ON THE ASYMPTOTIC BEHAVIOR OF GRADIENT DESCENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work has studied the asymptotic behavior of gradient descent (GD) near isolated minima, including the existence of convergence directions. In this work, it is investigated whether analogous directional behavior persists near non-isolated minima, where the minimum set locally forms a regular surface. This setting is particularly relevant to modern overparameterized learning models, where non-isolated minima arise naturally and the structure of the solution set can shape the local optimization dynamics. The non-isolated setting introduces an additional difficulty because the limiting minimizer is not fixed in advance, and the relevant Hessian eigendirections may vary along the minimum set. Under the stated local regularity assumptions, a forward-invariant neighborhood is first established, within which GD converges to a specific minimizer and satisfies a quantitative descent estimate. The asymptotic direction of approach is then characterized. Depending on the learning rate, the GD trajectory admits either a limiting convergence direction or an alternating convergence direction. These results extend the directional convergence phenomenon from isolated minima to non-isolated minima and provide a local description of the long-time behavior of GD near a regular surface of minima.

1 INTRODUCTION

Currently, gradient descent (GD) remains a cornerstone in the study of optimization algorithms within modern machine learning. For a loss function f and an initialization $\boldsymbol{x}_0$, the GD trajectory is defined by

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta \nabla f(\boldsymbol{x}_k), \tag{1}$$

where η is the constant learning rate.

The classical analysis of GD is typically established under the assumption of L-smoothness. Within this framework, significant theoretical attention is directed toward establishing convergence guarantees and descent properties. A canonical result is the descent lemma from Nesterov (2003), which guarantees a monotonic decrease of the objective function for conservative learning rates $\eta < 2/L$. Specifically, the descent property is formulated as

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \eta \left(1 - \frac{L\eta}{2}\right) \|\nabla f(\boldsymbol{x}_k)\|^2. \tag{2}$$

However, the optimization behavior of GD in modern learning models is not fully described by this classical framework. One representative phenomenon is the Edge of Stability (EoS), where full-batch GD can exhibit non-monotonic and oscillatory behavior under relatively large learning rates while still making optimization progress Cohen et al. (2021). These observations have motivated further investigation into the dynamical behavior of GD beyond conventional descent guarantees.

Furthermore, overparameterized neural networks provide a natural setting in which minima need not be isolated. Owing to the high dimensionality and redundancy of the parameter space, the loss function may possess a continuous set of minima rather than a single isolated minimizer Li and Liang (2018). Such non-isolated minimum sets introduce additional structure into the local optimization dynamics and motivate the study of GD near a regular surface of minima.

Recent work has established convergence-direction properties of GD near isolated minima Chen et al. (2026). In this setting, the limiting minimizer is fixed and the asymptotic behavior can be

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

characterized through the Hessian at that point. The non-isolated setting considered in this work introduces an additional difficulty: the limiting minimizer is not fixed in advance, and the relevant Hessian eigendirections may vary along the minimum set. Consequently, extending the convergence-direction analysis from isolated minima to a regular surface of minima requires additional control of the local GD dynamics.

Motivated by these considerations, fundamental properties of GD are investigated for smooth loss functions whose minimum set locally forms a regular surface. Specifically, the following results are established:

- **Forward-invariance:** Under the stated local conditions, GD trajectories that enter a suitable neighborhood of the minimum set remain within this neighborhood. In this regime, the GD trajectory converges to a specific minimizer and the objective satisfies a quantitative descent estimate.
- **Convergence Direction:** The asymptotic direction of the GD trajectory is characterized according to the learning rate η, yielding either a limiting convergence direction or an alternating convergence direction.

The forward-invariance property provides the local dynamical foundation for the convergence-direction analysis by ensuring that the trajectory remains in a neighborhood where the regular-surface structure can be used throughout the iterations. The resulting directional characterization extends the corresponding phenomenon from isolated minima to non-isolated minima and provides a local description of the long-time behavior of GD near a regular surface of minima.

2 Related Work

Optimization algorithms constitute the backbone of deep learning, fundamentally determining the efficiency of training and the generalization capability of the resulting models. While classical optimization theory provides rigorous convergence guarantees under assumptions such as global L-smoothness and conservative learning rates, these frameworks often diverge significantly from the practical settings of modern neural networks. In practice, training typically involves highly non-convex landscapes and learning rates that exceed standard stability bounds, challenging traditional theoretical understanding. To bridge this gap, a growing body of literature has investigated these dynamics through various lenses. In this section, we review these developments, focusing specifically on the geometry of the loss landscape, the dynamical regime known as the EoS, and the distinct optimization properties inherent in overparameterized networks.

The EoS phenomenon, first identified by Cohen et al. (2021), describes how full-batch GD with large learning rates exhibits sharpness growth until the largest Hessian eigenvalue approaches $2/\eta$, followed by oscillations near this boundary while the loss decreases. Extensions include the discovery of the "catapult phase" Lewkowycz et al. (2020), where training remains stable beyond classical bounds. In addition, oscillatory and monotonic modes have also been identified in matrix factorization Wang et al. (2022a). An important theoretical work Ahn et al. (2022) shows that GD can converge to a point with sharpness above the stability threshold only from a set of measure zero. Moreover, Arora et al. (2022); Damian et al. (2023) link EoS to implicit bias toward flat minima, with Damian et al. (2023); Zhu et al. (2023) providing simplified analyses of curvature dynamics in GD and SGD, and Chen and Bruna (2023) analyzing period-2 oscillations.

The interplay between overparameterization and optimization dynamics remains incompletely understood. Arora et al. (2018) shows that depth-induced symmetries transform isolated minima into connected manifolds, accelerating convergence via implicit preconditioning. This is extended by Chen and Chen (2020), which finds overparameterized matrix factorization behaves as if equipped with momentum and adaptive rates. Zhu et al. (2022) highlights how width and initialization affect training robustness and EoS-like behavior in shallow models, while Ghosh et al. (2025b) characterizes depth-dependent sharpness scaling in deep linear networks, though with local, model-specific analyses. Most closely related to our work, MacDonald et al. (2025) studies GD at EoS for overparameterized least squares, decomposing dynamics into tangential and orthogonal components and deriving precise convergence rates, including power-law convergence and period-two oscillations. Our work is inspired by theirs and relaxes their assumptions to propose the convergence direction property.

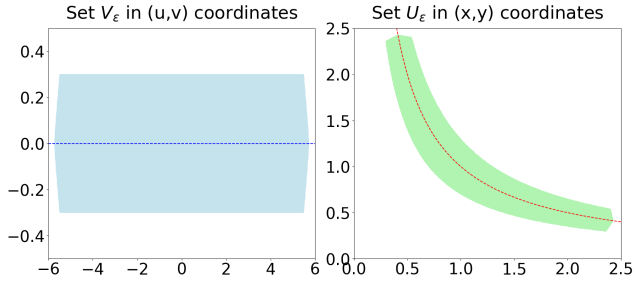
The structure of the loss landscape near minima also contributes to GD dynamics. Ma et al. (2022) observes that subquadratic curvature near minima plays a key role in sharpness measures and stability behavior, particularly in stochastic regimes. Normalization layers and their influence are studied by Dai et al. (2024), while Long and Bartlett (2024) demonstrates that Sharpness-Aware Minimization (SAM) can regulate Hessian sharpness through adaptive perturbations. Optimization methods with curvature adaptation or line search also display distinct characteristics. Curvature-adaptive learning rate methods Thomas et al. (2020); Wang and Ma (2024); Cohen et al. (2022) and non-monotonic line search strategies Fox et al. (2024) provide alternative ways to modulate sharpness and stability. Momentum methods, particularly Polyak's momentum, are shown to induce sharp transitions in curvature that favor flatter regions Phunyaphibarn et al. (2024). Parameter averaging strategies also affect the behavior of GD: Nitanda et al. (2024) shows that iteration smoothing can mitigate instability induced by large learning rates. The role of stochastic noise has been particularly insightful. Mulayoff et al. (2021) connected large learning rates with implicit regularization through noise geometry, a perspective further developed in Lee et al. (2016); Zhu et al. (2019); Martens (2020); Wang and Wu (2023). These works highlight how stochasticity can act as a stabilizing force and shape generalization. Multiple works have emphasized phase transition phenomena in optimization dynamics. Wang et al. (2022b); Cai et al. (2024) identify transitions between EoS-related regimes, and Yang et al. (2024) demonstrates how untuned SGD exhibits sensitive responses to hyperparameters. Complementing these insights, the implicit bias of gradient methods Arora et al. (2022); Nacson et al. (2023) continues to be a unifying thread in explaining the long-term optimization behavior.

Moreover, many studies on typical examples are also notable. The work by Zhu et al. (2023) analyzes a 4-scalar function $(xyzw-1)^2$, demonstrating oscillatory convergence near sharpness $2/\eta$. Subsequent studies explored simplified settings: Wang et al. (2022a) proves implicit balancing effects in matrix factorization, while Wang et al. (2023) connects EoS occurrence to function regularity in shallow models. For deeper architectures, Ghosh et al. (2025b) establishes depth-dependent sharpness scaling in linear networks, though their analysis remains local. Recent impossibility results by Crăciun and Ghoshdastidar (2024) further highlight the gap between empirical observations and rigorous convergence guarantees, particularly for large learning rates. Liu et al. (2025) investigate a deliberately minimalist neural network model consisting of a two-layer architecture with a low-dimensional input structure. A complementary theoretical analysis is provided by Ghosh et al. (2025a), who study GD dynamics in deep linear networks on matrix factorization losses. Although these works empirically observe loss oscillations, sharpness adaptation, and balancing, the theoretical foundations for general nonlinear systems remain unclear.

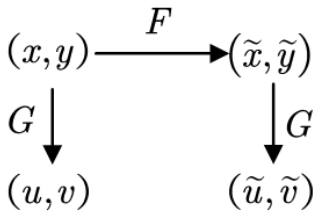
An important theoretical ingredient underlying our analysis is the resolution of the Gradient Conjecture, originally proposed by René Thom and established in Parusinski et al. (2000). This conjecture characterizes the asymptotic geometry of trajectories generated by gradient flows of real analytic functions near their critical points. It gives a positive result for the existence of a convergence direction, which inspires our investigation of the analogous property for GD. Related empirical and theoretical observations on directional behavior also appear in the optimization and deep learning literature. In Xing et al. (2018), the authors study SGD trajectories in neural networks and report that successive updates tend to align with narrow valley-like regions of the loss surface, often exhibiting alternating or oscillatory directions. Complementarily, Morchdi et al. (2023) analyzes correlations between consecutive gradient directions during training and finds that substantial optimization progress frequently coincides with phases of strong negative correlation, underscoring the nontrivial geometric structure of gradient dynamics in nonconvex settings. Moreover, the existence of convergence directions for discrete-time GD has been rigorously established near isolated local minima Chen et al. (2026). Motivated by these developments, we investigate whether analogous directional phenomena persist for discrete-time GD. Our results show that, under convergence, GD iterates display structured asymptotic behavior: the update directions either approach a fixed limiting direction or oscillate along a specific line related to an eigendirection of the Hessian of the loss function. This observation establishes a concrete link between the asymptotic geometry of continuous gradient flows and the dynamics of their discrete counterparts.

3 A Motivating Two-Dimensional Example

To illustrate the dynamical phenomena explored in this work, we begin with a canonical two-dimensional example that exhibits both structural properties central to our analysis: local forward-



(a) The sets U_ε and V_ε with $\eta = 0.3$ and $\varepsilon = 0.09$. The set V_ε covers some of the $v = 0$ axis, and U_ε is its preimage under the diffeomorphism G.



