Provable Accelerated Convergence of Nesterov's Momentum for Deep ReLU Neural Networks

Fangshuo Liao FANGSHUO.LIAO@RICE.EDU
Rice University

Anastasios Kyrillidis ANASTASIOS@RICE.EDU
Rice University

Editors: Claire Vernade and Daniel Hsu

Abstract

Current state-of-the-art analyses on the convergence of gradient descent for training neural networks focus on characterizing properties of the loss landscape, such as the Polyak-Łojaciewicz (PL) condition and the restricted strong convexity. While gradient descent converges linearly under such conditions, it remains an open question whether Nesterov's momentum enjoys accelerated convergence under similar settings and assumptions. In this work, we consider a new class of objective functions, where only a subset of the parameters satisfies strong convexity, and show Nesterov's momentum achieves acceleration in theory for this objective class. We provide two realizations of the problem class, one of which is deep ReLU networks, which constitutes this work as the first that proves an accelerated convergence rate for non-trivial neural network architectures.

Keywords: Momentum, provable acceleration, deep ReLU neural networks

1. Introduction

Training neural networks with gradient-based methods has shown surprising empirical success (Lecun et al., 1998; LeCun et al., 2015; Zhang et al., 2017; Goodfellow et al., 2016); yet, it has been a mystery why such a simple algorithm can consistently find a good minimum for these highly non-convex objectives (Zhang et al., 2018; Li et al., 2020; Yun et al., 2019; Auer et al., 1995; Safran and Shamir, 2018). As a consequence of this mysterious phenomenon, an equally, if not more, intriguing question is *why momentum methods, which are designed originally for accelerating the minimization of convex objectives, can achieve faster convergence speed when applied to complicated non-convex objectives, as that of neural network training.*

The advances of the Neural Tangent Kernel (NTK) (Jacot et al., 2020) promoted the theoretical understanding of neural network training. The use of NTK shows that when the width of the neural network approaches infinity, the training process can be treated as a kernel machine. Inspired by the NTK analysis, a large body of work has focused on showing the convergence of gradient descent for various neural network architectures under finite over-parameterization requirements (Du et al., 2019b,a; Allen-Zhu et al., 2019b; Zou and Gu, 2019; Zhang et al., 2019; Awasthi et al., 2021; Ling et al., 2023; Allen-Zhu et al., 2019a; Song and Yang, 2020; Su and Yang, 2019). Yet, this line of analysis is hard to extend to deeper and more complicated architectures and requires a significantly larger over-parameterization than what is used in practice.

To resolve this limitation, later work started to build connections between the understanding of neural network training and the widely studied optimization theory. In particular, a recent line of work characterizes the loss landscape of neural networks using the local Polyak-Łojaciewicz (PL)

condition (Song et al., 2021; Liu et al., 2020; Nguyen, 2021; Ling et al., 2023). Based on the well-established theory of how gradient descent converges under the PL condition (Karimi et al., 2020), this line of work decouples the neural network structure from the dynamics of the loss function along the optimization trajectory. This way, these works could perform a more fine-grained analysis of the relationship between regulatory conditions, such as the PL condition, and the neural network structure. Such analysis not only resulted in further relaxed over-parameterization requirements (Song et al., 2021; Nguyen, 2021; Liu et al., 2020) but was also shown to be easily extended to deep architectures (Ling et al., 2023), suggesting that it is more suitable in practice.

In contrast to the fast-growing research devoted to (stochastic) gradient descent, there is limited theoretical work on the convergence of momentum methods in deep learning. The acceleration of both the Heavy Ball method and Nesterov's momentum is shown only for shallow ReLU networks (Wang et al., 2021; Liu et al., 2022a) and deep linear networks (Wang et al., 2021; Liu et al., 2022b). It remains an open question to prove the acceleration for neural network training in a scenario closer to what is used in practice in terms of both the over-parameterization requirement and the depth and architecture of the neural network. As a result, we are interested in finding a regulatory condition for neural networks that enables the accelerated convergence for momentum methods.

Showing acceleration under only the PL condition has been a long-standing difficulty. For the Heavy Ball method, Danilova et al. (2018) established a linear convergence rate under the PL condition, but no acceleration is shown without assuming strong convexity. Wang et al. (2022) proved an accelerated convergence rate; yet, the authors assume the $\lambda^\star$-AVERAGE OUT condition, which cannot be easily justified for complicated objectives like neural networks. To our knowledge, the convergence proof for Nesterov's momentum under the PL condition in non-convex settings is currently missing. In the continuous limit, acceleration is proved in a limited scenario (Apidopoulos et al., 2022), which does not easily extend to the discrete case (Shi et al., 2022). Finally, Yue et al. (2022) shows that gradient descent already achieves an optimal convergence rate for functions satisfying smoothness and the PL condition. This suggests that we need to leverage properties beyond the PL condition to prove the acceleration of momentum methods in a broader class of neural networks.

Based on prior work (Liu et al., 2020) that shows over-parameterized systems are essentially non-convex in any neighborhood of the global minimum, we aim at developing a relaxation to the (strong) convexity in the non-convex setting that enables the momentum methods to achieve acceleration. In particular, we consider the minimization of a new class of objectives:

$$\min_{\mathbf{x}\in\mathbb{R}^{d_1},\mathbf{u}\in\mathbb{R}^{d_2}} f(\mathbf{x},\mathbf{u}), \tag{1}$$

where f satisfies the strong convexity with respect to $\mathbf{x}$, among other assumptions (c.f., Assumption 1-6). Intuitively, our construction assumes that the parameter space can be partitioned into two sets, and only one of the two sets enjoys rich properties, such as strong convexity. In this paper, we focus on Nesterov's momentum since it has been shown in a recent work that the Heavy Ball method cannot achieve acceleration even for smooth and strongly convex functions (Goujaud et al., 2023). Indeed, in previous empirical works, Nesterov's momentum is not only shown to achieve acceleration in neural network training (Sutskever et al., 2013), but also demonstrate better performance under large-scale testing than the Heavy Ball method (Dahl et al., 2023).

Our contribution. Our paper starts with an investigation of the properties of the problem class in (1) that satisfies Assumption 1-6. In particular, we show that this set of assumptions is stronger than the PL condition but weaker than strong convexity, and, as a consequence, gradient descent

converges linearly with rate $1-\Theta\left(1/\kappa\right)$ under these assumptions. Next, we prove that Nesterov's momentum enjoys an accelerated linear convergence with a convergence rate $1-\Theta\left(1/\sqrt{\kappa}\right)$. Under Assumption 1-6, our result holds even when f is non-convex and non-smooth:

Theorem 1 (Informal statement of Theorem 7) *Let $f\left(\mathbf{x},\mathbf{u}\right):\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}\to\mathbb{R}$ be L_1-smooth and μ-strongly convex with respect to $\mathbf{x}$ for all $\mathbf{u}\in\mathbb{R}^{d_2}$, and let $\kappa=L_1/\mu$. If $f\left(\mathbf{x},\mathbf{u}\right)$ also satisfies Assumption 3-6 with sufficiently small G_1,G_2 and sufficiently large $R_{\mathbf{x}},R_{\mathbf{u}}$, then the sequence $\left\{\left(\mathbf{x}_k,\mathbf{u}_k\right)\right\}_{k=0}^{\infty}$ generated by Nesterov's momentum satisfies:*

$$f\left(\mathbf{x}_k,\mathbf{u}_k\right)\leq 2\left(1-\Theta\left(\frac{1}{\sqrt{\kappa}}\right)\right)^k\left(f(\mathbf{x}_0,\mathbf{u}_0)-f^*\right).$$

Next, we provide two realizations of our problem class. In Section 4.1, we first consider fitting an additive model under the MSE loss. We prove the acceleration of Nesterov's momentum as long as the non-convex component of the additive model is small enough to guarantee the Lipschitz-type assumptions. Next, we turn to deep ReLU network training in Section 4.2. We show that when the width of the neural network trained with n samples is $\Omega\left(n^4\right)$, under proper initialization, Nesterov's momentum converges to zero training loss with rate $1-\Theta\left(1/\sqrt{\kappa}\right)$. To the best of our knowledge, *this is the first result that establishes accelerated convergence of deep ReLU networks*:

Theorem 2 (Informal statement of Theorem 15) *Given a dataset with n samples and d_0 features, we let $\mathbf{F}$ be a deep ReLU neural network with width $\Omega\left(n^4d_0^2\right)$ and let $\mathcal{L}_k\in\mathbb{R}$ be the MSE loss value at iteration k generated by training $\mathbf{F}$ with Nesterov's momentum. Then, for all $k\geq 0$, we have that:*

$$\mathcal{L}_k\leq 2\left(1-\Theta\left(\frac{1}{\sqrt{\kappa}}\right)\right)^k\mathcal{L}_0.$$

1.1. Related Works

Convergence in neural network training. The NTK-based analysis builds upon the idea that when the width approaches infinity, training neural networks behaves like training a kernel machine. Various techniques are developed to control the error when the width becomes finite. In particular, (Du et al., 2019b) tracks the change of activation patterns in ReLU-based neural networks and often requires a large over-parameterization. Later works improve the over-parameterization requirement by leveraging matrix concentration inequalities (Song and Yang, 2020), performing fine-grained analysis on the change of Jacobians (Oymak and Soltanolkotabi, 2019), analyzing the functional approximation property (Su and Yang, 2019), and building their analysis upon the separability assumption of the data in the reproducing Hilbert space of the neural network (Ji and Telgarsky, 2020). Going beyond two-layer neural networks, Allen-Zhu et al. (2019a); Du et al. (2019a) analyze the convergence of gradient descent on deep neural networks under a large over-parameterization. In the meantime, the analysis was also extended to other training algorithms and settings, such as stochastic gradient descent (Oymak and Soltanolkotabi, 2019; Ji and Telgarsky, 2020; Xu and Zhu, 2021; Zou et al., 2018), drop-out (Liao and Kyrillidis, 2022; Mianjy and Arora, 2020), federated training (Huang et al., 2021), and adversarial training (Li et al., 2022).

A noticeable line of work focuses on establishing that the PL condition is satisfied by neural networks, where the coefficient of the PL condition is based on the eigenvalue of the NTK matrix. Nguyen (2021) shows the PL condition is satisfied by deep ReLU neural networks by considering the dominance of the gradient with respect to the weight in the last layer. Liu et al. (2020) proves the PL condition by upper bounding the Hessian for deep neural networks with smooth activation

functions. Song et al. (2021) further reduces the over-parameterization while maintaining the PL condition via the expansion of the activation function with the Hermite polynomials. Lastly, Banerjee et al. (2023) establishes the restricted strong convexity of neural networks within a sequence of ball-shaped regions centered around the weights per iteration; yet, the coefficient of the strong convexity is not explicitly characterized in theory.

It should be noted that the above work relies on the over-parameterization of the neural network, which, in many cases, leads the training dynamic to be trapped in the so-called kernel regime (Woodworth et al., 2020; Yehudai and Shamir, 2022; Yang and Hu, 2021). While crucial to guarantee a favorable loss landscape (Safran and Shamir, 2018), it is also shown that even mild over-parameterization leads to an exponentially slower convergence rate (Xu and Du, 2023) and cannot explain the behavior of learning a single neuron (Yehudai and Shamir, 2022). However, the above work focuses solely on the training with gradient descent. While our analysis is based on the over-parameterization assumption, it is, to the best of our knowledge, the first to show the convergence of Nesterov's momentum on deep neural networks and opens up the possibility of studying Nesterov's momentum on neural networks in a more realistic scenario.

Convergence of Nesterov's Momentum. The original proof of Nesterov's momentum (Nesterov, 2018) builds upon the idea of estimating sequences for both convex smooth objectives and strongly convex smooth objectives. Later work in (Bansal and Gupta, 2019) provides an alternate proof within the same setting by constructing a Lyapunov function. In the non-convex setting, a large body of works focuses on variants of Nesterov's momentum that lead to a guaranteed convergence by employing techniques such as negative curvature exploitation (Carmon et al., 2017), cubic regularization (Carmon and Duchi, 2020), and restarting schemes (Li and Lin, 2022). For neural networks, (Liu et al., 2022a,b) are the only works that study the convergence of Nesterov's momentum. However, considering the over-parameterization requirement, the objective is similar to a quadratic function. Deviating from Nesterov's momentum, Wang et al. (2021) studies the convergence of the Heavy-ball method under similar over-parameterization requirements. A recent work (Wu et al., 2023) proves the convergence of the Heavy-ball method under the mean-field limit; such a limit is not the focus of our study in this paper. Lastly, Jelassi and Li (2022) shows that momentum-based methods improve the generalization ability of neural networks. However, there is no explicit convergence guarantee for the training loss.

2. Problem Setup and Assumptions

Notations Standard lower-case letters (e.g. a) denote scalars, bold lower-case letters (e.g. $\mathbf{a}$) denote vectors, and bold upper-case letters (e.g. $\mathbf{A}$) denote matrices. For a vector $\mathbf{a}$, we use a_i to denote its i-th entry and $\|\mathbf{a}\|_2$ its ℓ_2-norm. For a matrix $\mathbf{A}$, we use a_{ij} to denote its (i,j)-th entry and $\|\mathbf{A}\|_F$ its Frobenius norm. we use $(\mathbf{a}_1, \mathbf{a}_2)$ to denote the concatenation of $\mathbf{a}_1, \mathbf{a}_2$. For a matrix $\mathbf{A}$ with columns $\mathbf{a}_1, \ldots, \mathbf{a}_n$, we use $\mathtt{V}(\mathbf{A}) = (\mathbf{a}_1, \ldots, \mathbf{a}_n)$ to denote the vectorized form of $\mathbf{A}$.

Optimization literature often focuses on the constraint-free minimization of a function $\hat{f} : \mathbb{R}^d \to \mathbb{R}$. In this scenario, Nesterov's momentum with step size η and momentum parameter β for minimizing $\hat{f}(\mathbf{w})$ bears the form, as in Bansal and Gupta (2019) and (2.2.22) of Nesterov (2018)[1]

$$\mathbf{w}_{k+1} = \bar{\mathbf{w}}_k - \eta \nabla \hat{f}(\bar{\mathbf{w}}); \quad \bar{\mathbf{w}}_{k+1} = \mathbf{w}_{k+1} + \beta(\mathbf{w}_{k+1} - \mathbf{w}_k) \tag{2}$$

In this paper, we reformulate this problem using the following definition.

1. Despite a different choice of step size and momentum parameter.

Definition 1 *A function $f : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ is called a partitioned equivalence of $\hat{f} : \mathbb{R}^{d} \to \mathbb{R}$, if $i)$ $d_1 + d_2 = d$, and $ii)$ there exists a permutation function $\pi : \mathbb{R}^{d} \to \mathbb{R}^{d}$ over the parameters of $\hat{f}$, such that $\hat{f}(\mathbf{w}) = f(\mathbf{x}, \mathbf{u})$ if and only if $\pi(\mathbf{w}) = (\mathbf{x}, \mathbf{u})$. We say that $(\mathbf{x}, \mathbf{u})$ is a partition of $\mathbf{w}$.*

Despite the difference in the representation of their parameters, $\hat{f}$ and f share the same properties, and any algorithm for $\hat{f}$ would produce the same result for f. Therefore, we turn our focus from the minimization problem of $\hat{f}$ to the minimization problem in (1). We should clarify that when we study the property of f as an attempt to study the property of $\hat{f}$, *we only need to assume the existence of such a partitioned equivalence*, instead of requiring an efficient algorithm to identify this equivalence explicitly. We further assume that f is a composition of a loss function $g : \mathbb{R}^{\hat{d}} \to \mathbb{R}$ and a possibly non-smooth and non-convex model function $h : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}^{\hat{d}}$, for some dimension $\hat{d} \in \mathbb{Z}_+$; i.e., $f(\mathbf{x}, \mathbf{u}) = g(h(\mathbf{x}, \mathbf{u}))$. With this construction of functional composition, we can assume only a partial smoothness on f together with the smoothness of g, instead of the full smoothness property of f. We obey the following notation with respect to gradients of f:

$$\nabla_1 f(\mathbf{x}, \mathbf{u}) = \tfrac{\partial f(\mathbf{x},\mathbf{u})}{\partial \mathbf{x}};\ \nabla_2 f(\mathbf{x}, \mathbf{u}) = \tfrac{\partial f(\mathbf{x},\mathbf{u})}{\partial \mathbf{u}};\ \nabla f(\mathbf{x}, \mathbf{u}) = (\nabla_1 f(\mathbf{x}, \mathbf{u}), \nabla_2 f(\mathbf{x}, \mathbf{u}))\,. \tag{3}$$

We will consider Nesterov's momentum with constant step size η and momentum parameter β:

$$\begin{aligned}(\mathbf{x}_{k+1}, \mathbf{u}_{k+1}) &= (\mathbf{y}_k, \mathbf{v}_k) - \eta \nabla f(\mathbf{y}_k, \mathbf{v}_k)\\ (\mathbf{y}_{k+1}, \mathbf{v}_{k+1}) &= (\mathbf{x}_{k+1}, \mathbf{u}_{k+1}) + \beta\left((\mathbf{x}_{k+1}, \mathbf{u}_{k+1}) - (\mathbf{x}_k, \mathbf{u}_k)\right)\end{aligned} \tag{4}$$

with $\mathbf{y}_0 = \mathbf{x}_0$ and $\mathbf{v}_0 = \mathbf{u}_0$. The algorithm formulation in (4) is mathematically equivalent to (2) for optimizing $\hat{f}(\mathbf{x})$. Therefore, the execution of (4) is completely agnostic to the parameter partition. To state our assumptions, let $\mathcal{B}^{(1)}_{R_\mathbf{x}}$ and $\mathcal{B}^{(2)}_{R_\mathbf{u}}$ denote the balls centered as $\mathbf{x}_0$ and $\mathbf{u}_0$:

$$\mathcal{B}^{(1)}_{R_\mathbf{x}} = \{\mathbf{x} \in \mathbb{R}^{d_1} : \|\mathbf{x} - \mathbf{x}_0\|_2 \le R_\mathbf{x}\}; \quad \mathcal{B}^{(2)}_{R_\mathbf{u}} = \{\mathbf{u} \in \mathbb{R}^{d_2} : \|\mathbf{u} - \mathbf{u}_0\|_2 \le R_\mathbf{u}\}.$$

Next, we state the assumptions on the general class of objectives we consider.

Assumption 1 *f is μ-strongly convex with $\mu > 0$ with respect to the first part of its parameters:*

$$f(\mathbf{y}, \mathbf{u}) \ge f(\mathbf{x}, \mathbf{u}) + \langle \nabla_1 f(\mathbf{x}, \mathbf{u}), \mathbf{y} - \mathbf{x}\rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|_2^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^{d_1};\ \mathbf{u} \in \mathcal{B}^{(2)}_{R_\mathbf{u}}.$$

Assumption 2 *f is L_1-smooth with respect to the first part of its parameters:*

$$f(\mathbf{y}, \mathbf{u}) \le f(\mathbf{x}, \mathbf{u}) + \langle \nabla_1 f(\mathbf{x}, \mathbf{u}), \mathbf{y} - \mathbf{x}\rangle + \frac{L_1}{2}\|\mathbf{y} - \mathbf{x}\|_2^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^{d_1};\ \mathbf{u} \in \mathcal{B}^{(2)}_{R_\mathbf{u}}.$$

Based on Assumption 1 and 2, we define the condition number of f.

Definition 2 *(Condition Number) The condition number κ of f is given by $\kappa = {}^{L_1}\!/_{\mu}$.*

Assumption 3 *g satisfies $\min_{\mathbf{s}\in\mathbb{R}^{\hat{d}}} g(\mathbf{s}) = \min_{\mathbf{x}\in\mathbb{R}^{d_1}, \mathbf{u}\in\mathbb{R}^{d_2}} f(\mathbf{x}, \mathbf{u})$, and is L_2-smooth:*

$$g(\mathbf{s}_1) \le g(\mathbf{s}_2) + \langle \nabla g(\mathbf{s}_1), \mathbf{s}_2 - \mathbf{s}_1\rangle + \frac{L_2}{2}\|\mathbf{s}_2 - \mathbf{s}_1\|_2^2, \quad \forall \mathbf{s}_1, \mathbf{s}_2 \in \mathbb{R}^{\hat{d}}.$$

Assumptions 1 and 2 are relaxed versions of the smoothness and strong convexity. Instead of assuming that the objective is smooth and strongly convex over all parameters, we only assume such property to hold with respect to a subset of the parameters while the rest lie near initialization. Assumption 3 is standard in prior literature (Liu et al., 2020; Song et al., 2021) of neural network training, and holds for loss functions such as the MSE loss and the logistic loss.

Assumption 4 *h satisfies G_1-Lipschitzness with respect to the second part of its parameters:*

$$\|h(\mathbf{x},\mathbf{u}) - h(\mathbf{x},\mathbf{v})\|_2 \leq G_1 \|\mathbf{u}-\mathbf{v}\|_2, \quad \forall \mathbf{x} \in \mathcal{B}^{(1)}_{R_\mathbf{x}};\ \mathbf{u},\mathbf{v} \in \mathcal{B}^{(2)}_{R_\mathbf{u}}.$$

Assumption 5 *The gradient of f with respect to the first part of its parameter, $\nabla_1 f(\mathbf{x},\mathbf{u})$, satisfies G_2-Lipschitzness with respect to the second part of its parameters:*

$$\|\nabla_1 f(\mathbf{x},\mathbf{u}) - \nabla_1 f(\mathbf{x},\mathbf{v})\|_2 \leq G_2 \|\mathbf{u}-\mathbf{v}\|_2, \quad \forall \mathbf{x} \in \mathcal{B}^{(1)}_{R_\mathbf{x}};\ \mathbf{u},\mathbf{v} \in \mathcal{B}^{(2)}_{R_\mathbf{u}}.$$

Assumption 6 *Minimum values of f restricted to the optimization over $\mathbf{x}$ equal the global minimum value:*

$$\min_{\mathbf{x}\in\mathbb{R}^{d_1}} f(\mathbf{x},\mathbf{u}) = f^\star := \min_{\mathbf{x}\in\mathbb{R}^{d_1},\mathbf{u}\in\mathbb{R}^{d_2}} f(\mathbf{x},\mathbf{u}); \quad \forall \mathbf{u} \in \mathcal{B}^{(2)}_{R_\mathbf{u}}.$$

Since we do not assume f to be convex or smooth with respect to $\mathbf{u}$, we cannot guarantee that the updates in (4) on $\mathbf{u}$ will make positive progress towards finding the global minimum. Nevertheless, the updates on $\mathbf{u}$ are unavoidable since the execution of (4) is agnostic to the parameter partition. Therefore, we treat the change in the second part of the parameters as errors induced by the updates. Assumptions 4 and 5 are made to control the effect on the change of the model output $h(\mathbf{x},\mathbf{u})$ and the gradient with respect to $\mathbf{x}$ caused by the change of $\mathbf{u}$. Moreover, without Assumption 6, it is possible that the change of $\mathbf{u}$ will lead the optimization trajectory to some local minimum of $\mathbf{u}$, such that the global minimum value cannot be achieved even when $\mathbf{x}$ is fully optimized. We show that Assumptions 1-6 are satisfied by a smooth and strongly convex function:

Theorem 3 *Let $\tilde{f}$ be $\tilde{\mu}$-strongly convex and $\tilde{L}$-smooth. Then $\tilde{f}$ satisfies Assumptions 1-6 with:*

$$R_\mathbf{x} = R_\mathbf{u} = \infty;\ \mu = \tilde{\mu};\ L_1 = L_2 = \tilde{L};\ G_1 = G_2 = 0.$$

Theorem 3 shows that the combination of Assumptions 1-6 is no stronger than the assumption that the objective is smooth and strongly convex. Therefore, the minimization of the class of functions satisfying Assumptions 1-6 does not have a better lower complexity bound than the class of smooth and strong convex functions. That is, the best convergence rate we can achieve is $1 - \Theta\left(1/\sqrt{\kappa}\right)$.

3. Accelerated Convergence under Partial Strong Convexity

3.1. Warmup: Convergence of Gradient Descent

The previous section shows that f satisfying Assumption 1-6 is weaker than the combination of smoothness and strong convexity. Before diving into the convergence of gradient descent, we first show that Assumptions 1,6 imply the PL condition:

Lemma 4 *Suppose that Assumption 1, 6 hold. Then, for all $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{u} \in \mathcal{B}^{(2)}_{R_\mathbf{u}}$, we have:*

$$\|\nabla f(\mathbf{x},\mathbf{u})\|_2^2 \geq \|\nabla_1 f(\mathbf{x},\mathbf{u})\|_2^2 \geq 2\mu\left(f(\mathbf{x},\mathbf{u}) - f^\star\right).$$

Recall that, due to the minimum assumption made on the relationship between $f(\mathbf{x}, \mathbf{u})$ and $\mathbf{u}$, we treat the change of $\mathbf{u}$ during the iterates as an error. Thus, we need the following lemma, which bounds how much f is affected by the change of $\mathbf{u}$.

Lemma 5 *Let Assumptions 3, 4 hold. For any $\hat{\mathcal{Q}} > 0$ and $\mathbf{x} \in \mathcal{B}_{R_\mathbf{x}}^{(1)}, \mathbf{u}, \mathbf{v} \in \mathcal{B}_{R_\mathbf{u}}^{(2)}$, we have:*

$$f(\mathbf{x}, \mathbf{u}) - f(\mathbf{x}, \mathbf{v}) \leq \hat{\mathcal{Q}}^{-1} L_2 \left(f(\mathbf{x}, \mathbf{v}) - f^\star\right) + \frac{G_1^2}{2} \left(L_2 + \hat{\mathcal{Q}}\right) \|\mathbf{u} - \mathbf{v}\|_2^2 .$$

With the help of Lemmas 4 and 5, we can show the linear convergence of gradient descent:

Theorem 6 *Suppose that Assumptions 1-4 and 6 hold with $G_1^4 \leq \frac{\mu^2}{8L_2^2}$ and*

$$R_\mathbf{x} \geq 16\eta\kappa\sqrt{L_1}\left(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star\right)^{\frac{1}{2}}; \quad R_\mathbf{u} \geq 16\eta\kappa G_1\sqrt{L_2}\left(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star\right)^{\frac{1}{2}} .$$

Then there exists constant $c > 0$ such that gradient descent with $\eta = \frac{c}{L_1}$ converges according to:

$$f(\mathbf{x}_k, \mathbf{u}_k) - f^\star \leq \left(1 - \tfrac{c}{4\kappa}\right)^k \left(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star\right).$$

I.e., Theorem 6 shows that gradient descent applied to f converges linearly with a rate of $1 - \Theta(1/\kappa)$ within our settings. The proofs for Lemma 4 and 5, and Theorem 6 are deferred to Appendix B.

3.2. Acceleration of Nesterov's Momentum

We will now study the convergence property of Nesterov's momentum in (4) under only Assumptions 1-6. Our result shows an accelerated convergence rate compared with gradient descent.

Theorem 7 *Let Assumptions (1)-(6) hold. Consider Nesterov's momentum given by (4) with initialization $\{\mathbf{x}_0, \mathbf{u}_0\} = \{\mathbf{y}_0, \mathbf{v}_0\}$. There exists absolute constants $c, C_1, C_2 > 0$, such that, if μ, L_1, L_2, G_1, G_2 and $R_\mathbf{x}, R_\mathbf{u}$ satisfy:*

$$\begin{aligned}
&G_1^4 \leq \tfrac{C_1\mu^2}{L_2(L_2+1)^2}\left(\tfrac{1-\beta}{1+\beta}\right)^3; \quad G_1^2 G_2^2 \leq \tfrac{C_2\mu^3}{L_2(L_2+1)\sqrt{\kappa}}\left(\tfrac{1-\beta}{1+\beta}\right)^2; \\
&R_\mathbf{x} \geq \frac{36}{c}\sqrt{\kappa}\left(\tfrac{\eta(L_2+1)}{1-\beta}\right)^{\frac{1}{2}}\left(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star\right)^{\frac{1}{2}}; \\
&R_\mathbf{u} \geq \frac{36}{c}\sqrt{\kappa}\left(\tfrac{\eta G_1^2 L_2(L_2+1)(1+\beta)^3}{\mu\beta(1-\beta)^3}\right)^{\frac{1}{2}}\left(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star\right)^{\frac{1}{2}},
\end{aligned} \tag{5}$$

and, if we choose $\eta = c/L_1$, $\beta = (4\sqrt{\kappa} - \sqrt{c})/(4\sqrt{\kappa} + 7\sqrt{c})$, then $\mathbf{x}_k, \mathbf{y}_k \in \mathcal{B}_{R_\mathbf{x}}^{(1)}$ and $\mathbf{u}_k, \mathbf{v}_k \in \mathcal{B}_{R_\mathbf{u}}^{(2)}$ for all $k \in \mathbb{N}$, and Nesterov's recursion converges according to:

$$f(\mathbf{x}_k, \mathbf{u}_k) - f^\star \leq 2\left(1 - \tfrac{c}{4\sqrt{\kappa}}\right)^k \left(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star\right). \tag{6}$$

Theorem 7 shows that, under Assumptions 1-6 with a sufficiently small G_1 and G_2 as in (5), Nesterov's iteration enjoys an accelerated convergence, as in (6). Moreover, the iterates of Nesterov's momentum $\{(\mathbf{x}_k, \mathbf{y}_k)\}_{k=1}^\infty$ and $\{(\mathbf{u}_k, \mathbf{v}_k)\}_{k=1}^\infty$ stay in a ball around initialization with radius in (5).

