& 2\|\mathbf{v}^{(t)}\|^2 + 4\eta^2 L_{\max}^2(\|\mathbf{v}_1^{(t+1)}\|^2 + \|\mathbf{v}_1^{(t)}\|^2) + 8\eta^2 L_{\max}^2 \|\mathbf{v}_2^{(t)}\|^2
\end{aligned} \tag{81}$$

where (a) is true due to Lipschitz continuity.

We can express (81) as

$$(1 - 4\eta^2 L_{\max}^2)\|\mathbf{v}^{(t+1)}\| \leq (2 + 8\eta^2 L_{\max}^2)\|\mathbf{v}^{(t)}\|^2$$

which implies

$$\|\mathbf{v}^{(t+1)}\| \leq \sqrt{\frac{2 + \frac{8}{100}}{1 - \frac{4}{100}}}\|\mathbf{v}^{(t)}\| < \sqrt{2.2}\|\mathbf{v}^{(t)}\| < 1.5\|\mathbf{v}^{(t)}\|, \tag{82}$$

where we choose $\eta \leq c_{\max}/L_{\max}$ and $c_{\max} = 1/10$.

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

According to ρ-Hessian Lipschitz continuity and Lemma 6, we have the size of $\widetilde{\Delta}_u^{(t)}$ as the following.

$$\begin{aligned}
\|(\widetilde{\Delta}_u^{(t)})\| \leq& \int_0^1 \|\widetilde{\boldsymbol{\mathcal{H}}}_u^{(t)}(\theta) - \boldsymbol{\mathcal{H}}_u\| d\theta \\
\overset{(23)}{\leq}& \int_0^1 \rho \left(\|\mathbf{u}^{(t)} + \theta\mathbf{v}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| + \left\| \begin{bmatrix} \mathbf{u}_1^{(t+1)} + \theta\mathbf{v}_1^{(t+1)} \\ \mathbf{u}_2^{(t)} + \theta\mathbf{v}_2^{(t)} \end{bmatrix} - \widetilde{\mathbf{x}}^{(t)} \right\| \right) d\theta \\
\overset{(a)}{\leq}& \int_0^1 \rho \left(2\|\mathbf{u}^{(t)} + \theta\mathbf{v}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{u}^{(t+1)} + \theta\mathbf{v}^{(t+1)} - \widetilde{\mathbf{x}}^{(t)}\| \right) d\theta \\
\leq& \rho(\|\mathbf{u}^{(t+1)} - \widetilde{\mathbf{x}}^{(t)}\| + 2\|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\|) + \rho \int_0^1 \theta(\|\mathbf{v}^{(t+1)}\| + \|\mathbf{v}^{(t)}\|) d\theta \\
\leq& \rho \left(\|\mathbf{u}^{(t+1)} - \mathbf{u}^{(t)}\| + \|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| + 2\|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\|) + 0.5\|\mathbf{v}^{(t+1)}\| + 0.5\|\mathbf{v}^{(t)}\| \right) \\
\overset{(82)}{\leq}& \rho \left(\|\mathbf{u}^{(t+1)} - \mathbf{u}^{(t)}\| + 3\|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| + 1.25\|\mathbf{v}^{(t)}\| \right) \\
\leq& \rho(1 + 27.5\widehat{c})\mathcal{S}
\end{aligned} \tag{83}$$

where (a) is true because

$$\begin{aligned}
\left\| \begin{bmatrix} \mathbf{u}_1^{(t+1)} + \theta\mathbf{v}_1^{(t+1)} \\ \mathbf{u}_2^{(t)} + \theta\mathbf{v}_2^{(t)} \end{bmatrix} - \widetilde{\mathbf{x}}^{(t)} \right\| \leq& \left\| \mathbf{I}_1 \left(\mathbf{u}^{(t+1)} + \theta\mathbf{v}^{(t+1)} - \widetilde{\mathbf{x}}^{(t)} \right) \right\| + \left\| \mathbf{I}_2 \left(\mathbf{u}^{(t)} + \theta\mathbf{v}^{(t)} - \widetilde{\mathbf{x}}^{(t)} \right) \right\| \\
\overset{(26)}{\leq}& \|\mathbf{u}^{(t+1)} + \theta\mathbf{v}^{(t+1)} - \widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{u}^{(t)} + \theta\mathbf{v}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\|.
\end{aligned} \tag{84}$$

Applying Lemma 6, we can also get the upper bound of $\|\widetilde{\Delta}_l^{(t)}\|$, i.e.,

$$\begin{aligned}\|(\widetilde{\Delta}_l^{(t)})\| \leq & \int_0^1 \|\widetilde{\boldsymbol{\mathcal{H}}}_l^{(t)}(\theta) - \boldsymbol{\mathcal{H}}_l\| d\theta \\ \overset{(24)}{\leq} & \int_0^1 \rho \left\| \begin{bmatrix} \mathbf{u}_1^{(t+1)} + \theta \mathbf{v}_1^{(t+1)} \\ \mathbf{u}_2^{(t)} + \theta \mathbf{v}_2^{(t)} \end{bmatrix} - \widetilde{\mathbf{x}}^{(t)} \right\| d\theta \\ \leq & \int_0^1 \rho(\|\mathbf{u}^{(t)} + \theta \mathbf{v}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{u}^{(t+1)} + \theta \mathbf{v}^{(t+1)} - \widetilde{\mathbf{x}}^{(t)})\| d\theta \\ \leq & \rho(\|\mathbf{u}^{(t+1)} - \widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\|) + \rho \int_0^1 \theta(\|\mathbf{v}^{(t+1)}\| + \|\mathbf{v}^{(t)}\|) d\theta \\ \overset{(82)}{\leq} & \rho \left(\|\mathbf{u}^{(t+1)} - \mathbf{u}^{(t)}\| + 2\|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| + 1.25\|\mathbf{v}^{(t)}\| \right) \\ \leq & \rho(1 + 22.5\widehat{c})\mathcal{S}. \end{aligned} \tag{85}$$

With the upper bounds of $\|\mathbf{v}^{(t+1)}\|$, $\|\widetilde{\Delta}_u^{(t)}\|$, $\|\widetilde{\Delta}_l^{(t)}\|$ and relation between $\|\mathbf{v}^{(t+1)}\|$ and $\|\mathbf{v}^{(t)}\|$, we can further simply (76) and (77) as follows,

$$\begin{aligned} \phi^{(t+1)} & \overset{(76)}{\geq} (1 + \widehat{\delta})\phi^{(t)} - \eta(1.5\|\widetilde{\Delta}_l^{(t)}\| + \|\widetilde{\Delta}_u^{(t)}\|)\|\mathbf{M}^{-1}\|\|\mathbf{v}^{(t)}\| \\ \theta^{(t+1)} & \overset{(77)}{\leq} (1 + \widehat{\delta})\theta^{(t)} + \eta(1.5\|\widetilde{\Delta}_l^{(t)}\| + \|\widetilde{\Delta}_u^{(t)}\|)\|\mathbf{M}^{-1}\|\|\mathbf{v}^{(t)}\| \end{aligned}$$

and further we have

$$\begin{aligned} \phi^{(t+1)} & \geq (1 + \widehat{\delta})\phi^{(t)} - \eta(1.5\|\widetilde{\Delta}_l^{(t)}\| + \|\widetilde{\Delta}_u^{(t)}\|)\|\mathbf{M}^{-1}\|\sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2}, \\ \theta^{(t+1)} & \leq (1 + \widehat{\delta})\theta^{(t)} + \eta(1.5\|\widetilde{\Delta}_l^{(t)}\| + \|\widetilde{\Delta}_u^{(t)}\|)\|\mathbf{M}^{-1}\|\sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2}, \end{aligned}$$

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

Consequently, we can arrive at

$$\phi^{(t+1)} \geq (1 + \widehat{\delta})\phi^{(t)} - \mu\sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2}, \tag{86}$$

$$\theta^{(t+1)} \leq (1 + \widehat{\delta})\theta^{(t)} + \mu\sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2}, \tag{87}$$

where μ is the upper bound of $\eta(1.5\|\widetilde{\Delta}_l^{(t)}\| + \|\widetilde{\Delta}_u^{(t)}\|)\|\mathbf{M}^{-1}\|$ and can be obtained by

$$\mu \triangleq \eta\rho\mathcal{S}\mathcal{P}_2(2.5 + 62\widehat{c}). \tag{88}$$

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

$$\theta^{(t)} \leq 4\mu t\phi^{(t)}. \tag{89}$$

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

Assuming that equation (89) is true at the tth iteration, we need to prove

$$\theta^{(t+1)} \leq 4\mu(t + 1)\phi^{(t+1)}. \tag{90}$$

Applying (86) into RHS of (90), we have

$$4\mu(t + 1)\phi^{(t+1)} \geq 4\mu(t + 1)\left((1 + \widehat{\delta})\phi^{(t)} - \mu\sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2} \right) \tag{91}$$

and substituting (87) into LHS of (90), we have

$$\theta^{(t+1)} \leq (1+\widehat{\delta})(4\mu t\phi^{(t)}) + \mu\sqrt{(\phi^{(t)})^2+(\theta^{(t)})^2}. \tag{92}$$

Then, our goal is to prove RHS of (91) is greater than RHS of (92). After some manipulations, it is sufficient to show

$$(1+4\mu(t+1))\left(\sqrt{(\phi^{(t)})^2+(\theta^{(t)})^2}\right) \leq 4\phi^{(t)}. \tag{93}$$

In the following, we will show that the above relation is true.

First step : We know that

$$4\mu(t+1) \leq 4\mu T \overset{(88)}{\leq} 4\eta\rho\mathcal{S}\mathcal{P}_2(2.5+62\widehat{c})\widehat{c}\mathcal{T} \overset{(30d),(88)}{\leq} \frac{4\widehat{c}\eta^2 L_{\max}^2(2.5+62\widehat{c})}{\kappa\log(\frac{d\kappa}{\delta})\mathcal{P}_1} \overset{(a)}{\leq} 1 \tag{94}$$

where (a) is true because $\mathcal{P}_1 \geq 2$ and we choose $c'_{\max} = 1/(2\widehat{c}(2.5+62\widehat{c}))$ and $\eta \leq c'_{\max}/L_{\max}$.

Second step : Also, we know that

$$4\phi^{(t)} \geq 2\sqrt{2(\phi^{(t)})^2} \overset{(89),(94)}{\geq} (1+4\mu(t+1))\sqrt{(\phi^{(t)})^2+(\theta^{(t)})^2}. \tag{95}$$

With the above two steps, we have $\theta^{(t+1)} \leq 4\mu(t+1)\phi^{(t+1)}$, which completes the induction.

Recursion of $\phi^{(t)}$:Using (89), we have $\theta^{(t)} \overset{(89)}{\leq} 4\mu t\phi^{(t)} \overset{(94)}{\leq} \phi^{(t)}$, which implies

$$\begin{aligned}
\phi^{(t+1)} &\overset{(86)}{\geq} (1+\widehat{\delta})\phi^{(t)} - \mu\sqrt{(\phi^{(t)})^2+(\theta^{(t)})^2} \\
&\overset{(a)}{\geq} (1+\frac{\gamma\eta}{1+L/L_{\max}})\phi^{(t)} - \mu\sqrt{(\phi^{(t)})^2+(\theta^{(t)})^2} \\
&\overset{(b)}{\geq} (1+\frac{1}{1+L/L_{\max}}\frac{\gamma\eta}{2})\phi^{(t)}
\end{aligned} \tag{96}$$

where in (a) we used Lemma 1, and (b) is true because

$$\begin{aligned}
\mu =& \eta\rho\mathcal{S}\mathcal{P}_2(2.5+62\widehat{c}) \\
\leq& \frac{\gamma\eta}{1+L/L_{\max}}\frac{\eta^2 L_{\max}^2(2.5+62\widehat{c})}{\log^2(\frac{d\kappa}{\delta})\mathcal{P}_1} \\
\overset{(a)}{\leq}& \frac{1}{1+L/L_{\max}}\frac{\gamma\eta}{2\sqrt{2}}
\end{aligned}$$

where in (a) we choose $c''_{\max} = 1/(2\sqrt{2}(2.5+62\widehat{c}))$ and $\eta \leq c''_{\max}/L_{\max}$.

Quantifying Escaping Time: From (80), we have

$$\begin{aligned}
10\mathcal{S}\widehat{c} \geq& \|\mathbf{v}^{(t)}\| \geq \phi^{(t)} \\
\overset{(96)}{\geq}& (1+\frac{\gamma\eta}{2(1+L/L_{\max})})^t\phi^{(0)} \\
\overset{(a)}{\geq}& (1+\frac{\gamma\eta}{2(1+L/L_{\max})})^t\frac{\delta}{2\sqrt{d}}\frac{\eta L_{\max}\mathcal{S}}{\kappa}\log^{-1}(\frac{d\kappa}{\delta})\mathcal{P}_1^{-1} \\
\overset{(b)}{\geq}& (1+\frac{\gamma\eta}{2(1+L/L_{\max})})^t\frac{\delta}{2\sqrt{d}}\frac{c\mathcal{S}}{\kappa}\log^{-1}(\frac{d\kappa}{\delta})\mathcal{P}_1^{-1} \quad \forall t < T
\end{aligned} \tag{97}$$

where in (a) we use condition $\upsilon \in [\delta/(2\sqrt{d}), 1]$, in (b) we used $\eta = c/L_{\max}$.