(b) A single GD update visualized simultaneously in (x, y)–coordinates and in (u, v)–coordinates.

Figure 1: Visualization of the forward-invariant sets and a single GD update.

invariance and characterization of the convergence direction. This example serves as the primary source of intuition for the general theory, and its full technical treatment is given in Appendix 7. Here we present a concise but self-contained discussion of the main ideas, emphasizing the geometry of the dynamics and the role of the coordinate transformation.

We begin with the function

$$f(x, y) = \tfrac{1}{2}(xy - 1)^2, \tag{3}$$

defined on the first quadrant $\mathbb{R}^2_+$. The global minima of f form a smooth hyperbola

$$\mathcal{M} = \{(x, y) \in \mathbb{R}^2_+ : xy = 1\}. \tag{4}$$

The GD mapping with step size $\eta > 0$ is given by

$$F(x, y) = \big(x - \eta y(xy - 1),\ y - \eta x(xy - 1)\big), \tag{5}$$

and we denote its k-th step of GD by $(x_{k+1}, y_{k+1}) = F(x_k, y_k)$ inductively with initial point (x_0, y_0).

To illustrate the intrinsic structure of the example, we introduce the following new coordinates. Define the mapping

$$\begin{aligned} G\colon \mathbb{R}^2_+ &\to \mathbb{R} \times (-1, +\infty), \\ (x, y) &\mapsto (u, v) = (x^2 - y^2,\ xy - 1). \end{aligned} \tag{6}$$

The transformation separates the motion along the minimum set in the u component from the motion orthogonal to it in the v component. Under this transformation, the minimum set $\mathcal{M}$ is mapped precisely to the horizontal axis $\{v = 0\}$. Technically, the mapping is regular since its Jacobian is everywhere nonsingular on $\mathbb{R}^2_+$, and one checks that G is bijective onto its domain. Hence G is a diffeomorphism and provides a coordinate system adapted to $\mathcal{M}$.

Next, we state the main results for this example. To analyze GD dynamics, we introduce a family of regions in (u, v)-coordinates,

$$V_\varepsilon = \big\{(u, v) \in \mathbb{R}\times(-1, +\infty) :\ |v| < \varepsilon/\eta, u^2 < (2(1-\varepsilon)/\eta)^2 - 4, u^2 + 4(1+|v|)^2 < (2(1-\varepsilon)/\eta)^2\big\}, \tag{7}$$

for $0 < \varepsilon < \min\{1 - \eta, \eta/3\}$. Intuitively, each V_ε is a narrow tube surrounding the horizontal axis, with controlled width determined by ε and η. We pull this region back to (x, y)-coordinates by

$$U_\varepsilon = G^{-1}(V_\varepsilon). \tag{8}$$

To visualize the appearance of U_ε and V_ε, one can refer Figure 1a.

Next, one observes that the V_ε defined above is a forward-invariant region in (u, v)-coordinate under GD dynamics. Let $(u, v) = G(x, y)$ and $(\widetilde{u}, \widetilde{v}) = (G \circ F)(x, y)$, then a direct computation gives the transformed GD dynamics in (u, v)-coordinate:

$$(\widetilde{u}, \widetilde{v}) = \big(u(1 - \eta^2 v^2), v + \eta^2 v^2(v + 1) - \eta v\sqrt{u^2 + 4(v + 1)^2}\big). \tag{9}$$

These coordinates reveal the essential structure of the dynamics: u moves slowly along $\mathcal{M}$, while v is rapidly driven toward zero since the v component is contractive by the definition of V_ε. Figure 1b illustrates the relation between one GD step in the two coordinate systems.

For sufficiently small ε, it follows that V_ε is a forward-invariant set with

$$(\widetilde{u}, \widetilde{v}) \in V_\varepsilon \quad \text{whenever } (u, v) \in V_\varepsilon. \tag{10}$$

Transporting this back to (x, y)–space, we conclude that each set U_ε is forward invariant under GD. This ensures that when the GD dynamics get sufficiently close to the minimum manifold, they cannot escape from this region. To gather all permissible ε into one domain, we define

$$V = \cup_{0<\varepsilon<\min\{1-\eta,\eta/3\}} V_\varepsilon, \quad U = G^{-1}(V). \tag{11}$$

Then V is the full forward-invariant set in (u, v)-coordinate. Pulling this back to the (x, y)-coordinates, GD trajectories in U satisfy the forward-invariance property: once a trajectory falls into U, all subsequent iterates remain in U. This property also implies that (x_k, y_k) converges to a point on the minimum set $\mathcal{M}$, and that the function value decreases along the GD trajectory,

$$f(x_{k+1}, y_{k+1}) \leq C\, f(x_k, y_k), \qquad 0 < C < 1. \tag{12}$$

Furthermore, a particularly interesting feature is that the direction of convergence depends on the sharpness of the limiting minimizer. Let (x^*, y^*) be the convergent point starting from $(x_0, y_0) \in U$, and define

$$R = (x^*)^2 + (y^*)^2 \tag{13}$$

and

$$d_k = \frac{(x_k, y_k) - (x^*, y^*)}{\|(x_k, y_k) - (x^*, y^*)\|}. \tag{14}$$

One has the following result:

- If $R < 1/\eta$, the convergence direction $\lim_{k\to\infty} d_k$ exists.
- If $1/\eta < R < 2/\eta$, the alternative convergence direction $\lim_{k\to\infty}(-1)^k d_k$ exists.

Crucially, the analysis shows that for almost every initialization in U, the limiting behavior falls into one of the two cases above.

This example encapsulates the essential behavior of GD within overparameterized networks. The general results proved later follow the same blueprint, with this example providing the clearest and most explicit instance of the phenomena.

4 Main Theorems

This section establishes our main theoretical findings that characterize GD behavior for loss functions with minimizers forming a locally regular surface. The analysis builds upon the canonical example examined previously, whose techniques naturally generalize to arbitrary n-dimensional settings. We outline the structural assumptions, present our key components and theorems, and offer an intuitive description of the main results. Complete proofs appear in the Appendix 8.

To better understand the following results in this section, we provide an illustration of the convergence direction in GD dynamics near a regular surface of minima in Figure 2. The visualization summarizes our core theoretical finding: the qualitative behavior of the GD trajectory is determined by the learning rate η. As shown, the minimum set $\mathcal{M}$ is locally characterized by tangential directions $\boldsymbol{\gamma}$ and normal directions $\boldsymbol{\varphi}$. We prove that for a small learning rate (illustrated by the **green trajectory**), the GD trajectory will asymptotically align with the "flattest" normal direction $\boldsymbol{\varphi}_1(\boldsymbol{x}^*)$, corresponding to the smallest positive eigenvalue of the Hessian. Conversely, when η is large (illustrated by the **blue trajectory**), the dynamics enter an oscillatory regime. In this state, the iterates exhibit oscillations along the "sharpest" normal direction $\boldsymbol{\varphi}_m$ while slowly drifting along the manifold. The theorems in this section rigorously establish these distinct asymptotic convergence directions.

Consider a function $f \in C^\infty(\mathbb{R}^n)$ and without loss of generality, suppose that $\mathbf{0}$ is one of its minima. For a given radius $r > 0$, denote M as the set of all minima of f within $B_n(\mathbf{0}, r)$, which is an open ball centered at $\mathbf{0}$ with radius r in n dimension. In addition, we assume that M is a regular surface, i.e., there exists a diffeomorphism

$$\psi : U \to M, \tag{15}$$

where $U \subset \mathbb{R}^{n-m}$ denotes an open set containing $\mathbf{0}$, satisfying

$$\psi(\mathbf{0}) = \mathbf{0}. \tag{16}$$

Additionally, we require that for every minimum $\boldsymbol{x} \in M$,

$$\operatorname{rank}(\nabla^2 f(\boldsymbol{x})) = m > 1. \tag{17}$$

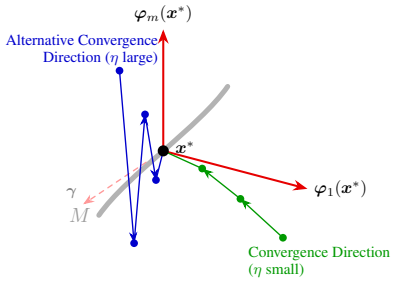


Figure 2: Visualization of the asymptotic convergence direction with different learning rates.

We further stipulate that the non-zero eigenvalues of $\nabla^2 f(\boldsymbol{x})$ are strictly positive and mutually distinct. Specifically, for any minimum $\boldsymbol{x} \in M$,

$$0 < \lambda_1(\boldsymbol{x}) < \lambda_2(\boldsymbol{x}) < \cdots < \lambda_m(\boldsymbol{x}), \tag{18}$$

where $\lambda_i(\boldsymbol{x})$ represents the i-th smallest positive eigenvalue of $\nabla^2 f(\boldsymbol{x})$. Moreover, we impose the condition that

$$\inf_{\boldsymbol{x} \in M} \lambda_1(\boldsymbol{x}) = \inf_{\boldsymbol{u} \in U} \lambda_1(\psi(\boldsymbol{u})) > 0. \tag{19}$$

We denote the GD dynamics by $F : \mathbb{R}^n \to \mathbb{R}^n$, defined as

$$F(\boldsymbol{x}) = \boldsymbol{x} - \eta \nabla f(\boldsymbol{x}), \tag{20}$$

where $\eta > 0$ represents the constant learning rate. From a technical standpoint, we also require that the preimage under F of any Lebesgue null set $\mathcal{N}$, denoted $F^{-1}(\mathcal{N})$, is itself a Lebesgue null set. Given an initial condition $\boldsymbol{x}_0 \in \mathbb{R}^n$, we define the GD trajectory recursively by setting, for $k \geq 0$,

$$\boldsymbol{x}_{k+1} = F(\boldsymbol{x}_k). \tag{21}$$

Paralleling the example presented in the preceding section and from Theorem 7.1, we restrict attention to learning rates η satisfying

$$\eta < 2/\lambda_m(\mathbf{0}), \tag{22}$$

where $\lambda_m(\mathbf{0})$ denotes the maximal eigenvalue of $\nabla^2 f(\mathbf{0})$. These hypotheses collectively ensure that M forms a regular surface, guaranteeing the minimizer set admits a well-defined local parametrization.

Having established the requisite framework, we now state the main contributions of this work through the following two theorems. Our central objective is to characterize the asymptotic behavior of the GD sequence $\{\boldsymbol{x}_k\}_{k \geq 0}$ when its limit lies in a neighborhood of the minimum $\mathbf{0}$.

Theorem 4.1 (Forward-invariance)**.** *Consider the f and η defined above, there exists an open set $\Omega \subset \mathbb{R}^n$ containing the minimum $\mathbf{0}$, such that for each initial point $\boldsymbol{x}_0 \in \Omega$, $\boldsymbol{x}_k \in \Omega$ for all $k \geq 0$, $\lim_{k\to\infty} \boldsymbol{x}_k$ exists and belongs to M, and there exists a constant $C > 0$ such that for all $k \geq 0$,*

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - C\|\nabla f(\boldsymbol{x}_k)\|^2. \tag{23}$$

Theorem 4.2 (Convergence direction)**.** *Consider the f and η defined above, if η satisfies*

$$\eta < \frac{2}{\lambda_1(\mathbf{0}) + \lambda_m(\mathbf{0})}, \tag{24}$$

then there exists an open set $\Omega \subset \mathbb{R}^n$ containing the minimum $\mathbf{0}$, such that for almost all $\boldsymbol{x}_0 \in \Omega$, $\lim_{k\to\infty} \boldsymbol{x}_k = \boldsymbol{x}^*$ exists and belongs to M. Moreover, the convergence direction $\lim_{k\to\infty} \boldsymbol{d}_k$ exists, where $\boldsymbol{d}_k$ is defined as

$$\boldsymbol{d}_k = \frac{\boldsymbol{x}_k - \boldsymbol{x}^*}{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|}. \tag{25}$$

On the other hand, if

$$\frac{2}{\lambda_1(\mathbf{0}) + \lambda_m(\mathbf{0})} < \eta < \frac{2}{\lambda_m(\mathbf{0})}, \tag{26}$$

then there exists an open set $\Omega \subset \mathbb{R}^n$ containing the minimum $\mathbf{0}$, such that for almost all $\boldsymbol{x}_0 \in \Omega$, $\lim_{k\to\infty} \boldsymbol{x}_k = \boldsymbol{x}^*$ exists and belongs to M. Moreover, the alternative convergence direction $\lim_{k\to\infty} (-1)^k \boldsymbol{d}_k$ exists, where $\boldsymbol{d}_k$ is defined as in equation 25.