To better interpret our result, we first focus on (5). By our choice of β, we have that $1 - \beta = \Theta\left(1/\sqrt{\kappa}\right)$ and $1 + \beta = \Theta\left(1\right)$. Therefore, the requirement of G_1, G_2 in (5) can be simplified to $G_1^4 \leq O\left(\mu^{7/2}/L_1^{3/2}L_2^3\right)$ and $G_1^2 G_2^2 \leq O\left(\mu^{9/2}/L_1^{3/2}L_2^2\right)$. This simplified condition implies that we need a smaller G_1 and G_2 if μ is small and L_1 and L_2 are large. For the requirement on $R_\mathbf{x}$ and $R_\mathbf{u}$

in (5), we can simplify with $\eta = O\left(1/L_1\right)$ and $\beta = \Theta\left(1\right)$. In this way, $R_{\mathbf{x}}$ and $R_{\mathbf{u}}$ reduce to $\Omega\left(L_1^{1/4}L_2^{1/2}/\mu^{3/4}\right) \cdot \left(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star\right)^{\frac{1}{2}}$ and $\Omega\left(G_1 L_1^{3/4} L_2/\mu^{7/4}\right) \cdot \left(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star\right)^{\frac{1}{2}}$, respectively. Both quantities grow with a larger L_1 and L_2 and a smaller μ. Noticeably $R_{\mathbf{u}}$ also scales with G_1. Focusing on the convergence property in (6), we can conclude that Nesterov's momentum achieves an accelerated convergence rate of $1 - \Theta\left(1/\sqrt{\kappa}\right)$ compared with the $1 - \Theta\left(1/\kappa\right)$ rate in Theorem 6. In more detail, we discuss the proof of Theorem 7 in the sections below.

3.3. Technical Difficulty

Similar to the previous work on showing the convergence of Nesterov's momentum (Bansal and Gupta, 2019; d'Aspremont et al., 2021), the core of our proof is the construction of a Lyapunov function that upper bounds the optimality gap $f(\mathbf{x}_k, \mathbf{u}_k) - f^*$ at each step k, and enjoys a linear convergence. However, the construction of this Lyapunov function faces the following difficulty.

Difficulty 1. *Most previous analyses of Nesterov's momentum use the global minimum as a reference point to construct the Lyapunov function; see Bansal and Gupta (2019); d'Aspremont et al. (2021). In the original proof of Nesterov, the construction of the estimating sequence also assumes the existence of a unique global minimum (Nesterov, 2018). However, in our scenario, the objective function is non-convex. It allows the existence of multiple global minima, which prevents us from directly applying the Lyapunov function or estimating sequence, as in previous works.*

While the non-convexity of f introduces the possibility of multiple global minima, Assumption 1 implies that, with a fixed $\mathbf{u}$, there exists a unique $\mathbf{x}^\star(\mathbf{u})$ that minimizes $f(\mathbf{x}, \mathbf{u})$. Moreover, Assumption 6 implies that, for all $\mathbf{u} \in \mathcal{B}^{(2)}_{R_{\mathbf{u}}}$, the local minimum $(\mathbf{x}^\star(\mathbf{u}), \mathbf{u})$ is also a global minimum. Thus, we resolve the difficulty by using the $\mathbf{x}^\star(\mathbf{u}_k)$ as the reference point for the Lyapunov function at the kth iteration and ensure the stability of the Lyapunov function by bounding the change of $\mathbf{x}^\star(\mathbf{u}_k)$. The following lemma gives a characterization of this property.

Lemma 8 *Let $\mathbf{x}^\star(\mathbf{u}) = \arg\min_{\mathbf{x}\in\mathbb{R}^{d_1}} f(\mathbf{x}, \mathbf{u})$. Suppose Assumptions 1 and 5 hold. Then, we have:*

$$\left\|\mathbf{x}^\star(\mathbf{u}_1) - \mathbf{x}^\star(\mathbf{u}_2)\right\|_2 \leq \frac{G_2}{\mu}\left\|\mathbf{u}_1 - \mathbf{u}_2\right\|_2, \quad \forall \mathbf{u}_1, \mathbf{u}_2 \in \mathcal{B}^{(2)}_{R_{\mathbf{u}}}.$$

Lemma 8 indicates that, if we view $\mathbf{x}^\star(\mathbf{u})$ as a function of $\mathbf{u}$, then this function is $\frac{G_2}{\mu}$-Lipschitz. For a fixed $\mathbf{u}$, given the nice properties on $\mathbf{x}$, the iterates of Nesterov's momentum will guide $\mathbf{x}$ to the minimum, based on the current $\mathbf{u}$. Lemma 8 guarantees that the progress toward the minimum induced by $\mathbf{u}_1$ does not deviate much from the progress toward the minimum induced by $\mathbf{u}_2$. However, to apply Lemma 8, we must control the change of $\mathbf{u}_k$ between iterations. Indeed, this bound is also necessary to apply the smoothness-like condition in Lemma 5. Unlike gradient descent, $\{\mathbf{u}_k\}_{k=1}^{\infty}$ generated by Nesterov's momentum introduces the following difficulty.

Difficulty 2. *Unrolling the Nesterov's momentum iterates shows that $\mathbf{u}_{k+1} - \mathbf{u}_k$ is a linear combination of previous gradients. Under the assumption of smoothness, the norm of the gradient is bounded by a factor times the optimality gap at the current point, namely*:[2]

$$\left\|\nabla_2 f\left(\mathbf{x}, \mathbf{u}\right)\right\|_2^2 \leq 2G_1 L_2\left(f\left(\mathbf{x}, \mathbf{u}\right) - f^*\right).$$

2. For the proof of this property, please see Lemma 20

In the case of gradient descent, the applied gradients are evaluated at steps $(\mathbf{x}_k, \mathbf{u}_k)$*, and thus* $\|\nabla_2 f(\mathbf{x}_k, \mathbf{u}_k)\|_2$ *can be controlled since* $(f(\mathbf{x}, \mathbf{u}) - f^*)$ *can be shown to enjoy a linear convergence by an induction-based argument (Du et al., 2019b; Nguyen, 2021). However, in the case of Nesterov's momentum, we cannot directly utilize this relationship since the applied gradient is evaluated at the intermediate step* $(\mathbf{y}_k, \mathbf{v}_k)$*, and while we know that the optimality gap at* $(\mathbf{x}_k, \mathbf{u}_k)$ *converges linearly, we have very little knowledge about the optimality gap at* $(\mathbf{y}_k, \mathbf{v}_k)$*.*

To tackle this difficulty, our analysis starts with a careful bound on $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2$ by utilizing the convexity with respect to $\mathbf{x}$ to characterize the inner product $\langle \nabla_1 f(\mathbf{y}_k, \mathbf{v}_k), \mathbf{x}_k - \mathbf{x}_{k-1} \rangle$. After that, we bound $\|\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k)\|_2^2$ using a combination of $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2$ and $\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_2^2$. Lastly, we relate $\|\nabla_2 f(\mathbf{y}_k, \mathbf{v}_k)\|_2^2$ to $\|\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k)\|_2^2$ using the following gradient dominance property.

Lemma 9 *Suppose that Assumptions 1, 3, 4, and 6 hold. Then, we have:*

$$\|\nabla_2 f(\mathbf{x}, \mathbf{u})\|_2^2 \leq \tfrac{G_1^2 L_2}{\mu} \|\nabla_1 f(\mathbf{x}, \mathbf{u})\|_2^2, \quad \forall \mathbf{x} \in \mathcal{B}_{R_\mathbf{x}}^{(1)};\ \mathbf{u} \in \mathcal{B}_{R_\mathbf{u}}^{(2)}.$$

Lemma 9 establishes the upper bound of $\|\nabla_2 f(\mathbf{x}, \mathbf{u})\|_2^2$ using $\|\nabla_1 f(\mathbf{x}, \mathbf{u})\|_2^2$. Direct application of this result will contribute to the bound on $\|\mathbf{u}_{k+1} - \mathbf{u}_k\|_2$. Intuitively, this lemma also implies that the effect of gradient update on $\mathbf{u}$ is less significant than that on $\mathbf{x}$.

3.4. Proof Overview

Our analysis is based on the Lyapunov function proof given by Bansal and Gupta (2019) for proving the accelerated convergence rate of Nesterov's momentum in (2) in minimizing a strongly convex and smooth objective $\hat{f}(\mathbf{w})$. In particular, Bansal and Gupta (2019)[3] uses the following Lyapunov function,

$$\hat{\phi}_k = f(\mathbf{w}_k) - f^* + \frac{\mu}{2} \|\hat{\mathbf{z}}_k - \mathbf{w}^\star\|_2 \tag{7}$$

where $\mathbf{w}^\star$ is the global minimum, and $\hat{\mathbf{z}}_k$ can be considered as a mixing of the sequences $\{\mathbf{w}_k\}_{k=1}^\infty$ and $\{\bar{\mathbf{w}}_k\}_{k=1}^\infty$ in (2). In particular, the second term in (7) computes the distance between the mixed variable $\hat{\mathbf{z}}_k$ and the reference point $\mathbf{w}^\star$ and is added to $\hat{\phi}_k$ to guarantee a linear convergence on $\hat{\phi}_k$. Following our discussion in the previous section, we construct our Lyapunov function using $\mathbf{x}^\star(\mathbf{u}) = \arg\min_{\mathbf{x} \in \mathbb{R}^{d_1}} f(\mathbf{x}, \mathbf{u})$ as a reference point. For the simplicity of notations, we define $\mathbf{x}_k^\star = \mathbf{x}^\star(\mathbf{u}_k)$. Similar to Bansal and Gupta (2019), we let $\mathbf{z}_k$ be the linear combination of $\mathbf{y}_k$ and $\mathbf{x}_k$, and choose a proper scaling factor $\mathcal{Q}_1$ for the distance between $\mathbf{z}_k$ and the previous reference point $\mathbf{x}_{k-1}^\star$. For some properly choose γ and λ, we define

$$\mathbf{z}_k = \frac{1 - \beta\lambda}{\beta\lambda}(\mathbf{y}_k - \mathbf{x}_k) + \mathbf{y}_k; \quad \mathcal{Q}_1 = \frac{\lambda^2}{2\eta(1+\gamma)^5}$$

Compared with the proof of Nesterov's momentum on smooth and strongly convex functions, the proof in our setting has to accommodate the errors caused by the change of $\mathbf{u}$. Our complicated scaling in the form of $\mathbf{z}_k$ and $\mathcal{Q}_1$ is to make sure the errors caused by $\mathbf{u}$ can be properly canceled out by the positive progress made by updating $\mathbf{x}$. Setting $\mathbf{y}_{-1} = \mathbf{y}_0$ and $\mathbf{v}_{-1} = \mathbf{v}_0$, we consider the following Lyapunov function:

$$\phi_k = f(\mathbf{x}_k, \mathbf{u}_k) - f^\star + \mathcal{Q}_1 \left\|\mathbf{z}_k - \mathbf{x}_{k-1}^\star\right\|_2^2 + \frac{\eta}{8} \|\nabla_1 f(\mathbf{y}_{k-1}, \mathbf{v}_{k-1})\|_2^2$$

3. after rephrasing in the notations used in our scenario

The last term in the expression of ϕ_k also eliminates the errors from updating $\mathbf{u}$. Our proof recursively establishes the following three properties:

$$\left(1-\tfrac{c}{2\sqrt{\kappa}}\right)^{-1}\phi_{k+1}-\phi_k \le \frac{c}{4\sqrt{\kappa}}\left(1-\tfrac{c}{4\sqrt{\kappa}}\right)^{k}\phi_0; \tag{8}$$

$$\|\mathbf{x}_k-\mathbf{x}_{k-1}\|_2^2 \le \mathcal{Q}_2\left(1-\tfrac{c}{4\sqrt{\kappa}}\right)^{k}\phi_0; \quad \|\mathbf{u}_k-\mathbf{u}_{k-1}\|_2^2 \le G_1^2\mathcal{Q}_3\left(1-\tfrac{c}{4\sqrt{\kappa}}\right)^{k}\phi_0. \tag{9}$$

Intuitively, (8) implies an accelerated linear convergence of $f(\mathbf{x}_{k'},\mathbf{u}_{k'})-f^\star$ up to $k'\le k+1$, which further implies the bound on $\|\mathbf{x}_{k'}-\mathbf{x}_{k'-1}\|_2$ and $\|\mathbf{u}_{k'}-\mathbf{u}_{k'-1}\|_2$ as in (9) up to $k'\le k+1$. In turn, (9) will guarantee that $\mathbf{x}_k\in\mathcal{B}_{R_\mathbf{x}}, \mathbf{u}\in\mathcal{B}_{R_\mathbf{u}}$, and control the error caused by updating $\mathbf{u}$. These two conditions combined will imply that (8) holds. This idea is further detailed below.

By our choice of ϕ_k, we must have that $f(\mathbf{x}_k,\mathbf{u}_k)-f^\star\le\phi_k$. Unrolling (8) implies that:

$$f(\mathbf{x}_k,\mathbf{u}_k)-f^\star\le\phi_k\le\left(1-\tfrac{c}{4\sqrt{\kappa}}\right)^{k}\phi_0, \tag{10}$$

Combined with the bound that $\phi_0\le 2\left(f(\mathbf{x}_0,\mathbf{u}_0)-f^\star\right)$[4], (8) further implies the convergence in (6). The following lemma shows that, if (6) holds for all $k\le\hat{k}$, we can guarantee (9) with $k=\hat{k}+1$.

Lemma 10 *Let the Assumptions of Theorem 7 hold. If (10) holds for all $k\le\hat{k}$, then (9) holds for $k=\hat{k}+1$ with $\mathcal{Q}_2=\frac{6\eta(L_2+1)}{1-\beta}$ and $\mathcal{Q}_3=\frac{6\eta L_2(L_2+1)(1+\beta)^3}{\mu\beta(1-\beta)^3}$.*

The proof of Lemma 10 utilizes how we resolve Difficulty 2 in the previous section. Lemma 10 implies that up to iteration $\hat{k}+1$, $\mathbf{x}_k$ and $\mathbf{u}_k$ stay in $\mathcal{B}^{(1)}_{R_\mathbf{x}}$ and $\mathcal{B}^{(2)}_{R_\mathbf{u}}$ with the lower bound of $R_\mathbf{x}$ and $R_\mathbf{u}$ assumed in Theorem 7. This allows us to apply Assumptions 1-6 in showing (8) for iteration $\hat{k}+1$. Moreover, the bound on $\|\mathbf{u}_k-\mathbf{u}_{k-1}\|_2$ connects the following lemma to (8).

Lemma 11 *Let the Assumptions of Theorem (7) hold. Then, we have:*

$$\left(1-\tfrac{c}{2\sqrt{\kappa}}\right)^{-1}\phi_{k+1}-\phi_k\le c\beta^2\sqrt{\kappa}\left(G_1^2L_2+\tfrac{8\mathcal{Q}_1G_1^2G_2^2}{\mu^2}\right)\|\mathbf{u}_k-\mathbf{u}_{k-1}\|_2^2$$

The proof of Lemma 11 has a similar idea to Bansal and Gupta (2019) when showing that the Lyapunov function converges. The additional difficulty of Lemma 11 lies in carefully controlling the errors caused by the change of $\mathbf{u}_k$. Plugging the form of $\mathcal{Q}_3$ into the upper bound in (9) and then plugging the resulting upper bound into Lemma 11 yields

$$\left(1-\tfrac{c}{2\sqrt{\kappa}}\right)^{-1}\phi_{k+1}-\phi_k\le\frac{6c\sqrt{\kappa}\eta L_2(L_2+1)(1+\beta)^3}{\mu(1-\beta)^3}\left(G_1^4L_2+\tfrac{8\mathcal{Q}_1G_1^4G_2^2}{\mu^2}\right)\left(1-\tfrac{c}{4\sqrt{\kappa}}\right)^{k}\phi_0$$

The equation above directly implies (8) after we plug in the upper bound on G_1 and G_2 assumed in Theorem 7. This establishes the inductive step and thus finishes the proof.

4. Realization of Assumption 1-6

4. For the detail please see Lemma 21

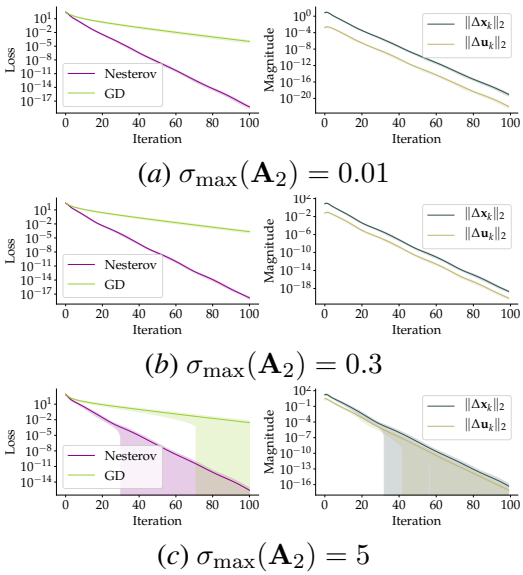
We consider two realizations of problems that satisfy Assumptions 1-6. By enforcing the requirement in Theorem 7 on the two models, we show that although nonconvex and possibly non-smooth, the two models enjoy accelerated convergence when trained with Nesterov's momentum.

4.1. Additive model

Given matrices $\mathbf{A}_1 \in \mathbb{R}^{m\times m}, \mathbf{A}_2 \in \mathbb{R}^{m\times d}$ and a non-linear function $\sigma : \mathbb{R}^m \to \mathbb{R}^m$, we consider $h(\mathbf{x},\mathbf{u}) = \mathbf{A}_1\mathbf{x} + \sigma(\mathbf{A}_2\mathbf{u})$ as the summation of a linear model and a non-linear model. If we train $h(\mathbf{x},\mathbf{u})$ over the loss $g(\mathbf{s}) = \frac{1}{2}\|\mathbf{s}-\mathbf{b}\|_2^2$ for some label $\mathbf{b}\in\mathbb{R}^m$, the objective can be written as

$$\begin{aligned} f(\mathbf{x},\mathbf{u}) &= g(\mathbf{A}_1\mathbf{x} + \sigma\left(\mathbf{A}_2\mathbf{u}\right)) \\ &= \frac{1}{2}\left\|\mathbf{A}_1\mathbf{x} + \sigma\left(\mathbf{A}_2\mathbf{u}\right) - \mathbf{b}\right\|_2^2. \end{aligned} \tag{11}$$

Due to the non-linearity of σ, $f(\mathbf{x},\mathbf{u})$ is generally non-convex. If we further choose σ to be some non-smooth function such as ReLU, i.e., $\sigma(\mathbf{x})_i = \max\{0, x_i\}$, the objective can also be non-smooth. Yet, assuming that σ is Lipschitz, we can show that $f(\mathbf{x},\mathbf{u})$ satisfies Assumptions 1-6.



(a) $\sigma_{\max}(\mathbf{A}_2) = 0.01$

(b) $\sigma_{\max}(\mathbf{A}_2) = 0.3$

(c) $\sigma_{\max}(\mathbf{A}_2) = 5$

Figure 1: Experiment of learning additive model with gradient descent and Nesterov's momentum.

Lemma 12 *Let σ be B-Lipschitz and $\sigma_{\min}(\mathbf{A}_1) > 0$. Then $f(\mathbf{x},\mathbf{u})$ satisfies Assumptions 1-6 with*

$$\begin{aligned} R_{\mathbf{x}} = R_{\mathbf{u}} = \infty;\ \mu = \sigma_{\min}\left(\mathbf{A}_1\right)^2;\ L_1 = \sigma_{\max}\left(\mathbf{A}_1\right)^2;\ L_2 = 1 \\ G_1 = B\sigma_{\max}\left(\mathbf{A}_2\right);\ G_2 = B\sigma_{\max}\left(\mathbf{A}_1\right)\sigma_{\max}\left(\mathbf{A}_2\right). \end{aligned} \tag{12}$$

Notice that while μ and L_1 depend entirely on the property of $\mathbf{A}_1$, both G_1 and G_2 can be made smaller by choosing $\mathbf{A}_2$ with a small enough $\sigma_{\max}(\mathbf{A}_2)$. Intuitively, this means that G_1 and G_2 can be controlled, as long as the component that introduces the non-convexity and non-smoothness $\sigma\left(\mathbf{A}_2\mathbf{u}\right)$ is small enough. Therefore, we can apply Theorem 7 to the minimization of (11).

Theorem 13 *Consider the problem in (11) and assume using Nesterov's momentum to minimize $f(\mathbf{x},\mathbf{u})$, where σ is a B-Lipschitz function. Let $\kappa = {\sigma_{\max}\left(\mathbf{A}_1\right)^2}/{\sigma_{\min}\left(\mathbf{A}_1\right)^2}$, and suppose:*

$$\sigma_{\min}\left(\mathbf{A}_1\right) \geq \tilde{C}\sigma_{\max}\left(\mathbf{A}_2\right)B\kappa^{0.75}, \tag{13}$$

for large enough constant $\tilde{C} > 0$. Then there exists constant $c > 0$ such that if we choose $\eta = {c}/{\sigma_{\max}\left(\mathbf{A}_1\right)^2}$ and $\beta = {(4\sqrt{\kappa}-\sqrt{c})}/{(4\sqrt{\kappa}+7\sqrt{c})}$, Nesterov's momentum in (4) converges according to:

$$f(\mathbf{x}_k,\mathbf{u}_k) \leq 2\left(1 - \frac{c}{4\sqrt{\kappa}}\right)^k f(\mathbf{x}_0,\mathbf{u}_0).$$

Notice that the requirement in (13) favors a larger $\sigma_{\min}(\mathbf{A}_1)$ and smaller $\sigma_{\max}(\mathbf{A}_2)$, B and κ. Using this example, we empirically verify the theoretical result of Theorem 7, as shown in Figure 1. The

three rows correspond to the cases of $\sigma_{\max}(\mathbf{A}_2) \in \{0.01, 0.3, 5\}$, respectively. The plot lines denote the average loss/distance among ten trials, while the shaded region denotes the standard deviation. Observing the plots in the left column, in all three cases, Nesterov's momentum achieves a faster convergence compared with gradient descent, while the case with the largest $\sigma_{\max}(\mathbf{A}_2)$ introduces the largest variance between results of the ten trials. Recall that $\sigma_{\max}(\mathbf{A}_2)$ controls the magnitude of G_1 and G_2. Thus, this phenomenon shows that when G_1 and G_2 become larger, the theoretical guarantee in Theorem 7 begins to break down, and the result depends more on the initialization. Figures in the right column plots the evolution of $\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_2^2$ and $\|\mathbf{u}_k - \mathbf{u}_{k-1}\|_2^2$. All three figures show that the two quantity decrease linearly. This phenomenon supports the linear decay of the two quantities, as shown in Lemma 10. Moreover, as $\sigma_{\max}(\mathbf{A}_2)$ increases, the relative magnitude of $\|\mathbf{u}_k - \mathbf{u}_{k-1}\|_2^2$ to $\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_2^2$ also increase (the line of "$\|\Delta \mathbf{u}_k\|_2$" get closer to "$\|\Delta \mathbf{x}_k\|_2$"). This supports that $\|\mathbf{u}_k - \mathbf{u}_{k-1}\|_2^2$ scales with G_1, as shown in Lemma 10.

4.2. Deep ReLU Neural Networks

Consider the Λ-layer ReLU neural network with layer widths $\{d_\ell\}_{\ell=0}^{\Lambda}$. Denoting the number of training samples with n, we consider the input and label of the training data given by $\mathbf{X} \in \mathbb{R}^{n \times d_0}$ and $\mathbf{Y} \in \mathbb{R}^{n \times d_\Lambda}$. Let the weight matrix in the ℓ-th layer be $\mathbf{W}_\ell$. We use $\boldsymbol{\theta} = \{\mathbf{W}_\ell\}_{\ell=1}^{\Lambda}$ to denote the collection of all weights and $\sigma(\mathbf{A})_{ij} = \max\{0, a_{ij}\}$ to denote the ReLU function. Then, the output of each layer is given by:

$$\mathbf{F}_\ell(\boldsymbol{\theta}) = \begin{cases} \mathbf{X}, & \text{if } \ell = 0; \\ \sigma\left(\mathbf{F}_{\ell-1}(\boldsymbol{\theta})\, \mathbf{W}_\ell\right), & \text{if } \ell \in [\Lambda - 1]; \\ \mathbf{F}_{\Lambda-1}(\boldsymbol{\theta})\, \mathbf{W}_\Lambda, & \text{if } \ell = \Lambda. \end{cases} \tag{14}$$

We consider the training of $\mathbf{F}_\Lambda(\boldsymbol{\theta})$ over the MSE loss, as in $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{F}_\Lambda(\boldsymbol{\theta}) - \boldsymbol{Y}\|_F^2$. We can interpret the scenario using our partition model in (1). Let $g(\mathbf{s}) = \frac{1}{2}\|\mathbf{s} - \mathbb{V}(\mathbf{Y})\|_2^2$. If we partition the parameter $\boldsymbol{\theta}$ into $\mathbf{x} = \mathbb{V}(\mathbf{W}_\Lambda)$ and $\mathbf{u} = (\mathbb{V}(\mathbf{W}_1), \ldots, \mathbb{V}(\mathbf{W}_{\Lambda-1}))$, then we can write:

$$h(\mathbf{x}, \mathbf{u}) = \mathbb{V}(\mathbf{F}_\Lambda(\boldsymbol{\theta}))\,; \quad f(\mathbf{x}, \mathbf{u}) = \frac{1}{2}\|\mathbb{V}(\mathbf{F}_\Lambda(\boldsymbol{\theta})) - \mathbb{V}(\mathbf{Y})\|_2^2 = \frac{1}{2}\|\mathbf{F}_\Lambda(\boldsymbol{\theta}) - \mathbf{Y}\|_F^2\,. \tag{15}$$

For some given $\mathbf{x}$ and $\mathbf{u}$, we let $\mathbf{W}_\Lambda(\mathbf{x})$ be the matrix such that $\mathbf{x} = \mathbb{V}(\mathbf{W}_\Lambda(\mathbf{x}))$; similarly, we let $\mathbf{W}_\ell(\mathbf{u})$ with $\ell \in [L-1]$ be the matrices such that $\mathbf{u} = (\mathbb{V}(\mathbf{W}_1(\mathbf{u})), \ldots, \mathbb{V}(\mathbf{W}_{\Lambda-1}(\mathbf{u})))$. Denote $\lambda_\Lambda = \sup_{\mathbf{x} \in \mathcal{B}_{R_\mathbf{x}}^{(1)}} \sigma_{\max}(\mathbf{W}_\Lambda(\mathbf{x}))$ and $\lambda_\ell = \sup_{\mathbf{u} \in \mathcal{B}_{R_\mathbf{x}}^{(2)}} \sigma_{\max}(\mathbf{W}_\ell(\mathbf{u}))$ for $\ell \in [\Lambda - 1]$. Moreover, denote $\lambda_{i \to j} = \prod_{\ell=i}^{j} \lambda_\ell$. Then we can show that $f(\mathbf{x}, \mathbf{u})$ defined in (15) satisfies Assumptions 1-6.

Lemma 14 *Let $\boldsymbol{\theta}(0)$ be the initialization of the ReLU network in (14) and $\alpha_0 = \sigma_{\min}(\mathbf{F}_{\Lambda-1}(\boldsymbol{\theta}(0)))$. Assume that each $\mathbf{W}_\ell$ is initialized such that $\|\mathbf{W}_\ell(0)\|_2 \le \frac{\lambda_\ell}{2}$ and $\alpha_0 > 0$. Then, $f(\mathbf{x}, \mathbf{u})$ satisfies Assumptions 1-6 with:*

$$R_\mathbf{x} = \frac{\lambda_\Lambda}{2}; \quad R_\mathbf{u} = \frac{1}{2}\left(\min_{\ell \in [\Lambda-1]} \lambda_\ell\right) \min\left\{1, \frac{\alpha_0}{2\sqrt{\Lambda}\,\|\mathbf{X}\|_F\, \lambda_{1 \to \Lambda-1}}\right\}^2; \quad \mu = \frac{\alpha_0^2}{2}$$

$$L_1 = \|\mathbf{X}\|_F^2\, \lambda_{1 \to \Lambda-1}^2; \quad L_2 = 1; \quad G_1 = (\lambda_\Lambda + R_\mathbf{u})\sqrt{\Lambda}\,\|\mathbf{X}\|_F\, \lambda_{1 \to \Lambda-1}\left(\min_{\ell \in [\Lambda-1]} \lambda_\ell\right)^{-1}$$

$$G_2 = ((2\lambda_\Lambda + R_\mathbf{u})\,\|\mathbf{X}\|_F\, \lambda_{1 \to \Lambda-1} + \|\mathbf{Y}\|_F)\sqrt{\Lambda}\,\|\mathbf{X}\|_F\, \lambda_{1 \to \Lambda-1}\left(\min_{\ell \in [\Lambda-1]} \lambda_\ell\right)^{-1}$$

As shown in Lemma 3.3 by Nguyen (2021), we can guarantee that $\alpha_0 > 0$ with sufficient over-parameterization. To show the acceleration of Nesterov's momentum when training (14), we need to $i)$ guarantee that the condition of $R_{\mathbf{x}}$ and $R_{\mathbf{u}}$ in (5) satisfies the upper bound in Lemma 14, and $ii)$ the quantities μ, L_1, L_2, G_1 and G_2 defined in Lemma 14 satisfy the requirement in (5). Enforcing the two conditions with sufficient over-parameterization gives us the following theorem.