Since (97) is true for all $t < T$, we can have

$$\begin{aligned} T-1 \leq & \frac{\log(20\frac{\widehat{c}}{c}(\frac{\kappa\sqrt{d}}{\delta})\log(\frac{d\kappa}{\delta})\mathcal{P}_1)}{\log(1+\frac{\eta\gamma}{2(1+L/L_{\max})})} \\ \overset{(a)}{<} & \frac{4(1+L/L_{\max})\log(20(\frac{\sqrt{d}\kappa}{\delta})\frac{\widehat{c}}{c}\log(\frac{d\kappa}{\delta})\mathcal{P}_1)}{\eta\gamma} \\ \overset{(b)}{<} & \frac{4(1+L/L_{\max})\log(20(\frac{d\kappa}{\delta})^2\frac{\widehat{c}}{c}\mathcal{P}_1)}{\eta\gamma} \\ \overset{(c)}{<} & 4(2+\log(20\frac{\widehat{c}}{c}))\mathcal{T} \end{aligned} \tag{98}$$

where (a) comes from inequality $\log(1+x) > x/2$ when $x < 1$, in (b) we used relation $\log(x) < x, x > 0$, and (c) is true because $\delta \in (0, \frac{d\kappa}{e}]$ and $\log(d\kappa/\delta) > 1$ and $\mathcal{P}_1 > 1$ we have

$$\log(\frac{d\kappa}{\delta}\mathcal{P}_1) \leq \log(\frac{d\kappa}{\delta}) + \log(1+\frac{L}{L_{\max}}) \leq \log(\frac{d\kappa}{\delta}) + \frac{L}{L_{\max}} \leq \log(\frac{d\kappa}{\delta})\mathcal{P}_1.$$

From (98), we know that

$$T < 4(2+\log(20\frac{\widehat{c}}{c}))\mathcal{T} + 1 \overset{(a)}{<} 4(2\frac{1}{4} + \log(20\frac{\widehat{c}}{c})\mathcal{T} \tag{99}$$

where (a) is true due to the fact that $\eta L_{\max} \geq 1$, $\log(d\kappa/\delta) > 1$ and $\mathcal{P}_1 > 1$ so we know $\mathcal{T} \geq 1$.

When

$$4(2.25 + \log(20\frac{\widehat{c}}{c})) \leq \widehat{c}, \tag{100}$$

we will have $T < \widehat{c}\mathcal{T}$ where $c_{\max}^{(2)} \triangleq \min\{c_{\max}, c'_{\max}, c''_{\max}\}$.

Since $\widehat{c} \geq 2$, we have $c_{\max} = \min\{c_{\max}^{(1)}, c_{\max}^{(2)}\} \leq 1/(5\widehat{c})^3$. Also, we know that $c \leq c_{\max}$. Combining with (100), we need

$$\frac{\widehat{c}}{2^{\frac{\widehat{c}}{4}-2.25-\log(20)}} \leq c \leq \frac{1}{(5\widehat{c})^3}, \tag{101}$$

meaning that

$$125(2^{2.25+\log(20)}\widehat{c}^4) \leq 2^{\frac{\widehat{c}}{4}}. \tag{102}$$

It can be observed that LHS of (102) is a polynomial with respect to $\widehat{c}$ and RHS of (102) is a exponential function in terms of $\widehat{c}$, implying there exists a universal $\widehat{c}$ such that (102) holds. The proof is complete. □

B.6 Proof of Lemma 10

Proof. The proof of Lemma 10 is similar as the one of proving convergence of PGD shown in [Jin et al., 2017a, Lemma 14,15]. Considering the completeness of the whole proof in this paper, here we give the following proof of this lemma in details.

First, after the random perturbation, the objective function value in the worst case is increased at most

by

$$
\begin{aligned}
f(\mathbf{u}^{(0)}) - f(\widetilde{\mathbf{x}}^{(t)}) \leq & \sum_{k=1}^{2} \nabla_k f(\widetilde{\mathbf{h}}_{-k}^{(t)}, \widetilde{\mathbf{x}}_k^{(t)})^\mathsf{T} \xi_k + \frac{L_k}{2}\|\xi_k\|^2 \\
\leq & \sum_{k=1}^{2} \|\nabla_k f(\widetilde{\mathbf{h}}_{-k}^{(t)}, \widetilde{\mathbf{x}}_k^{(t)})\|\|\xi_k\| + \frac{L_{\max}}{2}\|\xi\|^2 \\
\overset{(a)}{\leq} & \|\xi\| \sqrt{\sum_{k=1}^{2} 2\|\nabla_k f(\widetilde{\mathbf{h}}_{-k}^{(t)}, \widetilde{\mathbf{x}}_k^{(t)})\|^2} + \frac{L_{\max}}{2}\|\xi\|^2 \\
\overset{(b)}{\leq} & \frac{\mathcal{G}}{\kappa} \frac{\eta L_{\max} \mathcal{S}}{\kappa \log(\frac{d\kappa}{\delta})\mathcal{P}_1} + \frac{L_{\max}}{2}(\frac{\eta L_{\max} \mathcal{S}}{\kappa \log(\frac{d\kappa}{\delta})\mathcal{P}_1})^2 \leq \frac{3}{2}\mathcal{F}
\end{aligned} \tag{103}
$$

where $\mathbf{u}^{(0)}$ is a vector that follows uniform distribution within the ball $\mathbb{B}^{(d)}_{\widetilde{\mathbf{x}}^{(t)}}(r)$, $\mathbb{B}^{(d)}_{\widetilde{\mathbf{x}}^{(t)}}$ denotes the d-dimensional ball centered at $\widetilde{\mathbf{x}}^{(t)}$ with radius r, ξ_k represents the kth block of the vector which is the difference between random generated vector $\mathbf{u}^{(0)}$ and $\widetilde{\mathbf{x}}^{(t)}$, and (a) is true because $\xi \triangleq [\xi_1, \ldots, \xi_K]$, $\|\xi_k\| \leq \|\xi\|, \forall k$, and in (b) we used $\kappa > 1$, $\log(d\kappa/\delta) > 1$ and Condition 1.

Second, under Assumption 1, let $\widetilde{\mathbf{x}}^{(t)}$ satisfy conditions Condition 1, and two PA-GD iterates $\{\mathbf{u}^{(t)}\}$ $\{\mathbf{w}^{(t)}\}$ satisfy the conditions as in Lemma 9. Selecting $c_{\max} = \min\{c^{(1)}_{\max}, c^{(2)}_{\max}\}$, so we have that $\eta \leq c_{\max}/L_{\max}$ is small enough such that Lemma 8 and Lemma 9 can both hold.

Let $T^* \triangleq \widehat{c}\mathcal{T}$ and $T' \triangleq \inf_t\{t | \widehat{f}_{\mathbf{u}^{(0)}}(\mathbf{u}^{(t)}) - f(\mathbf{u}^{(0)}) \leq -3\mathcal{F}\}$. Then, we have the following two cases to analyze the decrease of the objective value after T iterations with the random perturbation.

1. Case $T' \leq T^*$:

$$
\begin{aligned}
f(\mathbf{u}^{(T')}) - f(\mathbf{u}^{(0)}) \leq & \nabla f(\mathbf{u}^{(0)})^\mathsf{T}(\mathbf{u}^{(T')} - \mathbf{u}^{(0)}) + \frac{1}{2}(\mathbf{u}^{(T')} - \mathbf{u}^{(0)})^\mathsf{T}\nabla^2 f(\mathbf{u}^{(0)})(\mathbf{u}^{(T')} - \mathbf{u}^{(0)}) + \frac{\rho}{6}\|\mathbf{u}^{(T')} - \mathbf{u}^{(0)}\|^3 \\
\leq & \widehat{f}_{\mathbf{u}^{(0)}}(\mathbf{u}^{(t)}) - f(\mathbf{u}^{(0)}) + \frac{\rho}{2}\|\mathbf{u}^{(0)} - \widetilde{\mathbf{x}}^{(t)}\|\|\mathbf{u}^{(T')} - \mathbf{u}^{(0)}\|^2 + \frac{\rho}{6}\|\mathbf{u}^{(T')} - \mathbf{u}^{(0)}\|^3 \\
\overset{(54)-(55)}{\leq} & -3\mathcal{F} + 0.5\rho\mathcal{S}^3 \overset{(30c)}{\leq} -2.5\mathcal{F}.
\end{aligned} \tag{104}
$$

   Based on Lemma 2, we know that AGD is always decreasing the objective function. For any $T \geq \mathcal{T}/c_{\max} \geq \widehat{c}\mathcal{T} = T^* \geq T'$, we have

$$
f(\mathbf{u}^{(T)}) - f(\mathbf{u}^{(0)}) \leq f(\mathbf{u}^{(T^*)}) - f(\mathbf{u}^{(0)}) \leq f(\mathbf{u}^{(T')}) - f(\mathbf{u}^{(0)}) \leq -2.5\mathcal{F}
$$

   where $c_{\max} = \min\{1, 1/\widehat{c}\}$.

2. Case $T' > T^*$: Applying Lemma 8, we know that $\|\mathbf{u}^{(t)} - \mathbf{u}^{(0)}\| \leq 5\widehat{c}\mathcal{S}$ for $t \leq T^*$. Define $T'' = \inf_t\{t | \widehat{f}_{\mathbf{w}^{(0)}}(\mathbf{w}^{(t)}) - f(\mathbf{w}^{(0)}) \leq -3\mathcal{F}\}$. Then, after applying Lemma 9, we know $T'' \leq T^*$. Similar as (104), for $T \geq 1/c_{\max}\mathcal{T}$, we also have $f(\mathbf{w}^{(T)}) - f(\mathbf{w}^{(0)}) \leq f(\mathbf{w}^{T^*}) - f(\mathbf{w}^{(0)}) \leq f(\mathbf{w}^{T''}) - f(\mathbf{w}^{(0)}) \leq -2.5\mathcal{F}$.

Combining the above two cases, we have

$$
\min\{f(\mathbf{u}^{(T)}) - f(\mathbf{u}^{(0)}), f(\mathbf{w}^{(T)}) - f(\mathbf{w}^{(0)})\} \leq -2.5\mathcal{F}, \tag{105}
$$

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 $\mathbf{u}^{(0)} \in \mathcal{X}_{\mathsf{stuck}}$, then $(\mathbf{u}^{(0)} \pm \upsilon r\vec{\mathbf{e}}) \notin \mathcal{X}_{\mathsf{stuck}}$ where $\upsilon \in [\frac{\delta}{2\sqrt{d}}, 1]$.

Finally, we give the upper bound of the volume of $\mathcal{X}_{\mathsf{stuck}}$,

$$\begin{aligned}\mathsf{Vol}(\mathcal{X}_{\mathsf{stuck}}) &= \int_{\mathbb{B}^{(d)}_{\widetilde{\mathbf{x}}^{(t)}}} d\mathbf{u} I_{\mathcal{X}_{\mathsf{stuck}}}(\mathbf{u}) = \int_{\mathbb{B}^{(d-1)}_{\widetilde{\mathbf{x}}^{(t)}}} du_{-1} \int_{\widetilde{x}_1^{(t)}-\sqrt{r^2-\|\widetilde{\mathbf{x}}_{-1}^{(t)}-u_{-1}\|^2}}^{\widetilde{x}_1^{(t)}+\sqrt{r^2-\|\widetilde{\mathbf{x}}_{-1}^{(t)}-u_{-1}\|^2}} du_1 I_{\mathcal{X}_{\mathsf{stuck}}}(\mathbf{u}) \\ &\leq \int_{\mathbb{B}^{(d-1)}_{\widetilde{\mathbf{x}}^{(t)}}} du_{-1} \left(2\frac{\delta}{2\sqrt{d}r}\right) = \mathsf{Vol}(\mathbb{B}^{(d-1)}_{\widetilde{\mathbf{x}}^{(t)}}(r))\frac{r\delta}{\sqrt{d}}\end{aligned}$$

where $I_{\mathsf{stuck}}(\mathbf{u})$ is an indicator function showing that $\mathbf{u}$ belongs to set $\mathcal{X}_{\mathsf{stuck}}$, and 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

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

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 $x \geq 0$.