To facilitate the analysis of the asymptotic behavior and build intuition, we construct a local coordinate system adapted to the regular surface. First, we establish the existence of a smooth local frame on M. Since the eigenvalues $\lambda_i(\boldsymbol{x})$ are distinct and strictly positive for $\boldsymbol{x} \in M$, the implicit function theorem guarantees that the corresponding unit normal eigenvectors $\boldsymbol{\varphi}_i(\boldsymbol{x})$ depend smoothly on $\boldsymbol{x}$, where $i = 1, \cdots, m$. Then $\{\boldsymbol{\varphi}_1(\boldsymbol{x}), \ldots, \boldsymbol{\varphi}_m(\boldsymbol{x})\}$ forms a smooth local frame in the orthogonal directions of M at $\boldsymbol{x}$. Moreover, by the assumption that M is locally a regular surface, $D\psi$ has full rank and provides a linearly independent smooth local frame in the tangential directions of M. Consequently, together with all columns of $D\psi$ and the orthogonal frame above, we establish a smooth local frame of M at $\boldsymbol{x}$.

Leveraging this smooth frame, we define a mapping $G : U \times V \to \mathbb{R}^n$ as follows:

$$G(\boldsymbol{u}, \boldsymbol{v}) = \psi(\boldsymbol{u}) + \sum_{i=1}^{m} v_i \boldsymbol{\varphi}_i(\psi(\boldsymbol{u})), \tag{27}$$

where U is defined in equation 15 and $V = B_m(\mathbf{0}, \varepsilon_0) \subset \mathbb{R}^m$. It is verified that the Jacobian of G is non-singular at $(\mathbf{0}, \mathbf{0})$, which implies that G acts as a diffeomorphism on a local neighborhood of the minima. This transformation is critical to our analysis as it effectively decouples the highly nonlinear dynamics of GD into two distinct components:

Tangential Component ($\boldsymbol{u}$): Parameterizes the motion along the minimum set M, corresponds to the eigen-directions of zero eigenvalues of the Hessian. This component corresponds to the minor drift in the GD dynamics.

Normal Component ($\boldsymbol{v}$): Represents the displacement orthogonal to M. This component is governed by the eigen-direction of strictly positive eigenvalues of the Hessian and is responsible for the fast contraction or oscillation observed in the GD dynamics.

Next, to ensure the stability of the optimization process, it is essential to establish that the GD trajectory remains confined within a specific neighborhood of the regular surface. This property, referred to as forward-invariance in this work, serves a critical theoretical function: it guarantees that the local coordinate system defined in equation 27 remains valid throughout the iterations. Without this confinement, the trajectory could escape the domain of the diffeomorphism, rendering the local analysis of the dynamics invalid.

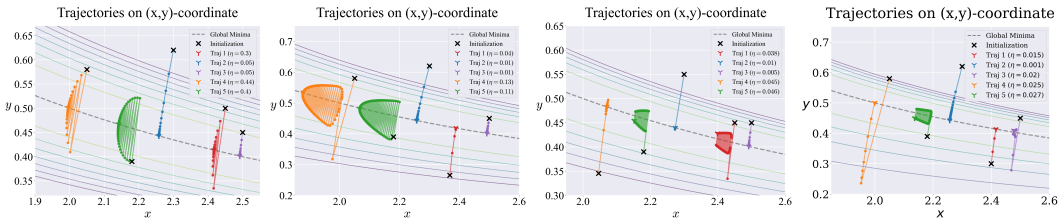
In addition, to provide intuition for Theorem 4.2, we briefly elucidate the mechanism driving the directional convergence, with formal proofs detailed in the appendix. As established by the local diffeomorphism G in equation 27, the GD update near the minimum set essentially operates on the normal components $\boldsymbol{v}$ via distinct contraction rates. For each eigen-direction $\boldsymbol{\varphi}_i(\boldsymbol{x}^*)$ with eigenvalue $\lambda_i(\boldsymbol{x}^*)$, the dynamics along this direction are governed approximately by the scaling factor $1 - \eta\lambda_i(\boldsymbol{x}^*)$.

The asymptotic convergence direction is determined by the specific orthogonal eigen-direction whose scaling factor possesses the largest absolute value, representing the slowest rate of contraction. Moreover, when the learning rate is small, the quantity $1 - \eta\lambda_1(\boldsymbol{x}^*)$ is positive and dominates the contraction rates of all other directions. Consequently, the trajectory aligns asymptotically with $\boldsymbol{\varphi}_1(\boldsymbol{x}^*)$, corresponding to the direction of minimal curvature orthogonal to the regular surface. Conversely, in the large learning rate regime, the factor $1-\eta\lambda_m(\boldsymbol{x}^*)$ becomes negative and dominates in absolute magnitude. This forces the trajectory to oscillate along $\boldsymbol{\varphi}_m(\boldsymbol{x}^*)$, the direction of maximal curvature, resulting in the alternative convergence direction characterized by the flipping behavior.

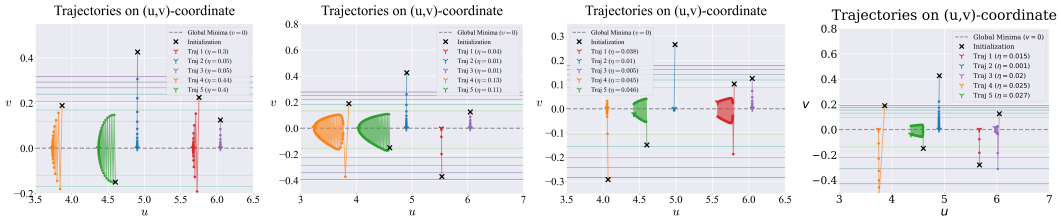
Thus, in the general n-dimensional setting, GD near the regular surface M remains in a forward-invariant neighborhood and approaches the minimum set along a single normal eigendirection, either monotonically or with sign alternation depending on the learning rate.

5 Numerical Experiments

In this section, numerical experiments are presented to illustrate directional behavior of GD across a range of models, from two-dimensional examples to deep neural networks.



(a) $d = 1$: (x, y)-coord (b) $d = 2$: (x, y)-coord (c) $d = 3$: (x, y)-coord (d) $d = 4$: (x, y)-coord



(e) $d = 1$: (u, v)-coord (f) $d = 2$: (u, v)-coord (g) $d = 3$: (u, v)-coord (h) $d = 4$: (u, v)-coord

Figure 3: **Evolution of GD dynamics across degrees $d = 1$ to $d = 4$.** The top row shows trajectories in (x, y)-coordinates; the relevant branch of the minimum set in the first quadrant is the curve $xy = 1$. The bottom row shows the same trajectories in (u, v)-coordinates, where the minimum set is flattened to the line $v = 0$. Across all degrees, the iterates contract rapidly in the normal direction while the tangential drift remains comparatively small.

5.1 Minimalist Examples

We first consider the canonical example

$$f_d(x, y) = \frac{1}{2}\big((xy)^d - 1\big)^2, \tag{28}$$

which has been studied in Zhu et al. (2023); Wang et al. (2023). The global minimum set of equation 28 is given by

$$\mathcal{M} := \{(x, y) \in \mathbb{R}^2 : (xy)^d = 1\}. \tag{29}$$

We perform GD with different learning rates from multiple initializations. The trajectories are visualized both in the original (x, y) coordinates and in the transformed coordinates (u, v) defined by

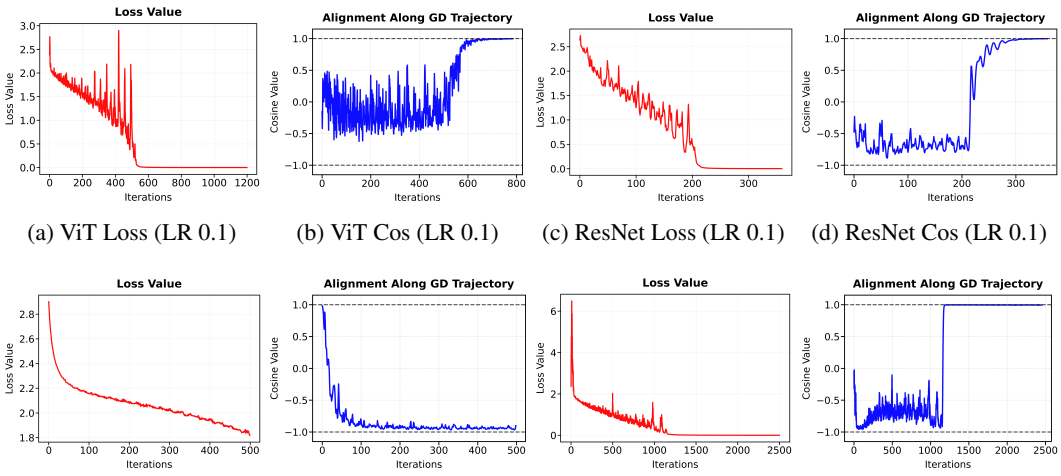
$$u = x^2 - y^2, \quad v = xy - 1, \tag{30}$$

which align with the analytical framework introduced in earlier sections. Figure 3 presents a comprehensive visualization of the optimization paths across different degrees d. In these experiments, although the iterates approach different points on $\mathcal{M}$ depending on initialization, the dynamics exhibit convergence directions toward $\mathcal{M}$. In particular, the iterates rapidly contract in the normal direction, while motion along the tangent direction is minor. Such behavior provides a clear numerical demonstration of the existence of a convergence direction, consistent with our theoretical results.

5.2 Experiments on Neural Networks

In this section, we present experiments on neural networks with different architectures, following the experimental setups in Xing et al. (2018); Morchdi et al. (2023). The experimental evaluation is performed on the CIFAR-10 dataset. To ensure controlled experimentation, a subset of the training data is utilized. Specifically, $N = 5000$ samples are randomly selected, ensuring a balanced distribution with 500 samples for each of the 10 classes. The neural network architectures employed in this study are Vision Transformer, ResNet-56, and ConvNeXt. The training is carried out using full-batch GD, where the model parameters are updated according to standard GD. See Figure 4 for more details.