Theorem 15 *Consider training the ReLU neural network in (14) using the MSE loss, or equivalently, minimizing $f(\mathbf{x}, \mathbf{u})$ defined in (15) with Nesterov's momentum with $\eta = c/L_1$ and $\beta = \frac{4\sqrt{\kappa}-\sqrt{c}}{4\sqrt{\kappa}+7\sqrt{c}}$, where $\kappa = \frac{2L_1}{\alpha_0^2}$ and α_0, L_1 defined in Lemma 14. If the width of the network satisfies:*

$$d_\ell = \Theta(m) \quad \forall \ell \in [\Lambda - 2]; \quad d_{\Lambda-1} = \Omega\left(n^{4.5} \max\left\{n, d^2\right\}\right), \tag{16}$$

for some $m \geq \max\{d_0, d_\Lambda\}$, and we initialize the weights according to:

$$[\mathbf{W}_\ell(0)]_{ij} \sim \mathcal{N}\left(0, d_{\ell-1}^{-1}\right), \quad \forall \ell \in [\Lambda - 1]; \quad [\mathbf{W}_\Lambda(0)]_{ij} \sim \mathcal{N}\left(0, d_{\Lambda-1}^{-\frac{3}{2}}\right).$$

Then, with high probability over the initialization, there exists an absolute constant $c > 0$ such that:

$$\mathcal{L}(\boldsymbol{\theta}(k)) \leq 2\left(1 - \tfrac{c}{4\sqrt{\kappa}}\right)^k \mathcal{L}(\boldsymbol{\theta}(0)). \tag{17}$$

As in prior work (Nguyen, 2021), we treat the depth of the neural network Λ to be a constant when computing the over-parameterization requirement. Next, we compare our result with Theorem 2.2 and Corollary 3.2 of Nguyen (2021). To deal with the additional complexity of Nesterov's momentum as introduced in Section 3.3, our over-parameterization is slightly larger than the over-parameterization of $d_{\Lambda-1} = \Theta\left(n^3 m^3\right)$ in Corollary 3.2 of Nguyen (2021). Moreover, in Theorem 2.2 of Nguyen (2021), since their choice of η is also $O\left(1/L_1\right)$, the convergence rate they achieve for gradient descent is $1 - \Theta\left(1/\kappa\right)$. Compared with this rate, Theorem 15 achieves a faster convergence of $1 - \Theta\left(1/\sqrt{\kappa}\right)$. This shows that Nesterov's momentum enjoys acceleration when training deep ReLU neural networks.

5. Conclusion and Broader Impact

We consider the minimization of a new class of objective functions, namely the partition model, where the function is smooth and strongly convex with respect to only a subset of its parameters. This class of objectives is more general than the class of smooth and strongly convex functions. We prove the convergence of gradient descent and Nesterov's momentum on this class of objectives under certain assumptions and show that Nesterov's momentum achieves an accelerated convergence rate of $1 - \Theta\left(1/\sqrt{\kappa}\right)$ compared to the $1 - \Theta\left(1/\kappa\right)$ convergence rate of gradient descent. Moreover, we considered the training of the additive model and deep ReLU networks as two realizations of the partition model. We showed the acceleration of Nesterov's momentum on these two realizations.

Future works can focus on three aspects. First, one can consider the case where Assumption 6 does not hold, and study whether Nesterov's momentum can still converge to up to some error with acceleration under a milder condition of this assumption. Second, one can extend the analysis to different neural network architectures by investigating whether Assumptions 1-6 hold on CNNs and ResNets. Lastly, since the weight selection process is extensively studied by literature on neural network pruning, one can study whether neural network pruning keeps weights with good optimization properties and potentially connects our result with the theory of pruning methods such as the Lottery Ticket Hypothesis.

Acknowledgments

This work is supported by NSF CMMI no. 2037545 and NSF CAREER award no. 2145629, Welch Foundation Grant #A22-0307, a Microsoft Research Award, and an Amazon Research Award.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks, 2019a.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 242–252. PMLR, 09–15 Jun 2019b. URL https://proceedings.mlr.press/v97/allen-zhu19a.html.

Vassilis Apidopoulos, Nicolò Ginatta, and Silvia Villa. Convergence rates for the heavy-ball continuous dynamics for non-convex optimization, under polyak–Łojasiewicz condition. *Journal of Global Optimization*, 84, 05 2022. doi: 10.1007/s10898-022-01164-w.

Peter Auer, Mark Herbster, and Manfred K. K Warmuth. Exponentially many local minima for single neurons. In D. Touretzky, M.C. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1995. URL https://proceedings.neurips.cc/paper_files/paper/1995/file/3806734b256c27e41ec2c6bffa26d9e7-Paper.pdf.

Pranjal Awasthi, Abhimanyu Das, and Sreenivas Gollapudi. A convergence analysis of gradient descent on graph neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 20385–20397. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/aaf2979785deb27864047e0ea40ef1b7-Paper.pdf.

Arindam Banerjee, Pedro Cisneros-Velarde, Libin Zhu, and Misha Belkin. Restricted strong convexity of deep learning models with smooth activations. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PINRbk7h01.

Nikhil Bansal and Anupam Gupta. Potential-function proofs for gradient methods. *Theory of Computing*, 15(4):1–32, 2019. doi: 10.4086/toc.2019.v015a004. URL https://theoryofcomputing.org/articles/v015a004.

Yair Carmon and John C. Duchi. First-order methods for nonconvex quadratic minimization. *SIAM Review*, 62(2):395–436, jan 2020. doi: 10.1137/20m1321759. URL https://doi.org/10.1137%2F20m1321759.

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for nonconvex optimization, 2017.

George E. Dahl, Frank Schneider, Zachary Nado, Naman Agarwal, Chandramouli Shama Sastry, Philipp Hennig, Sourabh Medapati, Runa Eschenhagen, Priya Kasimbeg, Daniel Suo, Juhan Bae, Justin Gilmer, Abel L. Peirson, Bilal Khan, Rohan Anil, Mike Rabbat, Shankar Krishnan, Daniel Snider, Ehsan Amid, Kongtao Chen, Chris J. Maddison, Rakshith Vasudev, Michal Badura, Ankush Garg, and Peter Mattson. Benchmarking neural network training algorithms, 2023.

Marina Danilova, Anastasiya Kulakova, and Boris Polyak. Non-monotone behavior of the heavy ball method, 2018.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1675–1685. PMLR, 09–15 Jun 2019a. URL https://proceedings.mlr.press/v97/du19c.html.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019b. URL https://openreview.net/forum?id=S1eK3i09YQ.

Alexandre d'Aspremont, Damien Scieur, and Adrien Taylor. Acceleration methods. *Foundations and Trends® in Optimization*, 5(1-2):1–245, 2021. ISSN 2167-3888. doi: 10.1561/2400000036. URL http://dx.doi.org/10.1561/2400000036.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Baptiste Goujaud, Adrien Taylor, and Aymeric Dieuleveut. Provable non-accelerations of the heavy-ball method, 2023.

Baihe Huang, Xiaoxiao Li, Zhao Song, and Xin Yang. Fl-ntk: A neural tangent kernel-based framework for federated learning analysis. In *International Conference on Machine Learning*, 2021.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks, 2020.

Samy Jelassi and Yuanzhi Li. Towards understanding how momentum improves generalization in deep learning, 2022.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HygegyrYwH.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition, 2020.

Yann Lecun, Leon Bottou, Y. Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278 – 2324, 12 1998. doi: 10.1109/5.726791.

Yann LeCun, Y. Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–44, 05 2015. doi: 10.1038/nature14539.

Huan Li and Zhouchen Lin. Restarted nonconvex accelerated gradient descent: No more polylogarithmic factor in the $o(\epsilon^{-7/4})$ complexity. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12901–12916. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/li22o.html.

Xiaoxiao Li, Zhao Song, and Jiaming Yang. Federated adversarial learning: A framework with convergence analysis, 2022.

Yuanzhi Li, Tengyu Ma, and Hongyang R. Zhang. Learning over-parametrized two-layer relu neural networks beyond ntk, 2020.

Fangshuo Liao and Anastasios Kyrillidis. On the convergence of shallow neural network training with randomly masked neurons. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=e7mYYMSyZH.

Zenan Ling, Xingyu Xie, Qiuhao Wang, Zongpeng Zhang, and Zhouchen Lin. Global convergence of over-parameterized deep equilibrium models, 2023.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. 2020.

Xin Liu, Zhisong Pan, and Wei Tao. Provable convergence of nesterov's accelerated gradient method for over-parameterized neural networks. *Knowledge-Based Systems*, 251:109277, 2022a. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2022.109277. URL https://www.sciencedirect.com/science/article/pii/S0950705122006402.

Xin Liu, Wei Tao, and Zhisong Pan. A convergence analysis of nesterov's accelerated gradient method in training deep linear neural networks. *Information Sciences*, 612:898–925, 2022b. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2022.08.090. URL https://www.sciencedirect.com/science/article/pii/S0020025522010003.

Poorya Mianjy and Raman Arora. On convergence and generalization of dropout training. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21151–21161. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f1de5100906f31712aaa5166689bfdf4-Paper.pdf.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer Publishing Company, Incorporated, 2nd edition, 2018. ISBN 3319915770.

Quynh Nguyen. On the proof of global convergence of gradient descent for deep relu networks with linear widths. *CoRR*, abs/2101.09612, 2021. URL https://arxiv.org/abs/2101.09612.

Quynh Nguyen and Marco Mondelli. Global convergence of deep networks with one wide layer followed by pyramidal topology, 2020.

Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks, 2019.

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks, 2018.

Bin Shi, Simon Du, Michael Jordan, and Weijie Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, 195:79–148, 01 2022. doi: 10.1007/s10107-021-01681-8.

Chaehwan Song, Ali Ramezani-Kebrya, Thomas Pethick, Armin Eftekhari, and Volkan Cevher. Subquadratic overparameterization for shallow neural networks. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=NhbFhfM960.

Zhao Song and Xin Yang. Quadratic suffices for over-parametrization via matrix chernoff bound, 2020.

Lili Su and Pengkun Yang. On learning over-parameterized neural networks: A functional approximation perspective, 2019.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/sutskever13.html.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Jun-Kun Wang, Chi-Heng Lin, and Jacob D Abernethy. A modular analysis of provable acceleration via polyak's momentum: Training a wide relu network and a deep linear network. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10816–10827. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/wang21n.html.

Jun-Kun Wang, Chi-Heng Lin, Andre Wibisono, and Bin Hu. Provable acceleration of heavy ball beyond quadratics for a class of polyak-lojasiewicz functions when the non-convexity is averaged-out. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 22839–22864. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wang22p.html.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models, 2020.

Diyuan Wu, Vyacheslav Kungurtsev, and Marco Mondelli. Mean-field analysis for heavy ball methods: Dropout-stability, connectivity, and global convergence, 2023.

Jiaming Xu and Hanjing Zhu. One-pass stochastic gradient descent in overparametrized two-layer neural networks, 2021.

Weihang Xu and Simon S. Du. Over-parameterization exponentially slows down gradient descent for learning a single neuron, 2023.

Greg Yang and Edward J. Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11727–11737. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/yang21c.html.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks, 2022.

Pengyun Yue, Cong Fang, and Zhouchen Lin. On the lower bound of minimizing polyak-łojasiewicz functions, 2022.

Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Small nonlinearities in activation functions create bad local minima in neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rke_YiRct7.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization, 2017.

Xiao Zhang, Yaodong Yu, Lingxiao Wang, and Quanquan Gu. Learning one-hidden-layer relu networks via gradient descent, 2018.

Xiao Zhang, Yaodong Yu, Lingxiao Wang, and Quanquan Gu. Learning one-hidden-layer relu networks via gradient descent. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1524–1534. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/zhang19g.html.

Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/6a61d423d02a1c56250dc23ae7ff12f3-Paper.pdf.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks, 2018.

Contents

Appendix A. Proofs for Section 2

A.1. Proof of Theorem 3

Let $\tilde{f} : \mathbb{R}^d \to \mathbb{R}$ be a $\tilde{L}$-smooth and $\tilde{\mu}$-strongly convex function. Let $\{\mathbf{0}\}$ be the zero-dimensional vector space. We define $h(\mathbf{x}, \mathbf{u}) : \mathbb{R}^d \times \{\mathbf{0}\} \to \mathbb{R}^d$ as $h(\mathbf{x}, \mathbf{u}) = \mathbf{x}$ for all $\mathbf{x} \in \mathbb{R}^d$. Moreover, we define $g : \mathbb{R} \to \mathbb{R}$ as $g(s) = \tilde{f}(\mathbf{s})$ for all $s \in \mathbb{R}$, and $f(\mathbf{x}, \mathbf{u}) = g(h(\mathbf{x}, \mathbf{u}))$. Then $f(\mathbf{x}, \mathbf{u})$ is a partitioned equivalence of $\tilde{f}(\mathbf{x})$ since

$$f(\mathbf{x}, \mathbf{u}) = g(\mathbf{x}) = \tilde{f}(\mathbf{x}) \quad \forall \mathbf{x} \in \mathbb{R}^d$$

Choosing $R_{\mathbf{x}} = R_{\mathbf{u}} = \infty$, we have that $\mathcal{B}^{(1)}_{R_{\mathbf{x}}} = \mathbb{R}^d$ and $\mathcal{B}^{(2)}_{R_{\mathbf{u}}} = \{\mathbf{0}\}$. Therefore, we have

$$\begin{aligned}
f(\mathbf{y}, \mathbf{u}) = \tilde{f}(\mathbf{y}) &\geq \tilde{f}(\mathbf{x}) + \left\langle \nabla \tilde{f}(\mathbf{x}), \mathbf{y} - \mathbf{x} \right\rangle + \frac{\tilde{\mu}}{2} \|\mathbf{y} - \mathbf{x}\|_2^2 \\
&= f(\mathbf{x}, \mathbf{u}) + \langle \nabla_1 f(\mathbf{x}, \mathbf{u}), \mathbf{y} - \mathbf{x} \rangle + \frac{\tilde{\mu}}{2} \|\mathbf{y} - \mathbf{x}\|_2^2 \\
f(\mathbf{y}, \mathbf{u}) = \tilde{f}(\mathbf{y}) &\leq \tilde{f}(\mathbf{x}) + \left\langle \nabla \tilde{f}(\mathbf{x}), \mathbf{y} - \mathbf{x} \right\rangle + \frac{\tilde{L}}{2} \|\mathbf{y} - \mathbf{x}\|_2^2 \\
&= f(\mathbf{x}, \mathbf{u}) + \langle \nabla_1 f(\mathbf{x}, \mathbf{u}), \mathbf{y} - \mathbf{x} \rangle + \frac{\tilde{L}}{2} \|\mathbf{y} - \mathbf{x}\|_2^2
\end{aligned}$$

where the first and second inequality follows from the strong convexity and smoothness of $\tilde{f}$. This shows that $f(\mathbf{x}, \mathbf{u})$ satisfies Assumption 1, 2 with $\mu = \tilde{\mu}$ and $L_1 = \tilde{L}$. Since $g(\mathbf{s}) = \tilde{f}(\mathbf{s})$, it must be L_2-smooth with $L_2 = \tilde{L}$ as well. Moreover, we must have $g^\star = \tilde{f}^\star = f^\star$. This shows that Assumption 3 holds. Also, since for all $\mathbf{u} \in \{\mathbf{0}\}$ we must have $\mathbf{u} = \mathbf{0}$, it holds naturally that

$$h(\mathbf{x}, \mathbf{u}) = h(\mathbf{x}, \mathbf{v}); \quad \nabla_1 f(\mathbf{x}, \mathbf{u}) = \nabla_1 f(\mathbf{x}, \mathbf{v})$$

Therefore, Assumption 4, 5 hold with $G_1 = G_2 = 0$. Lastly, since $\tilde{f}$ is strongly convex, there must exist a unique $\mathbf{x}^\star = \arg\min_{\mathbf{x} \in \mathbb{R}^d} \tilde{f}(\mathbf{x})$. Therefore, we must have that

$$\min_{x \in \mathbb{R}^d} f(\mathbf{x}, \mathbf{u}) \leq f(\mathbf{x}^\star, \mathbf{u}) = \tilde{f}(\mathbf{x}^\star) = f^\star \quad \forall \mathbf{u} \in \{\mathbf{0}\}$$

This shows that Assumption 6 is satisfied.

Appendix B. Proofs for Section 3.1

B.1. Proof of Lemma 4

Fix any $\mathbf{u} \in \mathcal{B}_{R_{\mathbf{u}}}^{(2)}$ and let $\mathbf{x}^\star = \arg\min_{\mathbf{x}\in\mathbb{R}^{d_1}} f(\mathbf{x},\mathbf{u})$. By Assumption 1, we have

$$\begin{aligned} f(\mathbf{x}^\star,\mathbf{u}) &\geq f(\mathbf{x},\mathbf{u}) + \langle \nabla_1 f(\mathbf{x},\mathbf{u}), \mathbf{x}^\star - \mathbf{u}\rangle + \frac{\mu}{2}\|\mathbf{x}^\star - \mathbf{x}\|_2^2 \\ &\geq \min_{\mathbf{y}\in\mathbb{R}^{d_1}} \Bigg(\underbrace{f(\mathbf{x},\mathbf{u}) + \langle \nabla_1 f(\mathbf{x},\mathbf{u}), \mathbf{y} - \mathbf{u}\rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|_2^2}_{f_{\mathbf{x},\mathbf{u}}(\mathbf{y})} \Bigg) \end{aligned}$$

Notice that $\nabla^2 f_{\mathbf{x},\mathbf{u}}(\mathbf{y}) = \mu$. Therefore, $f_{\mathbf{x},\mathbf{u}}(\mathbf{y})$ is strongly convex with respect to $\mathbf{y}$. Thus, $\mathbf{y}^\star = \arg\min_{\mathbf{y}\in\mathbb{R}^{d_1}} f_{\mathbf{x},\mathbf{u}}(\mathbf{y})$ must satisfy

$$\nabla f_{\mathbf{x},\mathbf{u}}(\mathbf{y}^\star) = \nabla_1 f(\mathbf{x},\mathbf{u}) + \mu(\mathbf{y}^\star - \mathbf{x}) = 0$$

which implies that $\mathbf{y}^\star = \mathbf{x} - \frac{1}{\mu}\nabla_1 f(\mathbf{x},\mathbf{u})$, and $\min_{\mathbf{y}\in\mathbb{R}^{d_1}} f_{\mathbf{x},\mathbf{u}}(\mathbf{y}) = f(\mathbf{x},\mathbf{u}) - \frac{1}{2\mu}\|\nabla_1 f(\mathbf{x},\mathbf{u})\|_2^2$. This implies that

$$f(\mathbf{x}^\star,\mathbf{u}) \geq f(\mathbf{x},\mathbf{u}) - \frac{1}{2\mu}\|\nabla_1 f(\mathbf{x},\mathbf{u})\|_2^2 \Rightarrow \|\nabla_1 f(\mathbf{x},\mathbf{u})\|_2^2 \geq 2\mu\left(f(\mathbf{x},\mathbf{u}) - f(\mathbf{x}^\star,\mathbf{u})\right)$$

By Assumption 6, we have that $f(\mathbf{x}^\star,\mathbf{u}) = f^\star$. Also, by definition of $\nabla f(\mathbf{x},\mathbf{u})$, we have $\|\nabla f(\mathbf{x},\mathbf{u})\|_2^2 = \|\nabla_1 f(\mathbf{x},\mathbf{u})\|_2^2 + \|\nabla_2 f(\mathbf{x},\mathbf{u})\|_2^2$. Therefore

$$\|\nabla f(\mathbf{x},\mathbf{u})\|_2^2 \geq \|\nabla_1 f(\mathbf{x},\mathbf{u})\|_2^2 \geq 2\mu\left(f(\mathbf{x},\mathbf{u}) - f^\star\right)$$

B.2. Proof of Lemma 5

Since Assumption 3 holds, we can invoke Lemma 19 to get that

$$\|\nabla g(\mathbf{s})\|_2^2 \leq 2L_2\left(g(\mathbf{s}) - f^\star\right)$$

Moreover, Assumption 3 also implies that

$$f(\mathbf{x},\mathbf{u}) = g(h(\mathbf{x},\mathbf{u})) \leq f(\mathbf{x},\mathbf{v}) + \langle \nabla g(h(\mathbf{x},\mathbf{v})), h(\mathbf{x},\mathbf{u}) - h(\mathbf{x},\mathbf{v})\rangle + \frac{L_2}{2}\|h(\mathbf{x},\mathbf{u}) - h(\mathbf{x},\mathbf{v})\|_2^2$$

Assumption 4 implies that

$$\|h(\mathbf{x},\mathbf{u}) - h(\mathbf{x},\mathbf{v})\|_2 \leq G_1\|\mathbf{u} - \mathbf{v}\|_2$$

Therefore, applying the triangle inequality to the inner-product term, we have

$$\begin{aligned} f(\mathbf{x},\mathbf{u}) &\leq f(\mathbf{x},\mathbf{v}) + G_1\|\nabla g(h(\mathbf{x},\mathbf{v}))\|_2 \cdot \|\mathbf{u} - \mathbf{v}\|_2 + \frac{G_1^2 L_2}{2}\|\mathbf{u} - \mathbf{v}\|_2^2 \\ &\leq f(\mathbf{x},\mathbf{v}) + \mathcal{Q}^{-1}\|\nabla g(h(\mathbf{x},\mathbf{v}))\|_2^2 + \frac{G_1^2}{2}(L_2 + \mathcal{Q})\|\mathbf{u} - \mathbf{v}\|_2^2 \\ &\leq f(\mathbf{x},\mathbf{v}) + \mathcal{Q}^{-1}\left(f(\mathbf{x},\mathbf{v}) - f^\star\right) + \frac{G_1^2}{2}(L_2 + \mathcal{Q})\|\mathbf{u} - \mathbf{v}\|_2^2 \end{aligned}$$

This implies that

$$f(\mathbf{x},\mathbf{u}) - f(\mathbf{x},\mathbf{v}) \leq \mathcal{Q}^{-1}\left(f(\mathbf{x},\mathbf{v}) - f^\star\right) + \frac{G_1^2}{2}(L_2 + \mathcal{Q})\|\mathbf{u} - \mathbf{v}\|_2^2$$

B.3. Proof of Theorem 6

We prove the following two conditions by induction.

Condition 1 *Let $k \in \mathbb{N}$, then for all $t \leq k$, we have $\|\mathbf{x}_t - \mathbf{x}_0\|_2 \leq R_{\mathbf{x}}$ and $\|\mathbf{u}_t - \mathbf{u}_0\|_2 \leq R_{\mathbf{u}}$.*

Condition 2 *Let $k \in \mathbb{N}$, then for all $t \leq k$, we have*

$$f(\mathbf{x}_t, \mathbf{u}_t) - f^\star \leq \left(1 - \frac{c}{\kappa}\right)^t (f(\mathbf{x}_0, \mathbf{u}_0) - f^\star)$$

B.3.1. BASE CASE: $k = 0$

When $k = 0$, Condition 1 reads $\|\mathbf{x}_t - \mathbf{x}_0\|_2 \leq R_{\mathbf{x}}$ and $\|\mathbf{u}_t - \mathbf{u}_0\|_2 \leq R_{\mathbf{u}}$ for $t = 0$. This is automatically true since $\|\mathbf{x}_0 - \mathbf{x}_0\|_2 = \|\mathbf{u}_0 - \mathbf{u}_0\|_2 = 0$. Condition 2 is also true since

$$f(\mathbf{x}_0, \mathbf{u}_0) - f^\star \leq \left(1 - \frac{c}{\kappa}\right)^0 (f(\mathbf{x}_0, \mathbf{u}_0) - f^\star)$$

B.3.2. INDUCTIVE STEP 1: CONDITION 1 $\Rightarrow$ CONDITION 2

Suppose that Condition 1 and Condition 2 holds for all $t \leq k$. We show that Condition 2 holds for all $t \leq k+1$. Since $\|\mathbf{x}_t - \mathbf{x}_0\|_2 \leq R_{\mathbf{x}}$ and $\|\mathbf{u}_t - \mathbf{u}_0\|_2 \leq R_{\mathbf{u}}$, we have that $\mathbf{x}_t \in \mathcal{B}^{(1)}_{R_{\mathbf{x}}}$ and $\mathbf{u}_t \in \mathcal{B}^{(2)}_{R_{\mathbf{u}}}$ for all $t \leq k$. Therefore, Assumption 2 holds for all $\mathbf{x} \in \mathbb{R}^{d_1}$ and $\mathbf{u} = \mathbf{u}_k$, which implies that

$$\begin{aligned} f(\mathbf{x}_{k+1}, \mathbf{u}_k) &\leq f(\mathbf{x}_k, \mathbf{u}_k) + \langle \nabla_1 f(\mathbf{x}_k, \mathbf{u}_k), \mathbf{x}_{k+1} - \mathbf{u}_k \rangle + \frac{L_1}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 \\ &= f(\mathbf{x}_k, \mathbf{u}_k) - \eta \|\nabla_1 f(\mathbf{x}_k, \mathbf{u}_k)\|_2^2 + \frac{L_2}{2}\eta^2 \|\nabla_1 f(\mathbf{x}_k, \mathbf{u}_k)\|_2^2 \\ &= f(\mathbf{x}_k, \mathbf{u}_k) - \eta \left(1 - \frac{\eta L_1}{2}\right) \|\nabla_1 f(\mathbf{x}_k, \mathbf{u}_k)\|_2^2 \end{aligned} \tag{18}$$

Since Assumption 1, 6 holds, we can apply Lemma 4 to (18) and choose $\eta = \frac{1}{L_1}$ to get that

$$f(\mathbf{x}_{k+1}, \mathbf{u}_k) - f^\star \leq \left(1 - \frac{1}{\kappa}\right) (f(\mathbf{x}_k, \mathbf{u}_k) - f^\star) \tag{19}$$

Moreover, since $\mathbf{u}_{k+1} = \mathbf{u}_k - \eta \nabla_2 f(\mathbf{x}_k, \mathbf{u})$, we have

$$\|\mathbf{u}_{k+1} - \mathbf{u}_k\|_2^2 = \eta \|\nabla_2 f(\mathbf{x}_k, \mathbf{u})\|_2^2 \leq \eta 2 G_1^2 L_2 (f(\mathbf{x}_k, \mathbf{u}_k) - f^\star)$$

where the last inequality follows from Lemma 20. Since Assumption 3, 4 holds, we can use Lemma 5 with $\mathcal{Q} = 2(\kappa - 1)L_2$ to get that

$$\begin{aligned} f(\mathbf{x}_{k+1}, \mathbf{u}_{k+1}) - f(\mathbf{x}_{k+1}, \mathbf{u}_k) &\leq \mathcal{Q}^{-1} L_2 (f(\mathbf{x}_{k+1}, \mathbf{u}_k) - f^\star) \\ &\quad + \frac{G_1^2}{2} (L_2 + \mathcal{Q}) \|\mathbf{u}_{k+1} - \mathbf{u}_k\|_2^2 \\ &\leq \mathcal{Q}^{-1} L_2 (f(\mathbf{x}_{k+1}, \mathbf{u}_k) - f^\star) \\ &\quad + \eta^2 G_1^4 L_2 (L_2 + \mathcal{Q}) (f(\mathbf{x}_k, \mathbf{u}_k) - f^\star) \\ &\leq \frac{1}{2(\kappa - 1)} (f(\mathbf{x}_{k+1}, \mathbf{u}_k) - f^\star) \\ &\quad + \eta^2 G_1^4 L_2^2 (2\kappa - 1) (f(\mathbf{x}_k, \mathbf{u}_k) - f^\star) \end{aligned} \tag{20}$$

Combining (18) with (19), we have

$$\begin{aligned} f(\mathbf{x}_{k+1},\mathbf{u}_{k+1})-f^\star &= f(\mathbf{x}_{k+1},\mathbf{u}_k)-f^\star+f(\mathbf{x}_{k+1},\mathbf{u}_{k+1})-f(\mathbf{x}_{k+1},\mathbf{u}_k)\\ &\le \left(1+\frac{1}{2(\kappa-1)}\right)(f(\mathbf{x}_{k+1},\mathbf{u}_k)-f^\star)\\ &\quad +\eta^2G_1^4L_2^2\,(2\kappa-1)\,(f(\mathbf{x}_k,\mathbf{u}_k)-f^\star)\\ &\le \left(\frac{2\kappa-1}{2(\kappa-1)}\left(1-\frac{1}{\kappa}\right)+\eta^2G_1^4L_2^2\,(2\kappa-1)\right)\cdot(f(\mathbf{x}_k,\mathbf{u}_k)-f^\star)\\ &\le \left(1-\frac{1}{2\kappa}+\frac{2G_1^4L_2^2\kappa}{L_1^2}\right)\cdot(f(\mathbf{x}_k,\mathbf{u}_k)-f^\star) \end{aligned}$$

As long as $G_1^4\le\frac{\mu^2}{8L_2^2}$, we can guarantee that

$$1-\frac{1}{2\kappa}+\frac{G_1^4L_2^2}{L_1^2}\,(2\kappa-1)\le 1-\frac{1}{2\kappa}+\frac{\mu^2\kappa}{4L_1^2}\le 1-\frac{1}{4\kappa}$$

Therefore, we have

$$f(\mathbf{x}_{k+1},\mathbf{u}_{k+1})-f^\star\le\left(1-\frac{1}{4\kappa}\right)(f(\mathbf{x}_k,\mathbf{u}_k)-f^\star)$$

which implies Condition 2 for all $t\le k+1$.