Combining (103) and (105), we can show that

$$f(\mathbf{x}^{(T)}) - f(\widetilde{\mathbf{x}}^{(t)}) = f(\mathbf{x}^{(T)}) - f(\mathbf{u}^{(0)}) + f(\mathbf{u}^{(0)}) - f(\widetilde{\mathbf{x}}^{(t)}) \leq -2.5\mathcal{F} + 1.5\mathcal{F} \leq -\mathcal{F} \tag{106}$$

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

C Proof of PA-PP

First, we need to introduce some constants defined as follows,

$$\mathcal{F} \triangleq \eta^5 L_{\max}^5 \frac{\gamma^3}{\kappa^3\rho^2} \log^{-6}\left(\frac{d\kappa}{\delta}\right)\mathcal{P}^{-2}, \qquad \mathcal{G} \triangleq \eta^2 L_{\max}^2 \frac{\gamma^2}{\rho} \log^{-3}\left(\frac{d\kappa}{\delta}\right)\mathcal{P}^{-1},$$
$$\mathcal{S} \triangleq \eta^2 L_{\max}^2 \frac{\gamma}{\kappa\rho} \log^{-2}\left(\frac{d\kappa}{\delta}\right)\mathcal{P}^{-1}, \qquad \mathcal{T} \triangleq \frac{\log\left(\frac{d\kappa}{\delta}\right)}{\eta\gamma}$$

where $\eta = 1/\nu$. In order to keep the completeness of the proof, the certain relations of these quantities are listed as follows, which are useful of simplifying the expressions in the proofs.

$$\sqrt{\mathcal{F}} = \frac{\sqrt{\eta}\mathcal{G}}{\kappa}, \tag{107a}$$
$$\frac{\eta\mathcal{G}\mathcal{T}}{\kappa} = \mathcal{S}, \tag{107b}$$
$$\rho\mathcal{S}^3 = \frac{\eta L_{\max}\mathcal{F}}{\mathcal{P}}, \tag{107c}$$
$$\eta\rho\mathcal{S}\mathcal{T} = \frac{\eta^2 L_{\max}^2}{\kappa\log(\frac{d\kappa}{\delta})\mathcal{P}}, \tag{107d}$$
$$\eta\rho\mathcal{S} = \eta L_{\max}\frac{\eta^2\gamma^2}{\log^2(\frac{d\kappa}{\delta})\mathcal{P}}. \tag{107e}$$

We also consider saddle point $\widetilde{\mathbf{x}}^{(t)}$ that satisfies the following condition.

Condition 2. *An ϵ-second order stationary point $\widetilde{\mathbf{x}}^{(t)}$ satisfies the following conditions:*

$$\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\| \leq g_{th}/\nu \quad \textit{and} \quad \lambda_{\min}(\nabla^2 f(\widetilde{\mathbf{x}}^{(t)})) \leq -\gamma \tag{108}$$

where $g_{th} = \frac{\mathcal{G}}{2\kappa}$.

Then, we have the following preliminary lemmas.

Lemma 11. *If function $f(\cdot)$ is L-smooth with Lipschitz constant, the we have*

$$\|\nabla f(\mathbf{x}^{(t)})\|^2 \leq 4\nu\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\|^2 \tag{109}$$

where sequence $\mathbf{x}_k^{(t)}, k = 1, 2$ is generated by the APP algorithm.

Lemma 12. *Under Assumption 1, we have block-wise Lipschitz continuity as the follows:*

$$\left\|\begin{bmatrix} \nabla_{11}^2 f(\mathbf{x}) & \mathbf{0} \\ \nabla_{21}^2 f(\mathbf{y}) & \nabla_{22}^2 f(\mathbf{y}) \end{bmatrix} - \begin{bmatrix} \nabla_{11}^2 f(\mathbf{z}) & \mathbf{0} \\ \nabla_{21}^2 f(\mathbf{z}) & \nabla_{22}^2 f(\mathbf{z}) \end{bmatrix}\right\| \leq \rho\left(\|\mathbf{x}-\mathbf{z}\| + \|\mathbf{y}-\mathbf{z}\|\right), \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \tag{110}$$

and

$$\left\|\begin{bmatrix} \mathbf{0} & \nabla_{21}^2 f(\mathbf{x}) \\ \mathbf{0} & \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{0} & \nabla_{12}^2 f(\mathbf{y}) \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right\| \leq \rho\|\mathbf{x}-\mathbf{y}\|, \forall \mathbf{x}, \mathbf{y}. \tag{111}$$

Second, we can have the descent lemma as the following

Lemma 13. *Under Assumption 1, for the APP algorithm with penalizer $\nu \geq 3L_{\max}$, we have*

$$f(\mathbf{x}^{(t+1)}) \leq f(\mathbf{x}^{(t)}) - \frac{\nu}{2}\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\|^2.$$

Third, we need to characterize the convergence behaviour of PA-PP when $\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\|$ is small. In this case, we need three steps to arrive the final results.

Step 1 : Quantify upper bound of the distance between generic iterate $\mathbf{u}^{(t)}$ and saddle point $\widetilde{\mathbf{x}}^{(t)}$.

Lemma 14. *Under Assumption 1, consider saddle point $\widetilde{\mathbf{x}}^{(t)}$ that satisfies Condition 2. For any constant $\widehat{c} \geq 2$, $\delta \in (0, \frac{d\kappa}{e}]$, when initial point $\mathbf{u}^{(0)}$ satisfies*

$$\|\mathbf{u}^{(0)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 2r, \tag{112}$$

then, with the definition of

$$r \triangleq \frac{\frac{L_{\max}}{\nu}\mathcal{S}}{\kappa \log(\frac{d\kappa}{\delta})\mathcal{P}_1} \quad \text{and} \quad T \triangleq \min\{\inf_t\{t|\widehat{f}_{\mathbf{u}^{(0)}}(\mathbf{u}^{(t)}) - f(\mathbf{u}^{(0)}) \leq -3\mathcal{F}\}, \widehat{c}\mathcal{T}\}, \tag{113}$$

there exits constants $c_{\max}^{(1)}, \widehat{c}$ such that for any $\nu \geq L_{\max}/c_{\max}^{(1)}$, the iterates generated by PA-PP satisfy $\|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 5\widehat{c}\mathcal{S}, \forall t < T$.

Step 2 : Quantify the escaping time of iterates near a strict saddle point.

Lemma 15. *Under Assumption 1, consider saddle point $\widetilde{\mathbf{x}}^{(t)}$ that satisfies satisfies Condition 2. There exist constants $c_{\max}^{(2)}$, $\widehat{c}$ such that: for any $\delta \in (0, \frac{d\kappa}{e}]$ and $\nu \geq L_{\max}/c_{\max}^{(2)}$, with the definition of*

$$T \triangleq \min\left\{\inf_t\{t|\widehat{f}_{\mathbf{w}_0}(\mathbf{w}^{(t)}) - f(\mathbf{w}^{(0)}) \leq -3\mathcal{F}\}, \widehat{c}\mathcal{T}\right\} \tag{114}$$

where two iterates $\{\mathbf{u}^{(t)}\}$ and $\{\mathbf{w}^{(t)}\}$ that are generated by PA-PP with initial points $\{\mathbf{u}^{(0)}, \mathbf{w}^{(0)}\}$ satisfying

$$\|\mathbf{u}^{(0)} - \widetilde{\mathbf{x}}^{(t)}\| \leq r, \quad \mathbf{w}^{(0)} = \mathbf{u}^{(0)} + \upsilon r\vec{\mathbf{e}}', \quad \upsilon \in [\delta/(2\sqrt{d}), 1], \tag{115}$$

where $\vec{\mathbf{e}}'$ denotes the eigenvector of $\mathbf{T}'^{-1}\mathbf{M}'$ whose corresponding positive eigenvalue is minimum, if $\|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 5\widehat{c}\mathcal{S}, \forall t < T$, we will have $T < \widehat{c}\mathcal{T}$.

Step 3 : Quantify sufficient decrease with random perturbation. With Lemma 14 and Lemma 15, we can apply Lemma 10 directly and obtain the following lemma.

Lemma 16. *Under Assumption 1, there exists a universal constant $c_{\max}$, for any $\delta \in (0, d\kappa/e]$: consider a saddle point $\widetilde{\mathbf{x}}^{(t)}$ which satisfies (3), let $\mathbf{x}^{(0)} = \widetilde{\mathbf{x}}^{(t)} + \xi$ where ξ is generated randomly which follows the uniform distribution over a ball with radius r, and let $\mathbf{x}^{(t)}$ be the iterates of PA-PP starting from $\mathbf{x}^{(0)}$. Then, when step size $\nu \geq L_{\max}/c_{\max}$, with at least probability $1 - \delta$, we have the following for any $T \geq \mathcal{T}/c_{\max}$*

$$f(\mathbf{x}^{(T)}) - f(\widetilde{\mathbf{x}}^{(t)}) \leq -\mathcal{F}. \tag{116}$$

Substituting $\nu = \frac{L_{\max}}{c}$,$\gamma = (L_{\max}\rho\epsilon)^{1/3}$, and $\delta = \frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{1/3}}e^{-\chi}$ in to Lemma 16, we can obtain the following lemma immediately.

Lemma 17. *Under Assumption 1, there exists a absolute constant $c_{\max}$. Let $c \leq c_{\max}$, $\chi \geq 1$, and η, r, g_{th}, t_{th} calculated as Algorithm 2 describes. Let $\widetilde{\mathbf{x}}^{(t)}$ be a strict saddle point, which satisfies*

$$\|\nabla f(\widetilde{\mathbf{x}}^{(t)})\|^2 \leq 4\nu\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\|^2 \leq 4g_{th}^2 \tag{117}$$

and

$$\lambda_{\min}(\nabla^2 f(\widetilde{\mathbf{x}}^{(t)})) \leq -\gamma.$$

Let $\mathbf{x}^{(t)} = \widetilde{\mathbf{x}}^{(t)} + \xi^{(t)}$ where $\xi^{(t)}$ is generated randomly which follows the uniform distribution over $\mathbb{B}_{\widetilde{\mathbf{x}}^{(t)}}(r)$, and let $\mathbf{x}^{(t+t_{th})}$ be the iterates of PA-PP. With at least probability $1 - \frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{1/3}}e^{-\chi}$, we have

$$f(\mathbf{x}^{(t+t_{th})}) - f(\widetilde{\mathbf{x}}^{(t)}) \leq -f_{th}. \tag{118}$$

Finally, we can get the convergence rate of PA-PP as the following.

C.1 Proof of Corollary 1

Next, we prove the main theorem.

Proof. Submitting $\nu = \frac{L_{\max}}{c}$,$\gamma = (L_{\max}\rho\epsilon)^{1/3}$, and $\delta = \frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{1/3}}e^{-\chi}$ into the definition of $\mathcal{F}, \mathcal{G}, \mathcal{T}$, we will have the following definitions.

$$\begin{aligned} f_{\mathsf{th}} &\triangleq \mathcal{F} = \frac{c^5\epsilon^2}{L_{\max}\chi^6\mathcal{P}^2}, \\ g_{\mathsf{th}} &\triangleq \frac{\mathcal{G}}{2\kappa} = \frac{c^2\epsilon}{2\chi\mathcal{P}}, \\ t_{\mathsf{th}} &\triangleq \frac{\mathcal{T}}{c} = \frac{L_{\max}\chi}{c^2(L_{\max}\rho\epsilon)^{\frac{1}{3}}}. \end{aligned} \tag{119}$$

After applying Lemma 7, we know that

$$\|\nabla f(\mathbf{x})\| \leq \frac{c}{\chi^3\mathcal{P}}\epsilon \tag{120}$$

where $c \leq 1, \chi, \mathcal{P} \geq 1$.

Similarly, 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 $\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| > g_{\mathsf{th}}/\nu$: According to Lemma 13, we have

$$\begin{aligned} f(\mathbf{x}^{(1)}) - f(\mathbf{x}^{(0)}) &\leq -\frac{\nu}{2}\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|^2 \leq -\frac{\nu}{2}g_{\mathsf{th}}^2 \\ &\overset{(a)}{=} -\frac{c^5}{8\chi^6\mathcal{P}^2}\frac{\epsilon^2}{L_{\max}} \end{aligned} \tag{121}$$

   where in (a) use the definition of g_{th}^2 and $\nu \geq L_{\max}/c$.

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

$$f(\mathbf{x}^{(t_{\mathsf{th}})}) - f(\mathbf{x}^{(0)}) \leq -f_{\mathsf{th}} = -\frac{c^5\epsilon^2}{L_{\max}\chi^6\mathcal{P}^2}, \tag{122}$$

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

$$\frac{f(\mathbf{x}^{(t_{\mathsf{th}})}) - f(\mathbf{x}^{(0)})}{t_{\mathsf{th}}} \leq -\frac{c^7}{\chi^7\mathcal{P}^2}\frac{\epsilon^2}{L_{\max}}\frac{(L_{\max}\rho\epsilon)^{\frac{1}{3}}}{L_{\max}}. \tag{123}$$

   Since $\kappa = L_{\max}/(L_{\max}\rho\epsilon)^{1/3} \geq 1$ and $c \leq 1/3$, we know that RHS of (123) is greater than RHS of (121).