To analyze the trajectory of the optimization process, the alignment between consecutive update vectors is monitored. Let $\boldsymbol{g}_k$ denote the update vector at step k, defined as $\boldsymbol{g}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k$. The



(a) ViT Loss (LR 0.1) (b) ViT Cos (LR 0.1) (c) ResNet Loss (LR 0.1) (d) ResNet Cos (LR 0.1)

(e) ResNet Loss (LR 0.001) (f) ResNet Cos (LR 0.001) (g) ConvNeXt Loss(LR0.1) (h) ConvNeXt Cos (LR 0.1)

Figure 4: Training dynamics of several architectures under full-batch GD on CIFAR-10. The four experiment pairs correspond to ViT (LR 0.1), ResNet-56 (LR 0.1), ResNet-56 (LR 0.001), and ConvNeXt (LR 0.1). In each pair, the left plot shows the loss and the right plot shows the cosine between consecutive updates. These cosine plots provide qualitative illustrations of aligned and oscillatory update behavior observed during training.

cosine value c_k between consecutive updates is calculated as

$$c_k = \frac{\boldsymbol{g}_{k+1} \cdot \boldsymbol{g}_k}{\|\boldsymbol{g}_{k+1}\| \|\boldsymbol{g}_k\|}. \tag{31}$$

This metric provides a qualitative indicator of directional behavior along the optimization trajectory. Values close to 1 indicate alignment between consecutive updates, whereas values close to -1 indicate alternating update directions. Since the theoretical results characterize the local asymptotic regime near the limiting minimizer, transient negative cosine values do not by themselves imply an alternating asymptotic convergence direction.

6 Conclusion

In this work, a local framework is developed for analyzing GD near non-isolated minima. A forward-invariant neighborhood near a regular surface of minima is first established, within which the GD trajectory converges to a specific minimizer and satisfies a quantitative descent estimate. Building on this local dynamical foundation, the asymptotic convergence direction is characterized, yielding either a limiting direction or an alternating limiting direction depending on the learning rate. These results extend the directional convergence phenomenon from isolated minima to the non-isolated setting. While this work establishes a theoretical framework for GD on regular surfaces, several avenues for future research remain to bridge the gap between theory and practice.

- First, the minimum set is currently assumed to be locally a regular surface. However, the loss landscapes of deep neural networks are known to be significantly more complex. Consequently, further investigations are needed to extend these results to more general settings.
- Second, it is significant to relax the current structural assumptions and investigate whether the forward-invariance and directional convergence properties established here persist for a broader class of functions, such as those satisfying the Kurdyka–Łojasiewicz (KL) inequality. Establishing such extensions could provide a more general understanding of the asymptotic behavior of GD beyond regular minimum manifolds.
- Finally, while this study focuses on full-batch GD, it remains to be determined whether other optimization algorithms exhibit similar directional convergence and invariant properties. Future work considers extensions of this analysis to methods such as Stochastic Gradient Descent (SGD) and momentum-based optimizers, which are ubiquitous in training modern deep learning models.

REFERENCES

Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. In *Proceedings of International Conference on Machine Learning*, pages 247–257, 2022.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *Proceedings of International Conference on Machine Learning*, pages 244–253, 2018.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *Proceedings of International Conference on Machine Learning*, pages 948–1024, 2022.

Yuhang Cai, Jingfeng Wu, Song Mei, Michael Lindsey, and Peter L. Bartlett. Large stepsize gradient descent for non-homogeneous two-layer networks: Margin improvement and fast optimization. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Lin Chen and Joan Bruna. Beyond the edge of stability via two-step gradient updates. In *Proceedings of International Conference on Machine Learning*, pages 4330–4391. PMLR, 2023.

Pu Chen and Hung-Hsuan Chen. Accelerating matrix factorization by overparameterization. In *Proceedings of International Conference on Deep Learning Theory and Applications*, pages 89–97, 2020.

Shuo Chen, Jiaying Peng, Xiaolong Li, and Yao Zhao. On the convergence direction of gradient descent. In *Proceedings of International Conference on Learning Representations*, 2026.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *Proceedings of International Conference on Learning Representations*, 2021.

Jeremy M Cohen, Behrooz Ghorbani, Shankar Krishnan, Naman Agarwal, Sourabh Medapati, Michal Badura, Daniel Suo, David Cardoze, Zachary Nado, George E Dahl, et al. Adaptive gradient methods at the edge of stability. In *Advances in Neural Information Processing Systems*, 2022.

Alexandru Crăciun and Debarghya Ghoshdastidar. On the convergence of gradient descent for large learning rates. In *arXiv preprint arXiv:2402.13108*, 2024.

Yan Dai, Kwangjun Ahn, and Suvrit Sra. The crucial role of normalization in sharpness-aware minimization. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Alex Damian, Eshaan Nichani, and Jason D Lee. Selfstabilization: The implicit bias of gradient descent at the edge of stability. In *Proceedings of International Conference on Learning Representations*, 2023.

Curtis Fox, Leonardo Galli, Mark Schmidt, and Holger Rauhut. Nonmonotone line searches operate at the edge of stability. In *Proceedings of the OPT 2024 Workshop on Optimization for Machine Learning*, 2024.

Avrajit Ghosh, Soo Min Kwon, Rongrong Wang, Saiprasad Ravishankar, and Qing Qu. Learning dynamics of deep linear networks beyond the edge of stability. In *arXiv preprint arXiv:2502.20531*, 2025a. doi: 10.48550/arXiv.2502.20531.

Avrajit Ghosh, Soo Min Kwon, Rongrong Wang, Saiprasad Ravishankar, and Qing Qu. Learning dynamics of deep matrix factorization beyond the edge of stability. In *Proceedings of International Conference on Learning Representations*, 2025b.

Nicholas J. Higham. *Accuracy and stability of numerical algorithms*. SIAM, 2 edition, 2002.

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

John M. Lee. *Introduction to smooth manifolds*. Springer, New York, NY, 2 edition, 2012.

Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. In *arXiv preprint arXiv:2003.02218*, 2020.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Liming Liu, Zixuan Zhang, Simon S. Du, and Tuo Zhao. A minimalist example of edge-of-stability and progressive sharpening. In *Annual Conference on Neural Information Processing Systems*, 2025.

Philip M. Long and Peter L. Bartlett. Sharpness-aware minimization and the edge of stability. *Journal of Machine Learning Research*, 25(179):1–20, 2024.

Chao Ma, Daniel Kunin, Lei Wu, and Lexing Ying. Beyond the quadratic approximation: The multiscale structure of neural network loss landscapes. *Journal of Machine Learning*, 2022.

Lachlan Ewen MacDonald, Hancheng Min, Leandro Palma, Salma Tarmoun, Ziqing Xu, and Rene Vidal. Convergence rates for gradient descent on the edge of stability in overparameterised least squares. In *Annual Conference on Neural Information Processing Systems*, 2025.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

Boris S. Mityagin. The zero set of a real analytic function. *Mathematical Notes*, 107:774–783, 2020.

Chedi Morchdi, Yi Zhou, Jie Ding, and Bei Wang. Exploring gradient oscillation in deep neural network training. In *Annual Allerton Conference on Communication, Control, and Computing*, pages 1–7. IEEE, 2023.

Rotem Mulayoff, Tomer Michaeli, and Daniel Soudry. The implicit bias of minima stability: A view from function space. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Mor Shpigel Nacson, Rotem Mulayoff, Greg Ongie, Tomer Michaeli, and Daniel Soudry. The implicit bias of minima stability in multivariate shallow ReLU networks. In *Proceedings of International Conference on Learning Representations*, 2023.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Atsushi Nitanda, Ryuhei Kikuchi, Shugo Maeda, and Denny Wu. Why is parameter averaging beneficial in SGD? An objective smoothing perspective. In *Proceedings of International Conference on Artificial Intelligence and Statistics*, pages 3565–3573, 2024.

Adam Parusinski, Tadeusz Mostowski, and Krzysztof Kurdyka. Proof of the gradient conjecture of R. Thom. *Annals of Mathematics*, 152(3):763–792, 2000.

Prin Phunyaphibarn, Junghyun Lee, Bohan Wang, Huishuai Zhang, and Chulhee Yun. Gradient descent with Polyak's momentum finds flatter minima via large catapults. In *Workshop on High-dimensional Learning Dynamics (HiLD) at ICML 2024*, 2024.

Valentin Thomas, Fabian Pedregosa, Bart van Merriënboer, Pierre-Antoine Manzagol, Yoshua Bengio, and Nicolas Le Roux. On the interplay between noise and curvature and its effect on optimization and generalization. In *Proceedings of International Conference on Artificial Intelligence and Statistics*, volume 108, pages 3503–3513, 2020.

Mingze Wang and Chao Ma. Understanding multi-phase optimization dynamics and rich nonlinear behaviors of ReLU networks. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Mingze Wang and Lei Wu. A theoretical analysis of noise geometry in stochastic gradient descent. In *NeurIPS Workshop on M3L*, 2023.

Yuqing Wang, Minshuo Chen, Tuo Zhao, and Molei Tao. Large learning rate tames homogeneity: Convergence and balancing effect. In *Proceedings of International Conference on Learning Representations*, 2022a.

Yuqing Wang, Zhenghao Xu, Tuo Zhao, and Molei Tao. Good regularity creates large learning rate implicit biases. In *NeurIPS Workshop*, 2023.

Zixuan Wang, Zhouzi Li, and Jian Li. Analyzing sharpness along GD trajectory: Progressive sharpening and edge of stability. In *Advances in Neural Information Processing Systems*, volume 35, 2022b.

Chen Xing, Devansh Arpit, Christos Tsirigotis, and Yoshua Bengio. A walk with SGD. *CoRR*, abs/1802.08770, 2018.

Junchi Yang, Xiang Li, Ilyas Fatkhullin, and Niao He. Two sides of one coin: the limits of untuned SGD and the power of adaptive methods. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Xingyu Zhu, Zixuan Wang, Xiang Wang, Mo Zhou, and Rong Ge. Understanding edge-of-stability training dynamics with a minimalist example. In *Proceedings of International Conference on Learning Representations*, 2023.

Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In *Proceedings of International Conference on Machine Learning*, pages 7654–7663, 2019.

Zhenyu Zhu, Fanghui Liu, Grigorios Chrysos, and Volkan Cevher. Robustness in deep learning: The good (width), the bad (depth), and the ugly (initialization). In *Advances in Neural Information Processing Systems*, volume 35, pages 36094–36107, 2022.

AI USE STATEMENT

In this work, we used generative AI tools, including ChatGPT and Gemini, to assist in the writing and checking of parts of mathematical proofs, to provide limited feedback on experimental design, and to assist with the implementation of the experimental pipeline. We did not use generative AI tools to generate synthetic datasets, develop the theoretical model or conceptual framework, formulate the mathematical claims, propose or refine the research hypotheses, assist with translation, clean or reformat datasets, or interpret experimental results. Qualitative and thematic data analysis is not applicable to this work.

Additionally, we used generative AI tools to improve the readability of parts of the manuscript, to create a schematic scientific figure, to create or edit experimental code, and to assist with preliminary reference-list preparation. The research problem, theoretical framework, mathematical claims, experimental objectives, and final scientific conclusions were determined by the authors.

All AI-assisted material was reviewed by the authors. In particular, all mathematical arguments and proofs were independently checked and revised as necessary; AI-assisted code was reviewed and tested to ensure that it correctly implemented the intended experimental design; the AI-assisted schematic figure was checked for mathematical correctness; and the final bibliography was selected and verified by the authors, with all cited references checked for validity and relevance. The final manuscript was carefully reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims, code, figures, or other artifacts produced with the aid of generative AI.

APPENDIX

This appendix contains the complete proofs of our main theoretical results. The appendix is organized as follows: Section 7 provides a comprehensive treatment of the motivating two-dimensional example, establishing the blueprint for the general case. Section 8 then extends these results to n-dimensional functions, beginning with the formal setup in Section 8.1 and culminating in the proofs in the later sections.