B.3.3. INDUCTIVE STEP 1: CONDITION 2 ⇒ CONDITION 1

Assume that Condition 2 holds for all $t\le k$. We prove that Condition 1 holds for $k+1$. Notice that the change of $\mathbf{x}$ and $\mathbf{u}$ from initialization can be bounded as

$$\begin{aligned} \|\mathbf{x}_{k+1}-\mathbf{x}_0\|_2&\le\sum_{t=0}^k\|\mathbf{x}_{t+1}-\mathbf{x}_t\|_2=\eta\sum_{t=0}^k\|\nabla_1 f(\mathbf{x}_t,\mathbf{u}_t)\|_2\\ \|\mathbf{u}_{k+1}-\mathbf{u}_0\|_2&\le\sum_{t=0}^k\|\mathbf{u}_{t+1}-\mathbf{u}_t\|_2=\eta\sum_{t=0}^k\|\nabla_2 f(\mathbf{x}_t,\mathbf{u}_t)\|_2 \end{aligned}\tag{21}$$

Since Assumption 2, 3, 4 holds, we can invoke Lemma 18, 20 to have that

$$\|\nabla_1 f(\mathbf{x}_t,\mathbf{u}_t)\|_2^2\le 2L_1\,(f(\mathbf{x}_t,\mathbf{u}_t)-f^\star)\,;\quad \|\nabla_2 f(\mathbf{x}_t,\mathbf{u}_t)\|_2^2\le 2G_1^2L_2\,(f(\mathbf{x}_t,\mathbf{u}_t)-f^\star)$$

Therefore, (21) boils down to

$$\begin{aligned} \|\mathbf{x}_{k+1}-\mathbf{x}_0\|_2&\le\eta\sqrt{2L_1}\sum_{t=0}^k(f(\mathbf{x}_t,\mathbf{u}_t)-f^\star)^{\frac12}\\ \|\mathbf{u}_{k+1}-\mathbf{u}_0\|_2&\le\eta G_1\sqrt{2L_2}\sum_{t=0}^k(f(\mathbf{x}_t,\mathbf{u}_t)-f^\star)^{\frac12} \end{aligned}\tag{22}$$

Moreover, Condition 2 implies that

$$(f(\mathbf{x}_t, \mathbf{u}_t) - f^\star)^{\frac{1}{2}} \leq \left(1 - \frac{1}{4\kappa}\right)^{\frac{t}{2}} (f(\mathbf{x}_0, \mathbf{u}_0) - f^\star)^{\frac{1}{2}} \leq \left(1 - \frac{1}{8\kappa}\right)^{t} (f(\mathbf{x}_0, \mathbf{u}_0) - f^\star)^{\frac{1}{2}}$$

Plugging this bound into (22) gives

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{x}_0\|_2 &\leq \eta\sqrt{2L_1}\,(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star)^{\frac{1}{2}} \sum_{t=0}^{k} \left(1 - \frac{1}{8\kappa}\right)^{t} \\
&\leq 16\eta\kappa\sqrt{L_1}\,(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star)^{\frac{1}{2}} \\
\|\mathbf{u}_{k+1} - \mathbf{u}_0\|_2 &\leq \eta G_1\sqrt{2L_2}\,(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star)^{\frac{1}{2}} \sum_{t=0}^{k} \left(1 - \frac{1}{8\kappa}\right)^{t} \\
&\leq 16\eta\kappa G_1\sqrt{L_2}\,(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star)^{\frac{1}{2}}
\end{aligned} \tag{23}$$

Therefore, as long as

$$R_{\mathbf{x}} \geq 16\eta\kappa\sqrt{L_1}\,(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star)^{1/2};\ R_{\mathbf{u}} \geq 16\eta\kappa G_1\sqrt{L_2}\,(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star)^{1/2}$$

we can guarantee that $\|\mathbf{x}_{k+1} - \mathbf{x}_0\|_2 \leq R_{\mathbf{x}}$ and $\|\mathbf{u}_{k+1} - \mathbf{u}_0\|_2 \leq R_{\mathbf{u}}$. Combining the two inductive steps and the base case completes the proof.

Appendix C. Proofs for Section 3.3

C.1. Proof of Lemma 8

By Assumption 1, we have that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\mathbf{v} \in \mathcal{B}^{(2)}_{R_{\mathbf{u}}}$, it holds that

$$\begin{aligned}
f(\mathbf{y}, \mathbf{u}) &\geq f(\mathbf{x}, \mathbf{u}) + \langle \nabla_1 f(\mathbf{x}, \mathbf{u}), \mathbf{y} - \mathbf{x}\rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 \\
f(\mathbf{x}, \mathbf{u}) &\geq f(\mathbf{y}, \mathbf{u}) + \langle \nabla_1 f(\mathbf{y}, \mathbf{u}), \mathbf{x} - \mathbf{y}\rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|_2^2
\end{aligned}$$

Summing the two inequality gives

$$\langle \nabla_1 f(\mathbf{y}, \mathbf{u}) - \nabla_1 f(\mathbf{x}, \mathbf{u}), \mathbf{y} - \mathbf{x}\rangle \geq \mu\|\mathbf{y} - \mathbf{x}\|_2^2$$

Applying Cauchy-Schwarz inequality to the left-hand side gives

$$\|\nabla_1 f(\mathbf{y}, \mathbf{u}) - \nabla_1 f(\mathbf{x}, \mathbf{u})\| \geq \mu\|\mathbf{y} - \mathbf{x}\|_2 \tag{24}$$

Let $\mathbf{u}, \mathbf{v} \in \mathcal{B}^{(2)}_{R_{\mathbf{u}}}$ be given. Then (24) implies that

$$\|\mathbf{x}^\star(\mathbf{u}) - \mathbf{x}^\star(\mathbf{v})\|_2 \leq \frac{1}{\mu}\|\nabla_1 f(\mathbf{x}^\star(\mathbf{u}), \mathbf{u}) - \nabla_1 f(\mathbf{x}^\star(\mathbf{v}), \mathbf{u})\|_2 \tag{25}$$

By the definition of $\mathbf{x}^\star(\mathbf{u})$ and $\mathbf{x}^\star(\mathbf{v})$, we have

$$\nabla_1 f(\mathbf{x}^\star(\mathbf{u}), \mathbf{u}) = \mathbf{0} = \nabla_1 f(\mathbf{x}^\star(\mathbf{v}), \mathbf{v})$$

Therefore, (25) reduces to

$$\|\mathbf{x}^\star(\mathbf{u}) - \mathbf{x}^\star(\mathbf{v})\|_2 \leq \frac{1}{\mu}\|\nabla_1 f(\mathbf{x}^\star(\mathbf{v}), \mathbf{v}) - \nabla_1 f(\mathbf{x}^\star(\mathbf{v}), \mathbf{u})\|_2 \leq \frac{G_2}{\mu}\|\mathbf{u} - \mathbf{v}\|_2$$

where the last inequality follows from Assumption 5.

C.2. Proof of Lemma 9

Since Assumption 1, 3, 4, 6 holds, we can invoke Lemma 4, 20 to get that

$$\|\nabla_1 f(\mathbf{x},\mathbf{u})\|_2^2 \geq 2\mu\left(f(\mathbf{x},\mathbf{u})-f^\star\right); \quad \|\nabla_2 f(\mathbf{x},\mathbf{u})\|_2^2 \leq 2G_1^2 L_2\left(f(\mathbf{x},\mathbf{u})-f^\star\right)$$

for all $\mathbf{x}\in\mathcal{B}^{(1)}_{R_\mathbf{x}}$ and $\mathbf{u}\in\mathcal{B}^{(2)}_{R_\mathbf{u}}$. Combining the two inequality gives

$$\|\nabla_2 f(\mathbf{x},\mathbf{u})\|_2^2 \leq \frac{G_1^2 L_2}{\mu}\|\nabla_1 f(\mathbf{x},\mathbf{u})\|_2^2$$

Appendix D. Proofs for Section 3.2

D.1. Proof of Theorem 7

Define $\gamma=\frac{c}{2\sqrt{\kappa}-c}$ such that $1+\gamma=\left(1-\frac{c}{2\sqrt{\kappa}}\right)^{-1}$. Let $\lambda=(1+\gamma)^3-1$. Denote $\mathbf{x}_{-1}=\mathbf{y}_{-1}=\mathbf{x}_0$ and $\mathbf{u}_{-1}=\mathbf{v}_{-1}=\mathbf{u}_0$. In this proof, we will focus on the following Lyapunov function

$$\phi_k = f(\mathbf{x}_k,\mathbf{u}_k)-f^\star+\mathcal{Q}_1\left\|\mathbf{z}_k-\mathbf{x}^\star_{k-1}\right\|_2^2+\frac{\eta}{8}\|\nabla_1 f(\mathbf{y}_{k-1},\mathbf{v}_{k-1})\|_2^2 \tag{26}$$

where $\mathbf{x}^\star_k,\mathbf{z}_k$ and $\mathcal{Q}_1$ are defined as

$$\mathbf{x}^\star_k=\arg\min_{\mathbf{x}\in\mathbb{R}^{d_1}} f(\mathbf{x},\mathbf{v}_k);\quad \mathbf{z}_k=\frac{1-\beta\lambda}{\beta\lambda}(\mathbf{y}_k-\mathbf{x}_k)+\mathbf{y}_k;\quad \mathcal{Q}_1=\frac{\lambda^2}{2\eta(1+\gamma)^5}$$

We will prove the following two conditions by induction

Condition 3 *For all $k\leq\hat{k}$, we have*

$$\phi_k\leq\left(1-\frac{c}{4\sqrt{\kappa}}\right)^k\phi_0$$

Condition 4 *For all $k\leq\hat{k}$, we have that $\mathbf{x}_k,\mathbf{y}_k\in\mathcal{B}^{(1)}_{R_\mathbf{x}/2}$ and $\mathbf{u}_k,\mathbf{v}_k\in\mathcal{B}^{(2)}_{R_\mathbf{u}/2}$, and*

$$\begin{aligned}\|\mathbf{x}_k-\mathbf{x}_{k-1}\|_2^2&\leq\frac{6\eta(L_2+1)}{1-\beta}\left(1-\frac{c}{4\sqrt{\kappa}}\right)^k\phi_0\\ \|\mathbf{u}_k-\mathbf{u}_{k-1}\|_2^2&\leq G_1^2\frac{6\eta L_2(L_2+1)(1+\beta)^3}{\mu\beta(1-\beta)^3}\left(1-\frac{c}{4\sqrt{\kappa}}\right)^k\phi_0\end{aligned} \tag{27}$$

Moreover, we have $\eta\|\nabla_1 f(\mathbf{y}_k,\mathbf{v}_k)\|_2\leq\beta\|\mathbf{x}_k-\mathbf{x}_{k-1}\|_2+\|\mathbf{x}_{k+1}-\mathbf{x}_k\|_2$.

Notice that $f(\mathbf{x}_k,\mathbf{u}_k)-f^\star\leq\phi_k$. By Lemma 21, we have $\phi_0\leq 2\left(f(\mathbf{x}_0,\mathbf{u}_0)-f^\star\right)$. Thus, Condition 3 implies that

$$f(\mathbf{x}_k,\mathbf{u}_k)-f^\star\leq 2\left(1-\frac{c}{4\sqrt{\kappa}}\right)^k\left(f(\mathbf{x}_0,\mathbf{u}_0)-f^\star\right)$$

Therefore, Condition 3 and 4 together implies Theorem 7. We now show that the two conditions hold by induction.

D.1.1. BASE CASE: $\hat{k} = 0$

When $\hat{k} = 0$, the only possible $k \leq \hat{k}$ is $k = 0$. In this case, Condition 3 reads $\phi_0 \leq \phi_0$, which is automatically true. Condition 4 also holds automatically since when $k = 0$, we have

$$\|\mathbf{x}_k - \mathbf{x}_{k-1}\|_2 = \|\mathbf{x}_0 - \mathbf{x}_{-1}\|_2 = 0; \quad \|\mathbf{u}_k - \mathbf{u}_{k-1}\|_2 = \|\mathbf{u}_0 - \mathbf{u}_{-1}\|_2 = 0$$

D.1.2. INDUCTIVE STEP 1: CONDITION 3 $\Rightarrow$ CONDITION 4

Assume that Condition 3 holds for all $k \leq \hat{k}$. We want to show that Condition 4 holds for all $k \leq \hat{k} + 1$ Notice that $f(\mathbf{x}_k, \mathbf{u}_k) - f^\star \leq \phi_k$. This implies that

$$f(\mathbf{x}_k, \mathbf{u}_k) - f^\star \leq \left(1 - \frac{c}{4\sqrt{\kappa}}\right)^k \phi_0$$

Combining with Assumption 1-6, and the condition that $G_1 \leq \frac{C_1 \mu^2}{L_2(L_2+1)^2} \left(\frac{1-\beta}{1+\beta}\right)^3$, we can invoke Lemma 16 to conclude directly that Condition 4 holds for all $k \leq \hat{k} + 1$.

D.1.3. INDUCTIVE STEP 2: CONDITION 4 $\Rightarrow$ CONDITION 3

Assume that Condition 3 and 4 holds for all $k \leq \hat{k}$. We show that Condition 3 holds for all $k \leq \hat{k}+1$. To start, we first show that $\mathbf{x}_{\hat{k}+1} \in \mathcal{B}^{(1)}_{R_\mathbf{x}}$ and $\mathbf{u}_{\hat{k}+1} \in \mathcal{B}^{(2)}_{R_\mathbf{u}}$. By the triangle inequality, we have

$$\left\|\mathbf{x}_{\hat{k}+1} - \mathbf{x}_0\right\|_2 \leq \left\|\mathbf{y}_{\hat{k}} - \mathbf{x}_0\right\|_2 + \left\|\mathbf{x}_{\hat{k}+1} - \mathbf{y}_{\hat{k}}\right\|_2; \ \left\|\mathbf{u}_{\hat{k}+1} - \mathbf{x}_0\right\|_2 \leq \left\|\mathbf{v}_{\hat{k}} - \mathbf{x}_0\right\|_2 + \left\|\mathbf{u}_{\hat{k}+1} - \mathbf{v}_{\hat{k}}\right\|_2$$

Since Condition 4 implies that $\mathbf{y}_{\hat{k}} \in \mathcal{B}^{(1)}_{R_\mathbf{x}/2}$ and $\mathbf{v}_{\hat{k}} \in \mathcal{B}^{(2)}_{R_\mathbf{u}/2}$, it suffice to show that $\left\|\mathbf{x}_{\hat{k}+1} - \mathbf{y}_{\hat{k}}\right\|_2 \leq R_\mathbf{x}/2$ and $\left\|\mathbf{u}_{\hat{k}+1} - \mathbf{v}_{\hat{k}}\right\|_2 \leq R_\mathbf{u}/2$. By Lemma 16, we have

$$\begin{aligned}
\left\|\mathbf{x}_{\hat{k}+1} - \mathbf{y}_{\hat{k}}\right\|_2 = \eta \left\|\nabla_1 f(\mathbf{y}_{\hat{k}}, \mathbf{v}_{\hat{k}})\right\|_2 &\leq \beta \left\|\mathbf{x}_{\hat{k}} - \mathbf{x}_{\hat{k}-1}\right\|_2 + \left\|\mathbf{x}_{\hat{k}+1} - \mathbf{x}_{\hat{k}}\right\|_2 \\
&\leq 2\left(6\frac{\eta(L_2+1)}{1-\beta}\phi_0\right)^{\frac{1}{2}} \\
&\leq 8\left(\frac{\eta(L_2+1)}{1-\beta}\right)^{\frac{1}{2}} (f(\mathbf{x}_0, \mathbf{u}_0) - f^\star)^{\frac{1}{2}} \\
&\leq \frac{R_\mathbf{x}}{2}
\end{aligned} \tag{28}$$

This shows that $\mathbf{x}_{\hat{k}+1} \in \mathcal{B}^{(1)}_{R_\mathbf{x}}$. Similarly, we have

$$\begin{aligned}
\left\|\mathbf{x}_{\hat{k}+1} - \mathbf{y}_{\hat{k}}\right\|_2 = \eta \left\|\nabla_2 f(\mathbf{y}_{\hat{k}}, \mathbf{v}_{\hat{k}})\right\|_2 &\leq G_1\sqrt{\frac{L_2}{\mu}} \left(\beta \left\|\mathbf{x}_{\hat{k}} - \mathbf{x}_{\hat{k}-1}\right\|_2 + \left\|\mathbf{x}_{\hat{k}+1} - \mathbf{x}_{\hat{k}}\right\|_2\right) \\
&\leq 2G_1\sqrt{\frac{L_2}{\mu}}\left(6\frac{\eta(L_2+1)}{1-\beta}\phi_0\right)^{\frac{1}{2}} \\
&\leq 8\left(\frac{\eta G_1^2 L_2(L_2+1)}{\mu(1-\beta)}\right)^{\frac{1}{2}} (f(\mathbf{x}_0, \mathbf{u}_0) - f^\star)^{\frac{1}{2}} \\
&\leq \frac{R_\mathbf{u}}{2}
\end{aligned} \tag{29}$$

Thus, $\mathbf{u}_{\hat{k}+1} \in \mathcal{B}^{(1)}_{R_{\mathbf{u}}}$. Now, we can invoke Lemma 17 to get that

$$(1+\gamma)\phi_{\hat{k}+1} \leq \phi_{\hat{k}} + \left(\frac{G_1^2 L_2}{2\gamma} + \frac{4\mathcal{Q}_1 G_2^2}{\gamma\mu^2}(1+\gamma)\right)\beta^2 \left\|\mathbf{u}_{\hat{k}} - \mathbf{u}_{\hat{k}-1}\right\|_2^2 \tag{30}$$

With Condition 4, we can write (30) as

$$\begin{aligned}(1+\gamma)\phi_{\hat{k}+1} - \phi_{\hat{k}} \leq \left(\frac{G_1^4 L_2}{2\gamma} + \frac{4\mathcal{Q}_1 G_1^2 G_2^2}{\gamma\mu^2}(1+\gamma)\right)\frac{6\eta\beta L_2(L_2+1)(1+\beta)^3}{\mu(1-\beta)^3} \\ \cdot \left(1 - \frac{c}{4\sqrt{\kappa}}\right)^{\hat{k}}\phi_0\end{aligned} \tag{31}$$

Since $G_1^4 \leq \frac{C_1\mu^2}{L_2(L_2+1)^2}\left(\frac{1-\beta}{1+\beta}\right)^3$, we have

$$\frac{G_1^4 L_2}{2\gamma} = \frac{C_1\mu^2}{2(L_2+1)^2\gamma}\left(\frac{1-\beta}{1+\beta}\right)^3 \leq \frac{C_1\mu L_1}{c\sqrt{\kappa}(L_2+1)^2}\left(\frac{1-\beta}{1+\beta}\right)^3 = \frac{C_1\mu}{\eta\sqrt{\kappa}(L_2+1)^2}\left(\frac{1-\beta}{1+\beta}\right)^3 \tag{32}$$

where the inequality follows from $\frac{1}{\gamma} \leq \frac{2}{c}\sqrt{\kappa} = \frac{2L_1}{c\mu\sqrt{\kappa}}$. Since $G_1^2 G_2^2 \leq \frac{C_2\mu^3}{L_2(L_2+1)\sqrt{\kappa}}\left(\frac{1-\beta}{1+\beta}\right)^2$, we have

$$\begin{aligned}\frac{4\mathcal{Q}_1 G_1^2 G_2^2}{\gamma\mu^2}(1+\gamma) &= \frac{100 C_2\gamma\mu}{\eta(1+\gamma)^4 L_2(L_2+1)\sqrt{\kappa}}\left(\frac{1-\beta}{1+\beta}\right)^2 \\ &\leq \frac{100 C_2\gamma\mu}{\eta\sqrt{\kappa}L_2(L_2+1)}\left(\frac{1-\beta}{1+\beta}\right)^2 \\ &\leq \frac{100 C_2\mu}{\eta\sqrt{\kappa}L_2(L_2+1)}\left(\frac{1-\beta}{1+\beta}\right)^3\end{aligned} \tag{33}$$

where the last inequality follows from

$$\gamma \leq \frac{c}{\sqrt{\kappa}} \leq \frac{\sqrt{c}}{\sqrt{\kappa}+\sqrt{c}} \leq \frac{1-\beta}{1+\beta}$$

Combining (32) and (33) gives

$$\frac{G_1^4 L_2}{2\gamma} + \frac{4\mathcal{Q}_1 G_1^2 G_2^2}{\gamma\mu^2}(1+\gamma) \leq \frac{(C_1 + 100C_2)\mu}{\eta\sqrt{\kappa}L_2(L_2+1)}\left(\frac{1-\beta}{1+\beta}\right)^3$$

Thus (31) becomes

$$(1+\gamma)\phi_{\hat{k}+1} - \phi_{\hat{k}} \leq \frac{6}{\sqrt{\kappa}}(C_1 + 100C_2)\left(1 - \frac{c}{4\sqrt{\kappa}}\right)^{\hat{k}}\phi_0 \tag{34}$$

Plugging in the value of $\gamma = \frac{c}{2\sqrt{\kappa}-c}$ and choose a small enough C_1 and C_2 gives

$$\left(1 - \frac{c}{2\sqrt{\kappa}}\right)^{-1}\phi_{\hat{k}+1} - \phi_{\hat{k}} \leq \frac{1}{4\sqrt{\kappa}}\left(1 - \frac{c}{4\sqrt{\kappa}}\right)^{\hat{k}}\phi_0 \tag{35}$$

Therefore

$$\begin{aligned}\phi_{\hat{k}+1} &\leq \left(1-\frac{c}{2\sqrt{\kappa}}\right)\phi_{\hat{k}} + \frac{1}{4\sqrt{\kappa}}\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\hat{k}+1}\phi_0 \\ &\leq \left(1-\frac{c}{2\sqrt{\kappa}}\right)\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\hat{k}}\phi_0 + \frac{1}{4\sqrt{\kappa}}\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\hat{k}}\phi_0 \\ &\leq \left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\hat{k}+1}\phi_0\end{aligned}$$

This proves Condition 3 for $\hat{k}+1$. Since we have established the induction for both conditions, we have completed the proof.

D.2. Proof of Lemma 10/16

We first restate an extended version of Lemma 10 here.

Lemma 16 *Suppose that Assumption 1-6 holds with* $G_1^4 \leq \frac{C_1\mu^2}{L_2(L_2+1)^2}\left(\frac{1-\beta}{1+\beta}\right)^3$ *and for all* $k \leq \hat{k}$

$$f(\mathbf{x}_k, \mathbf{u}_k) - f^\star \leq \left(1-\frac{c}{4\sqrt{\kappa}}\right)^k \phi_0 \tag{36}$$

Then we have $\mathbf{x}_k, \mathbf{y}_k \in \mathcal{B}_{R_\mathbf{x}}^{(1)}$ *and* $\mathbf{u}_k, \mathbf{v}_k \in \mathcal{B}_{R_\mathbf{u}}^{(2)}$ *for all* $k \leq \hat{k}+1$ *with*

$$\begin{aligned}R_\mathbf{x} &\geq \frac{18}{c}\sqrt{\kappa}\left(\frac{3\eta(L_2+1)}{1-\beta}\right)^{\frac{1}{2}}\left(f(\mathbf{x}_0,\mathbf{u}_0)-f^\star\right)^{\frac{1}{2}} \\ R_\mathbf{u} &\geq \frac{18}{c}\sqrt{\kappa}\left(\frac{6\eta G_1^2 L_2(L_2+1)(1+\beta)^3}{\mu\beta(1-\beta)^3}\right)\left(f(\mathbf{x}_0,\mathbf{u}_0)-f^\star\right)^{\frac{1}{2}}\end{aligned}$$

Moreover, for all $k \leq \hat{k}+1$, *we have* $\eta\left\|\nabla_1 f(\mathbf{y}_k,\mathbf{v}_k)\right\|_2 \leq \beta\left\|\mathbf{x}_k-\mathbf{x}_{k-1}\right\|_2 + \left\|\mathbf{x}_{k+1}-\mathbf{x}_k\right\|_2$ *and*

$$\begin{aligned}\left\|\mathbf{x}_k-\mathbf{x}_{k-1}\right\|_2^2 &\leq \frac{6\eta(L_2+1)}{1-\beta}\left(1-\frac{c}{4\sqrt{\kappa}}\right)^k\phi_0 \\ \left\|\mathbf{u}_k-\mathbf{u}_{k-1}\right\|_2^2 &\leq G_1^2\frac{6\eta L_2(L_2+1)(1+\beta)^3}{\mu\beta(1-\beta)^3}\left(1-\frac{c}{4\sqrt{\kappa}}\right)^k\phi_0\end{aligned} \tag{37}$$

Proof We prove Lemma 16 by induction. Notice that for $k=0$, all the statements are automatically true. Assume that 37 holds up to iteration k, we first show that $\mathbf{x}_k, \mathbf{y}_k \in \mathcal{B}_{R_\mathbf{x}}^{(1)}$ and $\mathbf{u}_k, \mathbf{v}_k \in \mathcal{B}_{R_\mathbf{u}}^{(2)}$.

Step 1: $\mathbf{x}_k, \mathbf{y}_k \in \mathcal{B}_{R_\mathbf{x}}^{(1)}$**.** By the triangle inequality, we have

$$\begin{aligned}\left\|\mathbf{x}_k-\mathbf{x}_0\right\|_2 &\leq \sum_{t=0}^{k-1}\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|_2 \\ &\leq \sqrt{\frac{6\eta(L_2+1)}{1-\beta}\phi_0}\sum_{t=0}^{k-1}\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{k}{2}} \\ &\leq \sqrt{\frac{6\eta(L_2+1)}{1-\beta}\phi_0}\cdot\frac{1}{1-\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{1}{2}}}\end{aligned} \tag{38}$$

Notice that $\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{1}{2}} \leq 1-\frac{c}{8\sqrt{\kappa}}$. Moreover, by lemma 21, we have $\phi_0 \leq 2\left(f(\mathbf{x}_0,\mathbf{u}_0)-f^\star\right)$ Thus (38) becomes

$$\|\mathbf{x}_k-\mathbf{x}_0\|_2 \leq \frac{8}{c}\sqrt{\kappa}\cdot\sqrt{\frac{6\eta(L_2+1)}{1-\beta}\phi_0} \leq \frac{16}{c}\sqrt{\kappa}\left(\frac{3\eta(L_2+1)}{1-\beta}\right)^{\frac{1}{2}}\left(f(\mathbf{x}_0,\mathbf{u}_0)-f^\star\right)^{\frac{1}{2}} \tag{39}$$

This shows that $\|\mathbf{x}_k-\mathbf{x}_0\|_2 \leq R_{\mathbf{x}}$. Moreover, since $\mathbf{y}_k-\mathbf{x}_k=\beta(\mathbf{x}_k-\mathbf{x}_{k-1})$, we have

$$\|\mathbf{y}_k-\mathbf{x}_0\|_2 \leq \|\mathbf{y}_k-\mathbf{x}_k\|_2+\|\mathbf{x}_k-\mathbf{x}_0\|_2 = \beta\|\mathbf{x}_k-\mathbf{x}_{k-1}\|+\|\mathbf{x}_k-\mathbf{x}_0\|_2$$

By the inductive hypothesis, we have $\|\mathbf{x}_k-\mathbf{x}_{k-1}\|_2^2 \leq \frac{6\eta(L_2+1)}{1-\beta}\left(1-\frac{c}{4\sqrt{\kappa}}\right)^k\phi_0$. Combining with the bound in (39), we have

$$\|\mathbf{y}_k-\mathbf{x}_0\|_2 \leq \left(\beta+\frac{8}{c}\sqrt{\kappa}\right)\sqrt{\frac{6\eta(L_2+1)}{1-\beta}\phi_0} \leq \frac{18}{c}\sqrt{\kappa}\left(\frac{3\eta(L_2+1)}{1-\beta}\right)^{\frac{1}{2}}\left(f(\mathbf{x}_0,\mathbf{u}_0)-f^\star\right)^{\frac{1}{2}}$$

This shows that $\|\mathbf{y}_k-\mathbf{x}_0\|_2 \leq R_{\mathbf{x}}$.