   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-PP can have a sufficient decrease of the objective function value by $\frac{c^7}{\chi^7\mathcal{P}^2}\frac{\epsilon^2}{L_{\max}}\frac{(L_{\max}\rho\epsilon)^{1/3}}{L_{\max}}$ per iteration on average. This means that Algorithm 1 must stop within a finite number of iterations, which is

$$\frac{f(\mathbf{h}_{-1}^{(0)}, \mathbf{x}_1^{(0)}) - f^*}{\frac{c^7}{\chi^7\mathcal{P}^2}\frac{\epsilon^2}{L_{\max}}\frac{(L_{\max}\rho\epsilon)^{1/3}}{L_{\max}}} = \frac{\chi^7\mathcal{P}^2}{c^7}\frac{L_{\max}^2\Delta f}{\epsilon^2(L_{\max}\rho\epsilon)^{1/3}} = \mathcal{O}\left(\frac{\Delta f\chi^7\mathcal{P}^2L_{\max}^{5/3}}{\rho^{1/3}\epsilon^{7/3}}\right) \tag{124}$$

   where $\Delta f \triangleq f(\mathbf{h}_{-1}^{(0)}, \mathbf{x}_1^{(0)}) - f^*$.

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

$$n' = \frac{1}{t_{\mathsf{th}}}\frac{\chi^7\mathcal{P}^2}{c^7}\frac{L_{\max}^2\Delta f}{\epsilon^2(L_{\max}\rho\epsilon)^{1/3}} = \frac{\chi^6\mathcal{P}^2}{c^5}\frac{L_{\max}\Delta_f}{\epsilon^2}. \tag{125}$$

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

$$1-n'\frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{\frac{1}{3}}}e^{-\chi} = 1-\frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{\frac{1}{3}}}e^{-\chi}\frac{\chi^6\mathcal{P}^2}{c^5}\frac{L_{\max}\Delta_f}{\epsilon^2} = 1-\underbrace{\frac{dL_{\max}}{(L_{\max}\rho\epsilon)^{\frac{1}{3}}}\frac{\mathcal{P}^2}{c^5}\frac{\Delta_f}{\epsilon^2}}_{\triangleq\mathcal{C}'}\chi^6e^{-\chi}. \tag{126}$$

With chosen $\chi = 6\max\{\ln(\mathcal{C}'/\delta),4\}$, we have $\chi^6e^{-\chi}\leq e^{-\chi/6}$, which implies $\chi^6e^{-\chi}\mathcal{C}'\leq e^{-\chi/6}\mathcal{C}'\leq\delta$. The proof is complete. □

C.2 Proof of Corollary 2

Proof. Recall the definitions:

$$\boldsymbol{\mathcal{H}}'_u = \begin{bmatrix} 0 & \nabla^2_{12}f(\widetilde{\mathbf{x}}^{(t)}) \\ 0 & 0 \end{bmatrix}, \quad \boldsymbol{\mathcal{H}}'_l = \begin{bmatrix} \nabla^2_{11}f(\widetilde{\mathbf{x}}^{(t)}) & 0 \\ \nabla^2_{21}f(\widetilde{\mathbf{x}}^{(t)}) & \nabla^2_{22}f(\widetilde{\mathbf{x}}^{(t)}) \end{bmatrix}, \tag{127}$$

where $\widetilde{\mathbf{x}}^{(t)}$ is an ϵ-second order stationary point, and

$$\mathbf{M}' \triangleq \mathbf{I}+\eta\boldsymbol{\mathcal{H}}'_l \quad \mathbf{T}' \triangleq \mathbf{I}-\eta\boldsymbol{\mathcal{H}}'_u. \tag{128}$$

Obviously, we also have $\boldsymbol{\mathcal{H}} = \boldsymbol{\mathcal{H}}'_l+\boldsymbol{\mathcal{H}}'_u$.

Note that $\det(\mathbf{T}')=1$, which implies that $\det(\mathbf{T}'^{-1}\mathbf{M}'-\lambda\mathbf{I}) = \det(\mathbf{M}'-\lambda\mathbf{T}')$, where λ denotes the eigenvalue. We can analyze the determinant of $\mathbf{M}'-\lambda\mathbf{T}'$. We have

$$\det[\mathbf{M}'-\lambda\mathbf{T}'] = \left[\underbrace{\begin{matrix}(1-\lambda)\mathbf{I}+\eta\nabla^2_{11}f(\widetilde{\mathbf{x}}^{(t)}) & \lambda\eta\nabla^2_{12}f(\widetilde{\mathbf{x}}^{(t)}) \\ \eta\nabla^2_{21}f(\widetilde{\mathbf{x}}^{(t)}) & (1-\lambda)\mathbf{I}+\eta\nabla^2_{22}f(\widetilde{\mathbf{x}}^{(t)})\end{matrix}}_{\triangleq\mathbf{Q}'(\lambda)}\right].$$

It can be observed that

$$\mathbf{Q}'(\lambda) = \begin{bmatrix}\mathbf{I} & \\ & \frac{1}{\sqrt{\lambda}}\end{bmatrix}\underbrace{\begin{bmatrix}(1-\lambda)\mathbf{I}+\eta\nabla^2_{11}f(\widetilde{\mathbf{x}}^{(t)}) & \eta\sqrt{\lambda}\nabla^2_{12}f(\widetilde{\mathbf{x}}^{(t)}) \\ \eta\sqrt{\lambda}\nabla^2_{21}f(\widetilde{\mathbf{x}}^{(t)}) & (1-\lambda)\mathbf{I}+\eta\nabla^2_{22}f(\widetilde{\mathbf{x}}^{(t)})\end{bmatrix}}_{\mathbf{G}'(\lambda)}\begin{bmatrix}\mathbf{I} & \\ & \sqrt{\lambda}\end{bmatrix},$$

meaning that $\mathbf{Q}'(\lambda)$ is similar to $\mathbf{G}'(\lambda)$. Consequently, we can conclude that $\mathbf{Q}'(\delta)$ has the same eigenvalues of $\mathbf{G}'(\delta)$. Furthermore, since matrix $\mathbf{G}'(\lambda)$ is symmetric, we know that all eigenvalues of $\mathbf{Q}'(\lambda)$ and $\mathbf{G}'(\lambda)$ are real. Then, we can need to show there exists λ such that $\det(\mathbf{Q}'(\lambda))=0$.

Consider $0\leq\delta\leq 1$. We have

$$\mathbf{G}'(1-\delta) = \begin{bmatrix}\delta\mathbf{I}+\eta\nabla^2_{11}f(\widetilde{\mathbf{x}}^{(t)}) & \eta\sqrt{1-\delta}\nabla^2_{12}f(\widetilde{\mathbf{x}}^{(t)}) \\ \eta\sqrt{1-\delta}\nabla^2_{21}f(\widetilde{\mathbf{x}}^{(t)}) & \delta\mathbf{I}+\eta\nabla^2_{22}f(\widetilde{\mathbf{x}}^{(t)})\end{bmatrix}. \tag{129}$$

Since we know that $\boldsymbol{\mathcal{H}}$ and $\mathbf{G}(1-\delta)$ 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 matrices

$$\max_{1\leq i\leq d}|\lambda_i(\eta\boldsymbol{\mathcal{H}})-\lambda_i(\mathbf{G}'(1-\delta))| \leq \|\eta\boldsymbol{\mathcal{H}}-\mathbf{G}'(1-\delta)\| \tag{130}$$

where $\lambda_i(\boldsymbol{\mathcal{H}})$ and $\lambda_i(\mathbf{G}'(1-\delta))$ denote the ith eigenvalue of $\boldsymbol{\mathcal{H}}$ and $\mathbf{G}'(1-\delta)$, which are listed in a decreasing order.

With the help of (130), we can check

$$\begin{aligned}
&\|\mathbf{G}'(1-\delta)-\eta\boldsymbol{\mathcal{H}}\| \\
=&\left\|\delta\mathbf{I}+\begin{bmatrix} 0 & (\sqrt{1-\delta}-1)\eta\nabla^2_{12}f(\widetilde{\mathbf{x}}^{(t)}) \\ (\sqrt{1-\delta}-1)\eta\nabla^2_{21}f(\widetilde{\mathbf{x}}^{(t)}) & 0 \end{bmatrix}\right\| \\
\leq&\delta+(\sqrt{1-\delta}-1)\eta\|\boldsymbol{\mathcal{H}}\|+(\sqrt{1-\delta}-1)\eta\left\|\begin{matrix} \nabla^2_{11}f(\widetilde{\mathbf{x}}^{(t)}) & 0 \\ 0 & \nabla^2_{22}f(\widetilde{\mathbf{x}}^{(t)}) \end{matrix}\right\| \\
\overset{(a)}{\leq}&\delta+(\sqrt{1-\delta}-1)(\frac{L}{L_{\max}}+1)
\end{aligned} \tag{131}$$

where (a) is true since we used $\eta \leq c_{\max}/L_{\max}$. Also, it can be observed that when $\delta = 0$, matrix $\mathbf{G}'(\delta)$ is reduced to $\eta\boldsymbol{\mathcal{H}}$.

We know that the minimum eigenvalue of $\eta\boldsymbol{\mathcal{H}}$ which is $-\eta\gamma$ and the maximum difference of the eigenvalues between $\eta\boldsymbol{\mathcal{H}}$ and $\mathbf{G}'(\delta)$ is upper bounded by (131). 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 $\delta \in [0,1]$ such that

$$\delta+(\sqrt{1-\delta}-1)(\frac{L}{L_{\max}}+1)<\eta\gamma. \tag{132}$$

In other words, if we choose

$$\delta^*=\frac{\eta\gamma}{2}$$

then we can conclude that $\mathbf{G}'(\delta^*)$ has a negative eigenvalue which is less than $-\eta\gamma+\delta^*=-\frac{\eta\gamma}{2}$. In the following, we will check that δ^* is a valid choice, meaning that equation (132) holds when $\delta^*=\frac{\eta\gamma}{2}$.

Actually, equation (132) can be rewritten as

$$\delta+\sqrt{1-\delta}(1+\frac{L}{L_{\max}})<\eta\gamma+(1+\frac{L}{L_{\max}}), \tag{133}$$

Since $\kappa = L_{\max}/\gamma \geq 1$ and $\eta \leq c_{\max}/L_{\max}$ where $c_{\max} \leq 1/2$, we have

$$\sqrt{1-\delta^*}=\sqrt{1-\eta\gamma/2}<1, \tag{134}$$

which implies that equation (132) is true with chosen δ^* Therefore, we can conclude that $\mathbf{Q}'(1+\delta^*)$ has a negative eigenvalue.

When δ is large, i.e., $\delta > 1$, we have

$$\mathbf{Q}'(1-\delta)=\begin{bmatrix} \mathbf{I} & \\ & \frac{j}{\sqrt{1-\delta}} \end{bmatrix}\underbrace{\begin{bmatrix} \delta\mathbf{I}+\eta\nabla^2_{11}f(\widetilde{\mathbf{x}}^{(t)}) & -j\eta\sqrt{1-\delta}\nabla^2_{12}f(\widetilde{\mathbf{x}}^{(t)}) \\ \eta\sqrt{1-\delta}\nabla^2_{21}f(\widetilde{\mathbf{x}}^{(t)}) & \delta\mathbf{I}+\eta\nabla^2_{22}f(\widetilde{\mathbf{x}}^{(t)}) \end{bmatrix}}_{\mathbf{G}'(1-\delta)}\begin{bmatrix} \mathbf{I} & \\ & j\sqrt{1-\delta} \end{bmatrix}, \tag{135}$$

where j denotes the imaginary number, so $\mathbf{Q}'(1-\delta)$ is similar to $\mathbf{G}'(1-\delta)$ when $\delta > 1$. Also, we know that $\mathbf{G}'(1-\delta)$ is a Hermitian matrix. It is easy to check $\mathbf{Q}'(1-\delta)$ has a positive eigenvalue, since term $\delta\mathbf{I}$ dominates the spectrum of matrix $\mathbf{Q}'(1-\delta)$ in (135). Considering the eigenvalue is continuous with respect to δ, we can conclude there exists a δ, i.e., $\widehat{\delta}'$, such that $\mathbf{Q}'(1-\widehat{\delta}')$ has a zero eigenvalue, i.e., $\det(\mathbf{Q}'(1-\widehat{\delta}'))=0$ where $1-\widehat{\delta}'$ is at least as small as

$$1-\delta^*=1-\frac{\eta\gamma}{2}, \tag{136}$$

meaning that $1-\widehat{\delta}' \leq 1-\frac{\eta\gamma}{2}$. □

In the following, we will give the proofs of Lemma 12–Lemma 16 in details.