7 PROOF OF THE TWO-DIMENSIONAL MOTIVATING EXAMPLE

In the theoretical analysis of deep learning optimization, understanding the trajectory of GD requires several fundamental properties. Specifically, we are concerned with:

- **Forward-invariance:** Ensuring that once the dynamics enter a specific region near the local minima, they remain confined within that region. With this property, the GD trajectory approaches a specific limit point rather than oscillating or diverging. Moreover, the loss function value decreases monotonically: there exists a constant $0 < C < 1$ such that, for sufficiently large $k > 0$, $f(x_{k+1}, y_{k+1}) \leq Cf(x_k, y_k)$.
- **Convergence Direction:** Characterizing the direction that the trajectory approaches the minimum, which demonstrates that the GD dynamics are controllable and simple.

To investigate these properties, in this part of the appendix we focus on the canonical nonconvex example defined on $\mathbb{R}^2$:

$$f(x, y) = \frac{1}{2}(xy - 1)^2. \tag{32}$$

This function serves as a minimalist, scalar proxy for the general matrix factorization or rank-1 approximation problem, which is often formulated as $\min_{U,V} \|UV^\top - Y\|_F^2$. While simplified, this example captures the essential geometric features of overparameterized networks, specifically the existence of non-isolated global minima.

A critical factor governing the convergence of GD is the sharpness of the loss landscape at the minimum, denoted by $\lambda_{\max}(\nabla^2 f(x^*, y^*))$. To establish the necessary conditions for convergence, we invoke a key impossibility result regarding the stability of GD dynamics as follows:

Theorem 7.1 (Theorem 1, Ahn et al. (2022))**.** *Let $f \in C^2(\mathcal{D})$ be defined on a subset $\mathcal{D} \subseteq \mathbb{R}^n$. Define the GD trajectory $F : \mathcal{D} \to \mathbb{R}^n$ as:*

$$F(\boldsymbol{\theta}) = \boldsymbol{\theta} - \eta \nabla f(\boldsymbol{\theta}). \tag{33}$$

Assume that the preimage $F^{-1}(\mathcal{S})$ of any measure-zero set $\mathcal{S} \subset \mathcal{D}$ also has measure zero. Furthermore, assume that for any stationary point $\boldsymbol{\theta}^* \in \mathcal{D}$, the value $1/\eta$ is not an eigenvalue of the Hessian $\nabla^2 f(\boldsymbol{\theta}^*)$. If $\boldsymbol{\theta}^*$ satisfies either $\lambda_{\max}(\nabla^2 f(\boldsymbol{\theta}^*)) > 2/\eta$ or $\lambda_{\min}(\nabla^2 f(\boldsymbol{\theta}^*)) < 0$, then the set of initializations $\boldsymbol{\theta}_0 \in \mathcal{D}$ for which the GD trajectory converges to $\boldsymbol{\theta}^*$ has Lebesgue measure zero.

This theorem implies that convergence is only possible if the sharpness satisfies the stability condition $\lambda_{\max} < 2/\eta$. For the case defined by equation 32, the gradient and Hessian of $f(x, y)$ are given by

$$\nabla f(x, y) = \begin{pmatrix} y(xy - 1) \\ x(xy - 1) \end{pmatrix}, \tag{34}$$

and

$$\nabla^2 f(x, y) = \begin{pmatrix} y^2 & 2xy - 1 \\ 2xy - 1 & x^2 \end{pmatrix} \tag{35}$$

respectively. Therefore, the GD dynamics for this example is

$$F(x, y) = (x - \eta y(xy - 1), y - \eta x(xy - 1)) \tag{36}$$

The set of global minima is the hyperbola defined by

$$\mathcal{M} = \{(x, y) \in \mathbb{R}^2 \mid xy = 1\}. \tag{37}$$

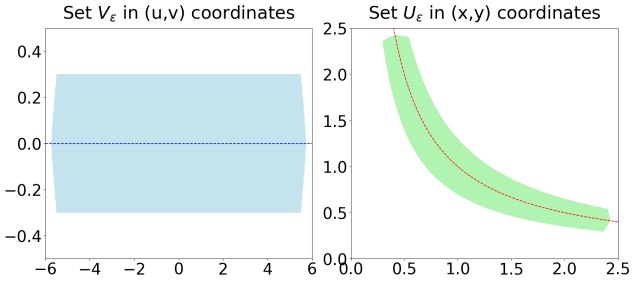


Figure 5: The sets U_ε and V_ε with $\eta = 0.3$ and $\varepsilon = 0.09$.

For each minimum $(x^*, y^*) \in \mathcal{M}$, the two eigenvalues of $\nabla^2 f$ are

$$\lambda_{\min} = 0, \quad \text{and} \quad \lambda_{\max} = (x^*)^2 + (y^*)^2. \tag{38}$$

The zero eigenvalue corresponds to the flat direction along the tangent of $\mathcal{M}$, while $\lambda_{\max}$ represents the sharpness in the orthogonal direction.

To ensure that there exists a set of initial points that converges to a minimum, we require the condition

$$\eta < 1 \tag{39}$$

since for all $(x^*, y^*) \in \mathcal{M}$,

$$\lambda_{\max} = (x^*)^2 + (y^*)^2 \geq 2. \tag{40}$$

With this context established, we now define the coordinate transformation and invariant sets used in our analysis. For the convenience of the reader and to ensure this section is self-contained, we formally restate the definitions of the key quantities introduced in the main text, including the coordinate transformation and the invariant sets. These definitions remain consistent with the main text but are reproduced here to facilitate the rigorous verification of the proofs.

Next, define the following change of coordinates:

$$G\colon (x, y) \in \mathbb{R}^2_+ \mapsto (u, v) \in \mathbb{R} \times (-1, +\infty), \tag{41}$$

with

$$G(x, y) = (x^2 - y^2,\ xy - 1). \tag{42}$$

As the Jacobian of G is

$$DG = \begin{pmatrix} 2x & -2y \\ y & x \end{pmatrix}, \tag{43}$$

and $\det(DG) = 2x^2 + 2y^2 > 0$ for all $(x, y) \in \mathbb{R}_+ \times \mathbb{R}_+$. It is easy to verify that G is a diffeomorphism from $\mathbb{R}_+ \times \mathbb{R}_+$ to $\mathbb{R} \times (-1, +\infty)$. With the coordinate change G, we can work in (u, v)-coordinates. For $0 < \eta < 1$ and $0 < \varepsilon < \min\{1 - \eta, \eta/3\}$, define V_ε as:

$$V_\varepsilon = \left\{ (u, v) \in \mathbb{R} \times (-1, +\infty) : |v| < \frac{\varepsilon}{\eta}, u^2 < \left(\frac{2 - 2\varepsilon}{\eta}\right)^2 - 4, u^2 + 4(1 + |v|)^2 < \left(\frac{2 - 2\varepsilon}{\eta}\right)^2 \right\}, \tag{44}$$

and

$$U_\varepsilon = G^{-1}(V_\varepsilon). \tag{45}$$

Since the mapping $G : (x, y) \mapsto (u, v)$ is a diffeomorphism, it establishes a one-to-one correspondence between points in the original space and the transformed space. This allows us to interpret the constraints defining V_ε directly in terms of the geometry of U_ε. Specifically, the condition on the v-coordinate, $|v| < \varepsilon/\eta$, translates directly via $v = xy - 1$ to the constraint $|xy - 1| < \varepsilon/\eta$ in the (x, y) coordinate system. This implies that the set U_ε is bounded within a hyperbolic strip around $\mathcal{M}$. Similarly, the constraints on u restrict the extent of the set along the hyperbolic curves, ensuring that the region U_ε remains bounded and well-defined for our stability analysis. To build intuition for U_ε and V_ε, see Figure 5. As a remark, if $\eta \to 0$, then these U_ε can cover the whole minimum set $\mathcal{M}$. Then we need the following technical lemma:

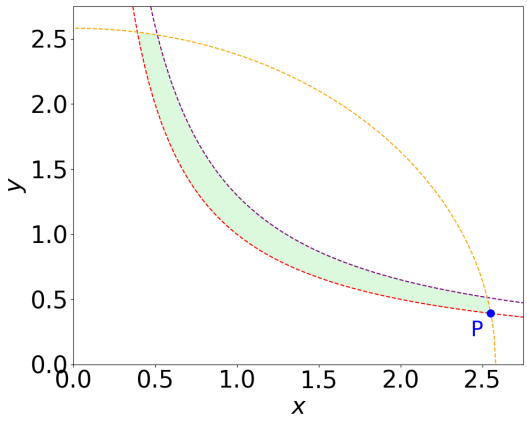


Figure 6: The point P with $\eta = 0.3$ and $\varepsilon = 0.09$.

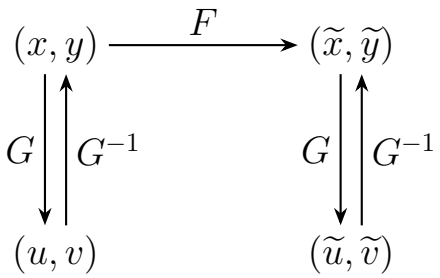


Figure 7: Illustration of the coordinates in (x, y) and (u, v).

Lemma 7.2. *Given $0 < \eta < 1$ and $0 < \varepsilon < \min\{1 - \eta, \eta/3\}$. Then, for each $(x, y) \in U_\varepsilon$, $F(x, y) \in \mathbb{R}_+ \times \mathbb{R}_+$, where U_ε and F are defined in equation 45 and equation 36 respectively.*

Proof. We prove this lemma in two cases.

Case 1: $(x, y) \in U_\varepsilon$ with $xy \leq 1$. In this case,

$$x - \eta y(xy - 1) \geq x > 0, \tag{46}$$

and

$$y - \eta x(xy - 1) \geq y > 0, \tag{47}$$

ensuring that $F(x, y) \in \mathbb{R}_+ \times \mathbb{R}_+$.

Case 2: $(x, y) \in U_\varepsilon$ with $xy > 1$. In this case, it is sufficient to prove that:

$$\eta(xy - 1) < \min\{y/x, x/y\}. \tag{48}$$

Without loss of generality, assume $x \geq y$, which implies

$$\min\{y/x, x/y\} = y/x. \tag{49}$$

By the definition of V_ε and U_ε, we have, in this case,

$$1 < xy = v + 1 < \varepsilon/\eta + 1 \tag{50}$$

and

$$(x^2 + y^2)^2 = u^2 + 4(1 + v)^2 < \left(\frac{2 - 2\varepsilon}{\eta}\right)^2 < \left(\frac{2}{\eta}\right)^2. \tag{51}$$

As a result, we know that (x, y) lies in the following set (see the green part of Figure 6).

$$\left\{(x, y) \in \mathbb{R}_+^2 : 1 \leq xy \leq \varepsilon/\eta + 1,\ x^2 + y^2 \leq 2/\eta\right\}. \tag{52}$$

Thus, the lemma holds if one can prove that

$$\eta(xy-1) < y_P/x_P \tag{53}$$

holds for the point $P=(x_P,y_P)$ satisfying $x_P > y_P$, and

$$\begin{cases} x_P y_P = 1, \\ (x_P)^2 + (y_P)^2 = 2/\eta. \end{cases} \tag{54}$$

In this case,

$$\frac{y_P}{x_P} = \frac{1-\sqrt{1-\eta^2}}{\eta} > \frac{1-\left(1-\eta^2/2\right)}{\eta} = \frac{\eta}{2} > \frac{\eta}{3} \tag{55}$$

and thus, as $\eta/3 > \varepsilon$, we have

$$y_P/x_P > \varepsilon > \eta(xy-1). \tag{56}$$

This proves the lemma. □

For $(x,y) \in U_\varepsilon$, according to Lemma 7.2, we know

$$(\widetilde{x},\widetilde{y}) = F(x,y) \in \mathbb{R}_+ \times \mathbb{R}_+. \tag{57}$$

Then we can define

$$(\widetilde{u},\widetilde{v}) = G(\widetilde{x},\widetilde{y}). \tag{58}$$

For a better understanding of the relations between the two coordinate systems, see Figure 7. Since

$$(\widetilde{u},\widetilde{v}) = (G \circ F \circ G^{-1})(u,v), \tag{59}$$

we have for each $(u,v) \in V_\varepsilon$,

$$\begin{cases} \widetilde{u} = u(1-\eta^2 v^2) \\ \widetilde{v} = v + \eta^2 v^2 (v+1) - \eta v \sqrt{u^2 + 4(v+1)^2} \end{cases} \tag{60}$$

To ensure the GD dynamics is well-defined, we also need the following lemma.