Step 2: $\mathbf{u}_k,\mathbf{v}_k \in \mathcal{B}^{(2)}_{R_{\mathbf{u}}}$. Now, we focus on $\mathbf{u}_k$ and $\mathbf{v}_k$. Similar to (38), we have

$$\begin{aligned}
\|\mathbf{u}_k-\mathbf{u}_0\|_2 &\leq \sum_{t=0}^{k-1}\|\mathbf{u}_{t+1}-\mathbf{u}_t\|_2 \\
&\leq G_1\sqrt{\frac{6\eta L_2(L_2+1)(1+\beta)^3}{\mu\beta(1-\beta)^3}\phi_0}\sum_{t=0}^{k-1}\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{k}{2}} \\
&\leq G_1\sqrt{\frac{6\eta L_2(L_2+1)(1+\beta)^3}{\mu\beta(1-\beta)^3}\phi_0}\cdot\frac{1}{1-\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{1}{2}}} \\
&\leq \frac{8}{c}\sqrt{\kappa}G_1\sqrt{\frac{6\eta L_2(L_2+1)(1+\beta)^3}{\mu\beta(1-\beta)^3}\phi_0} \\
&\leq \frac{16}{c}\sqrt{\kappa}\left(\frac{6\eta G_1^2 L_2(L_2+1)(1+\beta)^3}{\mu\beta(1-\beta)^3}\right)\left(f(\mathbf{x}_0,\mathbf{u}_0)-f^\star\right)^{\frac{1}{2}}
\end{aligned} \tag{40}$$

This shows that $\|\mathbf{u}_k-\mathbf{u}_0\|_2 \leq R_{\mathbf{u}}$. Moreover, for $\mathbf{v}_k$ we have

$$\|\mathbf{v}_k-\mathbf{u}_0\|_2 \leq \beta\|\mathbf{u}_k-\mathbf{u}_{k-1}\|_2+\|\mathbf{u}_k-\mathbf{u}_0\|_2$$

By the inductive hypothesis, we have $\|\mathbf{u}_k-\mathbf{u}_{k-1}\|_2^2 \leq G_1^2\frac{6\eta L_2(L_2+1)(1+\beta)^3}{\mu\beta(1-\beta)^3}\left(1-\frac{c}{4\sqrt{\kappa}}\right)^k\phi_0$. Therefore

$$\begin{aligned}
\|\mathbf{v}_k-\mathbf{u}_0\|_2 &\leq \left(\beta+\frac{8}{c}\sqrt{\kappa}\right)G_1\sqrt{\frac{6\eta L_2(L_2+1)(1+\beta)^3}{\mu\beta(1-\beta)^3}\phi_0} \\
&\leq \frac{18}{c}\sqrt{\kappa}\left(\frac{6\eta G_1^2 L_2(L_2+1)(1+\beta)^3}{\mu\beta(1-\beta)^3}\right)\left(f(\mathbf{x}_0,\mathbf{u}_0)-f^\star\right)^{\frac{1}{2}}
\end{aligned}$$

This shows that $\|\mathbf{v}_k - \mathbf{u}_0\|_2 \leq R_{\mathbf{u}}$. Thus, we have shown that $\mathbf{x}_k, \mathbf{y}_k \in \mathcal{B}^{(1)}_{R_{\mathbf{x}}}$ and $\mathbf{u}_k, \mathbf{v}_k \in \mathcal{B}^{(2)}_{R_{\mathbf{u}}}$.
Step 3: Bounding $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2$. Now, we show that (37) holds with $k+1$. We start by recalling the iterates of Nesterov's momentum (4). This iteration implies that

$$\mathbf{x}_{k+1} = \mathbf{y}_k - \eta \nabla_1 f(\mathbf{y}_k, \mathbf{v}_k) = \mathbf{x}_k + \beta(\mathbf{x}_k - \mathbf{x}_{k-1}) - \eta \nabla_1 f(\mathbf{y}_k, \mathbf{v}_k)$$
$$\mathbf{u}_{k+1} = \mathbf{v}_k - \eta \nabla_2 f(\mathbf{y}_k, \mathbf{v}_k) = \mathbf{u}_k + \beta(\mathbf{u}_k - \mathbf{u}_{k-1}) - \eta \nabla_2 f(\mathbf{y}_k, \mathbf{v}_k)$$

Therefore, we can conclude that

$$\mathbf{x}_{k+1} - \mathbf{x}_k = \beta(\mathbf{x}_k - \mathbf{x}_{k-1}) - \eta \nabla_1 f(\mathbf{y}_k, \mathbf{v}_k);\ \mathbf{u}_{k+1} - \mathbf{u}_k = \beta(\mathbf{u}_k - \mathbf{u}_{k-1}) - \eta \nabla_2 f(\mathbf{y}_k, \mathbf{v}_k)$$

For the convenience of our analysis, we will define

$$\mathcal{D}_{\mathbf{x},k} = \|\mathbf{x}_k - \mathbf{x}_{k-1}\|_2 ; \quad \mathcal{D}_{\mathbf{u},k} = \|\mathbf{u}_k - \mathbf{u}_{k-1}\|_2$$

To start, we notice that $\mathcal{D}_{\mathbf{x},k+1}$ can be expanded as

$$\begin{aligned} \mathcal{D}^2_{\mathbf{x},k+1} &= \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 \\ &= \|\beta(\mathbf{x}_k - \mathbf{x}_{k-1}) - \eta \nabla_1 f(\mathbf{y}_k, \mathbf{v}_k)\|_2^2 \\ &= \beta^2 \mathcal{D}^2_{\mathbf{x},k} - 2\eta\beta \langle \nabla_1 f(\mathbf{y}_k, \mathbf{v}_k), \mathbf{x}_k - \mathbf{x}_{k-1} \rangle + \eta^2 \|\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k)\|_2^2 \end{aligned} \tag{41}$$

Both the inner product term and the gradient norm in (41) need to be bounded carefully. For the inner product term, since $\mathbf{v}_k \in \mathcal{B}^{(2)}_{R_{\mathbf{u}}}$, we invoke Assumption 1

$$\begin{aligned} f(\mathbf{x}_k, \mathbf{v}_k) &\geq f(\mathbf{y}_k, \mathbf{v}_k) + \langle \nabla_1 f(\mathbf{y}_k, \mathbf{v}_k), \mathbf{x}_k - \mathbf{y}_k \rangle + \frac{\mu}{2} \|\mathbf{x}_k - \mathbf{y}_k\|_2^2 \\ &\geq f(\mathbf{y}_k, \mathbf{v}_k) - \beta \langle \nabla_1 f(\mathbf{y}_k, \mathbf{v}_k), \mathbf{x}_k - \mathbf{x}_{k-1} \rangle \end{aligned} \tag{42}$$

where the last inequality follows from $\mathbf{x}_k - \mathbf{y}_k = -\beta(\mathbf{x}_k - \mathbf{x}_{k-1})$ and $\|\mathbf{x}_k - \mathbf{y}_k\|_2^2 \geq 0$. Therefore (42) implies that

$$-\beta \langle \nabla_1 f(\mathbf{y}_k, \mathbf{v}_k), \mathbf{x}_k - \mathbf{x}_{k-1} \rangle \leq f(\mathbf{x}_k, \mathbf{v}_k) - f(\mathbf{y}_k, \mathbf{v}_k) \tag{43}$$

For the gradient norm, since $\mathbf{v}_k \in \mathcal{B}^{(2)}_{R_{\mathbf{u}}}$, Assumption 2 must hold. Therefore, we can invoke Lemma 18 to get that

$$\|\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k)\|_2^2 \leq 2L_1 (f(\mathbf{y}_k, \mathbf{v}_k) - f^\star) \tag{44}$$

Plugging (43) and (44) into (41), we can get that

$$\mathcal{D}^2_{\mathbf{x},k+1} \leq \beta^2 \mathcal{D}^2_{\mathbf{x},k} + 2\eta \left(f(\mathbf{x}_k, \mathbf{v}_k) - f(\mathbf{y}_k, \mathbf{v}_k)\right) + 2\eta^2 L_1 (f(\mathbf{y}_k, \mathbf{v}_k) - f^\star) \tag{45}$$

Since we choose $\eta = \frac{c}{L_1}$, as long as $c \leq 1$, we will have $2\eta^2 L_1 \leq 2\eta$. Moreover, since $f(\mathbf{y}_k, \mathbf{v}_k) - f^\star \geq 0$, (45) can be further bounded as

$$\begin{aligned} \mathcal{D}^2_{\mathbf{x},k+1} &\leq \beta^2 \mathcal{D}^2_{\mathbf{x},k} + 2\eta \left(f(\mathbf{x}_k, \mathbf{v}_k) - f(\mathbf{y}_k, \mathbf{v}_k)\right) + 2\eta (f(\mathbf{y}_k, \mathbf{v}_k) - f^\star) \\ &= \beta^2 \mathcal{D}^2_{\mathbf{x},k} + 2\eta \left(f(\mathbf{x}_k, \mathbf{v}_k) - f^\star\right) \end{aligned} \tag{46}$$

With $\mathbf{x}_k \in \mathcal{B}_{R_\mathbf{x}}^{(1)}$ and $\mathbf{u}_k, \mathbf{v}_k \in \mathcal{B}_{R_\mathbf{u}}^{(2)}$, we must have that Assumption 3, 4 holds. Therefore, we can invoke Lemma 5 with $\mathcal{Q} = 1$ to get that

$$f(\mathbf{x}_k, \mathbf{v}_k) \leq f(\mathbf{x}_k, \mathbf{u}_k) + L_2(f(\mathbf{x}_k, \mathbf{u}_k) - f^\star) + \frac{G_1^2}{2}(L_2+1)\|\mathbf{u}_k - \mathbf{v}_k\|_2^2$$

Subtracting $f^\star$ from both sides, and use the fact that $\mathbf{u}_k - \mathbf{v}_k = -\beta(\mathbf{u}_k - \mathbf{u}_{k-1})$, we have

$$f(\mathbf{x}_k, \mathbf{v}_k) - f^\star \leq (L_2+1)\left(f(\mathbf{x}_k, \mathbf{u}_k) - f^\star + \frac{G_1^2\beta^2}{2}\|\mathbf{u}_k - \mathbf{u}_{k-1}\|_2^2\right) \tag{47}$$

Plugging (47) into (46), and recalling that $\|\mathbf{u}_k - \mathbf{u}_{k-1}\|_2^2 = \mathcal{D}_{\mathbf{u},k}^2$ gives

$$\mathcal{D}_{\mathbf{x},k+1}^2 \leq \beta^2\mathcal{D}_{\mathbf{x},k}^2 + 2\eta(L_2+1)\left(f(\mathbf{x}_k, \mathbf{u}_k) - f^\star + \frac{G_1^2\beta^2}{2}\mathcal{D}_{\mathbf{u},k}^2\right) \tag{48}$$

Recall that $f(\mathbf{x}_k, \mathbf{u}_k) - f^\star$ satisfies (36). Moreover, by the inductive assumption, we have

$$\mathcal{D}_{\mathbf{x},k}^2 \leq \frac{6\eta(L_2+1)}{1-\beta}\left(1 - \frac{c}{4\sqrt{\kappa}}\right)^k \phi_0; \quad \mathcal{D}_{\mathbf{u},k}^2 \leq G_1^2\frac{6\eta L_2(L_2+1)(1+\beta)^3}{\mu\beta(1-\beta)^3}\left(1 - \frac{c}{4\sqrt{\kappa}}\right)^k \phi_0$$

Therefore, (47) becomes

$$\mathcal{D}_{\mathbf{x},k+1}^2 \leq \left(\underbrace{\frac{6\beta^2\eta(L_2+1)}{1-\beta} + 4\eta(L_2+1) + \frac{6\eta^2\beta G_1^4 L_2(L_2+1)^2(1+\beta)^3}{\mu(1-\beta)^3}}_{\zeta_1}\right)\left(1 - \frac{c}{4\sqrt{\kappa}}\right)^k \phi_0 \tag{49}$$

With $G_1^4 \leq \frac{C_1\mu^2}{L_2(L_2+1)^2}\left(\frac{1-\beta}{1+\beta}\right)^3$ and $\eta = \frac{c}{L_1} \leq \frac{1}{\mu}$, we can derive that

$$\frac{6\eta^2\beta G_1^4 L_2(L_2+1)^2(1+\beta)^3}{\mu(1-\beta)^3} \leq 6C_1\eta^2\beta\mu \leq 6C_1\eta\beta \leq 2\eta$$

with a small enough C_1. Therefore, ζ_1 in (49) can be simplified to

$$\begin{aligned}
\zeta_1 &\leq \frac{6\beta^2\eta(L_2+1)}{1-\beta} + 4\eta(L_2+1) + 2\eta \\
&\leq \frac{6\eta(L_2+1)}{1-\beta}\left(\beta^2 + \frac{2}{3}(1-\beta) + \frac{1-\beta}{3(L_2+1)}\right) \\
&\leq \frac{6\eta(L_2+1)}{1-\beta}\left(1 - \beta + \beta^2\right)
\end{aligned}$$

By the choice of $\beta = \frac{4\sqrt{\kappa}-\sqrt{c}}{4\sqrt{\kappa}+7\sqrt{c}}$, we have

$$\beta - \beta^2 = \frac{8\sqrt{c}(4\sqrt{\kappa}-\sqrt{c})}{(4\sqrt{\kappa}+7\sqrt{c})^2} \leq \frac{c}{4\sqrt{\kappa}}$$

Thus, ζ_1 in (49) satisfies $\zeta_1 \leq \frac{6\eta(L_2+1)}{1-\beta}\left(1-\frac{c}{4\sqrt{\kappa}}\right)$ and (49) becomes

$$\mathcal{D}^2_{\mathbf{x},k+1} \leq \frac{6\eta(L_2+1)}{1-\beta}\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{k+1}\phi_0 \tag{50}$$

This proves the inductive step for $\mathcal{D}^2_{\mathbf{x},k+1}$.

Step 4: Bounding $\|\mathbf{u}_{k+1}-\mathbf{u}_k\|_2$. Next, we focus on $\mathcal{D}_{\mathbf{u},k}$. We first need a bound on $\|\nabla_2 f(\mathbf{y}_k,\mathbf{v}_k)\|_2$. To start, (41) implies that

$$\begin{aligned}\eta^2\|\nabla_1 f(\mathbf{y}_k,\mathbf{v}_k)\|_2^2 &= \mathcal{D}^2_{\mathbf{x},k+1}-\beta^2\mathcal{D}^2_{\mathbf{x},k}+2\eta\beta\langle\nabla_1 f(\mathbf{y}_k,\mathbf{v}_k),\mathbf{x}_k-\mathbf{x}_{k-1}\rangle\\ &\leq \mathcal{D}^2_{\mathbf{x},k+1}-\beta^2\mathcal{D}^2_{\mathbf{x},k}+2\eta\beta\mathcal{D}_{\mathbf{x},k}\|\nabla_1 f(\mathbf{y}_k,\mathbf{v}_k)\|_2\end{aligned}$$

which implies that

$$\eta^2\|\nabla_1 f(\mathbf{y}_k,\mathbf{v}_k)\|_2^2-2\eta\beta\mathcal{D}_{\mathbf{x},k}\|\nabla_1 f(\mathbf{y}_k,\mathbf{v}_k)\|_2+\beta^2\mathcal{D}^2_{\mathbf{x},k}-\mathcal{D}^2_{\mathbf{x},k+1}\leq 0$$

Therefore, for all $\mathcal{G}=\eta\|\nabla_1 f(\mathbf{y}_k,\mathbf{v}_k)\|_2$, $\mathcal{G}$ must satisfy

$$\mathcal{G}^2-2\beta\mathcal{D}_{\mathbf{x},k}\mathcal{G}+\beta^2\mathcal{D}^2_{\mathbf{x},k}-\mathcal{D}^2_{\mathbf{x},k+1}\leq 0 \tag{51}$$

When (51) takes equality, the two solutions of $\mathcal{G}$ are

$$\mathcal{G}_{\text{lb}}=\beta\mathcal{D}_{\mathbf{x},k}-\mathcal{D}_{\mathbf{x},k+1};\quad \mathcal{G}_{\text{ub}}=\beta\mathcal{D}_{\mathbf{x},k}+\mathcal{D}_{\mathbf{x},k+1}$$

Therefore, $\|\nabla_1 f(\mathbf{y}_k,\mathbf{v}_k)\|_2$ must be bounded by

$$\eta\|\nabla_1 f(\mathbf{y}_k,\mathbf{v}_k)\|_2\leq\mathcal{G}_{\text{ub}}=\beta\mathcal{D}_{\mathbf{x},k}+\mathcal{D}_{\mathbf{x},k+1} \tag{52}$$

Moreover, since Assumption 1,3,4, and 6 holds, and that $\mathbf{y}_k\in\mathcal{B}^{(1)}_{R_{\mathbf{x}}}$ and $\mathbf{v}_k\in\mathcal{B}^{(2)}_{R_{\mathbf{u}}}$, we can apply Lemma 9 to (49) get that

$$\eta\|\nabla_2 f(\mathbf{y}_k,\mathbf{v}_k)\|_2\leq G_1\sqrt{\frac{L_2}{\mu}}\left(\beta\mathcal{D}_{\mathbf{x},k}+\mathcal{D}_{\mathbf{x},k+1}\right) \tag{53}$$

Recall that $\mathbf{u}_{k+1}-\mathbf{u}_k=\beta(\mathbf{u}_k-\mathbf{u}_{k-1})-\eta\nabla_2 f(\mathbf{y}_k,\mathbf{v}_k)$. Unrolling this recursion gives

$$\mathbf{u}_{k+1}-\mathbf{u}_k=\eta\sum_{t=0}^{k}\beta^{k-t}\nabla_2 f(\mathbf{y}_t,\mathbf{v}_t)$$

Now, we can bound $\mathcal{D}_{\mathbf{u},k+1}$ as

$$\begin{aligned}\mathcal{D}_{\mathbf{u},k+1}&\leq\eta\left\|\sum_{t=0}^{k}\beta^{k-t}\nabla_2 f(\mathbf{y}_t,\mathbf{v}_t)\right\|_2\leq\eta\sum_{t=0}^{k}\beta^{k-t}\|\nabla_2 f(\mathbf{y}_t,\mathbf{v}_t)\|_2\\ &\leq G_1\sqrt{\frac{L_2}{\mu}}\sum_{t=0}^{k}\beta^{k-t}\left(\beta\mathcal{D}_{\mathbf{x},t}+\mathcal{D}_{\mathbf{x},t+1}\right)\\ &=G_1\sqrt{\frac{L_2}{\mu}}\left(\sum_{t=0}^{k}\beta^{k-t+1}\mathcal{D}_{\mathbf{x},t}+\sum_{t=0}^{k}\beta^{k-t}\mathcal{D}_{\mathbf{x},t+1}\right)\end{aligned} \tag{54}$$

Recall the inductive hypothesis

$$\mathcal{D}_{\mathbf{x},k}^2 \leq \frac{6\eta(L_2+1)}{1-\beta}\left(1-\frac{c}{4\sqrt{\kappa}}\right)^k \phi_0$$

Since we have shows that this bound also hold for $k+1$ in (50), we can write (54) as

$$\mathcal{D}_{\mathbf{u},k+1} = G_1\sqrt{\frac{6\eta L_2(L_2+1)}{\mu(1-\beta)}}\phi_0^{\frac{1}{2}}\left(\sum_{t=0}^{k}\beta^{k-t+1}\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{t}{2}} + \sum_{t=0}^{k}\beta^{k-t}\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{t+1}{2}}\right) \tag{55}$$

By the standard geometric series result, we have

$$\sum_{t=0}^{k}\beta^{k-t+1}\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{t}{2}} = \beta\cdot\frac{\beta^{k+1}-\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{k+1}{2}}}{\beta-\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{1}{2}}}$$

$$\sum_{t=0}^{k}\beta^{k-t}\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{t+1}{2}} = \left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{1}{2}}\cdot\frac{\beta^{k+1}-\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{k+1}{2}}}{\beta-\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{1}{2}}}$$

By our choice of β, we must have that $\beta \leq 1-\frac{c}{4\sqrt{\kappa}} \leq \left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{1}{2}}$. Therefore (55) becomes

$$\begin{aligned}\mathcal{D}_{\mathbf{u},k+1} &\leq G_1\sqrt{\frac{6\eta L_2(L_2+1)}{\mu(1-\beta)}}\phi_0\cdot\left(\beta+\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{1}{2}}\right)\frac{\beta^{k+1}-\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{k+1}{2}}}{\beta-\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{1}{2}}} \\ &= G_1\sqrt{\frac{6\eta L_2(L_2+1)}{\mu(1-\beta)}}\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{k+1}{2}}\phi_0\cdot\frac{\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{1}{2}}+\beta}{\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{1}{2}}-\beta}\end{aligned}$$

Which implies that

$$\mathcal{D}_{\mathbf{u},k+1}^2 \leq \frac{6\eta G_1^2 L_2(L_2+1)}{\mu(1-\beta)}\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{k+1}\phi_0\cdot\left(\frac{\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{1}{2}}+\beta}{\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{1}{2}}-\beta}\right)^2 \tag{56}$$

We notice that, since $\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{1}{2}} \leq 1-\frac{c}{8\sqrt{\kappa}}$, we have

$$\begin{aligned}\frac{\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{1}{2}}+\beta}{\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{1}{2}}-\beta} &= \frac{1-\frac{c}{4\sqrt{\kappa}}+\beta^2+2\beta\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{1}{2}}}{1-\frac{c}{4\sqrt{\kappa}}+\beta^2+2\beta\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{1}{2}}} \\ &\leq \frac{1-\frac{c}{4\sqrt{\kappa}}+\beta^2-\beta\left(2-\frac{c}{4\sqrt{\kappa}}\right)}{1-\frac{c}{4\sqrt{\kappa}}+\beta^2-\beta\left(2-\frac{c}{4\sqrt{\kappa}}\right)} \\ &= \frac{\left(1-\frac{c}{4\sqrt{\kappa}}+\beta\right)(1+\beta)}{\left(1-\frac{c}{4\sqrt{\kappa}}-\beta\right)(1-\beta)}\end{aligned}$$

Since $\beta-\beta^2 \leq \frac{c}{4\sqrt{\kappa}}$, we have

$$1-\frac{c}{4\sqrt{\kappa}}+\beta \leq 1+\beta; \quad 1-\frac{c}{4\sqrt{\kappa}}-\beta \geq 1-\beta+\beta^2 \geq \beta(1-\beta)$$

This gives

$$\frac{\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{1}{2}}+\beta}{\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{\frac{1}{2}}-\beta} \leq \frac{(1+\beta)^2}{\beta(1-\beta)^2}$$

Thus, (56) becomes the following form, which proves the inductive step for $\mathcal{D}^2_{\mathbf{u},k+1}$.

$$\mathcal{D}^2_{\mathbf{u},k+1} \leq \frac{6\eta G_1^2 L_2(L_2+1)(1+\beta)^3}{\mu\beta(1-\beta)^3}\left(1-\frac{c}{4\sqrt{\kappa}}\right)^{k+1}\phi_0 \tag{57}$$

■

D.3. Proof of Lemma 11/17

We first restate an extended version of Lemma 10 here.

Lemma 17 *Suppose that Assumption 1-6 holds with* $G_1^4 \leq \frac{C_1\mu^2}{L_2(L_2+1)^2}\left(\frac{1-\beta}{1+\beta}\right)^3$ *and* $G_1^2G_2^2 \leq \frac{C_2\mu^3}{L_2(L_2+1)\sqrt{\kappa}}\left(\frac{1-\beta}{1+\beta}\right)^2$. *If* $\mathbf{x}_{k+1}, \mathbf{x}_k, \mathbf{y}_k \in \mathcal{B}^{(1)}_{R_\mathbf{x}}$ *and* $\mathbf{u}_{k+1}, \mathbf{u}_k, \mathbf{v}_k \in \mathcal{B}^{(2)}_{R_\mathbf{u}}$ *for all* $k \leq \hat{k}$, *then for all* $k \leq \hat{k}$ *we have*

$$(1+\gamma)\phi_{k+1} \leq \phi_k + \left(\frac{G_1^2L_2}{2\gamma} + \frac{4\mathcal{Q}_1G_2^2}{\gamma\mu^2}(1+\gamma)\right)\beta^2\left\|\mathbf{u}_k-\mathbf{u}_{k-1}\right\|_2^2$$

where $\gamma = \frac{c}{2\sqrt{\kappa}-c}$ *and* $\mathcal{Q}_1 = \frac{\lambda^2}{2\eta(1+\gamma)^5}$ *with* $\lambda = (1+\gamma)^3-1$.

Proof We will derive the bound for $(1+\gamma)\phi_{k+1} - \phi_k$. This quantity can be written as

$$\begin{aligned}(1+\gamma)\phi_{k+1} - \phi_k = &\underbrace{(1+\gamma)\left(f(\mathbf{x}_{k+1}, \mathbf{u}_{k+1}) - f^\star\right) - \left(f(\mathbf{x}_k, \mathbf{u}_k) - f^\star\right)}_{\Delta_1} \\ &+ \mathcal{Q}_1 \left(\underbrace{(1+\gamma)\left\|\mathbf{z}_{k+1} - \mathbf{x}_k^\star\right\|_2^2 - \left\|\mathbf{z}_k - \mathbf{x}_{k-1}^\star\right\|_2^2}_{\Delta_2} \right) \\ &+ \frac{\eta}{8}(1+\gamma)\left\|\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k\right\|_2^2 - \frac{\eta}{8}\left\|\nabla_1 f(\mathbf{y}_{k-1}, \mathbf{v}_{k-1})\right\|_2^2\end{aligned} \tag{58}$$

■

In the following parts of the proof, we will bound Δ_1 and Δ_2 respectively. Throughout the proof, we will define $\Delta\mathbf{y}_k = \mathbf{y}_k - \mathbf{x}_k^\star$, $\Delta\mathbf{z}_k = \mathbf{z}_k - \mathbf{x}_k^\star$, and $\Delta\mathbf{x}_k = \mathbf{x}_k - \mathbf{x}_k^\star$

Bound on Δ_1. We first study $f(\mathbf{x}_{k+1}, \mathbf{u}_{k+1}) - f^\star$. It can be decomposed as

$$f(\mathbf{x}_{k+1}, \mathbf{u}_{k+1}) - f^\star = f(\mathbf{x}_{k+1}, \mathbf{v}_k) - f^\star + f(\mathbf{x}_{k+1}, \mathbf{u}_{k+1}) - f(\mathbf{x}_{k+1}, \mathbf{v}_k) \tag{59}$$

Since $\mathbf{x}_{k+1} \in \mathcal{B}_{R_\mathbf{x}}^{(1)}$, and $\mathbf{u}_{k+1}, \mathbf{v}_k \in \mathcal{B}_{R_\mathbf{u}}^{(2)}$, we can invoke Lemma 5 with $\mathcal{Q} = {}^{L_2}\!/_{\gamma}$ to get that

$$f(\mathbf{x}_{k+1}, \mathbf{u}_{k+1}) - f(\mathbf{x}_{k+1}, \mathbf{v}_k) \leq \gamma\left(f(\mathbf{x}_{k+1}, \mathbf{v}_k) - f^\star\right) + \frac{G_1^2 L_2}{2\gamma}(1+\gamma)\left\|\mathbf{u}_{k+1} - \mathbf{v}_k\right\|_2^2$$

Notice that $\left\|\mathbf{u}_{k+1} - \mathbf{v}_k\right\|_2^2 = \eta^2\left\|\nabla_2 f(\mathbf{y}_k, \mathbf{v}_k)\right\|_2^2 \leq \eta^2 G_1^2 \cdot \frac{L_2}{\mu}\left\|\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k)\right\|_2^2$. Thus

$$\frac{G_1^2 L_2}{2\gamma}(1+\gamma)\left\|\mathbf{u}_{k+1} - \mathbf{v}_k\right\|_2^2 \leq \frac{\eta^2 G_1^4 L_2^2}{2\gamma\mu}(1+\gamma)\left\|\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k)\right\|_2^2 \leq \frac{\eta}{4}(1+\gamma)\left\|\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k)\right\|_2^2$$

for the choice of $G_1^4 \leq \frac{C_1\mu^2}{L_2(L_2+1)^2}\left(\frac{1-\beta}{1+\beta}\right)^3 \leq \frac{\mu L_1 \gamma}{2cL_2^2} = \frac{\mu\gamma}{2\eta L_2^2}$. In this way, (59) becomes

$$f(\mathbf{x}_{k+1}, \mathbf{u}_{k+1}) - f^\star \leq (1+\gamma)(f(\mathbf{x}_{k+1}, \mathbf{v}_k) - f^\star) + \frac{\eta}{4}(1+\gamma)\left\|\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k)\right\|_2^2 \tag{60}$$

Again, since $\mathbf{x}_{k+1} \in \mathcal{B}_{R_\mathbf{x}}^{(1)}$, and $\mathbf{u}_{k+1}, \mathbf{v}_k \in \mathcal{B}_{R_\mathbf{u}}^{(2)}$, we can use Assumption 2 and the iterate $\mathbf{x}_{k+1} = \mathbf{y}_k - \eta\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k)$ to get the

$$\begin{aligned}f(\mathbf{x}_{k+1}, \mathbf{v}_k) - f^\star &\leq f(\mathbf{y}_k, \mathbf{v}_k) - f^\star + \langle\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k, \mathbf{x}_{k+1} - \mathbf{y}_k\rangle + \frac{L_1}{2}\left\|\mathbf{x}_{k+1} - \mathbf{y}_k\right\|_2^2 \\ &= f(\mathbf{y}_k, \mathbf{v}_k) - f^\star - \eta\left\|\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k)\right\|_2^2 + \frac{\eta^2 L_1}{2}\left\|\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k)\right\|_2^2 \\ &= f(\mathbf{y}_k, \mathbf{v}_k) - f^\star - \eta\left(1 - \frac{c}{2}\right)\left\|\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k)\right\|_2^2\end{aligned} \tag{61}$$

Combining (61) with (60) gives

$$f(\mathbf{x}_{k+1}, \mathbf{u}_{k+1}) - f^\star \leq (1+\gamma)(f(\mathbf{y}_k, \mathbf{v}_k) - f^\star) - \eta(1+\gamma)\left(\frac{3}{4} - \frac{c}{2}\right)\left\|\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k)\right\|_2^2 \tag{62}$$

Next, we study $f(\mathbf{x}_k, \mathbf{u}_k) - f^\star$. It can be decomposed as

$$f(\mathbf{x}_k, \mathbf{u}_k) - f^\star = f(\mathbf{x}_k, \mathbf{v}_k) - f^\star - (f(\mathbf{x}_k, \mathbf{v}_k) - f(\mathbf{x}_k, \mathbf{u}_k)) \tag{63}$$

Since $\mathbf{x}_k \in \mathcal{B}^{(1)}_{R_\mathbf{x}}$, and $\mathbf{u}_k, \mathbf{v}_k \in \mathcal{B}^{(2)}_{R_\mathbf{u}}$, we can invoke Lemma 5 with $\mathcal{Q} = {}^{L_2}/_{\gamma}$ to get that

$$f(\mathbf{x}_k, \mathbf{v}_k) - f(\mathbf{x}_k, \mathbf{u}_k) \leq \gamma(f(\mathbf{x}_k, \mathbf{u}_k) - f^\star) + \frac{G_1^2 L_2}{2\gamma}(1+\gamma) \left\|\mathbf{u}_k - \mathbf{v}_k\right\|_2^2$$

Therefore, (63) becomes

$$f(\mathbf{x}_k, \mathbf{u}_k) - f^\star \geq f(\mathbf{x}_k, \mathbf{v}_k) - f^\star - \gamma(f(\mathbf{x}_k, \mathbf{u}_k) - f^\star) - \frac{G_1^2 L_2}{2\gamma}(1+\gamma) \left\|\mathbf{u}_k - \mathbf{v}_k\right\|_2^2$$

which implies that

$$f(\mathbf{x}_k, \mathbf{u}_k) - f^\star \geq \frac{1}{1+\gamma}(f(\mathbf{x}_k, \mathbf{v}_k) - f^\star) - \frac{G_1^2 L_2}{2\gamma} \left\|\mathbf{u}_k - \mathbf{v}_k\right\|_2^2 \tag{64}$$