D Proofs of Lemma 11–Lemma 16

D.1 Proof of Lemma 11

Proof. First, we have

$$\begin{aligned}\|\nabla_1 f(\mathbf{x}_1^{(t)},\mathbf{x}_2^{(t)})\|^2 \leq& 2\|\nabla_1 f(\mathbf{x}_1^{(t+1)},\mathbf{x}_2^{(t)})-\nabla_1 f(\mathbf{x}_1^{(t)},\mathbf{x}_2^{(t)})\|^2+2\|\nabla_1 f(\mathbf{x}_1^{(t+1)},\mathbf{x}_2^{(t)})\|^2\\ \overset{(a)}{\leq}& 2L_{\max}^2\|\mathbf{x}_1^{(t+1)}-\mathbf{x}_1^{(t)}\|^2+2\|\nabla_1 f(\mathbf{x}_1^{(t+1)},\mathbf{x}_2^{(t)})\|^2\\ \overset{(8)}{\leq}& 2L_{\max}^2\eta^2\|\nabla_1 f(\mathbf{x}_1^{(t+1)},\mathbf{x}_2^{(t)})\|^2+2\|\nabla_1 f(\mathbf{x}_1^{(t+1)},\mathbf{x}_2^{(t)})\|^2\\ \overset{(b)}{\leq}& 3\|\nabla_1 f(\mathbf{x}_1^{(t+1)},\mathbf{x}_2^{(t)})\|^2\end{aligned} \tag{137}$$

where in (a) we used block-wise Lipschitz continuity, in (b) we choose $\eta \leq 1/(2L_{\max})$.

$$\begin{aligned}\|\nabla_2 f(\mathbf{x}_1^{(t)},\mathbf{x}_2^{(t)})\|^2 \leq& 2\|\nabla_2 f(\mathbf{x}_1^{(t+1)},\mathbf{x}_2^{(t+1)})-\nabla_2 f(\mathbf{x}_1^{(t)},\mathbf{x}_2^{(t)})\|^2+2\|\nabla_2 f(\mathbf{x}_1^{(t+1)},\mathbf{x}_2^{(t+1)})\|^2\\ \leq& 4(\|\nabla_2 f(\mathbf{x}_1^{(t+1)},\mathbf{x}_2^{(t+1)})-\nabla_2 f(\mathbf{x}_1^{(t+1)},\mathbf{x}_2^{(t)})\|^2+\|\nabla_2 f(\mathbf{x}_1^{(t+1)},\mathbf{x}_2^{(t)})-\nabla_2 f(\mathbf{x}_1^{(t)},\mathbf{x}_2^{(t)})\|^2)\\ &+2\|\nabla_2 f(\mathbf{x}_1^{(t+1)},\mathbf{x}_2^{(t+1)})\|^2\\ \overset{(8)}{\leq}& 4(L_{\max}^2\|\mathbf{x}_2^{(t+1)}-\mathbf{x}_2^{(t)}\|^2+\|\mathbf{x}_1^{(t+1)}-\mathbf{x}_1^{(t)}\|^2)+2\|\nabla_2 f(\mathbf{x}_1^{(t+1)},\mathbf{x}_2^{(t+1)})\|^2\\ \overset{(a)}{\leq}& \|\nabla_1 f(\mathbf{x}_1^{(t+1)},\mathbf{x}_2^{(t)})\|^2+3\|\nabla_2 f(\mathbf{x}_1^{(t+1)},\mathbf{x}_2^{(t+1)})\|^2\end{aligned} \tag{138}$$

where (a) we also choose $\eta \leq 1/(2L_{\max})$.

Summing (137) and (138), we have

$$\|\nabla f(\mathbf{x}^{(t)})\|^2 \leq \sum_{k=1}^{2}\|\nabla_k f(\mathbf{x}_k^{(t)})\|^2 \leq 4\sum_{k=1}^{2}\|\nabla_k f(\mathbf{h}_{-k}^{(t)},\mathbf{x}_k^{(t+1)})\|^2 \overset{(8)}{=} 4\nu\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}\|^2 \tag{139}$$

where $\mathbf{h}_{-1}^{(t)}=\mathbf{x}_2^{(t)}$ and $\mathbf{h}_{-2}^{(t)}=\mathbf{x}_1^{(t+1)}$. □

D.2 Proof of Lemma 12

There proof involves two parts:

Upper Triangular Matrix: Consider three different vectors $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$. We can have

$$\begin{aligned}&\left\|\begin{bmatrix}\nabla_{11}^2 f(\mathbf{x}) & 0\\ \nabla_{21}^2 f(\mathbf{y}) & \nabla_{22}^2 f(\mathbf{y})\end{bmatrix}-\begin{bmatrix}\nabla_{11}^2 f(\mathbf{z}) & 0\\ \nabla_{21}^2 f(\mathbf{z}) & \nabla_{22}^2 f(\mathbf{z})\end{bmatrix}\right\|\\ \leq&\left\|\mathbf{I}_1\left(\begin{bmatrix}\nabla_{11}^2 f(\mathbf{x}) & \nabla_{12}^2 f(\mathbf{x})\\ \nabla_{21}^2 f(\mathbf{x}) & \nabla_{22}^2 f(\mathbf{x})\end{bmatrix}-\begin{bmatrix}\nabla_{11}^2 f(\mathbf{z}) & \nabla_{12}^2 f(\mathbf{z})\\ \nabla_{21}^2 f(\mathbf{z}) & \nabla_{22}^2 f(\mathbf{z})\end{bmatrix}\right)\mathbf{I}_1\right\|\\ &+\left\|\mathbf{I}_2\left(\begin{bmatrix}\nabla_{11}^2 f(\mathbf{y}) & \nabla_{12}^2 f(\mathbf{y})\\ \nabla_{21}^2 f(\mathbf{y}) & \nabla_{22}^2 f(\mathbf{y})\end{bmatrix}-\begin{bmatrix}\nabla_{11}^2 f(\mathbf{z}) & \nabla_{12}^2 f(\mathbf{z})\\ \nabla_{21}^2 f(\mathbf{z}) & \nabla_{22}^2 f(\mathbf{z})\end{bmatrix}\right)\right\|\\ \overset{(a)}{\leq}&\left\|\begin{bmatrix}\nabla_{11}^2 f(\mathbf{x}) & \nabla_{12}^2 f(\mathbf{x})\\ \nabla_{21}^2 f(\mathbf{x}) & \nabla_{22}^2 f(\mathbf{x})\end{bmatrix}-\begin{bmatrix}\nabla_{11}^2 f(\mathbf{z}) & \nabla_{12}^2 f(\mathbf{z})\\ \nabla_{21}^2 f(\mathbf{z}) & \nabla_{22}^2 f(\mathbf{z})\end{bmatrix}\right\|+\left\|\begin{bmatrix}\nabla_{11}^2 f(\mathbf{y}) & \nabla_{12}^2 f(\mathbf{y})\\ \nabla_{21}^2 f(\mathbf{y}) & \nabla_{22}^2 f(\mathbf{y})\end{bmatrix}-\begin{bmatrix}\nabla_{11}^2 f(\mathbf{z}) & \nabla_{12}^2 f(\mathbf{z})\\ \nabla_{21}^2 f(\mathbf{z}) & \nabla_{22}^2 f(\mathbf{z})\end{bmatrix}\right\|\\ \leq&\rho\left(\|\mathbf{x}-\mathbf{z}\|+\|\mathbf{y}-\mathbf{z}\|\right)\end{aligned}$$

where in (a) we use

$$\mathbf{I}_1=\begin{bmatrix}\mathbf{I} & 0\\ 0 & 0\end{bmatrix}\qquad \mathbf{I}_2=\begin{bmatrix}0 & 0\\ 0 & \mathbf{I}\end{bmatrix} \tag{140}$$

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

Lower Triangular Matrix:

$$\begin{aligned}
&\left\|\begin{bmatrix} 0 & \nabla^2_{21} f(\mathbf{x}) \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} 0 & \nabla^2_{21} f(\mathbf{y}) \\ 0 & 0 \end{bmatrix}\right\| \\
&= \left\|\mathbf{I}_1 \left(\begin{bmatrix} \nabla^2_{11} f(\mathbf{x}) & \nabla^2_{12} f(\mathbf{x}) \\ \nabla^2_{21} f(\mathbf{x}) & \nabla^2_{22} f(\mathbf{x}) \end{bmatrix} - \begin{bmatrix} \nabla^2_{11} f(\mathbf{y}) & \nabla^2_{12} f(\mathbf{y}) \\ \nabla^2_{21} f(\mathbf{y}) & \nabla^2_{22} f(\mathbf{y}) \end{bmatrix} \right) \mathbf{I}_2 \right\| \\
&\overset{(a)}{\leq} \rho \|\mathbf{x} - \mathbf{y}\|
\end{aligned}$$

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

D.3 Proof of Lemma 13

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

$$\begin{aligned}
f(\mathbf{x}^{(t+1)}) \leq & f(\mathbf{x}^{(t)}) + \sum_{k=1}^{2} \nabla_k f(\mathbf{h}^{(t)}_{-k}, \mathbf{x}^{(t)}_k)^\mathsf{T} (\mathbf{x}^{(t+1)}_k - \mathbf{x}^{(t)}_k) + \sum_{k=1}^{2} \frac{L_k}{2} \|\mathbf{x}^{(t+1)}_k - \mathbf{x}^{(t)}_k\|^2 \\
\leq & f(\mathbf{x}^{(t)}) + \sum_{k=1}^{2} \nabla_k f(\mathbf{h}^{(t)}_{-k}, \mathbf{x}^{(t+1)}_k)^\mathsf{T} (\mathbf{x}^{(t+1)}_k - \mathbf{x}^{(t)}_k) + \sum_{k=1}^{2} (\nabla_k f(\mathbf{h}^{(t)}_{-k}, \mathbf{x}^{(t)}_k) - \nabla_k f(\mathbf{h}^{(t)}_{-k}, \mathbf{x}^{(t+1)}_k))^\mathsf{T} (\mathbf{x}^{(t+1)}_k - \mathbf{x}^{(t)}_k) \\
& + \sum_{k=1}^{2} \frac{L_k}{2} \|\mathbf{x}^{(t+1)}_k - \mathbf{x}^{(t)}_k\|^2 \\
\overset{(a)}{\leq} & f(\mathbf{x}^{(t)}) - \sum_{k=1}^{2} \eta \|\nabla_k f(\mathbf{h}^{(t)}_{-k}, \mathbf{x}^{(t+1)}_k)\|^2 + \sum_{k=1}^{2} \frac{3\eta^2 L_k}{2} \|\nabla_k f(\mathbf{h}^{(t)}_{-k}, \mathbf{x}^{(t+1)}_k)\|^2 \\
\overset{(b)}{\leq} & f(\mathbf{x}^{(t+1)}) - \sum_{k=1}^{2} \frac{\eta}{2} \|\nabla_k f(\mathbf{h}^{(t)}_{-k}, \mathbf{x}^{(t+1)}_k)\|^2 \\
= & f(\mathbf{x}^{(t+1)}) - \frac{\nu}{2} \|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\|^2
\end{aligned} \tag{141}$$

where (a) is true because of the update rule of APP in each block and Assumption 1 and block-wise Lipschitz continuity, in (b) we choose $\eta \leq 1/(3L_{\max})$ and $\nu = 1/\eta$. □

D.4 Proof of Lemma 14

Proof. Without loss of generality, let $\mathbf{u}^{(0)}$ be the origin, i.e., $\mathbf{u}^{(0)} = 0$. According to the APP update rule of variables, we have

$$\mathbf{u}^{(t+1)} = \mathbf{u}^{(t)} - \eta \begin{bmatrix} \nabla_1 f(\mathbf{u}^{(t+1)}_1, \mathbf{u}^{(t)}_2) \\ \nabla_2 f(\mathbf{u}^{(t+1)}_1, \mathbf{u}^{(t+1)}_2) \end{bmatrix}. \tag{142}$$

It can be observed that the update rule of PA-PP is very similar as the one of PA-GD. The proof of this lemma is also similar as Lemma 8. We only need to replace $\nabla_1 f(\mathbf{u}^{(t)}_1, \mathbf{u}^{(t)}_2)$ as $\nabla_1 f(\mathbf{u}^{(t+1)}_1, \mathbf{u}^{(t)}_2)$ and $\nabla_2 f(\mathbf{u}^{(t+1)}_1, \mathbf{u}^{(t)}_2)$ as $\nabla_2 f(\mathbf{u}^{(t+1)}_1, \mathbf{u}^{(t+1)}_2)$, which can give us the claimed result after following the proof of Lemma 8. Hence, we ignore the repeated part with the proof of Lemma 8 for simplicity of expressions. □