Lemma 7.3. *Given $0<\eta<1$, and $0<\varepsilon<\min\{1-\eta,\eta/3\}$. For each $(u,v) \in V_\varepsilon$, consider $(\widetilde{u},\widetilde{v})$ defined in equation 58, we have $|\widetilde{u}| \le |u|$ and $|\widetilde{v}| \le (1-\varepsilon)|v|$, and thus $(\widetilde{u},\widetilde{v}) \in V_\varepsilon$. Moreover, for all $(x,y) \in U_\varepsilon$, we have $F(x,y) \in U_\varepsilon$. Here, V_ε and U_ε are defined in equation 44 and equation 45 respectively.*

Proof. By definition equation 44, we have

$$\begin{aligned} & 1+\eta^2 v(v+1) - \eta\sqrt{u^2+4(v+1)^2} \\ \le\ & 1+\eta^2 v(v+1) - 2\eta(v+1) \\ <\ & 1+\eta^2 \cdot \frac{\varepsilon}{\eta}\left(\frac{\varepsilon}{\eta}+1\right) - 2\eta\left(-\frac{\varepsilon}{\eta}+1\right) \\ =\ & 1+\varepsilon(\eta+\varepsilon) - 2\eta + 2\varepsilon \\ <\ & 1+\varepsilon - 2\eta + 2\varepsilon \\ <\ & 1-2\eta+3\varepsilon \\ <\ & 1-\eta \\ <\ & 1-\varepsilon. \end{aligned} \tag{61}$$

Similarly,

$$\begin{aligned} & 1+\eta^2 v(1+v) - \eta\sqrt{u^2+4(v+1)^2} \\ \ge\ & 1+\eta^2 v - \eta\sqrt{u^2+4(|v|+1)^2} \\ >\ & 1-\eta^2\cdot\frac{\varepsilon}{\eta} - \eta\cdot\frac{2-2\varepsilon}{\eta} \\ =\ & -1+2\varepsilon-\eta\varepsilon \\ >\ & -1+2\varepsilon-\varepsilon \\ =\ & -1+\varepsilon. \end{aligned} \tag{62}$$

Therefore, $|\widetilde{v}| \leq (1-\varepsilon)|v|$ is proved. Moreover, once again by equation 44, since $\eta^2 v^2 < \varepsilon^2$, we have

$$0 < 1 - \eta^2 v^2 \leq 1, \tag{63}$$

and thus $|\widetilde{u}| \leq |u|$ is obvious. In this case, we have $(\widetilde{u}, \widetilde{v}) \in V_\varepsilon$. Finally, for $(x, y) \in U_\varepsilon$, since we have just proved that $(\widetilde{u}, \widetilde{v}) \in V_\varepsilon$, it follows that $(\widetilde{x}, \widetilde{y}) \in G^{-1}(V_\varepsilon) = U_\varepsilon$. This completes the proof. □

To formally characterize the trajectory, we define the sequence of iterates as follows. Let (x_0, y_0) be an initialization and define the GD trajectory recursively by

$$(x_{k+1}, y_{k+1}) = F(x_k, y_k) \tag{64}$$

for $k \geq 0$. Correspondingly, in the transformed coordinates, we denote

$$(u_k, v_k) = G(x_k, y_k). \tag{65}$$

Furthermore, Lemma 7.3 guarantees forward-invariance via induction: if the initialization satisfies $(x_0, y_0) \in U_\varepsilon$, then $(x_k, y_k) \in U_\varepsilon$ for all $k \geq 0$. Moreover, since G is a diffeomorphism from U_ε to V_ε, then every (x_k, y_k) corresponds to a unique (u_k, v_k). Thus, $(x_k, y_k) \in U_\varepsilon$ implies that $(u_k, v_k) \in V_\varepsilon$. In this situation, by considering equation 59 with $(u, v) = (u_k, v_k)$, we have

$$(\widetilde{u}, \widetilde{v}) = (G \circ F \circ G^{-1})(u_k, v_k) = (G \circ F)(x_k, y_k) = G(x_{k+1}, y_{k+1}) = (u_{k+1}, v_{k+1}). \tag{66}$$

As a result of Lemma 7.3, we know that

$$\begin{cases} |u_{k+1}| \leq |u_k| \\ |v_{k+1}| \leq (1-\varepsilon)|v_k| \end{cases}. \tag{67}$$

Moreover, by equation 60, as u_k and u_{k+1} have the same sign, we know that $\{u_k\}_{k\geq 0}$ and $\{v_k\}_{k\geq 0}$ are convergent, and we denote

$$u^* = \lim_{k\to\infty} u_k \quad \text{and} \quad v^* = \lim_{k\to\infty} v_k. \tag{68}$$

Clearly, $(u^*, v^*) \in V_\varepsilon$ and $v^* = 0$. Next, we define

$$V = \bigcup_{0<\varepsilon<\min\{1-\eta, \eta/3\}} V_\varepsilon \tag{69}$$

and

$$U = \bigcup_{0<\varepsilon<\min\{1-\eta, \eta/3\}} U_\varepsilon. \tag{70}$$

Clearly,

$$U = G^{-1}(V). \tag{71}$$

Here, we take the union over all suitable ε to obtain a larger set that covers all minima whose sharpness satisfies $\lambda_{\max} < 2/\eta$. By the above preparation, we state our main theorem as follows.

Theorem 7.4. *Given $0 < \eta < 1$ and U defined above, for each $(x_0, y_0) \in U$, the following properties hold:*

- ***Forward-invariance:*** *For all $k \geq 0$, $(x_k, y_k) \in U$.*
- ***Descent property:*** *There exists a constant $0 < C < 1$ such that for all $k \geq 0$, $f(x_{k+1}, y_{k+1}) \leq C f(x_k, y_k)$.*
- ***Convergence property:*** *The limit $\lim_{k\to\infty} (x_k, y_k)$ exists and lies in U.*

Proof. First, for $(x_0, y_0) \in U$, by definition, there exists $0 < \varepsilon < \min\{1-\eta, \eta/3\}$ such that $(x_0, y_0) \in U_\varepsilon$. From Lemma 7.3, for all $k \geq 0$, $(x_k, y_k) \in U_\varepsilon$ holds inductively. Thus, $(x_k, y_k) \in U$ holds for all $k \geq 0$. Moreover, by Lemma 7.3, we have

$$f(x_{k+1}, y_{k+1}) \leq (1-\varepsilon)^2 f(x_k, y_k). \tag{72}$$

Thus, taking $C = (1-\varepsilon)^2 < 1$, we complete the proof of the descent property. Next,

$$\lim_{k\to\infty} (x_k, y_k) = \lim_{k\to\infty} G^{-1}(u_k, v_k) = G^{-1}(u^*, v^*). \tag{73}$$

As $G^{-1}(u^*, v^*) \in U_\varepsilon \subset U$, the convergence property is proved. □

Finally, we will prove the existence of convergence direction and alternative convergence direction. We first introduce some technical lemmas.

Lemma 7.5. *Consider η and U defined in Theorem 7.4. For a given $(x_0, y_0) \in U$ with*

$$\lim_{k\to\infty}(x_k, y_k) = (x^*, y^*), \tag{74}$$

define

$$R = (x^*)^2 + (y^*)^2, \tag{75}$$

and suppose that $v_k \neq 0$ holds for each $k \geq 0$. Then the following result holds:

- *If $R < 1/\eta$, then there exists $k_0 \geq 0$, such that both $\{x_k\}_{k\geq k_0}$ and $\{y_k\}_{k\geq k_0}$ are strictly decreasing.*
- *If $1/\eta < R < 2/\eta$, then there exists $k_0 \geq 0$, such that $\{x_{k_0+2k}\}_{k\geq 0}$ and $\{y_{k_0+2k}\}_{k\geq 0}$ are strictly decreasing, $\{x_{k_0+2k+1}\}_{k\geq 0}$ and $\{y_{k_0+2k+1}\}_{k\geq 0}$ are strictly increasing.*

Proof. First, recall that $v_k = x_k y_k - 1$ with

$$\lim_{k\to\infty}\frac{v_{k+1}}{v_k} = 1 - \eta R. \tag{76}$$

If $R < 1/\eta$, by equation 76, there exists $k_0 \geq 0$, such that for all $k \geq k_0$,

$$\frac{v_{k+1}}{v_k} > 0. \tag{77}$$

Therefore, v_k preserves its sign for all $k \geq k_0$. As a result, by equation 36 and equation 64, we know that

$$\begin{cases} x_{k+1} = x_k - \eta y_k v_k \\ y_{k+1} = y_k - \eta x_k v_k. \end{cases} \tag{78}$$

Thus, $\{x_k\}_{k\geq k_0}$ and $\{y_k\}_{k\geq k_0}$ are strictly decreasing or increasing, depending on the sign of v_{k_0}.

On the other hand, if $1/\eta < R < 2/\eta$, similarly, in this case, by equation 76, there exists $k_0 \geq 0$, such that for all $k \geq k_0$,

$$\frac{v_{k+1}}{v_k} < 0. \tag{79}$$

Thus, v_k alternates its sign for $k \geq k_0$. Without loss of generality, we assume that $v_{k_0} > 0$. By equation 36 and equation 64, we know that

$$x_{k+2} = (x_k - \eta y_k v_k) - \eta(y_k - \eta x_k v_k)v_{k+1}. \tag{80}$$

Rearranging the terms yields:

$$x_{k+2} = x_k - \eta v_k\left(\left(1 + \frac{v_{k+1}}{v_k}\right)y_k - \eta x_k v_{k+1}\right). \tag{81}$$

Then for any k such that $k - k_0$ is even, we have $v_k > 0$ and $v_{k+1} < 0$. In this case, since $|v_{k+1}| < (1-\varepsilon)|v_k|$, we have

$$1 + \frac{v_{k+1}}{v_k} > \varepsilon. \tag{82}$$

Therefore,

$$\left(1 + \frac{v_{k+1}}{v_k}\right)y_k - \eta x_k v_{k+1} > \varepsilon y_k > 0. \tag{83}$$

Then, equation 81 implies $x_{k+2} < x_k$. Similarly, by an analogous argument, we have $y_{k+2} < y_k$. As a result, $\{x_{k_0+2k}\}_{k\geq 0}$ and $\{y_{k_0+2k}\}_{k\geq 0}$ are strictly decreasing. Moreover, by the same deduction, one can prove that $\{x_{k_0+2k+1}\}_{k\geq 0}$ and $\{y_{k_0+2k+1}\}_{k\geq 0}$ are strictly increasing. This completes the proof. □

Next, we need to exclude some sets with bad behavior that have measure zero in the following lemma.