Now, recall the definition of Δ_1 in (58). Combined with (61) and (64), we can write Δ_1 as

$$\begin{aligned}
\Delta_1 &= (1+\gamma)(f(\mathbf{x}_{k+1}, \mathbf{u}_{k+1}) - f^\star) - (f(\mathbf{x}_k, \mathbf{u}_k) - f^\star) \\
&\leq (1+\gamma)^2(f(\mathbf{y}_k, \mathbf{v}_k) - f^\star) - \eta(1+\gamma)^2 \left(\frac{3}{4} - \frac{c}{2}\right) \left\|\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k)\right\|_2^2 \\
&\qquad - \frac{1}{1+\gamma}(f(\mathbf{x}_k, \mathbf{v}_k) - f^\star) + \frac{G_1^2 L_2}{2\gamma} \left\|\mathbf{u}_k - \mathbf{v}_k\right\|_2^2 \\
&= \frac{1}{1+\gamma}(f(\mathbf{y}_k, \mathbf{v}_k) - f(\mathbf{x}_k, \mathbf{v}_k)) + \frac{\lambda}{1+\gamma}(f(\mathbf{y}_k, \mathbf{v}_k) - f^\star) \\
&\qquad - \eta(1+\gamma)^2 \left(\frac{3}{4} - \frac{c}{2}\right) \left\|\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k)\right\|_2^2 + \frac{G_1^2 L_2}{2\gamma} \left\|\mathbf{u}_k - \mathbf{v}_k\right\|_2^2
\end{aligned} \tag{65}$$

Since $\mathbf{v}_k \in \mathcal{B}^{(2)}_{R_\mathbf{u}}$, we can apply Assumption 1 to get that

$$\begin{aligned}
f(\mathbf{x}_k, \mathbf{v}_k) &\geq f(\mathbf{y}_k, \mathbf{v}_k) + \langle \nabla_1 f(\mathbf{y}_k, \mathbf{v}_k), \mathbf{x}_k - \mathbf{y}_k \rangle \\
f^\star = f(\mathbf{x}_k^\star, \mathbf{v}_k) &\geq f(\mathbf{y}_k, \mathbf{v}_k) + \langle \nabla_1 f(\mathbf{y}_k, \mathbf{v}_k), \mathbf{x}_k^\star - \mathbf{y}_k \rangle + \frac{\mu}{2} \left\|\mathbf{x}_k^\star - \mathbf{y}_k\right\|
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\frac{1}{1+\gamma}(f(\mathbf{y}_k, \mathbf{v}_k) - f(\mathbf{x}_k, \mathbf{v}_k)) + \frac{\lambda}{1+\gamma}(f(\mathbf{y}_k, \mathbf{v}_k) - f^\star) \\
&\quad \leq \frac{1}{1+\gamma} \langle \nabla_1 f(\mathbf{y}_k, \mathbf{v}_k), \mathbf{y}_k - \mathbf{x}_k \rangle + \frac{\lambda}{1+\gamma} \langle \nabla_1 f(\mathbf{y}_k, \mathbf{v}_k), \mathbf{y}_k - \mathbf{x}_k^\star \rangle_2^2 \\
&\qquad - \frac{\mu\lambda}{2(1+\gamma)} \left\|\mathbf{x}_k^\star - \mathbf{y}_k\right\| \\
&\quad = \frac{1}{1+\gamma} \langle \nabla_1 f(\mathbf{y}_k, \mathbf{v}_k), \mathbf{y}_k - \mathbf{x}_k + \lambda(\mathbf{y}_k - \mathbf{x}_k^\star) \rangle - \frac{\mu\lambda}{2(1+\gamma)} \left\|\mathbf{x}_k^\star - \mathbf{y}_k\right\|_2^2
\end{aligned} \tag{66}$$

Recall that $\mathbf{z}_k = \frac{1-\beta\lambda}{\beta\lambda}(\mathbf{y}_k - \mathbf{x}_k) + \mathbf{y}_k$. This implies that $\mathbf{y}_k - \mathbf{x}_k = \frac{\beta\lambda}{1-\beta\lambda}(\mathbf{z}_k - \mathbf{y}_k)$. Therefore

$$\begin{aligned}\mathbf{y}_k - \mathbf{x}_k + \lambda(\mathbf{y}_k - \mathbf{x}_k^\star) &= \frac{\beta\lambda}{1-\beta\lambda}(\mathbf{z}_k - \mathbf{y}_k) + \lambda(\mathbf{y}_k - \mathbf{x}_k^\star) \\ &= \frac{\lambda}{1-\beta\lambda}\left(\beta(\mathbf{z}_k - \mathbf{y}_k) + (1-\beta\lambda)(\mathbf{y}_k - \mathbf{x}_k^\star)\right) \\ &= \frac{\lambda}{1-\beta\lambda}\left(\beta(\mathbf{z}_k - \mathbf{x}_k^\star) + (1-\beta\lambda-\beta)(\mathbf{y}_k - \mathbf{x}_k^\star)\right)\end{aligned}$$

Recall that $\Delta\mathbf{z}_k = \mathbf{z}_k - \mathbf{x}_k^\star$ and $\Delta\mathbf{y}_k = \mathbf{y}_k - \mathbf{x}_k^\star$. Thus (66) becomes

$$\begin{aligned}&\frac{1}{1+\gamma}(f(\mathbf{y}_k, \mathbf{v}_k) - f(\mathbf{x}_k, \mathbf{v}_k)) + \frac{\lambda}{1+\gamma}(f(\mathbf{y}_k, \mathbf{v}_k) - f^\star) \\ &\quad\leq \frac{\lambda}{(1+\gamma)(1-\beta\lambda)}\left\langle \nabla_1 f(\mathbf{y}_k, \mathbf{v}_k), \beta\Delta\mathbf{z}_k + (1-\beta\lambda-\beta)\Delta\mathbf{y}_k\right\rangle - \frac{\mu\lambda}{2(1+\gamma)}\left\|\Delta\mathbf{y}_k\right\|_2^2\end{aligned}$$

Combining with (65), the bound on Δ_1 becomes

$$\begin{aligned}\Delta_1 \leq \frac{\lambda}{(1+\gamma)(1-\beta\lambda)}\left\langle \nabla_1 f(\mathbf{y}_k, \mathbf{v}_k), \beta\Delta\mathbf{z}_k + (1-\beta\lambda-\beta)\Delta\mathbf{y}_k\right\rangle - \frac{\mu\lambda}{2(1+\gamma)}\left\|\Delta\mathbf{y}_k\right\|_2^2 \\ - \eta(1+\gamma)^2\left(\frac{3}{4} - \frac{c}{2}\right)\left\|\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k)\right\|_2^2 + \frac{G_1^2 L_2}{2\gamma}\left\|\mathbf{u}_k - \mathbf{v}_k\right\|_2^2\end{aligned} \tag{67}$$

Bound on Δ_2. Now we turn to the bound on Δ_2. Recall that Δ_2 is defined as

$$\Delta_2 = (1+\gamma)\left\|\mathbf{z}_{k+1} - \mathbf{x}_k^\star\right\|_2^2 - \left\|\mathbf{z}_k - \mathbf{x}_{k-1}^\star\right\|_2^2$$

To start, we notice that, since $\mathbf{v}_k, \mathbf{v}_{k-1} \in \mathcal{B}_{R_\mathbf{u}}^{(2)}$, we can invoke Lemma 8 to get that

$$\left\|\mathbf{x}_k^\star - \mathbf{x}_{k-1}^\star\right\|_2 \leq \frac{G_2}{\mu}\left\|\mathbf{v}_k - \mathbf{v}_{k-1}\right\|_2$$

Therefore, we can lower-bound $\left\|\mathbf{z}_k - \mathbf{x}_{k-1}^\star\right\|_2^2$ as

$$\begin{aligned}\left\|\mathbf{z}_k - \mathbf{x}_{k-1}^\star\right\|_2^2 &\geq \left\|\Delta\mathbf{z}_k\right\| + 2\left\langle\Delta\mathbf{z}_k, \mathbf{x}_k - \mathbf{x}_{k-1}^\star\right\rangle \\ &\geq \left(1 - \frac{\gamma}{2(1+\gamma)}\right)\left\|\Delta\mathbf{z}_k\right\|_2^2 - \frac{2}{\gamma}(1+\gamma)\left\|\mathbf{x}_k - \mathbf{x}_{k-1}^\star\right\|_2^2 \\ &\geq \frac{2+\gamma}{2+2\gamma}\left\|\Delta\mathbf{z}_k\right\|_2^2 - \frac{2G_2^2}{\gamma\mu^2}(1+\gamma)\left\|\mathbf{v}_k - \mathbf{v}_{k-1}\right\|_2^2\end{aligned} \tag{68}$$

Next, we provide an upper bound for $\|\mathbf{z}_{k+1} - \mathbf{x}_k^\star\|_2^2$. Recall that $\mathbf{y}_{k+1} = \mathbf{x}_{k+1} + \beta(\mathbf{x}_{k+1} - \mathbf{x}_k)$ and $\mathbf{x}_{k+1} = \mathbf{y}_k - \eta\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k)$. We can re-write $\mathbf{z}_{k+1}$ as

$$\begin{aligned}
\mathbf{z}_{k+1} &= \frac{1-\beta\lambda}{\beta\lambda}(\mathbf{y}_{k+1} - \mathbf{x}_{k+1}) + \mathbf{y}_{k+1} \\
&= \frac{1}{\beta\lambda}\mathbf{y}_{k+1} - \frac{1-\beta\lambda}{\beta\lambda}\mathbf{x}_{k+1} \\
&= \frac{1+\beta}{\beta\lambda}\mathbf{x}_{k+1} - \frac{1}{\lambda}\mathbf{x}_k - \frac{1-\beta\lambda}{\beta\lambda}\mathbf{x}_{k+1} \\
&= \frac{1+\lambda}{\lambda}\mathbf{x}_{k+1} - \frac{1}{\lambda}\mathbf{x}_k \\
&= \frac{1+\lambda}{\lambda}\mathbf{y}_k - \frac{1}{\lambda}\mathbf{x}_k - \frac{1+\lambda}{\lambda}\eta\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k) \\
&= \frac{\beta}{1-\beta\lambda}\mathbf{z}_k + \left(1 - \frac{\beta}{1-\beta\lambda}\right)\mathbf{y}_k - \frac{\eta}{\lambda}(1+\lambda)\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k)
\end{aligned}$$

Therefore, we can write $\|\mathbf{z}_{k+1} - \mathbf{x}_k^\star\|_2^2$ as

$$\begin{aligned}
\|\mathbf{z}_{k+1} - \mathbf{x}_k^\star\|_2^2 &= \left\| \frac{\beta}{1-\beta\lambda}\Delta\mathbf{z}_k + \left(1 - \frac{\beta}{1-\beta\lambda}\right)\Delta\mathbf{y}_k - \frac{\eta}{\lambda}(1+\lambda)\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k) \right\|_2^2 \\
&= \frac{\beta^2}{(1-\beta\lambda)^2}\|\Delta\mathbf{z}_k\|_2^2 + \frac{(1-\beta\lambda-\beta)^2}{(1-\beta\lambda)^2}\|\Delta\mathbf{y}_k\|_2^2 \\
&\quad - \frac{2\eta(1+\lambda)}{\lambda(1-\beta\lambda)}\langle\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k), \beta\Delta\mathbf{z}_k + (1-\beta\lambda-\beta)\Delta\mathbf{y}_k\rangle \\
&\quad + \frac{2\beta(1-\beta\lambda-\beta)}{(1-\beta\lambda)^2}\langle\Delta\mathbf{z}_k, \Delta\mathbf{y}_k\rangle + \frac{\eta^2}{\lambda^2}(1+\lambda)^2\|\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k)\|_2^2 \qquad (69) \\
&= \frac{\beta^2}{(1-\beta\lambda)^2}\|\Delta\mathbf{z}_k\|_2^2 + \frac{(1-\beta\lambda-\beta)^2}{(1-\beta\lambda)^2}\|\Delta\mathbf{y}_k\|_2^2 \\
&\quad - \frac{\lambda}{\mathcal{Q}_1(1+\gamma)^2(1-\beta\lambda)}\langle\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k), \beta\Delta\mathbf{z}_k + (1-\beta\lambda-\beta)\Delta\mathbf{y}_k\rangle \\
&\quad + \frac{2\beta(1-\beta\lambda-\beta)}{(1-\beta\lambda)^2}\langle\Delta\mathbf{z}_k, \Delta\mathbf{y}_k\rangle + \frac{\eta(1+\gamma)}{2\mathcal{Q}_1}\|\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k)\|_2^2
\end{aligned}$$

where the last inequality follows from $\mathcal{Q}_1 = \frac{\lambda^2}{2\eta(1+\gamma)^2}$ and $\lambda = (1+\gamma)^3 - 1$. Therefore, combining (69) and (68) gives

$$\begin{aligned}
\Delta_2 \leq &\left(\frac{\beta^2(1+\gamma)}{(1-\beta\lambda)^2} - \frac{2+\gamma}{2+2\gamma}\right)\|\Delta\mathbf{z}_k\|_2^2 + (1+\gamma)\cdot\frac{(1-\beta\lambda-\beta)^2}{(1-\beta\lambda)^2}\|\Delta\mathbf{y}_k\|_2^2 \\
&- \frac{\lambda}{\mathcal{Q}_1(1+\gamma)(1-\beta\lambda)}\langle\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k), \beta\Delta\mathbf{z}_k + (1-\beta\lambda-\beta)\Delta\mathbf{y}_k\rangle \\
&+ \frac{2(1+\gamma)\beta(1-\beta\lambda-\beta)}{(1-\beta\lambda)^2}\langle\Delta\mathbf{z}_k, \Delta\mathbf{y}_k\rangle + \frac{\eta(1+\gamma)^2}{2\mathcal{Q}_1}\|\nabla_1 f(\mathbf{y}_k, \mathbf{v}_k)\|_2^2 \qquad (70) \\
&+ \frac{2G_2^2}{\gamma\mu^2}(1+\gamma)\|\mathbf{v}_k - \mathbf{v}_{k-1}\|_2^2
\end{aligned}$$

Putting things together. Now, going back to (58). With the help of (67) and (70), we have

$$\begin{aligned}
(1+\gamma)\phi_{k+1}-\phi_k &= \Delta_1+\mathcal{Q}_1\Delta_2+\frac{\eta}{8}(1+\gamma)\left\|\nabla_1 f(\mathbf{y}_k,\mathbf{v}_k)\right\|_2^2-\frac{\eta}{8}\left\|\nabla_1 f(\mathbf{y}_{k-1},\mathbf{v}_{k-1})\right\|_2^2\\
&= \frac{\lambda}{(1+\gamma)(1-\beta\lambda)}\left\langle\nabla_1 f(\mathbf{y}_k,\mathbf{v}_k),\beta\Delta\mathbf{z}_k+(1-\beta\lambda-\beta)\Delta\mathbf{y}_k\right\rangle\\
&\quad-\frac{\mu\lambda}{2(1+\gamma)}\left\|\Delta\mathbf{y}_k\right\|_2^2-\eta(1+\gamma)^2\left(\frac{3}{4}-\frac{c}{2}\right)\left\|\nabla_1 f(\mathbf{y}_k,\mathbf{v}_k)\right\|_2^2\\
&\quad+\frac{G_1^2L_2}{2\gamma}\left\|\mathbf{u}_k-\mathbf{v}_k\right\|_2^2+\mathcal{Q}_1\left(\frac{\beta^2(1+\gamma)}{(1-\beta\lambda)^2}-\frac{2+\gamma}{2+2\gamma}\right)\left\|\Delta\mathbf{z}_k\right\|_2^2\\
&\quad+\mathcal{Q}_1(1+\gamma)\cdot\frac{(1-\beta\lambda-\beta)^2}{(1-\beta\lambda)^2}\left\|\Delta\mathbf{y}_k\right\|_2^2\\
&\quad-\frac{\lambda}{(1+\gamma)(1-\beta\lambda)}\left\langle\nabla_1 f(\mathbf{y}_k,\mathbf{v}_k),\beta\Delta\mathbf{z}_k+(1-\beta\lambda-\beta)\Delta\mathbf{y}_k\right\rangle\\
&\quad+\frac{2\mathcal{Q}_1(1+\gamma)\beta(1-\beta\lambda-\beta)}{(1-\beta\lambda)^2}\left\langle\Delta\mathbf{z}_k,\Delta\mathbf{y}_k\right\rangle\\
&\quad+\frac{\eta}{2}(1+\gamma)^2\left\|\nabla_1 f(\mathbf{y}_k,\mathbf{v}_k)\right\|_2^2+\frac{2\mathcal{Q}_1G_2^2}{\gamma\mu^2}(1+\gamma)\left\|\mathbf{v}_k-\mathbf{v}_{k-1}\right\|_2^2\\
&\quad+\frac{\eta}{8}(1+\gamma)\left\|\nabla_1 f(\mathbf{y}_k,\mathbf{v}_k)\right\|_2^2-\frac{\eta}{8}\left\|\nabla_1 f(\mathbf{y}_{k-1},\mathbf{v}_{k-1})\right\|_2^2\\
&= -\eta(1+\gamma)^2\left(\frac{1}{4}-\frac{c}{2}\right)\left\|\nabla_1 f(\mathbf{y}_k,\mathbf{v}_k)\right\|_2^2-\frac{\eta}{8}\left\|\nabla_1 f(\mathbf{y}_{k-1},\mathbf{v}_{k-1})\right\|_2^2\\
&\quad+\mathcal{E}_{1,k}+\mathcal{E}_{2,k}
\end{aligned}\tag{71}$$

where $\mathcal{E}_{1,k}$ and $\mathcal{E}_{2,k}$ are defined as

$$\begin{aligned}
\mathcal{E}_{1,k} &= \mathcal{Q}_1\left(\frac{\beta^2(1+\gamma)}{(1-\beta\lambda)^2}-\frac{2+\gamma}{2+2\gamma}\right)\left\|\Delta\mathbf{z}_k\right\|_2^2+\frac{2\mathcal{Q}_1(1+\gamma)\beta(1-\beta\lambda-\beta)}{(1-\beta\lambda)^2}\left\langle\Delta\mathbf{z}_k,\Delta\mathbf{y}_k\right\rangle\\
&\quad+\left(\mathcal{Q}_1(1+\gamma)\cdot\frac{(1-\beta\lambda-\beta)^2}{(1-\beta\lambda)^2}-\frac{\mu\lambda}{2(1+\gamma)}\right)\left\|\Delta\mathbf{y}_k\right\|_2^2\\
\mathcal{E}_{2,k} &= \frac{G_1^2L_2}{2\gamma}\left\|\mathbf{u}_k-\mathbf{v}_k\right\|_2^2+\frac{2\mathcal{Q}_1G_2^2}{\gamma\mu^2}(1+\gamma)\left\|\mathbf{v}_k-\mathbf{v}_{k-1}\right\|_2^2
\end{aligned}\tag{72}$$

Our first step is to show that $\mathcal{E}_{1,k}\le 0$. To start, notice that

$$\frac{\beta^2(1+\gamma)}{(1-\beta\lambda)^2}-\frac{2+\gamma}{2+2\gamma}\le 0\Leftrightarrow\beta\le\frac{\sqrt{1+\gamma/2}}{1+\gamma+\lambda\sqrt{1+\gamma/2}}$$

By definition of β and using $\lambda=(1+\gamma)^3-1\ge 3\gamma$, we have

$$\beta=\frac{4\sqrt{\kappa}-\sqrt{c}}{4\sqrt{\kappa}+7\sqrt{c}}\le\frac{4\sqrt{\kappa}-2c}{4\sqrt{\kappa}+8c}=\frac{1}{1+5\gamma}\le\frac{\sqrt{1+\gamma/2}}{1+\gamma+\lambda\sqrt{1+\gamma/2}}$$

for a small enough constant c. Thus, we can guarantee that $\frac{\beta^2(1+\gamma)}{(1-\beta\lambda)^2} - \frac{2+\gamma}{2+2\gamma} \leq 0$. Therefore

$$\frac{2\mathcal{Q}_1(1+\gamma)\beta(1-\beta\lambda-\beta)}{(1-\beta\lambda)^2}\langle \Delta\mathbf{z}_k, \Delta\mathbf{y}_k\rangle \leq \mathcal{Q}_1\left(\frac{2+\gamma}{2+2\gamma} - \frac{\beta^2(1+\gamma)}{(1-\beta\lambda)^2}\right)\|\Delta\mathbf{z}_k\|_2^2 + \frac{\mathcal{Q}_1(1+\gamma)^2\beta^2(1-\beta\lambda-\beta)^2}{\left(\frac{2+\gamma}{2+2\gamma} - \frac{\beta^2(1+\gamma)}{(1-\beta\lambda)^2}\right)(1-\beta\lambda)^4}\|\Delta\mathbf{y}_k\|_2^2$$

This implies that

$$\mathcal{E}_{1,k} \leq \left(\frac{\mathcal{Q}_1(1+\gamma)^2\beta^2(1-\beta\lambda-\beta)^2}{\left(\frac{2+\gamma}{2+2\gamma} - \frac{\beta^2(1+\gamma)}{(1-\beta\lambda)^2}\right)(1-\beta\lambda)^4} + \mathcal{Q}_1(1+\gamma)\cdot\frac{(1-\beta\lambda-\beta)^2}{(1-\beta\lambda)^2} - \frac{\mu\lambda}{2(1+\gamma)}\right)\|\Delta\mathbf{y}_k\|_2^2$$

To show that $\mathcal{E}_{1,k} \leq 0$, it suffice to show that

$$\frac{\mathcal{Q}_1(1+\gamma)^2\beta^2(1-\beta\lambda-\beta)^2}{\left(\frac{2+\gamma}{2+2\gamma} - \frac{\beta^2(1+\gamma)}{(1-\beta\lambda)^2}\right)(1-\beta\lambda)^4} + \mathcal{Q}_1(1+\gamma)\cdot\frac{(1-\beta\lambda-\beta)^2}{(1-\beta\lambda)^2} \leq \frac{\mu\lambda}{2(1+\gamma)}$$

Moving $\mathcal{Q}_1(1+\gamma)$ to the right-hand side gives

$$\frac{(1+\gamma)\beta^2(1-\beta\lambda-\beta)^2}{\left(\frac{2+\gamma}{2+2\gamma} - \frac{\beta^2(1+\gamma)}{(1-\beta\lambda)^2}\right)(1-\beta\lambda)^4} + \frac{(1-\beta\lambda-\beta)^2}{(1-\beta\lambda)^2} \leq \frac{\mu\lambda}{2\mathcal{Q}_1(1+\gamma)^2}$$

By the definition of $\mathcal{Q}_1$, we have $\frac{\mu\lambda}{2\mathcal{Q}_1(1+\gamma)^2} = \frac{\eta\mu(1+\gamma)^3}{\lambda} \geq \frac{\eta\mu(1+\gamma)^3}{7\gamma} \geq \frac{c}{7\kappa\gamma}$. Moreover,

$$\begin{aligned}
&\frac{(1+\gamma)\beta^2(1-\beta\lambda-\beta)^2}{\left(\frac{2+\gamma}{2+2\gamma} - \frac{\beta^2(1+\gamma)}{(1-\beta\lambda)^2}\right)(1-\beta\lambda)^4} + \frac{(1-\beta\lambda-\beta)^2}{(1-\beta\lambda)^2} \\
&\quad\leq \frac{(1+\gamma)\beta^2(1-\beta\lambda-\beta)^2 + (1-\beta\lambda-\beta)^2(1-\beta\lambda)^2 - (1+\gamma)\beta^2(1-\beta\lambda-\beta)^2}{\left(\frac{2+\gamma}{2+2\gamma} - \frac{\beta^2(1+\gamma)}{(1-\beta\lambda)^2}\right)(1-\beta\lambda)^4} \\
&\quad= \frac{(1-\beta\lambda-\beta)^2}{\left(\frac{2+\gamma}{2+2\gamma} - \frac{\beta^2(1+\gamma)}{(1-\beta\lambda)^2}\right)(1-\beta\lambda)^2} \\
&\quad= \frac{\left(1-\frac{\beta}{1-\beta\lambda}\right)^2}{1-\frac{\gamma}{2(1+\gamma)} - \left(\frac{\beta}{1-\beta\lambda}\right)^2}
\end{aligned}$$

Therefore, to make $\mathcal{E}_{1,k} \leq 0$, we just need

$$\frac{\left(1-\frac{\beta}{1-\beta\lambda}\right)^2}{1-\frac{\gamma}{2(1+\gamma)} - \left(\frac{\beta}{1-\beta\lambda}\right)^2} \leq \frac{c}{7\kappa\gamma}$$

Recall that $\beta = \frac{4\sqrt{\kappa}-\sqrt{c}}{4\sqrt{\kappa}+\sqrt{c}} = \frac{2\sqrt{c}-(1-2\sqrt{c})\gamma}{2\sqrt{c}+(7+2\sqrt{c})\gamma} \geq \frac{2\sqrt{c}-\gamma}{2\sqrt{c}+8\gamma}$. This implies that

$$\frac{\beta}{1-\beta\lambda} \geq \frac{\beta}{1-7\beta\gamma} \geq \frac{2\sqrt{c}-\gamma}{2\sqrt{c}+8\gamma}$$

Therefore

$$\begin{aligned}\frac{\left(1-\frac{\beta}{1-\beta\lambda}\right)^2}{1-\frac{\gamma}{2(1+\gamma)}-\left(\frac{\beta}{1-\beta\lambda}\right)^2} &\leq \frac{\left(\frac{9\gamma}{2\sqrt{c}+8\gamma}\right)^2}{1-\frac{\gamma}{2}-\left(\frac{2\sqrt{c}-\gamma}{2\sqrt{c}+8\gamma}\right)^2} \\ &\leq \frac{81\gamma^2}{\left(1-\frac{\gamma}{2}\right)(2\sqrt{c}+8\gamma)^2-(2\sqrt{c}-\gamma)^2} \\ &\leq \frac{81\gamma}{36\sqrt{c}-2c} \leq \frac{81\gamma}{35\sqrt{c}}\end{aligned}$$

Thus, to make $\mathcal{E}_{1,k} \leq 0$, we just need

$$\frac{81\gamma}{35\sqrt{c}} \leq \frac{c}{7\kappa\gamma} \Rightarrow \gamma \leq \frac{5c^{\frac{3}{4}}}{81\sqrt{\kappa}}$$

Since our choice of γ is $\gamma = \frac{c}{2\sqrt{\kappa}-c} \leq \frac{5c^{\frac{3}{4}}}{81\kappa}$ for a smalle enough c, we can guarantee that $\mathcal{E}_{1,k} \leq 0$. Thus, (71) becomes

$$(1+\gamma)\phi_{k+1}-\phi_k \leq -\eta(1+\gamma)^2\left(\frac{1}{8}-\frac{c}{2}\right)\|\nabla_1 f(\mathbf{y}_k,\mathbf{v}_k)\|_2^2-\frac{\eta}{8}\|\nabla_1 f(\mathbf{y}_{k-1},\mathbf{v}_{k-1})\|_2^2+\mathcal{E}_{2,k} \quad (73)$$

With $c \leq \frac{1}{4}$, it becomes

$$\begin{aligned}(1+\gamma)\phi_{k+1}-\phi_k \leq \frac{G_1^2L_2}{2\gamma}\|\mathbf{u}_k-\mathbf{v}_k\|_2^2+\frac{2\mathcal{Q}_1G_2^2}{\gamma\mu^2}(1+\gamma)\|\mathbf{v}_k-\mathbf{v}_{k-1}\|_2^2 \\ -\frac{\eta}{8}\|\nabla_1 f(\mathbf{y}_{k-1},\mathbf{v}_{k-1})\|_2^2\end{aligned} \quad (74)$$

By the iterates of Nesterov's momentum, we have

$$\mathbf{u}_k-\mathbf{v}_k=-\beta(\mathbf{u}_k-\mathbf{u}_{k-1});\ \mathbf{v}_k-\mathbf{v}_{k-1}=-\eta\nabla_2 f(\mathbf{y}_{k-1},\mathbf{v}_{k-1})+\beta(\mathbf{u}_k-\mathbf{u}_{k-1})$$

Therefore, $\|\mathbf{u}_k-\mathbf{v}_k\|_2^2=\beta^2\|\mathbf{u}_k-\mathbf{u}_{k-1}\|_2^2$ and

$$\begin{aligned}\|\mathbf{v}_k-\mathbf{v}_{k-1}\|_2^2 &\leq 2\eta^2\|\nabla_2 f(\mathbf{y}_{k-1},\mathbf{v}_{k-1})\|_2^2+2\beta^2\|\mathbf{u}_k-\mathbf{u}_{k-1}\|_2^2 \\ &\leq \frac{2\eta^2G_1^2L_2}{\mu}\|\nabla_1 f(\mathbf{y}_{k-1},\mathbf{v}_{k-1})\|_2^2+2\beta^2\|\mathbf{u}_k-\mathbf{u}_{k-1}\|_2^2\end{aligned}$$

where in the last inequality we invoke Lemma 9. In this way, (74) becomes

$$\begin{aligned}(1+\gamma)\phi_{k+1}-\phi_k \leq \beta^2\left(\frac{G_1^2L_2}{2\gamma}+\frac{4\mathcal{Q}_1G_2^2}{\gamma\mu^2}(1+\gamma)\right)\|\mathbf{u}_k-\mathbf{u}_{k-1}\|_2^2+ \\ +\eta\left(\frac{4\eta\mathcal{Q}_1G_1^2G_2^2L_2}{\gamma\mu^3}(1+\gamma)-\frac{1}{8}\right)\|\nabla_1 f(\mathbf{y}_{k-1},\mathbf{v}_{k-1})\|_2^2\end{aligned} \quad (75)$$

Plugging in the requirement $G_1^2G_2^2 \leq \frac{C_2\mu^3}{L_2(L_2+1)\sqrt{\kappa}}\left(\frac{1-\beta}{1+\beta}\right)^2$ gives that $\frac{4\eta\mathcal{Q}_1G_1^2G_2^2L_2}{\gamma\mu^3}(1+\gamma) \leq \frac{1}{8}$. Therefore, (75) becomes

$$(1+\gamma)\phi_{k+1}-\phi_k \leq \beta^2\left(\frac{G_1^2L_2}{2\gamma}+\frac{4\mathcal{Q}_1G_2^2}{\gamma\mu^2}(1+\gamma)\right)\|\mathbf{u}_k-\mathbf{u}_{k-1}\|_2^2$$

which completes the proof.