D.5 Proof of Lemma 15

Proof. Let $\mathbf{u}^{(0)} = 0$ and define $\mathbf{v}^{(t)} \triangleq \mathbf{w}^{(t)} - \mathbf{u}^{(t)}$. According to the assumption of Lemma 9, we know that $\mathbf{v}^{(0)} = \upsilon[\eta L_{\max} \mathcal{S}/(\kappa \log(\frac{d\kappa}{\delta}) \mathcal{P}_1)] \vec{\mathbf{e}}'$ when $\upsilon \in [\delta/(2\sqrt{d}), 1]$. First, we define the following auxiliary function

$$h(\theta) \triangleq \begin{bmatrix} \nabla_1 f(\mathbf{u}^{(t+1)}_1 + \theta \mathbf{v}^{(t+1)}_1, \mathbf{u}^{(t)}_2 + \theta \mathbf{v}^{(t)}_2) \\ \nabla_2 f(\mathbf{u}^{(t+1)}_1 + \theta \mathbf{v}^{(t+1)}_1, \mathbf{u}^{(t+1)}_2 + \theta \mathbf{v}^{(t+1)}_2) \end{bmatrix},$$

then have

$$h(0)=\left[\begin{array}{c}\nabla_1 f(\mathbf{u}_1^{(t+1)},\mathbf{u}_2^{(t)})\\ \nabla_2 f(\mathbf{u}_1^{(t+1)},\mathbf{u}_2^{(t+1)})\end{array}\right],\quad h(1)=\left[\begin{array}{c}\nabla_1 f(\mathbf{u}_1^{(t+1)}+\mathbf{v}_1^{(t+1)},\mathbf{u}_2^{(t)}+\mathbf{v}_2^{(t)})\\ \nabla_2 f(\mathbf{u}_1^{(t+1)}+\mathbf{v}_1^{(t+1)},\mathbf{u}_2^{(t+1)}+\mathbf{v}_2^{(t+1)})\end{array}\right],$$

$$g(\theta)=\frac{dh(\theta)}{d\theta}=\underbrace{\left[\begin{array}{cc}\nabla_{11}^2 f(\mathbf{u}_1^{(t+1)}+\theta\mathbf{v}_1^{(t+1)},\mathbf{u}_2^{(t)}+\theta\mathbf{v}_2^{(t)}) & 0\\ \nabla_{21}^2 f(\mathbf{u}_1^{(t+1)}+\theta\mathbf{v}_1^{(t+1)},\mathbf{u}_2^{(t+1)}+\theta\mathbf{v}_2^{(t+1)}) & \nabla_{22}^2 f(\mathbf{u}_1^{(t+1)}+\theta\mathbf{v}_1^{(t+1)},\mathbf{u}_2^{(t+1)}+\theta\mathbf{v}_2^{(t+1)})\end{array}\right]}_{\widetilde{\boldsymbol{\mathcal{H}}}_l^{\prime(t)}(\theta)}\mathbf{v}^{(t+1)}$$

$$+\underbrace{\left[\begin{array}{cc}0 & \nabla_{12}^2 f(\mathbf{u}_1^{(t+1)}+\theta\mathbf{v}_1^{(t+1)},\mathbf{u}_2^{(t)}+\theta\mathbf{v}_2^{(t)})\\ 0 & 0\end{array}\right]}_{\widetilde{\boldsymbol{\mathcal{H}}}_u^{\prime(t)}(\theta)}\mathbf{v}^{(t)},$$

$$\left[\begin{array}{c}\nabla_1 f(\mathbf{w}_1^{(t+1)},\mathbf{w}_2^{(t)})\\ \nabla_2 f(\mathbf{w}_1^{(t+1)},\mathbf{w}_2^{(t+1)})\end{array}\right]=\int_0^1 g(\theta)d\theta+\left[\begin{array}{c}\nabla_1 f(\mathbf{u}_1^{(t+1)},\mathbf{u}_2^{(t)})\\ \nabla_2 f(\mathbf{u}_1^{(t+1)},\mathbf{u}_2^{(t+1)})\end{array}\right].$$

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

$$\mathbf{u}^{(t+1)}+\mathbf{v}^{(t+1)}=\mathbf{w}^{(t+1)}=\mathbf{w}^{(t)}-\eta\left[\begin{array}{c}\nabla_1 f(\mathbf{w}_1^{(t+1)},\mathbf{w}_2^{(t)})\\ \nabla_2 f(\mathbf{w}_1^{(t+1)},\mathbf{w}_2^{(t+1)})\end{array}\right] \tag{143}$$

$$=\mathbf{u}^{(t)}+\mathbf{v}^{(t)}-\eta\left[\begin{array}{c}\nabla_1 f(\mathbf{u}_1^{(t+1)}+\mathbf{v}_1^{(t+1)},\mathbf{u}_1^{(t)}+\mathbf{v}_1^{(t)})\\ \nabla_2 f(\mathbf{u}_1^{(t+1)}+\mathbf{v}_1^{(t+1)},\mathbf{u}_2^{(t+1)}+\mathbf{v}_2^{(t+1)})\end{array}\right]$$

$$=\mathbf{u}^{(t)}+\mathbf{v}^{(t)}-\eta\left[\begin{array}{c}\nabla_1 f(\mathbf{u}_1^{(t+1)},\mathbf{u}_2^{(t)})\\ \nabla_2 f(\mathbf{u}_1^{(t+1)},\mathbf{u}_2^{(t+1)})\end{array}\right]-\int_0^1 g(\theta)d\theta \tag{144}$$

$$\overset{(a)}{=}\mathbf{u}^{(t)}+\mathbf{v}^{(t)}-\eta\left[\begin{array}{c}\nabla_1 f(\mathbf{u}_1^{(t+1)},\mathbf{u}_2^{(t)})\\ \nabla_2 f(\mathbf{u}_1^{(t+1)},\mathbf{u}_2^{(t+1)})\end{array}\right]-\eta\widetilde{\Delta}_u^{\prime(t)}\mathbf{v}^{(t)}-\boldsymbol{\mathcal{H}}_u^{\prime}\mathbf{v}^{(t)}-\eta\widetilde{\Delta}_l^{\prime(t)}\mathbf{v}^{(t+1)}-\eta\boldsymbol{\mathcal{H}}_l^{\prime}\mathbf{v}^{(t+1)} \tag{145}$$

where in (a) we used the following definitions

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

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

and

$$\boldsymbol{\mathcal{H}}_u^{\prime}=\left[\begin{array}{cc}0 & \nabla_{12}^2 f(\widetilde{\mathbf{x}}^{(t)})\\ 0 & 0\end{array}\right]\quad\boldsymbol{\mathcal{H}}_l^{\prime}=\left[\begin{array}{cc}\nabla_{11}^2 f(\widetilde{\mathbf{x}}^{(t)}) & 0\\ \nabla_{21}^2 f(\widetilde{\mathbf{x}}^{(t)}) & \nabla_{22}^2 f(\widetilde{\mathbf{x}}^{(t)})\end{array}\right]. \tag{146}$$

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

Dynamics of $\mathbf{v}^{(t)}$: Since the first two terms at RHS of (145) combined with $\mathbf{u}^{(t)}$ at LHS of (145) are exactly the same as (142). It can be observed that equation (145) gives the dynamic of $\mathbf{v}^{(t)}$, i.e.,

$$\mathbf{v}^{(t+1)}=\mathbf{v}^{(t)}-\eta\widetilde{\Delta}_u^{\prime(t)}\mathbf{v}^{(t)}-\eta\boldsymbol{\mathcal{H}}_u^{\prime}\mathbf{v}^{(t)}-\eta\widetilde{\Delta}_l^{\prime(t)}\mathbf{v}^{(t+1)}-\eta\boldsymbol{\mathcal{H}}_l^{\prime}\mathbf{v}^{(t+1)}, \tag{147}$$

which can be equivalently expressed by

$$\underbrace{(\mathbf{I}+\eta\boldsymbol{\mathcal{H}}_l^{\prime})}_{\triangleq\mathbf{M}^{\prime}}\mathbf{v}^{(t+1)}=\underbrace{(\mathbf{I}-\eta\boldsymbol{\mathcal{H}}_u^{\prime})}_{\triangleq\mathbf{T}^{\prime}}\mathbf{v}^{(t)}-\eta\widetilde{\Delta}_l^{\prime(t)}\mathbf{v}^{(t+1)}-\eta\widetilde{\Delta}_u^{\prime(t)}\mathbf{v}^{(t)}. \tag{148}$$

It is worth noting that matrix $\mathbf{T}'$ is an upper triangular matrix where the diagonal entries are all 1s, so it is invertible. Taking the inverse of $\mathbf{T}'$ on both sides of (148), we can obtain

$$\mathbf{T}'^{-1}\mathbf{M}'\mathbf{v}^{(t+1)} \overset{(145)}{=} \mathbf{v}^{(t)} - \mathbf{T}'^{-1}\eta\widetilde{\Delta}_l'^{(t)}\mathbf{v}^{(t+1)} - \mathbf{T}'^{-1}\eta\widetilde{\Delta}_u'^{(t)}\mathbf{v}^{(t)}. \tag{149}$$

Let $\mathbb{P}'_{\texttt{left}}$ denote the projection operator that projects the vector onto the space spanned by the eigenvector of $\mathbf{T}'^{-1}\mathbf{M}$ whose corresponding positive eigenvalue is minimum. Taking the projection on both sides of (149), we have

$$\mathbb{P}'_{\texttt{left}}(\mathbf{T}'^{-1}\mathbf{M}')\mathbf{v}^{(t+1)} + \mathbb{P}'_{\texttt{left}}\mathbf{T}'^{-1}\eta\widetilde{\Delta}_l'^{(t)}\mathbf{v}^{(t+1)} = \mathbb{P}'_{\texttt{left}}\mathbf{v}^{(t)} - \mathbb{P}'_{\texttt{left}}\mathbf{T}'^{-1}\eta\widetilde{\Delta}_u'^{(t)}\mathbf{v}^{(t)}. \tag{150}$$

Relationship of the Norm of $\mathbf{v}^{(t)}$ Projected onto the Two Subspaces: Let $\phi^{(t)}$ denote the norm of $\mathbf{v}^{(t)}$ projected onto the space spanned by the eigenvector of $\mathbf{T}'^{-1}\mathbf{M}'$ whose positive minimum eigenvalue of $\mathbf{M}'^{-1}\mathbf{T}'$ is $1-\widehat{\delta}' > 0$ and $\theta^{(t)}$ denote the norm of $\mathbf{v}^{(t)}$ projected onto the remaining space. From (150), we can have

$$(1-\widehat{\delta}')\phi^{(t+1)} \overset{(a)}{\geq} \phi^{(t)} - \eta\|\mathbf{T}'^{-1}\|\|\widetilde{\Delta}_l'^{(t)}\|\|\mathbf{v}^{(t+1)}\| - \eta\|\mathbf{T}'^{-1}\|\|\widetilde{\Delta}_u'^{(t)}\|\|\mathbf{v}^{(t)}\|, \tag{151}$$

$$(1-\widehat{\delta}')\theta^{(t+1)} \leq \theta^{(t)} + \eta\|\mathbf{T}'^{-1}\|\|\widetilde{\Delta}_l'^{(t)}\|\|\mathbf{v}^{(t+1)}\| + \eta\|\mathbf{T}'^{-1}\|\|\widetilde{\Delta}_u'^{(t)}\|\|\mathbf{v}^{(t)}\|. \tag{152}$$

where (a) is true because we applied the triangle inequality since η is sufficiently small.

Since $\|\mathbf{w}^{(0)} - \widetilde{\mathbf{x}}^{(t)}\| \leq \|\mathbf{u}^{(0)} - \widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{v}^{(0)}\| \leq 2r$, we can apply Lemma 14. Then, we know $\|\mathbf{w}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 5\widehat{c}\mathcal{S}, \forall t < T$. According to the assumptions of Lemma 15, we have $\|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 5\widehat{c}\mathcal{S}$, and

$$\|\mathbf{v}^{(t)}\| = \|\mathbf{w}^{(t)} - \mathbf{u}^{(t)}\| \leq \|\mathbf{u}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{w}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\| \leq 10\widehat{c}\mathcal{S}. \tag{153}$$

From (62), we know that

$$\|\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}\| \leq \frac{4.3\eta\mathcal{G}}{\kappa} = \frac{4.3\eta^3 L_{\max}^3 \frac{\gamma}{\rho}}{\kappa^2 \log^3 \frac{d\kappa}{\delta}\mathcal{P}} \leq \mathcal{S},$$

where we choose $\eta \leq c_{\max}/L_{\max}$ and $c_{\max} = 1/10$. Similarly, we also have $\|\mathbf{u}^{(t+1)} - \mathbf{u}^{(t)}\| \leq \mathcal{S}$.

Then, we need to quantify the upper bounds of $\|\mathbf{M}'^{-1}\|$, $\|\mathbf{v}^{(t+1)}\|$, $\|\widetilde{\Delta}_u'^{(t)}\|$ and $\|\widetilde{\Delta}_l'^{(t)}\|$.