Lemma 7.6. *Let U be the set defined in Theorem 7.4, for each $k \geq 0$, define*

$$S_k = \big\{(x_0, y_0) \in U : x_k y_k = 1\big\}, \tag{84}$$

Then S_k has measure zero. Moreover, for each $(x_0, y_0) \in U \setminus (\cup_{k\geq 0} S_k)$, there exists $k_1 \geq 0$ such that for all $k \geq k_1$, $z_k \neq 0$, where z_k is defined as

$$z_k = y^*(x_k - x^*) + x^*(y_k - y^*). \tag{85}$$

Proof. Clearly, $\{(x, y) \in \mathbb{R}^2 : xy = 1\}$ is a zero measure set Mityagin (2020). Moreover, since each S_k is actually a vanish set of a polynomial with respect to (x_0, y_0). By Mityagin (2020), S_k is measure zero as well. Next, if $R < 1/\eta$, by Lemma 7.5, we know that $z_k = 0$ only holds for at most one $k \geq k_0$. On the other hand, for $1/\eta < R < 2/\eta$, by Lemma 7.5, we know that there exists $k_0 \geq 0$ such that $\{z_{2k+k_0}\}_{k\geq 0}$ is strictly decreasing, and $\{z_{2k+k_0+1}\}_{k\geq 0}$ is strictly increasing. Thus, $z_k = 0$ holds for at most two $k \geq k_0$. Then the lemma is finally proved. □

Lemma 7.6 means that S_k contains all possible inital points that would cause $v_k = 0$ infinitely many times or its GD trajectory converges within finitely many steps. However, excluding these measure-zero sets is a fundamental step in proving the existence of a convergence direction.

Theorem 7.7. ***Convergence direction:*** *For η, U defined in Theorem 7.4, S_k defined in Lemma 7.6, then for each $(x_0, y_0) \in U \setminus (\cup_{k\geq 0} S_k)$, consider (x^*, y^*) and R defined in equation 74 and equation 75 respectively, define*

$$d_k = \frac{(x_k, y_k) - (x^*, y^*)}{\|(x_k, y_k) - (x^*, y^*)\|}. \tag{86}$$

The following result holds:

- *If $R < 1/\eta$, then the convergence direction of GD exists, i.e., $\lim_{k\to\infty} d_k$ exists.*
- *If $1/\eta < R < 2/\eta$, then the alternative convergence direction of GD exists, i.e., $\lim_{k\to\infty} (-1)^k d_k$ exists.*

Proof. We take the vector

$$\pi_k = (x_k, y_k) \tag{87}$$

and

$$\begin{cases} \alpha = (x^*, -y^*)\,, \\ \beta = (y^*, x^*)\,, \\ \gamma = (x^*, y^*)\,. \end{cases} \tag{88}$$

With these notations, γ is the limit of π_k, α and β are orthogonal with the same norm. We focus on the limit of the unit vector d_k, whose direction is parallel to $\pi_k - \gamma$. Knowing that

$$\pi_k - \gamma = \frac{\langle \pi_k - \gamma, \alpha\rangle \alpha + \langle \pi_k - \gamma, \beta\rangle \beta}{\|\alpha\|^2}, \tag{89}$$

our goal is to show that the vector d_k eventually has only the orthogonal component, and thus its convergence direction or alternating convergence direction is parallel to β and determined by the sign of v_k.

For a given $(x_0, y_0) \in U \setminus (\cup_{k\geq 0} S_k)$, by Lemma 7.5 and 7.6, there exists $k_2 \geq 0$ such that for all $k \geq k_2$, $x_k - x^*$ and $y_k - y^*$ has the same sign, and $z_k \neq 0$, where z_k is defined in equation 85. Then

$$\left|\frac{v_k}{z_k}\right| = \left|1 + \frac{(x_k - x^*)\,(y_k - y^*)}{z_k}\right| \leq 1 + \left|\frac{(x_k - x^*)\,(y_k - y^*)}{z_k}\right| \leq 1 + \left|\frac{x_k - x^*}{x^*}\right|. \tag{90}$$

Moreover, as $z_k = \langle \pi_k - \gamma, \beta\rangle$, it implies that

$$\left|\frac{v_k}{\langle \pi_k - \gamma, \beta\rangle}\right| \leq 1 + \left|\frac{x_k - x^*}{x^*}\right|. \tag{91}$$

Next, consider the term $\langle \pi_k - \gamma, \alpha \rangle$, by equation 91 and equation 67,

$$\begin{aligned} |\langle \pi_k - \gamma, \alpha \rangle| \leq & \sum_{p=0}^{\infty} |x^* (x_{k+p} - x_{k+p+1}) - y^* (y_{k+p} - y_{k+p+1})| \\ = & \eta \left(\sum_{p=0}^{\infty} |x^* y_{k+p} - y^* x_{k+p}| \, |v_{k+p}| \right) \\ \leq & \eta \sum_{p=0}^{\infty} (1-\varepsilon)^p \, |x^* y_{k+p} - y^* x_{k+p}| \, |v_k| \\ \leq & \frac{\eta}{\varepsilon} |v_k| \sup_{p \geq k} \{ |x^* y_p - y^* x_p| \} . \end{aligned} \tag{92}$$

As a result, we have

$$\lim_{k \to \infty} \frac{\langle \pi_k - \gamma, \alpha \rangle}{v_k} = 0. \tag{93}$$

Next, combining equation 91 and equation 93,

$$\frac{\langle \pi_k - \gamma, \alpha \rangle}{\langle \pi_k - \gamma, \beta \rangle} = \frac{\langle \pi_k - \gamma, \alpha \rangle}{v_k} \frac{v_k}{\langle \pi_k - \gamma, \beta \rangle} \longrightarrow 0. \tag{94}$$

Therefore,

$$\lim_{k \to \infty} |\cos \langle \pi_k - \gamma, \beta \rangle| = 1. \tag{95}$$

The last step is to discuss the sign of $\cos \langle \pi_k - \gamma, \beta \rangle$. If $R < 1/\eta$, by Lemmas 7.5 and 7.6, we have for all $k \geq k_2$, $z_k = \langle \pi_k - \gamma, \beta \rangle$ will preserve its sign and non-zero. Thus, this implies that, depending on the sign of v_{k_2},

$$\lim_{k \to \infty} d_k = \pm \beta / \|\beta\|. \tag{96}$$

Otherwise, if $1/\eta < R < 2/\eta$, similarly, by Lemmas 7.5 and 7.6, we have for all $k \geq k_2$, $z_k = \langle \pi_k - \gamma, \beta \rangle$ will alternate its sign and non-zero. Hence, this implies that, depending on the sign of v_{k_2},

$$\lim_{k \to \infty} (-1)^k d_k = \pm \beta / \|\beta\|. \tag{97}$$

This completes the proof. □

8 Proof of the General n-Dimensional Theorem

This appendix provides complete proofs of the main theoretical results presented in Section 4, extending the analysis from the motivating two-dimensional example to the general n-dimensional setting. Our objective is to establish two key properties of GD dynamics for loss functions whose minima form a regular surface.

We organize this appendix as follows. First, Section 8.1 introduces the main theorems discussed in the main text. Next, Section 8.2 develops the necessary technical machinery, including the coordinate transformation that separates tangential and normal directions relative to the minimum manifold. The subsequent sections then provide detailed proofs: Section 8.3 establishes crucial analytical estimates, Section 8.4 constructs the forward-invariant sets, and Sections 8.5 through 8.10 complete the proofs of Theorems 8.1 and 8.2. Throughout, we maintain the geometric intuition developed in the two-dimensional case while addressing the additional technical challenges that arise in higher dimensions.

8.1 Setup and Main Results

We begin by specifying the assumptions on the loss function and the minimum set. These conditions capture the essential geometric features observed in overparameterized neural networks and ensure that the coordinate transformation introduced in the main text extends naturally to the general setting.

Consider a function $f \in C^\infty(\mathbb{R}^n)$, suppose that $\mathbf{0}$ is one of its minima. For a given radius $r > 0$, denote M as the set of all minima of f within $B_n(\mathbf{0}, r)$, which is an n-dimensional open ball centered at $\mathbf{0}$ with radius r. In addition, we assume that M is a regular surface, i.e., there exists a diffeomorphism

$$\psi : U \to M, \tag{98}$$

where $U \subset \mathbb{R}^{n-m}$ is an open set containing $\mathbf{0}$, with

$$\psi(\mathbf{0}) = \mathbf{0}. \tag{99}$$

Moreover, for each minimum $\boldsymbol{x} \in M$, suppose that

$$\text{rank}(\nabla^2 f(\boldsymbol{x})) = m > 1. \tag{100}$$

Besides, assume that the non-zero eigenvalues of $\nabla^2 f(\boldsymbol{x})$ are strictly positive and distinct, i.e., for each minimum $\boldsymbol{x} \in M$,

$$0 < \lambda_1(\boldsymbol{x}) < \lambda_2(\boldsymbol{x}) < \cdots < \lambda_m(\boldsymbol{x}), \tag{101}$$

where $\lambda_i(\boldsymbol{x})$ denote the i-th smallest non-zero eigenvalue of $\nabla^2 f(\boldsymbol{x})$. Finally, we assume that

$$\inf_{\boldsymbol{x} \in M} \lambda_1(\boldsymbol{x}) = \inf_{\boldsymbol{u} \in U} \lambda_1(\psi(\boldsymbol{u})) > 0. \tag{102}$$

Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be the mapping of the GD dynamics given by

$$F(\boldsymbol{x}) = \boldsymbol{x} - \eta \nabla f(\boldsymbol{x}), \tag{103}$$

where $\eta > 0$ is a given learning rate. Technically, we also assume that the preimage of every Lebesgue null set $\mathcal{N}$, i.e., $F^{-1}(\mathcal{N})$, is a Lebesgue null set. Then define GD trajectory with initial point $\boldsymbol{x}_0 \in \mathbb{R}^n$ recursively as, for $k \geq 0$,

$$\boldsymbol{x}_{k+1} = F(\boldsymbol{x}_k). \tag{104}$$

As in the example introduced in the previous section, we consider learning rates η that satisfy

$$\eta < 2/\lambda_m(\mathbf{0}), \tag{105}$$

where $\lambda_m(\mathbf{0})$ is the largest eigenvalue of $\nabla^2 f(\mathbf{0})$.

By equation 105 and the continuity of λ_m, there exists $\varepsilon' > 0$, such that $\eta < 2/\lambda_m(\boldsymbol{x})$ holds for all $\boldsymbol{u} \in B_{n-m}(\mathbf{0}, \varepsilon')$. Without loss of generality, we assume that the learning rate η satisfies

$$\eta < \frac{2}{\lambda}, \tag{106}$$

where

$$\lambda = \sup_{\boldsymbol{x} \in G(U \times V)} \lambda_m(\boldsymbol{x}). \tag{107}$$

These assumptions ensure that the set of minima M is well-behaved and free of singularities, allowing for the construction of a local coordinate system.

Based on the above preparations, we now introduce the main result of this work, consisting of the following two theorems. Our focus is the asymptotic behavior of the GD trajectory $\{\boldsymbol{x}_k\}_{k \geq 0}$ when its limit is close to the minimum $\mathbf{0}$.