Appendix E. Proofs for Section 4.1

E.1. Proof of Lemma 12

To start, recall that our objective function is defined as

$$f(\mathbf{x},\mathbf{u}) = \frac{1}{2}\left\|\mathbf{A}_1\mathbf{x} + \sigma\left(\mathbf{A}_2\mathbf{u}\right) - \mathbf{b}\right\|_2^2 \tag{76}$$

We first compute its gradient and its Hessian

$$\begin{aligned}\nabla_1 f(\mathbf{x},\mathbf{u}) &= \mathbf{A}_1^\top\left(\mathbf{A}_1\mathbf{x} + \sigma\left(\mathbf{A}_2\mathbf{u}\right) - \mathbf{b}\right)\\ \nabla_{11} f(\mathbf{x},\mathbf{u}) &= \mathbf{A}_1^\top\mathbf{A}_1\end{aligned} \tag{77}$$

Thus, $\mathbf{a}^\top\nabla_{11}f(\mathbf{x},\mathbf{u})\mathbf{a} = \|\mathbf{A}_1\mathbf{a}\|_2^2$. This implies that $\lambda_{\max}\left(\nabla_{11}f(\mathbf{x},\mathbf{u})\right) \le \sigma_{\max}\left(\mathbf{A}_1\right)^2$. Moreover, since $\mathbf{A}_1 \in \mathbb{R}^{m\times m}$, we can also know that $\lambda_{\min}\left(\nabla_{11}f(\mathbf{x},\mathbf{u})\right) \ge \sigma_{\min}\left(\mathbf{A}_1\right)^2$. Thus, Assumption 1, 2 holds with $\mu = \sigma_{\min}\left(\mathbf{A}_1\right)^2$ and $L_1 = \sigma_{\max}\left(\mathbf{A}_1\right)^2$. Since g is defined as $g(\mathbf{s}) = \frac{1}{2}\|\mathbf{s}-\mathbf{b}\|_2^2$, it must be 1-smooth. Moreover, its minimum values is 0. Going back to f, we notice that since $\sigma_{\min}(\mathbf{A}_1) > 0$, choosing $\mathbf{x}^\star(\mathbf{u}) = \left(\mathbf{A}_1^\top\mathbf{A}_1\right)^{-1}\mathbf{A}_1^\top\left(\mathbf{b} - \sigma\left(\mathbf{A}_2\mathbf{u}\right)\right)$ gives $f(\mathbf{x},\mathbf{u}) = 0$. This shows that Assumption 3,6 hold with $L_2 = 1$. For Assumption 4, we can compute that

$$\begin{aligned}\|h(\mathbf{x},\mathbf{u}) - h(\mathbf{x},\mathbf{v})\|_2 &= \|\sigma\left(\mathbf{A}_2\mathbf{u}\right) - \sigma\left(\mathbf{A}_2\mathbf{v}\right)\|_2\\ &\le B\left\|\mathbf{A}_2\mathbf{u} - \mathbf{A}_2\mathbf{v}\right\|_2\\ &\le B\sigma_{\max}\left(\mathbf{A}_2\right)\|\mathbf{u}-\mathbf{v}\|_2\end{aligned}$$

where in the first inequality we use the B-Lipschitzness of σ. This shows that Assumption 4 holds with $G_1 = B\sigma_{\max}\left(\mathbf{A}_2\right)$. Lastly, for Assumption 5, we can compute that

$$\begin{aligned}\|\nabla_1 f(\mathbf{x},\mathbf{u}) - \nabla_1 f(\mathbf{x},\mathbf{v})\|_2 &= \left\|\mathbf{A}_1^\top\left(\sigma\left(\mathbf{A}_2\mathbf{u}\right) - \sigma\left(\mathbf{A}_2\mathbf{v}\right)\right)\right\|_2\\ &\le \sigma_{\max}\left(\mathbf{A}_1\right)\|\sigma\left(\mathbf{A}_2\mathbf{u}\right) - \sigma\left(\mathbf{A}_2\mathbf{v}\right)\|_2\\ &\le B\sigma_{\max}\left(\mathbf{A}_1\right)\sigma_{\max}\left(\mathbf{A}_2\right)\|\mathbf{u}-\mathbf{v}\|_2\end{aligned}$$

Therefore, Assumption 5 holds with $G_2 = B\sigma_{\max}\left(\mathbf{A}_1\right)\sigma_{\max}\left(\mathbf{A}_2\right)$.

E.2. Proof of Theorem 13

We want to invoke Theorem 7 to prove Theorem 13. Thus, it suffices to check the requirements in (5) for the coefficients in Lemma 12:

$$\begin{aligned}R_{\mathbf{x}} = R_{\mathbf{u}} = \infty;\ \mu = \sigma_{\min}\left(\mathbf{A}_1\right)^2;\ L_1 = \sigma_{\max}\left(\mathbf{A}_1\right)^2;\ L_2 = 1\\ G_1 = B\sigma_{\max}\left(\mathbf{A}_2\right);\ G_2 = B\sigma_{\max}\left(\mathbf{A}_1\right)\sigma_{\max}\left(\mathbf{A}_2\right).\end{aligned} \tag{78}$$

Since $\beta = \frac{4\sqrt{\kappa}-\sqrt{c}}{4\sqrt{\kappa}+7\sqrt{c}}$, we have

$$\frac{1-\beta}{1+\beta} = \frac{8\sqrt{c}}{6\sqrt{\kappa}+6\sqrt{c}} \ge \frac{c'}{\sqrt{\kappa}}$$

for some small enough constant c'. Treating $L_2 = 1$ as a constant, it suffices to guarantee that

$$G_1^4 \leq \frac{\tilde{C}_1 \mu^2}{\kappa^{\frac{3}{2}}}; \quad G_1^2 G_2^2 \leq \frac{\tilde{C}_2 \mu^3}{\kappa^{\frac{3}{2}}} \tag{79}$$

for some small enough constants $\tilde{C}_1$ and $\tilde{C}_2$. Plugging in the coefficients in (78)yields

$$B^4 \sigma_{\max}(\mathbf{A}_2)^4 \leq \frac{\tilde{C}_1 \sigma_{\min}(\mathbf{A}_1)^4}{\kappa^{\frac{3}{2}}}; \quad B^4 \sigma_{\max}(\mathbf{A}_2)^4 \sigma_{\max}(\mathbf{A}_1)^2 \leq \frac{\tilde{C}_2 \sigma_{\min}(\mathbf{A}_1)^6}{\kappa^{\frac{3}{2}}} \tag{80}$$

The second condition can also be written as

$$B^4 \sigma_{\max}(\mathbf{A}_2)^4 \leq \frac{\tilde{C}_2 \sigma_{\min}(\mathbf{A}_1)^4}{\kappa^{\frac{5}{2}}}$$

Therefore, we can guarantee the requirements in (5) as long as

$$\sigma_{\min}(\mathbf{A}_1) \geq \tilde{C} \sigma_{\max}(\mathbf{A}_2) B \kappa^{0.75}$$

for some large enough constant $\tilde{C}$. In this way, we can invoke Theorem 7 and notice that $f^\star = 0$ to get that

$$f(\mathbf{x}_k, \mathbf{u}_k) \leq 2\left(1 - \frac{c}{4\sqrt{\kappa}}\right)^k f(\mathbf{x}_0, \mathbf{u}_0)$$

Appendix F. Proofs for Section 4.2

F.1. Proof of Lemma 14

Orthogonal Transformation of the Parameters and Equivalence of Nesterov's Momentum. To obtain a parameter partition that achieves partial strong convexity, we cannot directly partition the parameters in the standard basis. Instead, we first apply an orthogonal transformation to all the parameters and then partition the transformed parameters. In particular, let $\mathbf{O} \in \mathbb{R}^{d \times d}$ be an orthogonal matrix, and for any objective $\hat{f} : \mathbb{R}^d \to \mathbb{R}$, we consider a new function $\tilde{f}(\mathbf{w}) = \hat{f}(\mathbf{O}\mathbf{w})$ for $\mathbf{x} \in \mathbb{R}^d$. Intuitively, $\tilde{f}$ is the equivalence of $\hat{f}$ on the orthogonally transformed parameter defined by $\mathbf{O}$. Ideally, using Nesterov's momentum to minimize $\tilde{f}$ executes

$$\mathbf{w}_{k+1} = \bar{\mathbf{w}}_k - \eta \nabla \tilde{f}(\bar{\mathbf{w}}_k); \quad \bar{\mathbf{w}}_{k+1} = \mathbf{w}_{k+1} + \beta(\mathbf{w}_{k+1} - \mathbf{w}_k)$$

Notice that, by the chain rule, $\nabla \tilde{f}(\bar{\mathbf{w}}_k) = \mathbf{O}^\top \nabla \hat{f}(\mathbf{O}\mathbf{w})$. If we multiply both sides of the two equations in the updates of Nesterov's momentum by $\mathbf{O}$, then we get

$$\mathbf{O}\mathbf{w}_{k+1} = \mathbf{O}\bar{\mathbf{w}}_k - \eta \nabla f(\mathbf{O}\bar{\mathbf{w}}_k); \quad \mathbf{O}\bar{\mathbf{w}}_{k+1} = \mathbf{O}\mathbf{w}_{k+1} + \beta(\mathbf{O}\mathbf{w}_{k+1} - \mathbf{O}\mathbf{w}_k)$$

which is precisely the update rule of Nesterov's momentum for minimizing $f(\mathbf{O}\mathbf{w})$. Therefore, we can conclude that orthogonal transformation preserves the property of the algorithm of interest.

Computation of the Coefficients. We let $\mathbf{w} = (\mathbb{V}(\mathbf{W}_1), \ldots, \mathbb{V}(\mathbf{W}_\Lambda)) \in \mathbb{R}^{\sum_{\ell=1}^{\Lambda} d_\ell d_{\ell-1}}$. Let $\boldsymbol{\theta}_0$ be the initialized parameter, and consider the SVD of $\mathbf{F}_{\Lambda-1}(\boldsymbol{\theta}_0)$ as $\mathbf{F}_{\Lambda-1}(\boldsymbol{\theta}_0) = \mathbf{U}\boldsymbol{\Sigma}_0\hat{\mathbf{V}}$ with $\mathbf{U} \in \mathbb{R}^{n \times n}$ and $\hat{\mathbf{V}} \in \mathbb{R}^{d_{\Lambda-1} \times d_{\Lambda-1}}$. Let $\hat{\mathbf{V}}_1 \in \mathbb{R}^{n \times d_{\Lambda-1}}$ be the top-n rows of $\hat{\mathbf{V}}$ and $\hat{\mathbf{V}}_2$ be the rest

rows. We define $\mathbf{V}_1 \in \mathbb{R}^{d_\Lambda n \times \sum_{\ell=1}^{\Lambda} d_\ell d_{\ell-1}}$, $\mathbf{V}_2 \in \mathbb{R}^{\left(\sum_{\ell=1}^{\Lambda} d_\ell d_{\ell-1} - d_\Lambda n\right) \times \sum_{\ell=1}^{\Lambda} d_\ell d_{\ell-1}}$ in the following sense ($\otimes$ denotes the Kronecker product):

$$\mathbf{V}_1 = \begin{bmatrix} \mathbf{0}_{d_\Lambda n \times \sum_{\ell=1}^{\Lambda-1} d_\ell d_{\ell-1}} & \mathbf{I}_{d_\Lambda} \otimes \hat{\mathbf{V}}_1 \end{bmatrix}$$
$$\mathbf{V}_2 = \begin{bmatrix} \mathbf{0}_{d_\Lambda (d_{\Lambda-1} - n) \times \sum_{\ell=1}^{\Lambda-1} d_\ell d_{\ell-1}} & \mathbf{I}_{d_\Lambda} \otimes \hat{\mathbf{V}}_2 \\ \mathbf{I}_{\sum_{\ell=1}^{\Lambda-1} d_\ell d_{\ell-1}} & \mathbf{0}_{\sum_{\ell=1}^{\Lambda-1} d_\ell d_{\ell-1} \times d_{\Lambda-1} d_\Lambda} \end{bmatrix}$$

Together, $\begin{bmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{bmatrix}$ is an orthogonal matrix. Under this orthogonal transformation, we partition the aggregation of all parameters $\mathbf{w}$ into $\mathbf{x} = \mathbf{V}_1 \mathbf{w}$ and $\mathbf{u} = \mathbf{V}_2 \mathbf{w}$. Moreover, we let $\hat{\mathbf{W}}_{\Lambda,1} = \hat{\mathbf{V}}_1 \mathbf{W}_\Lambda$ and $\hat{\mathbf{W}}_{\Lambda,2} = \hat{\mathbf{V}}_2 \mathbf{W}_\Lambda$, and observe that

$$\mathbf{x} = \mathbb{V}\left(\hat{\mathbf{W}}_{\Lambda,1}\right); \mathbf{u} = \left(\mathbb{V}\left(\hat{\mathbf{W}}_{\Lambda,2}\right), \mathbb{V}\left(\mathbf{W}_1\right), \ldots, \mathbb{V}\left(\mathbf{W}_{\Lambda-1}\right)\right)$$

Notice that $\mathbf{F}_\Lambda\left(\boldsymbol{\theta}\right)$ can be written as

$$\mathbf{F}_\Lambda\left(\boldsymbol{\theta}\right) = \mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}\right)\left(\hat{\mathbf{V}}_1^\top \hat{\mathbf{W}}_{\Lambda,1} + \hat{\mathbf{V}}_2^\top \hat{\mathbf{W}}_{\Lambda,2}\right)$$

Therefore, since $\mathbf{x} = \mathbb{V}\left(\hat{\mathbf{W}}_{\Lambda,1}\right)$, we have

$$\nabla_{11} f(\mathbf{x}, \mathbf{u}) = \mathbf{I}_{d_\Lambda} \otimes \hat{\mathbf{V}}_1 \mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}\right)^\top \mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}\right) \hat{\mathbf{V}}_1^\top$$

namely, $\nabla_{11} f\left(\mathbf{x}, \mathbf{u}\right)$ is a block-diagonal matrix. Therefore, its eigenvalues are given by

$$\lambda_{\max}\left(\nabla_{11} f(\mathbf{x}, \mathbf{u})\right) = \lambda_{\max}\left(\hat{\mathbf{V}}_1 \mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}\right)^\top \mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}\right) \hat{\mathbf{V}}_1^\top\right) = \sigma_1\left(\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}\right) \hat{\mathbf{V}}_1^\top\right)^2$$
$$\lambda_{\min}\left(\nabla_{11} f(\mathbf{x}, \mathbf{u})\right) = \lambda_{\min}\left(\hat{\mathbf{V}}_1 \mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}\right)^\top \mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}\right) \hat{\mathbf{V}}_1^\top\right) = \sigma_n\left(\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}\right) \hat{\mathbf{V}}_1^\top\right)^2$$

Based on our assumption, for all $\mathbf{x} \in \mathcal{B}_{R_\mathbf{x}}^{(1)}$ and $\mathbf{u} \in \mathcal{B}_{R_\mathbf{u}}^{(2)}$, we must have that

$$\|\mathbf{W}_\Lambda(\mathbf{x}) - \mathbf{W}_\Lambda(\mathbf{x}_0)\|_2 \leq \|\mathbf{x} - \mathbf{x}_0\|_2 \leq R_\mathbf{x};\ \|\mathbf{W}_\ell(\mathbf{u}) - \mathbf{W}_\ell(\mathbf{u}_0)\|_2 \leq \|\mathbf{u} - \mathbf{u}_0\|_2 \leq R_\mathbf{u}; \quad (81)$$

Moreover

$$\sum_{\ell=1}^{\Lambda-1} \|\mathbf{W}_\ell(\mathbf{u}) - \mathbf{W}_\ell(\mathbf{u}_0)\|_2 \leq \sqrt{\Lambda}\, \|\mathbf{u} - \mathbf{u}_0\|_2 \leq \sqrt{\Lambda} R_\mathbf{u} \quad (82)$$

Therefore, by (81), we have

$$\|\mathbf{W}_\Lambda(\mathbf{x})\|_2 \leq \|\mathbf{W}_\Lambda(\mathbf{x}_0)\|_2 + \|\mathbf{W}_\Lambda(\mathbf{x}) - \mathbf{W}_\Lambda(\mathbf{x}_0)\|_2 \leq \frac{\lambda_\Lambda}{2} + R_\mathbf{x} \leq \lambda_\Lambda$$
$$\|\mathbf{W}_\ell(\mathbf{u})\|_2 \leq \|\mathbf{W}_\ell(\mathbf{u}_0)\|_2 + \|\mathbf{W}_\ell(\mathbf{u}) - \mathbf{W}_\ell(\mathbf{u}_0)\|_2 \leq \frac{\lambda_\ell}{2} + R_\mathbf{u} \leq \lambda_\ell$$

by the initialization property. This shows that requiring $R_{\mathbf{u}} \leq \frac{1}{2}\min_{\ell\in[\Lambda-1]}\lambda_\ell$ and $R_{\mathbf{x}} \leq \frac{\lambda_\Lambda}{2}$ suffice for making the definition of λ_ℓ's valid. By Lemma 2.1 in (Nguyen, 2021), we have

$$\begin{aligned}\left\|\mathbf{F}_{\Lambda-1}(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}) - \mathbf{F}_{\Lambda-1}(\boldsymbol{\theta}(0))\right\|_F &\leq \|\mathbf{X}\|_F \lambda_{1\to\Lambda-1}\sum_{\ell=1}^{\Lambda-1}\lambda_\ell^{-1}\left\|\mathbf{W}_\ell(\mathbf{x},\mathbf{u}) - \mathbf{W}_\ell(\mathbf{x}_0,\mathbf{u}_0)\right\|_2 \\ &\leq \sqrt{\Lambda}\|\mathbf{X}\|_F \lambda_{1\to\Lambda-1} R_{\mathbf{u}}\left(\min_{\ell\in[\Lambda-1]}\lambda_\ell\right)^{-1} \\ &\leq \frac{\alpha_0}{4}\end{aligned}$$

where the second-to-last inequality follows from (82), and the last inequality follows from the upper bound on $R_{\mathbf{u}}$. Therefore, we have

$$\begin{aligned}\sigma_n\left(\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}\right)\hat{\mathbf{V}}_1^\top\right) &\leq \sigma_n\left(\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}(0)\right)\hat{\mathbf{V}}_1^\top\right) - \left\|\left(\mathbf{F}_{\Lambda-1}(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}) - \mathbf{F}_{\Lambda-1}(\boldsymbol{\theta}(0))\right)\hat{\mathbf{V}}_1^\top\right\|_2 \\ &\leq \sigma_n\left(\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}(0)\right)\right) - \left\|\mathbf{F}_{\Lambda-1}(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}) - \mathbf{F}_{\Lambda-1}(\boldsymbol{\theta}(0))\right\|_F \\ &\leq \alpha_0 - \frac{\alpha_0}{4} \\ &= \frac{3}{4}\alpha_0\end{aligned}$$

where the second inequality follows from the fact that $\left\|\hat{\mathbf{V}}_1\right\|_2 \leq 1$. This implies that for all $\mathbf{x} \in R_{\mathbf{x}}$ and $\mathbf{u} \in R_{\mathbf{u}}$, we have

$$\lambda_{\min}\left(\nabla_{11} f\left(\mathbf{x},\mathbf{u}\right)\right) \geq \left(\frac{3}{4}\alpha_0\right)^2 \geq \frac{\alpha_0^2}{2} =: \mu \tag{83}$$

This shows the partial strong convexity. To prove the partial smoothness, we have

$$\sigma_1\left(\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}\right)\hat{\mathbf{V}}_1^\top\right) \leq \left\|\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}\right)\right\|_2 \leq \|\mathbf{X}\|_F \lambda_{1\to\Lambda-1} \tag{84}$$

where the first inequality follows from the fact that $\left\|\hat{\mathbf{V}}_1\right\|_2 \leq 1$. Therefore,

$$\lambda_{\max}\left(\nabla_{11} f\left(\mathbf{x},\mathbf{u}\right)\right) \leq \|\mathbf{X}\|_F^2 \lambda_{1\to\Lambda-1}^2 =: L_1$$

Now, we proceed to compute $G_1$by bounding $\|h(\mathbf{x},\mathbf{u}) - h(\mathbf{x},\mathbf{v})\|_2$. We notice that

$$\begin{aligned}h(\mathbf{x},\mathbf{u}) - h(\mathbf{x},\mathbf{v}) &= \mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}\right)\left(\hat{\mathbf{V}}_1^\top\hat{\mathbf{W}}_{\Lambda,1}\left(\mathbf{x}\right) + \hat{\mathbf{V}}_2^\top\hat{\mathbf{W}}_{\Lambda,2}\left(\mathbf{u}\right)\right) \\ &\quad - \mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{v}}\right)\left(\hat{\mathbf{V}}_1^\top\hat{\mathbf{W}}_{\Lambda,1}\left(\mathbf{x}\right) + \hat{\mathbf{V}}_2^\top\hat{\mathbf{W}}_{\Lambda,2}\left(\mathbf{v}\right)\right) \\ &= \left(\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}\right) - \mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{v}}\right)\right)\left(\hat{\mathbf{V}}_1^\top\hat{\mathbf{W}}_{\Lambda,1}\left(\mathbf{x}\right) + \hat{\mathbf{V}}_2^\top\hat{\mathbf{W}}_{\Lambda,2}\left(\mathbf{u}\right)\right) \\ &\quad + \mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{v}}\right)\hat{\mathbf{V}}_2^\top\left(\hat{\mathbf{W}}_{\Lambda,2}(\mathbf{u}) - \hat{\mathbf{W}}_{\Lambda,2}(\mathbf{v})\right)\end{aligned}$$

Now, for the first term, we have

$$\begin{aligned}\left\|\left(\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}\right)-\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{v}}\right)\right)\hat{\mathbf{V}}_1^\top\hat{\mathbf{W}}_{\Lambda,1}\right\|_F &\leq \lambda_\Lambda\left\|\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}\right)-\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{v}}\right)\right\|_F \\ &\leq \|\mathbf{X}\|_F\,\lambda_{1\to\Lambda-1}\sum_{\ell=1}^{\Lambda}\lambda_\ell^{-1}\left\|\mathbf{W}_\ell(\mathbf{u})-\mathbf{W}_\ell(\mathbf{v})\right\|_2 \\ &\leq \sqrt{\Lambda}\,\|\mathbf{X}\|_F\,\lambda_{1\to\Lambda}\left(\min_{\ell\in[\Lambda-1]}\lambda_\ell\right)^{-1}\|\mathbf{u}-\mathbf{v}\|_2\end{aligned}$$

For the second term, we have

$$\begin{aligned}&\left\|\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{v}}\right)\hat{\mathbf{V}}_2^\top\left(\hat{\mathbf{W}}_{\Lambda,2}(\mathbf{u})-\hat{\mathbf{W}}_{\Lambda,2}(\mathbf{v})\right)\right\|_F \\ &\quad\leq\left\|\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{v}}\right)\hat{\mathbf{V}}_2^\top\right\|_2\left\|\hat{\mathbf{W}}_{\Lambda,2}(\mathbf{u})-\hat{\mathbf{W}}_{\Lambda,2}(\mathbf{v})\right\| \\ &\quad\leq\left\|\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{v}}\right)-\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}(0)\right)\right\|_F\|\mathbf{u}-\mathbf{v}\|_2 \\ &\quad\leq\sqrt{\Lambda}\,\|\mathbf{X}\|_F\,\lambda_{1\to\Lambda-1}R_{\mathbf{u}}\left(\min_{\ell\in[\Lambda-1]}\lambda_\ell\right)^{-1}\|\mathbf{u}-\mathbf{v}\|_2\end{aligned}$$

where the second inequality follows from

$$\begin{aligned}\left\|\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{v}}\right)\hat{\mathbf{V}}_2^\top\right\|_2 &\leq \left\|\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}(0)\right)\hat{\mathbf{V}}_2^\top\right\|_2+\left\|\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{v}}\right)-\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}(0)\right)\right\|_2 \\ &= \left\|\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{v}}\right)-\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}(0)\right)\right\|_2\end{aligned}$$

by noticing that $\left\|\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}(0)\right)\hat{\mathbf{V}}_2^\top\right\|_2=0$ by the definition of $\hat{\mathbf{V}}_2$. Combining the two, we have

$$\|h(\mathbf{x},\mathbf{u})-h(\mathbf{x},\mathbf{v})\|_2\leq\left(\lambda_\Lambda+R_{\mathbf{u}}\right)\sqrt{\Lambda}\,\|\mathbf{X}\|_F\,\lambda_{1\to\Lambda-1}\left(\min_{\ell\in[\Lambda-1]}\lambda_\ell\right)^{-1}\|\mathbf{u}-\mathbf{v}\|_2$$

This implies that

$$G_1=\left(\lambda_\Lambda+R_{\mathbf{u}}\right)\sqrt{\Lambda}\,\|\mathbf{X}\|_F\,\lambda_{1\to\Lambda-1}\left(\min_{\ell\in[\Lambda-1]}\lambda_\ell\right)^{-1}$$

Next, we proceed to compute G_1 by bounding $\|\nabla_1 f\left(\mathbf{x},\mathbf{u}\right)-\nabla_1 f\left(\mathbf{x},\mathbf{v}\right)\|_2$. Computing the gradient, we have

$$\nabla_{\hat{\mathbf{W}}_{\Lambda,1}}\mathcal{L}\left(\boldsymbol{\theta}\right)=\hat{\mathbf{V}}_1\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}\right)^\top\left(\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}\right)\left(\hat{\mathbf{V}}_1^\top\hat{\mathbf{W}}_{\Lambda,1}+\hat{\mathbf{V}}_2^\top\hat{\mathbf{W}}_{\Lambda,2}\right)-\mathbf{Y}\right)$$

Therefore

$$\begin{aligned}&\nabla_1 f(\mathbf{x},\mathbf{u})-\nabla_1 f\left(\mathbf{x},\mathbf{v}\right) \\ &\quad=\underbrace{\hat{\mathbf{V}}_1\left(\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}\right)^\top\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}\right)-\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{v}}\right)^\top\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{v}}\right)\right)\mathbf{W}_\Lambda(\mathbf{x},\mathbf{u})}_{\boldsymbol{\delta}_1} \\ &\quad-\hat{\mathbf{V}}_1\left(\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}\right)-\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{v}}\right)\right)\mathbf{Y} \\ &\quad+\underbrace{\hat{\mathbf{V}}_1^\top\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{v}}\right)^\top\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{v}}\right)\hat{\mathbf{V}}_2\left(\hat{\mathbf{W}}_{\Lambda,2}(\mathbf{u})-\hat{\mathbf{W}}_{\Lambda,2}(\mathbf{u})\right)}_{\boldsymbol{\delta}_2}\end{aligned}$$

We bound the magnitude of $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ separately. For $\boldsymbol{\delta}_1$, we have

$$\begin{aligned}
\|\boldsymbol{\delta}_1\|_F &= \left\|\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}\right)^\top \mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}\right) - \mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{v}}\right)^\top \mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{v}}\right)\right\|_F \\
&\leq \left(\|\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}\right)\|_F + \|\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{v}}\right)\|_F\right) \|\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}\right) - \mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{v}}\right)\|_F \|\mathbf{W}_\Lambda(\mathbf{x},\mathbf{u})\|_2 \\
&\leq 2\|\mathbf{X}\|_F \lambda_{1\to\Lambda-1} \cdot \sqrt{\Lambda}\|\mathbf{X}\|_F \lambda_{1\to\Lambda-1}\left(\min_{\ell\in[\Lambda-1]}\lambda_\ell\right)^{-1}\|\mathbf{u}-\mathbf{v}\|_2 \cdot \lambda_\Lambda \\
&= 2\sqrt{\Lambda}\|\mathbf{X}\|_F^2 \lambda_\Lambda \lambda_{1\to\Lambda-1}^2\left(\min_{\ell\in[\Lambda-1]}\lambda_\ell\right)^{-1}\|\mathbf{u}-\mathbf{v}\|_2
\end{aligned}$$

For the second term, we have

$$\begin{aligned}
\left\|\hat{\mathbf{V}}_1\left(\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}\right) - \mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{v}}\right)\right)\mathbf{Y}\right\| &\leq \|\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{u}}\right) - \mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{v}}\right)\|_F \|\mathbf{Y}\|_F \\
&\leq \sqrt{\Lambda}\|\mathbf{X}\|_F \|\mathbf{Y}\|_F \lambda_{1\to\Lambda-1}\left(\min_{\ell\in[\Lambda-1]}\lambda_\ell\right)^{-1}\|\mathbf{u}-\mathbf{v}\|_2
\end{aligned}$$

Lastly, for $\boldsymbol{\delta}_2$, we have

$$\begin{aligned}
\|\boldsymbol{\delta}_2\|_F &\leq \|\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{v}}\right)\|_2 \|\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}_{\mathbf{x},\mathbf{v}}\right) - \mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}(0)\right)\|_2 \left\|\hat{\mathbf{W}}_{\Lambda,2}(\mathbf{u}) - \hat{\mathbf{W}}_{\Lambda,2}(\mathbf{u})\right\|_F \\
&\leq \sqrt{\Lambda}\|\mathbf{X}\|_F^2 \lambda_{1\to\Lambda-1}^2\left(\min_{\ell\in[\Lambda-1]}\lambda_\ell\right)^{-1} R_{\mathbf{u}}\|\mathbf{u}-\mathbf{v}\|_2
\end{aligned}$$

Putting things together gives

$$G_2 = \left(\left(2\lambda_\Lambda + R_{\mathbf{u}}\right)\|\mathbf{X}\|_F \lambda_{1\to\Lambda-1} + \|\mathbf{Y}\|_F\right)\sqrt{\Lambda}\|\mathbf{X}\|_F \lambda_{1\to\Lambda-1}\left(\min_{\ell\in[\Lambda-1]}\lambda_\ell\right)^{-1}$$

Now that we have shown that Assumption 1,2,4,5 holds, we proceed to prove Assumption 3 and Assumption 6. Simple decomposition gives

$$\mathbf{F}_\Lambda\left(\boldsymbol{\theta}\right) = \mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}\right)\left(\hat{\mathbf{V}}_1^\top \hat{\mathbf{W}}_{\Lambda,1} + \hat{\mathbf{V}}_2^\top \hat{\mathbf{W}}_{\Lambda,2}\right)$$

Therefore, to set $\mathbf{F}_\Lambda\left(\boldsymbol{\theta}\right) = \mathbf{Y}$, we can simply let $\hat{\mathbf{W}}_{\Lambda,1}$ to be

$$\hat{\mathbf{W}}_{\Lambda,1} = \left(\hat{\mathbf{V}}_1\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}\right)^\top \mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}\right)\hat{\mathbf{V}}_1^\top\right)^{-1}\hat{\mathbf{V}}_1\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}\right)^\top\left(\mathbf{Y} - \hat{\mathbf{V}}_2^\top\hat{\mathbf{W}}_{\Lambda,2}\right)$$

since $\hat{\mathbf{V}}_1\mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}\right)^\top \mathbf{F}_{\Lambda-1}\left(\boldsymbol{\theta}\right)\hat{\mathbf{V}}_1^\top \in \mathbb{R}^{n\times n}$ has full rank. This shows that $\min_{\mathbf{x}} f\left(\mathbf{x},\mathbf{u}\right) = 0$ for any $\mathbf{u}$. Since $f\left(\mathbf{x},\mathbf{u}\right) \geq 0$ by the property of the MSE, we can conclude that Assumption 6 holds. Moreover, Assumption 3 also holds with $L_2 = 1$ since MSE is by itself 1-smooth.