1. Upper bound of $\|\mathbf{M}'^{-1}\|$: applying the steps of deriving (78), we can quantify the inverse of matrix $\mathbf{T}'$ as follows

$$\begin{aligned}\|\mathbf{T}'^{-1}\| \leq& 1 + \eta\|\boldsymbol{\mathcal{H}}_u'\| = 1 + \eta\|\boldsymbol{\mathcal{H}}_u'^{\mathsf{T}}\| \\ =& 1 + \|\eta\boldsymbol{\mathcal{H}} \odot \mathbf{D} - \eta\boldsymbol{\mathcal{H}}_d\| \\ <& 2(1 + \frac{L\log(2d)}{L_{\max}}).\end{aligned}$$

2. Relation between $\|\mathbf{v}^{(t)}\|$ and $\|\mathbf{v}^{(t+1)}\|$: We also know that

$$\begin{aligned}\|\mathbf{v}^{(t+1)}\|^2 =& \|\mathbf{w}^{(t+1)} - \mathbf{u}^{(t+1)}\|^2 = \left\|\mathbf{w}^{(t)} - \eta\begin{bmatrix}\nabla_1 f(\mathbf{w}_1^{(t+1)}, \mathbf{w}_2^{(t)}) \\ \nabla_2 f(\mathbf{w}_1^{(t+1)}, \mathbf{w}_2^{(t+1)})\end{bmatrix} - \left(\mathbf{u}^{(t)} - \eta\begin{bmatrix}\nabla_1 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t+1)})\end{bmatrix}\right)\right\|^2 \\ \leq& 2\|\mathbf{v}^{(t)}\|^2 + 4\eta^2\left\|\begin{bmatrix}\nabla_1 f(\mathbf{w}_1^{(t+1)}, \mathbf{w}_2^{(t)}) \\ \nabla_2 f(\mathbf{w}_1^{(t+1)}, \mathbf{w}_2^{(t+1)})\end{bmatrix} - \begin{bmatrix}\nabla_1 f(\mathbf{u}_1^{(t+1)}, \mathbf{w}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{w}_2^{(t+1)})\end{bmatrix}\right\|^2 \\ &+ 4\eta^2\left\|\begin{bmatrix}\nabla_1 f(\mathbf{u}_1^{(t+1)}, \mathbf{w}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{w}_2^{(t+1)})\end{bmatrix} - \begin{bmatrix}\nabla_1 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t)}) \\ \nabla_2 f(\mathbf{u}_1^{(t+1)}, \mathbf{u}_2^{(t+1)})\end{bmatrix}\right\|^2 \\ \overset{(a)}{\leq}& 2\|\mathbf{v}^{(t)}\|^2 + 8\eta^2 L_{\max}^2\|\mathbf{v}_1^{(t)}\|^2 + 4\eta^2 L_{\max}^2(\|\mathbf{v}_2^{(t+1)}\|^2 + \|\mathbf{v}_2^{(t)}\|^2)\end{aligned} \tag{154}$$

where (a) is true due to Lipschitz continuity.

We can express (154) as

$$(1-4\eta^2L_{\max}^2)\|\mathbf{v}^{(t+1)}\| \le (2+8\eta^2L_{\max}^2)\|\mathbf{v}^{(t)}\|^2,$$

which implies

$$\|\mathbf{v}^{(t+1)}\| \le \sqrt{\frac{2+\frac{8}{100}}{1-\frac{4}{100}}}\|\mathbf{v}^{(t)}\| < \sqrt{2.2}\|\mathbf{v}^{(t)}\| < 1.5\|\mathbf{v}^{(t)}\| \tag{155}$$

where we choose $\eta \le c_{\max}/L_{\max}$ and $c_{\max} = 1/10$.

3. Upper bound of $\|\widetilde{\Delta}_l'^{(t)}\|$: applying Lemma 12, we can also get the upper bound of $\|\widetilde{\Delta}_l'^{(t)}\|$, i.e.,

$$\begin{aligned}
\|(\widetilde{\Delta}_l'^{(t)})\| \le& \int_0^1 \|\widetilde{\boldsymbol{\mathcal{H}}}_l'^{(t)}(\theta) - \boldsymbol{\mathcal{H}}_l'\|d\theta \\
\overset{(110)}{\le}& \int_0^1 \rho\left(\|\mathbf{u}^{(t+1)}+\theta\mathbf{v}^{(t+1)}-\widetilde{\mathbf{x}}^{(t)}\| + \left\|\begin{bmatrix}\mathbf{u}_1^{(t+1)}+\theta\mathbf{v}_1^{(t+1)}\\ \mathbf{u}_2^{(t)}+\theta\mathbf{v}_2^{(t)}\end{bmatrix} - \widetilde{\mathbf{x}}^{(t)}\right\|\right)d\theta \\
\le& \int_0^1 \rho\left(2\|\mathbf{u}^{(t+1)}+\theta\mathbf{v}^{(t+1)}-\widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{u}^{(t)}+\theta\mathbf{v}^{(t)}-\widetilde{\mathbf{x}}^{(t)}\|\right)d\theta \\
\le& \rho(2\|\mathbf{u}^{(t+1)}-\widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{u}^{(t)}-\widetilde{\mathbf{x}}^{(t)}\|) + \rho\int_0^1 \theta(2\|\mathbf{v}^{(t+1)}\| + \|\mathbf{v}^{(t)}\|)d\theta \\
\le& \rho\left(2\|\mathbf{u}^{(t+1)}-\mathbf{u}^{(t)}\| + 2\|\mathbf{u}^{(t)}-\widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{u}^{(t)}-\widetilde{\mathbf{x}}^{(t)}\|) + 0.5\|\mathbf{v}^{(t+1)}\| + 0.5\|\mathbf{v}^{(t)}\|\right) \\
\overset{(155)}{\le}& \rho\left(2\|\mathbf{u}^{(t+1)}-\mathbf{u}^{(t)}\| + 3\|\mathbf{u}^{(t)}-\widetilde{\mathbf{x}}^{(t)}\| + 1.25\|\mathbf{v}^{(t)}\|\right) \\
\le& \rho(2+27.5\widehat{c})\mathcal{S}.
\end{aligned}$$

4. Upper bound of $\|\widetilde{\Delta}_u'^{(t)}\|$: according to ρ-Hessian Lipschitz continuity and Lemma 12, we have the size of $\widetilde{\Delta}_u'^{(t)}$ as the following.

$$\begin{aligned}
\|(\widetilde{\Delta}_u'^{(t)})\| \le& \int_0^1 \|\widetilde{\boldsymbol{\mathcal{H}}}_u'^{(t)}(\theta) - \boldsymbol{\mathcal{H}}_u'\|d\theta \\
\overset{(111)}{\le}& \int_0^1 \rho\left\|\begin{bmatrix}\mathbf{u}_1^{(t+1)}+\theta\mathbf{v}_1^{(t+1)}\\ \mathbf{u}_2^{(t)}+\theta\mathbf{v}_2^{(t)}\end{bmatrix} - \widetilde{\mathbf{x}}^{(t)}\right\|d\theta \\
\le& \int_0^1 \rho(\|\mathbf{u}^{(t)}+\theta\mathbf{v}^{(t)}-\widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{u}^{(t+1)}+\theta\mathbf{v}^{(t+1)}-\widetilde{\mathbf{x}}^{(t)})\|d\theta \\
\le& \rho(\|\mathbf{u}^{(t+1)}-\widetilde{\mathbf{x}}^{(t)}\| + \|\mathbf{u}^{(t)}-\widetilde{\mathbf{x}}^{(t)}\|) + \rho\int_0^1 \theta(\|\mathbf{v}^{(t+1)}\| + \|\mathbf{v}^{(t)}\|)d\theta \\
\overset{(155)}{\le}& \rho\left(\|\mathbf{u}^{(t+1)}-\mathbf{u}^{(t)}\| + 2\|\mathbf{u}^{(t)}-\widetilde{\mathbf{x}}^{(t)}\| + 1.25\|\mathbf{v}^{(t)}\|\right) \\
\le& \rho(1+22.5\widehat{c})\mathcal{S}.
\end{aligned} \tag{156}$$

With the bounds of $\|\mathbf{v}^{(t+1)}\|$, $\|\widetilde{\Delta}_u'^{(t)}\|$, $\|\widetilde{\Delta}_l'^{(t)}\|$ and relation between $\|\mathbf{v}^{(t+1)}\|$ and $\|\mathbf{v}^{(t)}\|$, we can further simply (151) and (152) as follows,

$$\begin{aligned}
(1-\widehat{\delta}')\phi^{(t+1)} &\overset{(151)}{\ge} \phi^{(t)} - \eta(1.5\|\widetilde{\Delta}_l'^{(t)}\| + \|\widetilde{\Delta}_u'^{(t)}\|)\|\mathbf{T}'^{-1}\|\sqrt{(\phi^{(t)})^2+(\theta^{(t)})^2}, \\
(1-\widehat{\delta}')\theta^{(t+1)} &\overset{(152)}{\le} \theta^{(t)} + \eta(1.5\|\widetilde{\Delta}_l'^{(t)}\| + \|\widetilde{\Delta}_u'^{(t)}\|)\|\mathbf{T}'^{-1}\|\sqrt{(\phi^{(t)})^2+(\theta^{(t)})^2},
\end{aligned}$$

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

Consequently, we can arrive at

$$(1-\widehat{\delta}')\phi^{(t+1)} \geq \phi^{(t)} - \mu\sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2}, \tag{157}$$

$$(1-\widehat{\delta}')\theta^{(t+1)} \leq \theta^{(t)} + \mu\sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2}, \tag{158}$$

where μ is the upper bound of term $\eta(1.5\|\widetilde{\Delta}_l'^{(t)}\| + \|\widetilde{\Delta}_u'^{(t)}\|)\|\mathbf{T}'^{-1}\|$ and can be obtained by

$$\mu \triangleq \eta\rho\mathcal{S}\mathcal{P}(4+62\widehat{c}). \tag{159}$$

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

$$\theta^{(t)} \leq 4\mu t\phi^{(t)}. \tag{160}$$

It is true when $t=0$ since $\|\theta^{(0)}\| \overset{(115)}{=} 0$.

Assuming that equation (160) is true at the tth iteration, we need to prove

$$\theta^{(t+1)} \leq 4\mu(t+1)\phi^{(t+1)}. \tag{161}$$

Applying (157) into RHS of (161), we have

$$4\mu(t+1)\phi^{(t+1)} \geq \frac{4\mu(t+1)}{1-\widehat{\delta}'}\left(\phi^{(t)} - \mu\sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2}\right), \tag{162}$$

and substituting (158) into LHS of (161), we have

$$\theta^{(t+1)} \leq \frac{(4\mu t\phi^{(t)}) + \mu\sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2}}{1-\widehat{\delta}'}. \tag{163}$$

Then, our goal is to prove RHS of (162) is greater than RHS of (163). After some manipulations, it is sufficient to show

$$(1+4\mu(t+1))\left(\sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2}\right) \leq 4\phi^{(t)}. \tag{164}$$

In the following, we will show that the above relation is true.

First step : We know that

$$4\mu(t+1) \leq 4\mu T \overset{(159)}{\leq} 4\eta\rho\mathcal{S}\mathcal{P}(4+62\widehat{c})\widehat{c}\mathcal{T} \overset{(107\text{d})(159)}{\leq} \frac{4\widehat{c}\eta^2 L_{\max}^2(4+62\widehat{c})}{\kappa\log(\frac{d\kappa}{\delta})} \overset{(a)}{\leq} 1 \tag{165}$$

where (a) is true because we choose $c'_{\max} = 1/(2\widehat{c}(4+62\widehat{c}))$ and $\eta \leq c'_{\max}/L_{\max}$.

Second step : Also, we know that

$$4\phi^{(t)} \geq 2\sqrt{2(\phi^{(t)})^2} \overset{(160),(165)}{\geq} (1+4\mu(t+1))\sqrt{(\phi^{(t)})^2 + (\theta^{(t)})^2}.$$

With the above two steps, we have $\theta^{(t+1)} \leq 4\mu(t+1)\phi^{(t+1)}$, which completes the induction.

Recursion of $\phi^{(t)}$:Using (160), we have $\theta^{(t)} \overset{(160)}{\leq} 4\mu t\phi^{(t)} \overset{(165)}{\leq} \phi^{(t)}$, and have

$$(1-\widehat{\delta}')\phi^{(t+1)} \overset{(157)}{\geq} \phi^{(t)} - \mu\sqrt{(\phi^{(t)})^2+(\theta^{(t)})^2},$$

which implies

$$\begin{aligned}\phi^{(t+1)} &\overset{(a)}{\geq} \frac{1}{1-\widehat{\delta}'}\left(\phi^{(t)} - \mu\sqrt{(\phi^{(t)})^2+(\theta^{(t)})^2}\right)\\ &\overset{(b)}{\geq} \frac{1}{1-\frac{\eta\gamma}{2}}\left(\phi^{(t)} - \mu\sqrt{(\phi^{(t)})^2+(\theta^{(t)})^2}\right)\\ &\overset{(c)}{\geq} \frac{1-\frac{\gamma^2\eta^2}{4}}{1-\frac{\eta\gamma}{2}}\phi^{(t)} = (1+\frac{\eta\gamma}{2})\phi^{(t)} \end{aligned} \tag{166}$$

where (a) is true because $1-\widehat{\delta}' > 0$, in (b) we used Corollary 2, i.e., $0 < 1-\widehat{\delta}' \leq 1-\frac{\eta\gamma}{2}$, and (c) is true because $\theta^{(t)} \leq \phi^{(t)}$ and

$$\mu = \eta\rho\mathcal{S}\mathcal{P}(4+62\widehat{c}) \overset{(107e)}{\leq} \gamma^2\eta^2\frac{\eta L_{\max}(4+62\widehat{c})}{\log^2(\frac{d\kappa}{\delta})} \overset{(a)}{\leq} \frac{\gamma^2\eta^2}{4\sqrt{2}}$$

where in (a) we choose $c''_{\max} = 1/(4\sqrt{2}(4+62\widehat{c}))$ and $\eta \leq c''_{\max}/L_{\max}$.