F.2. Proof of Theorem 15

We want to invoke Theorem 7 to prove Theorem 15. Thus it suffices to check the requirements in (5) (which we restate below):

$$G_1^4 \leq \frac{C_1\mu^2}{L_2(L_2+1)^2}\left(\frac{1-\beta}{1+\beta}\right)^3; \quad G_1^2G_2^2 \leq \frac{C_2\mu^3}{L_2(L_2+1)\sqrt{\kappa}}\left(\frac{1-\beta}{1+\beta}\right)^2;$$
$$R_{\mathbf{x}} \geq \frac{36}{c}\sqrt{\kappa}\left(\frac{\eta(L_2+1)}{1-\beta}\right)^{\frac{1}{2}}(f(\mathbf{x}_0,\mathbf{u}_0)-f^\star)^{\frac{1}{2}};$$
$$R_{\mathbf{u}} \geq \frac{36}{c}\sqrt{\kappa}\left(\frac{\eta G_1^2L_2(L_2+1)(1+\beta)^3}{\mu\beta(1-\beta)^3}\right)^{\frac{1}{2}}(f(\mathbf{x}_0,\mathbf{u}_0)-f^\star)^{\frac{1}{2}},$$

for the coefficients in Lemma 14 (which we restate below as well):

$$R_{\mathbf{x}} = \frac{\lambda_\Lambda}{2}; \quad R_{\mathbf{u}} = \frac{1}{2}\left(\min_{\ell\in[\Lambda-1]}\lambda_\ell\right)\min\left\{1, \frac{\alpha_0}{2\sqrt{\Lambda}\left\|\mathbf{X}\right\|_F\lambda_{1\to\Lambda-1}}\right\}^2; \quad \mu = \frac{\alpha_0^2}{2}$$
$$L_1 = \left\|\mathbf{X}\right\|_F^2\lambda_{1\to\Lambda-1}^2; \quad L_2 = 1; \quad G_1 = \left(\lambda_\Lambda + R_{\mathbf{u}}\right)\sqrt{\Lambda}\left\|\mathbf{X}\right\|_F\lambda_{1\to\Lambda-1}\left(\min_{\ell\in[\Lambda-1]}\lambda_\ell\right)^{-1}$$
$$G_2 = \left(\left(2\lambda_\Lambda + R_{\mathbf{u}}\right)\left\|\mathbf{X}\right\|_F\lambda_{1\to\Lambda-1} + \left\|\mathbf{Y}\right\|_F\right)\sqrt{\Lambda}\left\|\mathbf{X}\right\|_F\lambda_{1\to\Lambda-1}\left(\min_{\ell\in[\Lambda-1]}\lambda_\ell\right)^{-1}$$

Recall that $1-\beta = O\left(1/\sqrt{\kappa}\right) = O\left(\sqrt{\mu}/\sqrt{L_1}\right)$ and $1+\beta = O(1)$. Since $L_2 = 1$, we treat it as a constant. Moreover, we also treat Λ as a constant. We have shown in Lemma 14 that $f^\star = 0$. Thus $f(\mathbf{x}_0,\mathbf{u}_0) - f^\star = \mathcal{L}(\boldsymbol{\theta}(0))$. With $\eta = \frac{c}{L_1}$, the requirement in (5) can be simplified to

$$\underbrace{G_1^4 \leq \frac{\hat{C}_1\mu^{\frac{7}{2}}}{L_1^{\frac{3}{2}}}}_{\mathcal{R}_1}; \quad \underbrace{G_1^2G_2^2 \leq \frac{\hat{C}_2\mu^{\frac{9}{2}}}{L_1^{\frac{3}{2}}}}_{\mathcal{R}_2}; \quad \underbrace{R_{\mathbf{x}} \geq \frac{\hat{C}_3L_1^{\frac{1}{4}}}{\mu^{\frac{3}{4}}}\mathcal{L}(\boldsymbol{\theta}(0))^{\frac{1}{2}}}_{\mathcal{R}_3}; \quad \underbrace{R_{\mathbf{u}} \geq \frac{\hat{C}_4G_1L_1^{\frac{3}{4}}}{\mu^{\frac{7}{4}}}\mathcal{L}(\boldsymbol{\theta}(0))^{\frac{1}{2}}}_{\mathcal{R}_4} \tag{85}$$

In the following parts of the proof, we will use $\gtrsim$ and $\lesssim$ to denote the inequality hiding constants. We will analyze each requirement separately.

Calculation for $\mathcal{R}_1$. Notice that

$$G_1^4 \lesssim \left(\lambda_\Lambda^4 + R_{\mathbf{u}}^4\right)\left\|\mathbf{X}\right\|_F^4\lambda_{1\to\Lambda-1}^4\left(\min_{\ell\in[\Lambda-1]}\lambda_\ell\right)^{-4} = \left(\lambda_\Lambda^4 + R_{\mathbf{u}}^4\right)\left(\min_{\ell\in[\Lambda-1]}\lambda_\ell\right)^{-4}L_1^2$$

It suffices to show that

$$\max\left\{\lambda_\Lambda^4, R_{\mathbf{u}}^4\right\} \lesssim \min_{\ell\in[\Lambda-1]}\lambda_\ell^4\cdot\frac{\mu^{\frac{7}{2}}}{L_1^{\frac{7}{2}}}$$

Notice that, by definition,

$$R_{\mathbf{u}} \leq \frac{1}{2}\min_{\ell\in[\Lambda-1]}\lambda_\ell\cdot\frac{\mu}{L_1} \Rightarrow R_{\mathbf{u}}^4 \leq \left(\frac{1}{2}\right)^4\min_{\ell\in[\Lambda-1]}\lambda_\ell^4\cdot\frac{\mu^4}{L_1^4} \leq \left(\frac{1}{2}\right)^4\min_{\ell\in[\Lambda-1]}\lambda_\ell^4\cdot\frac{\mu^{\frac{7}{2}}}{L_1^{\frac{7}{2}}}$$

where the last inequality follows from $\mu \leq L_1$. Therefore the condition on $R_{\mathbf{u}}$ is satisfied automatically. It suffice to consider the condition on λ_Λ^4, which boils down to

$$\lambda_\Lambda^4 \lesssim \min_{\ell \in [\Lambda-1]} \lambda_\ell^4 \cdot \frac{\alpha^7}{\|\mathbf{X}\|_F^7 \lambda_{1\to\Lambda-1}^7}$$

Rearranging gives

$$\alpha^7 \gtrsim \frac{\|\mathbf{X}\|_F^7 \lambda_{1\to\Lambda}^7}{\lambda_\Lambda^3 \min_{\ell\in[\Lambda-1]} \lambda_\ell^4} \tag{86}$$

Calculation for $\mathcal{R}_2$. With the condition that $\mathcal{R}_1$ is satisfied, it suffice to show that

$$G_2^4 \lesssim \frac{\mu^{\frac{11}{2}}}{L_1^{\frac{3}{2}}}$$

For G_2, we have

$$G_2^4 \lesssim \max\{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3\}$$

with

$$\mathcal{T}_1 = \|\mathbf{X}\|_F^8 \lambda_\Lambda^4 \lambda_{1\to\Lambda-1}^8 \left(\min_{\ell\in[\Lambda-1]} \lambda_\ell\right)^{-4}$$
$$\mathcal{T}_2 = R_{\mathbf{u}}^4 \|\mathbf{X}\|_F^8 \lambda_{1\to\Lambda-1}^8 \left(\min_{\ell\in[\Lambda-1]} \lambda_\ell\right)^{-4}$$
$$\mathcal{T}_3 = \|\mathbf{Y}\|_F^4 \|\mathbf{X}\|_F^4 \lambda_{1\to\Lambda-1}^4 \left(\min_{\ell\in[\Lambda-1]} \lambda_\ell\right)^{-4}$$

Notice that since

$$R_{\mathbf{u}}^4 \leq \left(\frac{C^\perp}{2}\right)^4 \min_{\ell\in[\Lambda-1]} \lambda_\ell^4 \cdot \frac{\mu^4}{L_1^4}$$

we must have

$$\mathcal{T}_2 \leq \left(\frac{C^\perp}{2}\right)^4 \min_{\ell\in[\Lambda-1]} \lambda_\ell^4 \cdot \frac{\mu^4}{L_1^4} \cdot L_1^4 \left(\min_{\ell\in[\Lambda-1]} \lambda_\ell\right)^{-4} \leq \left(\frac{C^\perp}{2}\right)^4 \mu^4 \lesssim \frac{\mu^{\frac{11}{2}}}{L_1^{\frac{3}{2}}}$$

since $\mu \leq L_1$. Thus, we only need to consider $\mathcal{T}_1$ and $\mathcal{T}_3$. Combining the two conditions, $\mathcal{R}_2$ boils down to

$$\alpha_0^{11} \gtrsim \frac{\|\mathbf{X}\|_F^7 \lambda_{1\to\Lambda-1}^7}{\min_{\ell\in[\Lambda-1]} \lambda_\ell^4} \left(\|\mathbf{X}\|_F^4 \lambda_{1\to\Lambda}^4 + \|\mathbf{Y}\|_F^4\right) \tag{87}$$

Calculation for $\mathcal{R}_3$. We first notice that since $\mu \leq L_1$, $\mathcal{R}_3$ can be restricted to

$$R_{\mathbf{x}} \geq \frac{\hat{C}_3 L_1^{\frac{1}{2}}}{\mu} \mathcal{L}(\boldsymbol{\theta}(0))^{\frac{1}{2}}$$

Plugging in $R_{\mathbf{x}} = \frac{\lambda_\Lambda}{2}$ and μ, L_1 gives

$$\lambda_\Lambda \gtrsim \frac{\|\mathbf{X}\|_F \lambda_{1\to\Lambda-1}}{\alpha_0^2} \mathcal{L}(\boldsymbol{\theta}(0))^{\frac{1}{2}}$$

Rearranging the terms gives

$$\alpha_0^2 \gtrsim \frac{\|\mathbf{X}\|_F \lambda_{1\to L}}{\lambda_\Lambda^2} \mathcal{L}(\boldsymbol{\theta}(0))^{\frac{1}{2}} \tag{88}$$

Calculation for $\mathcal{R}_4$ Notice that $\alpha_0 = \sqrt{2\mu} \leq \sqrt{2L_1} \leq 2\sqrt{\Lambda}\,\|\mathbf{X}\|_F \lambda_{1\to\Lambda-1}$. Therefore

$$R_{\mathbf{u}} \lesssim \min_{\ell\in[\Lambda-1]} \lambda_\ell \frac{\alpha_0^2}{\|\mathbf{X}\|_F^2 \lambda_{1\to\Lambda-1}^2}$$

To satisfy $\mathcal{R}_4$, we need

$$R_{\mathbf{u}} \gtrsim R_{\mathbf{u}} \|\mathbf{X}\|_F \lambda_{1\to\Lambda-1} \mathcal{L}\left(\boldsymbol{\theta}(0)\right)^{\frac{1}{2}} \frac{L_1^{\frac{3}{4}}}{\mu^{\frac{7}{4}}} \left(\min_{\ell\in[\Lambda-1]} \lambda_\ell\right)^{-1}$$

$$R_{\mathbf{u}} \gtrsim \|\mathbf{X}\|_F \lambda_{1\to\Lambda} \mathcal{L}\left(\boldsymbol{\theta}(0)\right)^{\frac{1}{2}} \frac{L_1^{\frac{3}{4}}}{\mu^{\frac{7}{4}}} \left(\min_{\ell\in[\Lambda-1]} \lambda_\ell\right)^{-1}$$

To analyze the first, we simply remove $R_{\mathbf{u}}$ from both sides to get that

$$\alpha_0^7 \gtrsim \frac{\|\mathbf{X}\|_F^5 \lambda_{1\to\Lambda-1}^5 \mathcal{L}\left(\boldsymbol{\theta}(0)\right)^2}{\min_{\ell\in[\Lambda-1]} \lambda_\ell^4} \tag{89}$$

For the second, we plug in the upper bound on $R_{\mathbf{u}}$ to have

$$\alpha_0^{11} \gtrsim \frac{\|\mathbf{X}\|_F^9 \lambda_{1\to\Lambda}^9}{\lambda_\Lambda^7 \min_{\ell\in[\Lambda-1]} \lambda_\ell^4} \mathcal{L}\left(\boldsymbol{\theta}(0)\right) \tag{90}$$

Initialization Scheme. Recall our initialization scheme

$$d_\ell = \Theta\left(m\right) \quad \forall \ell \in [\Lambda-1]; \quad d_{L-1} = \Theta\left(n^{4.5} \max n, d_0^2\right)$$
$$[\mathbf{W}_\ell(0)]_{ij} \sim \mathcal{N}\left(0, d_{\ell-1}^{-1}\right) \quad \forall \ell \in [\Lambda-1]; \quad [\mathbf{W}_\Lambda(0)]_{ij} \sim \mathcal{N}\left(0, d_{\Lambda-1}^{-\frac{3}{2}}\right)$$

We will show that this initialization scheme satisfies (86)-(90), which we restate below

$$\alpha^7 \gtrsim \frac{\|\mathbf{X}\|_F^7 \lambda_{1\to\Lambda}^7}{\lambda_\Lambda^3 \min_{\ell\in[\Lambda-1]} \lambda_\ell^4}$$

$$\alpha_0^{11} \gtrsim \frac{\|\mathbf{X}\|_F^7 \lambda_{1\to\Lambda-1}^7}{\min_{\ell\in[\Lambda-1]} \lambda_\ell^4} \left(\|\mathbf{X}\|_F^4 \lambda_{1\to\Lambda}^4 + \|\mathbf{Y}\|_F^4\right)$$

$$\alpha_0^2 \gtrsim \frac{\|\mathbf{X}\|_F \lambda_{1\to\Lambda}}{\lambda_\Lambda^2} \mathcal{L}(\boldsymbol{\theta}(0))^{\frac{1}{2}}$$

$$\alpha_0^7 \gtrsim \frac{\|\mathbf{X}\|_F^5 \lambda_{1\to\Lambda-1}^5 \mathcal{L}\left(\boldsymbol{\theta}(0)\right)^2}{\min_{\ell\in[\Lambda-1]} \lambda_\ell^4}$$

$$\alpha_0^{11} \gtrsim \frac{\|\mathbf{X}\|_F^9 \lambda_{1\to\Lambda}^9}{\lambda_\Lambda^7 \min_{\ell\in[\Lambda-1]} \lambda_\ell^4} \mathcal{L}\left(\boldsymbol{\theta}(0)\right)$$

To start, we first compute λ_ℓ's. Recall that we required initializing $\|\mathbf{W}_\ell(0)\|_2 = \frac{\lambda_\ell}{2}$. This implies that $\lambda_\ell \leq 2\|\mathbf{W}_\ell(0)\|_2$ for all $\ell \in [\Lambda]$. By Theorem 4.4.5 in (Vershynin, 2018), we have

$$\|\mathbf{W}_\ell(0)\|_2 = \begin{cases} O\left(1 + \frac{\sqrt{m}}{\sqrt{d_0}}\right) & \text{if } \ell = 1 \\ O\left(1\right) & \text{if } \ell = 2, \dots \Lambda - 2 \\ O\left(1 + \frac{\sqrt{d_{\Lambda-1}}}{\sqrt{m}}\right) & \text{if } \ell = \Lambda - 1 \\ O\left(d_{\Lambda-1}^{-\frac{1}{4}} + \frac{\sqrt{d_\Lambda}}{d_{\Lambda-1}^{3/4}}\right) & \text{if } \ell = \Lambda \end{cases}$$

Since λ_ℓ satisfies the same scaling, plugging in the width, and notice that $d_{\Lambda-1} \geq m \geq \max\{d_\Lambda, d_0\}$ gives

$$\lambda_\ell = \begin{cases} O\left(\frac{\sqrt{m}}{\sqrt{d_0}}\right) & \text{if } \ell = 1 \\ O\left(1\right) & \text{if } \ell = 2, \dots, \Lambda - 2 \\ O\left(\frac{n^{9/4}}{\sqrt{m}} \max\{\sqrt{n}, d_0\}\right) & \text{if } \ell = \Lambda - 1 \\ O\left(\frac{1}{n^{9/8}\max\left\{n^{1/4}, \sqrt{d_0}\right\}}\right) & \text{if } \ell = \Lambda \end{cases}$$

Therefore

$$\min_{\ell \in [\Lambda-1]} \lambda_\ell = O\left(1\right); \quad \lambda_{1\to\Lambda-1} = O\left(\frac{n^{9/4}}{\sqrt{d_0}} \max\{\sqrt{n}, d_0\}\right)$$

Moreover, by Assumption 3.1 in (Nguyen, 2021), we have $\|\mathbf{X}\|_F = O\left(\sqrt{nd_0}\right)$ and $\|\mathbf{Y}\|_F = O\left(\sqrt{n}\right)$. By Lemma 3.3 in (Nguyen, 2021), we have that $\alpha_0 = \Omega\left(d_{\Lambda-1}^{\frac{1}{2}}\right) = \Omega\left(n^{9/4}\max\{\sqrt{n}, d_0\}\right)$. Lastly, by Lemma C.1 in (Nguyen and Mondelli, 2020) and (Nguyen, 2021), we have $\mathcal{L}(\boldsymbol{\theta}(0))^{\frac{1}{2}} = O\left(\sqrt{nd_0}\right)$. With these preparations, let's check each requirement. For (86), we have

$$\alpha_0^7 = \Omega\left(\max\left\{n^{77/4}, n^{63/4}d_0^7\right\}\right); \quad \frac{\|\mathbf{X}\|_F^7 \lambda_{1\to\Lambda}^7}{\lambda_\Lambda^3 \min_{\ell\in[\Lambda-1]}\lambda_\ell^4} = O\left(\max\left\{n^{69/4}, n^{59/4}d_0^5\right\}\right)$$

Therefore, we have that (86) is satisfied. For (87), we have

$$\alpha_0^{11} = \Omega\left(\max\left\{n^{121/4}, n^{99/4}d_0^9\right\}\right)$$
$$\frac{\|\mathbf{X}\|_F^7 \lambda_{1\to\Lambda-1}^7}{\min_{\ell\in[\Lambda-1]}\lambda_\ell^4}\left(\|\mathbf{X}\|_F^4 \lambda_{1\to\Lambda}^4 + \|\mathbf{Y}\|_F^4\right) = O\left(\max\left\{n^{165/8}, n^{143/8}d_0^{11/2}\right\}\right)$$

Therefore, we have that (87) is satisfied. For (88), we have

$$\alpha_0^2 = \Omega\left(\max\left\{n^{\frac{11}{2}}, n^{\frac{9}{2}}d_0^2\right\}\right); \quad \frac{\|\mathbf{X}\|_F \lambda_{1\to\Lambda}}{\lambda_\Lambda^2}\mathcal{L}(\boldsymbol{\theta}(0))^{\frac{1}{2}} = O\left(\max\left\{n^{\frac{33}{8}}, n^{\frac{31}{8}}d_0\right\}\right)$$

Therefore, we have that (88) is satisfied. For (89), we have

$$\alpha_0^7 = \Omega\left(\max\left\{n^{77/4}, n^{63/4}d_0^7\right\}\right); \quad \frac{\|\mathbf{X}\|_F^5 \lambda_{1\to\Lambda-1}^5 \mathcal{L}\left(\boldsymbol{\theta}(0)\right)^2}{\min_{\ell\in[\Lambda-1]}\lambda_\ell^4} = O\left(\max\left\{n^{73/4}d_0^2, n^{63/4}d_0^7\right\}\right)$$

Notice that $n^{73/4}d_0^2 \geq n^{63/4}d_0^7$ only when $d_0 \leq \sqrt{n}$. In this case, we must have that $n^{\frac{77}{4}} \geq n^{73/4}d_0^2$. Therefore, we have that (89) is satisfied. For (90), we have

$$\alpha_0^{11} = \Omega\left(\max\left\{n^{121/4}, n^{99/4}d_0^9\right\}\right); \quad \frac{\|\mathbf{X}\|_F^9 \lambda_{1\to\Lambda}^9}{\lambda_\Lambda^7 \min_{\ell\in[\Lambda-1]}\lambda_\ell^4}\mathcal{L}\left(\boldsymbol{\theta}(0)\right) = O\left(\max\left\{n^{51/4}d_0, n^{10}d_0^{13/2}\right\}\right)$$

Similarly, when $n^{51/4}d_0 \geq n^{10}d_0^{13/2}$, we must have $d_0 \leq \sqrt{n}$. This implies that $n^{\frac{121}{4}} \geq n^{51/4}d_0$. Therefore, we have that (90) is also satisfied. Now, all requirements in Theorem 7 can be satisfied by the initialization scheme with our over-parameterization. Thus, we can invoke Theorem 7 to get that

$$f(\mathbf{x}_k, \mathbf{u}_k) - f^\star \leq 2\left(1 - \frac{c}{4\sqrt{\kappa}}\right)\left(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star\right)$$

Noting that $f^\star = 0$ and $f(\mathbf{x}_k, \mathbf{u}_k) = \mathcal{L}(\boldsymbol{\theta}(k))$, we have

$$\mathcal{L}(\boldsymbol{\theta}(k)) \leq 2\left(1 - \frac{c}{4\sqrt{\kappa}}\right)\mathcal{L}(\boldsymbol{\theta}(0))$$

which completes the proof.

Appendix G. Auxiliary Lemma

Lemma 18 *Suppose that Assumption 2 holds. Then for all* $\mathbf{x} \in \mathbb{R}^{d_1}$ *and* $\mathbf{u} \in \mathcal{B}_{R_\mathbf{u}}^{(2)}$ *we have*

$$\|\nabla_1 f(\mathbf{x}, \mathbf{u})\|_2^2 \leq 2L_1\left(f(\mathbf{x}, \mathbf{u}) - f^\star\right)$$

Proof Assumption 2 implies that, for all $\mathbf{x} \in \mathbb{R}^{d_1}$ and $\mathbf{u} \in \mathcal{B}_{R_\mathbf{u}}^{(2)}$

$$\begin{aligned} f^\star \leq f\left(\mathbf{x} - \frac{1}{L_1}\nabla_1 f(\mathbf{x}, \mathbf{u}), \mathbf{u}\right) &\leq f(\mathbf{x}, \mathbf{u}) - \frac{1}{L_1}\|\nabla_1 f(\mathbf{x}, \mathbf{u})\|_2^2 + \frac{1}{2L_1}\|\nabla_1 f(\mathbf{x}, \mathbf{u})\|_2^2 \\ &= f(\mathbf{x}, \mathbf{u}) - \frac{1}{2L_1}\|\nabla_1 f(\mathbf{x}, \mathbf{u})\|_2^2 \end{aligned}$$

which implies that, for all $\mathbf{x} \in \mathbb{R}^{d_1}$ and $\mathbf{u} \in \mathcal{B}_{R_\mathbf{u}}^{(2)}$

$$\|\nabla_1 f(\mathbf{x}, \mathbf{u})\|_2^2 \leq 2L_1\left(f(\mathbf{x}, \mathbf{u}) - f^\star\right)$$

■

Lemma 19 *Suppose that Assumption 3 holds. Then for all* $\mathbf{s} \in \mathbb{R}^{\hat{d}}$ *we have*

$$\|\nabla g(\mathbf{s})\|_2^2 \leq 2L_2\left(g(\mathbf{s}) - f^\star\right)$$

Proof By Assumption 3, for all $\mathbf{s} \in \mathbb{R}^{\hat{d}}$ we have

$$f^\star = g^\star \leq g\left(\mathbf{s} - \frac{1}{L_2}\nabla g(\mathbf{s})\right) = g(\mathbf{s}) - \frac{1}{L_2}\|\nabla g(\mathbf{s})\|_2^2 + \frac{1}{2L_2}\|\nabla g(\mathbf{s})\|_2^2 = g(\mathbf{s}) - \frac{1}{2L_2}\|\nabla g(\mathbf{s})\|_2^2$$

Therefore, for all $\mathbf{s} \in \mathbb{R}^{\hat{d}}$, it holds that

$$\|\nabla g(\mathbf{s})\|_2^2 \leq 2L_2 \left(g(\mathbf{s}) - f^\star\right)$$

■

Lemma 20 *Suppose that Assumption 3, 4 holds. Then for all* $\mathbf{x} \in \mathcal{B}_{R_\mathbf{x}}^{(1)}$ *and* $\mathbf{u} \in \mathcal{B}_{R_\mathbf{u}}^{(2)}$ *we have*

$$\|\nabla_2 f(\mathbf{x}, \mathbf{u})\|_2^2 \leq 2G_1 L_2 \left(f(\mathbf{x}, \mathbf{u}) - f^\star\right)$$

Proof Since Assumption 3 holds, we can invoke Lemma 19 to get that for all $\mathbf{x} \in \mathcal{B}_{R_\mathbf{x}}^{(1)}$ and $\mathbf{u} \in \mathcal{B}_{R_\mathbf{u}}^{(2)}$, we have

$$\|\nabla g(h(\mathbf{x}, \mathbf{u}))\|_2^2 \leq 2L_2 \left(f(\mathbf{x}, \mathbf{u}) - f^\star\right)$$

By Assumption 4, we must have that $\|\nabla_2 h(\mathbf{x}, \mathbf{u})\|_2 \leq G_1$. Therefore, using the chain rule, we have

$$\|\nabla_2 f(\mathbf{x}, \mathbf{u})\|_2^2 \leq \|\nabla_2 h(\mathbf{x}, \mathbf{u})\|_2 \|\nabla g(h(\mathbf{x}, \mathbf{u}))\|_2^2 \leq 2G_1^2 L_2 \left(f(\mathbf{x}, \mathbf{u}) - f^\star\right)$$

■

Lemma 21 *Let* ϕ_0 *be defined in (26). Suppose that Assumption holds. Then we have that* $\phi_0 \leq 2(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star)$.

Proof To start, for all $c \leq 1$, we must have that

$$\gamma = \frac{c}{2\sqrt{\kappa} - c} \leq \frac{c}{\sqrt{\kappa}} \leq 1$$

Thus, for λ, we have $\lambda = (1+\gamma)^3 - 1 \leq 7\gamma$. This implies that

$$\mathcal{Q}_1 = \frac{\lambda^2}{2\eta(1+\gamma)^5} \leq \frac{25\gamma^2}{\eta} = \frac{25cL_1}{\kappa} = 25c\mu$$

When $k = 0$, we have $\mathbf{z}_0 = \mathbf{y}_0 = \mathbf{x}_0$. Moreover, $\mathbf{x}_{-1}^\star = \arg\min_{\mathbf{x} \in \mathbb{R}^{d_1}} f(\mathbf{x}, \mathbf{u}_0)$. At $\mathbf{u}_0$, Assumption 1 must hold, which implies that

$$f(\mathbf{x}_0, \mathbf{u}_0) \geq f(\mathbf{x}_{-1}^\star, \mathbf{u}_0) + \frac{\mu}{2} \left\|\mathbf{x}_0 - \mathbf{x}_{-1}^\star\right\|_2^2$$

This implies that $\left\|\mathbf{x}_0 - \mathbf{x}_{-1}^\star\right\|_2^2 \leq \frac{2}{\mu} \left(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star\right)$. Thus

$$\mathcal{Q}_1 \left\|\mathbf{z}_0 - \mathbf{x}_{-1}^\star\right\|_2^2 \leq 50c \left(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star\right)$$

Moreover, by Assumption 2, we have

$$\frac{\eta}{8} \|\nabla_1 f(\mathbf{y}_{-1}, \mathbf{v}_{-1}\|_2 = \frac{\eta}{8} \|\nabla_1 f(\mathbf{x}_0, \mathbf{u}_0)\|_2^2 \leq \frac{\eta L_1}{4} \left(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star\right) = \frac{c}{4} \left(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star\right)$$

Thus, putting things together, we have

$$\phi_0 \leq (1 + 50.25c) \left(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star\right) \leq 2 \left(f(\mathbf{x}_0, \mathbf{u}_0) - f^\star\right)$$

as long as $c \leq \frac{1}{51}$. ■