Quantifying Escaping Time: From (153), we have

$$\begin{aligned}10\mathcal{S}\widehat{c} \geq &\|\mathbf{v}^{(t)}\| \geq \phi^{(t)} \overset{(166)}{\geq} (1+\frac{\gamma\eta}{2})^t\phi^{(0)} \overset{(a)}{\geq} (1+\frac{\gamma\eta}{2})^t\frac{\delta}{2\sqrt{d}}\frac{\eta L_{\max}\mathcal{S}}{\kappa}\log^{-1}(\frac{d\kappa}{\delta})\\ &\overset{(b)}{\geq} (1+\frac{\gamma\eta}{2})^t\frac{\delta}{2\sqrt{d}}\frac{c\mathcal{S}}{\kappa}\log^{-1}(\frac{d\kappa}{\delta}) \quad \forall t < T\end{aligned} \tag{167}$$

where in (a) we use condition $\upsilon \in [\delta/(2\sqrt{d}), 1]$, in (b) we used $\eta = c/L_{\max}$.

Since (167) is true for all $t < T$, we can have

$$\begin{aligned}T-1 \leq &\frac{\log(20\frac{\widehat{c}}{c}(\frac{\kappa\sqrt{d}}{\delta})\log(\frac{d\kappa}{\delta}))}{\log(1+\frac{\eta\gamma}{2})} \overset{(a)}{<} \frac{4\log(20(\frac{\sqrt{d}\kappa}{\delta})\frac{\widehat{c}}{c}\log(\frac{d\kappa}{\delta}))}{\eta\gamma}\\ &\overset{(b)}{<} \frac{4\log(20(\frac{d\kappa}{\delta})^2\frac{\widehat{c}}{c})}{\eta\gamma} \overset{(c)}{<} 4(2+\log(20\frac{\widehat{c}}{c}))\mathcal{T}\end{aligned} \tag{168}$$

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

From (168), we know that

$$T < 4(2+\log(20\frac{\widehat{c}}{c}))\mathcal{T}+1 \overset{(a)}{<} 4(2\frac{1}{4}+\log(20\frac{\widehat{c}}{c}))\mathcal{T} \tag{169}$$

where (a) is true due to the fact that $\eta L_{\max} \geq 1$ and $\log(d\kappa/\delta) > 1$ so we know $\mathcal{T} \geq 1$.

Applying the proof from (100) to (102), we can also conclude that there exists a universal $\widehat{c}$ such that (169) holds. The proof is complete.

□

D.6 Proof of Lemma 16

First, after the random perturbation, the objective function value in the worst case is increased at most by

$$
\begin{aligned}
f(\mathbf{u}^{(0)}) - f(\widetilde{\mathbf{x}}^{(t)}) &\leq \sum_{k=1}^{2} \nabla_k f(\widetilde{\mathbf{h}}_{-k}^{(t)}, \widetilde{\mathbf{x}}_k^{(t)})^\mathsf{T} \xi_k + \frac{L_k}{2}\|\xi_k\|^2 \\
&\leq \sum_{k=1}^{2} \left(\nabla_k f(\widetilde{\mathbf{h}}_{-k}^{(t)}, \widetilde{\mathbf{x}}_k^{(t)}) - \nabla_k f(\widetilde{\mathbf{h}}_{-k}^{(t)}, \widetilde{\mathbf{x}}_k^{(t+1)})\right)^\mathsf{T} \xi_k + \sum_{k=1}^{2} \nabla_k f(\widetilde{\mathbf{h}}_{-k}^{(t)}, \widetilde{\mathbf{x}}_k^{(t+1)})^\mathsf{T} \xi_k + \frac{L_k}{2}\|\xi_k\|^2 \\
&\leq \sum_{k=1}^{2} L_{\max} \left\|\mathbf{x}_k^{(t+1)} - \mathbf{x}_k^{(t)}\right\| \|\xi_k\| + \sum_{k=1}^{2} \|\nabla_k f(\widetilde{\mathbf{h}}_{-k}^{(t)}, \widetilde{\mathbf{x}}_k^{(t+1)})\| \|\xi_k\| + \frac{L_{\max}}{2}\|\xi\|^2 \\
&\overset{(a)}{\leq} 1.25 \sum_{k=1}^{2} \|\nabla_k f(\widetilde{\mathbf{h}}_{-k}^{(t)}, \widetilde{\mathbf{x}}_k^{(t+1)})\| \|\xi_k\| + \frac{L_{\max}}{2}\|\xi\|^2 \\
&\overset{(b)}{\leq} 1.25\|\xi\| \sqrt{\sum_{k=1}^{2} 2\|\nabla_k f(\widetilde{\mathbf{h}}_{-k}^{(t)}, \widetilde{\mathbf{x}}_k^{(t+1)})\|^2} + \frac{L_{\max}}{2}\|\xi\|^2 \\
&\overset{(c)}{\leq} 1.25 \frac{\mathcal{G}}{\kappa} \frac{\eta L_{\max} \mathcal{S}}{\kappa \log(\frac{d\kappa}{\delta})\mathcal{P}} + \frac{L_{\max}}{2} (\frac{\eta L_{\max} \mathcal{S}}{\kappa \log(\frac{d\kappa}{\delta})\mathcal{P}})^2 \leq \frac{3}{2}\mathcal{F}
\end{aligned} \tag{170}
$$

where $\mathbf{u}^{(0)}$ is a vector that follows uniform distribution within the ball $\mathbb{B}_{\widetilde{\mathbf{x}}^{(t)}}^{(d)}(r)$, $\mathbb{B}_{\widetilde{\mathbf{x}}^{(t)}}^{(d)}$ denotes the d-dimensional ball centered at $\widetilde{\mathbf{x}}^{(t)}$ with radius r, ξ_k represents the kth block of the vector which is the difference between random generated vector $\mathbf{u}^{(0)}$ and saddle point $\widetilde{\mathbf{x}}^{(t)}$, and in (a) we choose $\eta \leq 1/(4L_{\max})$ and (b) is true because $\xi \triangleq [\xi_1, \ldots, \xi_K]$, $\|\xi_k\| \leq \|\xi\|, \forall k$, and in (c) we used $\kappa > 1$, $\log(d\kappa/\delta) > 1$, $\mathcal{P} \geq 2$ and Condition 2 where g_{th} is defined in (119).

Then, the rest of proof of Lemma 16 is the same as the rest of Lemma 10, therefore ignored for simplicity.

E Numerical Results

E.1 Proof of Lemma 4

Proof. Consider function

$$f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} + \frac{1}{4}\|\mathbf{x}\|_4^4 \tag{171}$$

where $\mathbf{x} \in \mathcal{S}$, $\mathcal{S} = \{\mathbf{x} | \|\mathbf{x}\|^2 \leq \tau\}$ and $\tau \geq \lambda_{\max}(\mathbf{A})$.

To prove L-smooth Lipschitz continuity :

$$\begin{aligned}
\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| =& \left\| 2(\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{y}) + \begin{bmatrix} x_1^3 - y_1^3 \\ \vdots \\ x_d^3 - y_d^3 \end{bmatrix} \right\|, \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{S} \\
\leq & 2\lambda_{\max}(\mathbf{A})\|\mathbf{x} - \mathbf{y}\| + \left\| \begin{bmatrix} (x_1 - y_1)(x_1^2 + x_1 y_1 + y_1^2) \\ \vdots \\ (x_d - y_d)(x_d^2 + x_d y_d + y_d^2) \end{bmatrix} \right\| \\
\overset{(a)}{\leq} & 2\tau\|\mathbf{x} - \mathbf{y}\| + 3\tau\|\mathbf{x} - \mathbf{y}\| \leq 5\tau\|\mathbf{x} - \mathbf{y}\|
\end{aligned}$$

where x_i denotes the ith entry of vector $\mathbf{x}$, and (a) is true because

$$x_i^2 \leq \tau, \quad y_i^2 \leq \tau, \quad x_i y_i \leq (x_i^2 + y_i^2)/2 \leq \tau, \forall i. \tag{172}$$

To prove block-wise Lipschitz continuity : Without loss of generality, consider first block $\mathbf{x}_1 \in \mathcal{S}'$ where $\mathcal{S}' = \{\mathbf{x}_1 | \|\mathbf{x}_1\|^2 \leq \tau', \mathbf{x}_1 \in \mathbb{R}^{d' \times 1}\}$ and d' denotes the dimension of $\mathbf{x}_1$. Consider $\tau' \geq \lambda_{\max}(\mathbf{A}')$ where $\mathbf{A}' \in \mathbb{R}^{d' \times d'}$ is the leading principal minor of matrix $\mathbf{A}$ of order d'. Obviously, we have $\tau' \leq \tau$.

$$\begin{aligned}
\|\nabla_1 f(\mathbf{x}_{-1}, \mathbf{x}_1) - \nabla_1 f(\mathbf{x}_{-1}, \mathbf{x}_1')\| =& \left\| 2\mathbf{I}_1' \left(\mathbf{A} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_{-1} \end{bmatrix} - \mathbf{A} \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_{-1} \end{bmatrix} \right) + \begin{bmatrix} x_1^3 - x_1'^3 \\ \vdots \\ x_{d'}^3 - x_{d'}'^3 \end{bmatrix} \right\|, \quad \forall \mathbf{x}, \mathbf{x}' \in \mathcal{S}' \\
\leq & 2\|\mathbf{I}_1' \left(\mathbf{A} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_{-1} \end{bmatrix} - \mathbf{A} \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_{-1} \end{bmatrix} \right) \| + \left\| \begin{bmatrix} (x_1 - x_1')(x_1^2 + x_1 x_1' + x_1'^2) \\ \vdots \\ (x_{d'} - x_{d'}')(x_{d'}^2 + x_{d'} x_{d'}' + x_{d'}'^2) \end{bmatrix} \right\| \\
\overset{(a)}{\leq} & 2\lambda_{\max}(\mathbf{A}')\|\mathbf{x}_1 - \mathbf{x}_1'\| + 3\tau'\|\mathbf{x}_1 - \mathbf{y}_1\| \\
\leq & 5\tau'\|\mathbf{x}_1 - \mathbf{x}_1'\|, \quad \forall \mathbf{x}, \mathbf{x}'
\end{aligned}$$

where (a) is true because we used $\mathbf{I}_1' \triangleq \begin{bmatrix} \mathbf{I}_{d'} & 0 \\ 0 & 0 \end{bmatrix}$ which selects the first d' rows of $\mathbf{A} \left(\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_{-1} \end{bmatrix} - \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_{-1} \end{bmatrix} \right)$.

To prove Hessian Lipschitz continuity :

$$\begin{aligned}
\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\| =& 3 \left\| \begin{bmatrix} x_1^2 - y_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & x_d^2 - y_d^2 \end{bmatrix} \right\| \\
\leq & 6\sqrt{\tau} \left\| \begin{bmatrix} x_1 - y_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & x_d - y_d \end{bmatrix} \right\| = 6\sqrt{\tau}\|\mathbf{x} - \mathbf{y}\|
\end{aligned}$$

where (a) is true because $x_i + y_i \leq \sqrt{(x_i + y_i)^2} = \sqrt{x_1^2 + 2x_i y_i + y_i^2} \overset{(172)}{\leq} 2\sqrt{\tau}, \forall i$.

□

E.2 Additional Simulation

Random matrix A : we also test the algorithms with a randomly generated symmetric matrix $\mathbf{A}$ by the following steps: 1) randomly generate a diagonal matrix $\mathbf{D}$ whose entries follow *i.i.d.* Gaussian distribution with zero mean and variance two; 2) generate an orthogonal matrix $\mathbf{U} \in \mathbb{R}^{d \times d}$; 3) obtain matrix $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^\mathsf{T}$. We initialize the PA-GD/AGD algorithms around the saddle point which is at the origin. The results are shown in Figure 3 where $d = 100$. It can be observed that PA-GD can still escape from the strict saddle point faster than ordinary AGD, illustrating the benefit of adding the random perturbation when the gradient size is small.

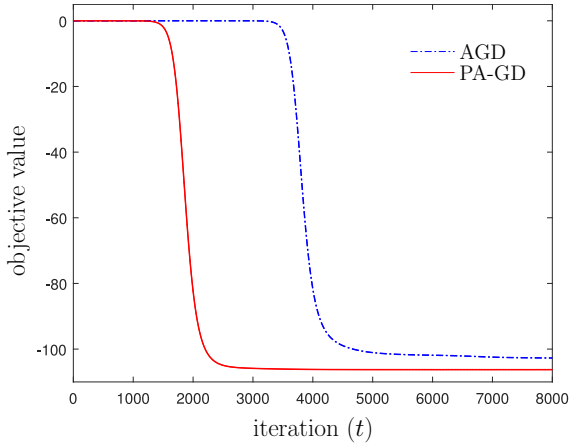


Figure 3: Convergence comparison between AGD and PA-GD, where $d = 100$, $\epsilon = 10^{-4}$, $g_{\mathsf{th}} = \epsilon/10$, $\eta = 1 \times 10^{-3}$, $t_{\mathsf{th}} = 10/\epsilon^{1/3}$, $r = \epsilon/10$.