Theorem 8.1 (Forward-invariance)**.** *Consider the f and η defined above, there exists an open set $\Omega \subset \mathbb{R}^n$ containing the minimum $\mathbf{0}$, such that for each initial point $\boldsymbol{x}_0 \in \Omega$, $\boldsymbol{x}_k \in \Omega$ for all $k \geq 0$, $\lim_{k \to \infty} \boldsymbol{x}_k$ exists and belongs to M, and there exists a constant $C > 0$ such that for all $k \geq 0$,*

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - C\|\nabla f(\boldsymbol{x}_k)\|^2. \tag{108}$$

Theorem 8.2 (Convergence direction)**.** *Consider the f and η defined above, if η satisfies*

$$\eta < \frac{2}{\lambda_1(\mathbf{0}) + \lambda_m(\mathbf{0})}, \tag{109}$$

then there exists an open set $\Omega \subset \mathbb{R}^n$ containing the minimum $\mathbf{0}$, such that for almost all $\boldsymbol{x}_0 \in \Omega$, $\lim_{k \to \infty} \boldsymbol{x}_k = \boldsymbol{x}^*$ exists and belongs to M. Moreover, the convergence direction $\lim_{k \to \infty} \boldsymbol{d}_k$ exists, where $\boldsymbol{d}_k$ is defined as

$$\boldsymbol{d}_k = \frac{\boldsymbol{x}_k - \boldsymbol{x}^*}{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|}. \tag{110}$$

On the other hand, if

$$\frac{2}{\lambda_1(\mathbf{0}) + \lambda_m(\mathbf{0})} < \eta < \frac{2}{\lambda_m(\mathbf{0})}, \tag{111}$$

then there exists an open set $\Omega \subset \mathbb{R}^n$ containing the minimum $\mathbf{0}$, such that for almost all $\boldsymbol{x}_0 \in \Omega$, $\lim_{k\to\infty} \boldsymbol{x}_k = \boldsymbol{x}^*$ exists and belongs to M. Moreover, the alternative convergence direction $\lim_{k\to\infty}(-1)^k \boldsymbol{d}_k$ exists, where $\boldsymbol{d}_k$ is defined as in equation 110.

Theorem 8.1 establishes the fundamental properties of the GD dynamics. It guarantees that trajectories starting sufficiently close to the minimum set M remain confined within a local neighborhood, ensuring monotonic descent and eventual convergence to a unique minimum. Meanwhile, Theorem 8.2 characterizes the asymptotic geometry of this convergence. It demonstrates that the approach to the limit is directional. Depending on the learning rate, the trajectory either aligns monotonically with a limiting direction or exhibits oscillations along an alternating direction.

8.2 Local Coordinate System

Under the assumptions established in the preceding section, the asymptotic behavior analysis requires a coordinate system adapted to the geometry of the minimum set M. The objective of this section is to construct a local diffeomorphism that decouples the dynamics into components tangential and normal to M. To achieve this, the smoothness of eigenvectors corresponding to $\lambda_i(\boldsymbol{x})$ defined in equation 101 is first verified for $\boldsymbol{x} \in M$. Subsequently, a coordinate transformation is defined to separate the dynamics into orthogonal and tangential components.

By the implicit function theorem, clearly, each $\lambda_i(\boldsymbol{x})$ is smooth for $\boldsymbol{x} \in M$. Moreover, define $\boldsymbol{\varphi}_i(\boldsymbol{x})$ as the unit eigenvector with respect to $\lambda_i(\boldsymbol{x})$, and we now prove that $\boldsymbol{\varphi}_i(\boldsymbol{x})$ is also smooth for $\boldsymbol{x} \in M$. Take

$$H(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x}) \tag{112}$$

for clarity in the following context.

For a given $\boldsymbol{x} \in M$, we first prove that $\boldsymbol{\varphi}_1$ is smooth on a neighborhood of $\boldsymbol{x}$. Without loss of generality, assume that

$$H(\boldsymbol{x}) = \operatorname{diag}(\lambda_1(\boldsymbol{x}), \cdots, \lambda_m(\boldsymbol{x}), 0, \cdots, 0). \tag{113}$$

In this case, we know that

$$\boldsymbol{\varphi}_1(\boldsymbol{x}) = (\pm 1, 0, \cdots, 0)^\top. \tag{114}$$

Take then, for instance,

$$\boldsymbol{\varphi}_1(\boldsymbol{x}) = (1, 0, \cdots, 0)^\top. \tag{115}$$

Let

$$A(\boldsymbol{y}) = \lambda_1(\boldsymbol{y}) I - H(\boldsymbol{y}), \tag{116}$$

which is smooth for $\boldsymbol{y} \in M$. Specifically, $A(\boldsymbol{x})$ is a diagonal matrix with

$$A(\boldsymbol{x}) = (0, \lambda_1(\boldsymbol{x}) - \lambda_2(\boldsymbol{x}), \cdots, \lambda_1(\boldsymbol{x}) - \lambda_m(\boldsymbol{x}), \lambda_1(\boldsymbol{x}), \cdots, \lambda_1(\boldsymbol{x})). \tag{117}$$

Define

$$g : M \times \mathbb{R}^n \to \mathbb{R}^n \tag{118}$$

by

$$g(\boldsymbol{y}, \boldsymbol{z}) = A(\boldsymbol{y})\boldsymbol{z} + (\|\boldsymbol{z}\|^2 - 1, 0, \ldots, 0)^\top. \tag{119}$$

We then have,

$$g(\boldsymbol{x}, \boldsymbol{\varphi}_1(\boldsymbol{x})) = 0, \tag{120}$$

and

$$\det\left(\frac{\partial g}{\partial \boldsymbol{z}}(\boldsymbol{x}, \boldsymbol{\varphi}_1(\boldsymbol{x}))\right) \neq 0. \tag{121}$$

By the implicit function theorem, there exists $\bar{\varepsilon} > 0$ and a smooth mapping

$$\widetilde{\boldsymbol{\varphi}} : B_n(\boldsymbol{x}, \bar{\varepsilon}) \to \mathbb{R}^n \tag{122}$$

such that $\widetilde{\boldsymbol{\varphi}}(\boldsymbol{x}) = \boldsymbol{\varphi}_1(\boldsymbol{x})$, and for each $\boldsymbol{y} \in B_n(\boldsymbol{x}, \bar{\varepsilon})$,

$$g(\boldsymbol{y}, \widetilde{\boldsymbol{\varphi}}(\boldsymbol{y})) = 0. \tag{123}$$

This implies

$$A(\boldsymbol{y})\widetilde{\boldsymbol{\varphi}}(\boldsymbol{y}) + (\|\widetilde{\boldsymbol{\varphi}}(\boldsymbol{y})\|^2 - 1, 0, \ldots, 0)^\top = \mathbf{0}. \tag{124}$$

Let

$$A(\boldsymbol{y}) = \begin{pmatrix} a_1(\boldsymbol{y})^\top \\ \vdots \\ a_n(\boldsymbol{y})^\top \end{pmatrix}. \tag{125}$$

Then we have

$$a_1(\boldsymbol{y})^\top \widetilde{\boldsymbol{\varphi}}(\boldsymbol{y}) + \|\widetilde{\boldsymbol{\varphi}}(\boldsymbol{y})\|^2 - 1 = 0, \tag{126}$$

and for each $i \in \{2, \ldots, n\}$,

$$a_i(\boldsymbol{y})^\top \widetilde{\boldsymbol{\varphi}}(\boldsymbol{y}) = 0. \tag{127}$$

In addition, as $H(\boldsymbol{y})$ is symmetric, there exists an orthonormal matrix P, such that

$$H(\boldsymbol{y}) = P^{-1}\text{diag}(\lambda_1(\boldsymbol{y}), \cdots, \lambda_m(\boldsymbol{y}), 0, \cdots, 0)P, \tag{128}$$

and thus $PA(\boldsymbol{y})P^{-1}$ is actually the following matrix

$$\lambda_1(\boldsymbol{y})I - \text{diag}(\lambda_1(\boldsymbol{y}), \cdots, \lambda_m(\boldsymbol{y}), 0, \cdots, 0). \tag{129}$$

As a result,

$$\text{rank }(A(\boldsymbol{y})) = n - 1. \tag{130}$$

Moreover, as $a_2(\boldsymbol{x}), \cdots, a_n(\boldsymbol{x})$ are independent, by the continuity of A, there exists $\widetilde{\varepsilon} < \overline{\varepsilon}$ such that for all $\boldsymbol{y} \in B_n(\boldsymbol{x}, \widetilde{\varepsilon})$, $a_2(\boldsymbol{y}), \cdots, a_n(\boldsymbol{y})$ are independent as well. Therefore, by equation 130, for a given $\boldsymbol{y} \in B_n(\boldsymbol{x}, \widetilde{\varepsilon})$, there exists $\theta_1, \cdots, \theta_n$ with $\theta_1 \neq 0$ such that

$$\sum_{i=1}^{n} \theta_i a_i(\boldsymbol{y}) = 0. \tag{131}$$

Combining this result with equation 126 and equation 127, we have

$$\|\widetilde{\boldsymbol{\varphi}}(\boldsymbol{y})\|^2 - 1 = 0. \tag{132}$$

Then, by equation 126, $\boldsymbol{a}_1(\boldsymbol{y})^\top \widetilde{\boldsymbol{\varphi}}(\boldsymbol{y}) = 0$ and $\|\widetilde{\boldsymbol{\varphi}}(\boldsymbol{y})\| = 1$. It implies that, by equation 124, $A(\boldsymbol{y})\widetilde{\boldsymbol{\varphi}}(\boldsymbol{y}) = 0$. Thus, the unit vector $\widetilde{\boldsymbol{\varphi}}(\boldsymbol{y})$ satisfies

$$H(\boldsymbol{y})\widetilde{\boldsymbol{\varphi}}(\boldsymbol{y}) = \lambda_1(\boldsymbol{y})\widetilde{\boldsymbol{\varphi}}(\boldsymbol{y}). \tag{133}$$

Consequently, the unit vector $\widetilde{\varphi}(\boldsymbol{y})$ is the eigenvector corresponding to $\lambda_1(\boldsymbol{y})$, and thus by taking $\boldsymbol{\varphi}_1(\boldsymbol{y}) = \widetilde{\boldsymbol{\varphi}}(\boldsymbol{y})$ for every $\boldsymbol{y} \in B_n(\boldsymbol{x}, \widetilde{\varepsilon})$, we then prove that $\boldsymbol{\varphi}_1$ is smooth at $\boldsymbol{x}$. Therefore, $\boldsymbol{\varphi}_1$ is smooth on M. Similarly, the smoothness for $\boldsymbol{\varphi}_i$ on M can be proved simultaneously.

Next, for each $\boldsymbol{u} \in U$, since 0 is an eigenvalue of $H(\psi(\boldsymbol{u}))$ with multiplicity $n - m$, we now prove that its eigenvectors are exactly the column of $D\psi(\boldsymbol{u})$. Since $f(\psi(\boldsymbol{u}))$ is a constant for $\boldsymbol{u} \in U$, then its Hessian matrix is zero, i.e.,

$$D\psi(\boldsymbol{u})^\top H(\psi(\boldsymbol{u}))D\psi(\boldsymbol{u}) = 0. \tag{134}$$

As $H(\psi(\boldsymbol{u}))$ is positive semidefinite, by Cholesky factorization theorem (Theorem 10.9 Higham (2002)), there exists a semidefinite matrix P such that

$$H(\psi(\boldsymbol{u})) = P^\top P. \tag{135}$$

As a result, by equation 134, $PD\psi(\boldsymbol{u}) = 0$, and thus

$$H(\psi(\boldsymbol{u}))D\psi(\boldsymbol{u}) = 0. \tag{136}$$

Therefore, denote $\boldsymbol{\gamma}_i(\boldsymbol{u})$ as the i-th column of $D\psi(\boldsymbol{u})$,

$$H(\psi(\boldsymbol{u}))\boldsymbol{\gamma}_i(\boldsymbol{u}) = \mathbf{0}. \tag{137}$$

On the other hand, as ψ is a diffeomorphism, $D\psi(\boldsymbol{u})$ is full rank. Therefore, by equation 137, $\boldsymbol{\gamma}_1(\boldsymbol{u}), \cdots, \boldsymbol{\gamma}_{n-m}(\boldsymbol{u})$ are independent eigenvectors of 0, and are orthogonal to each $\boldsymbol{\varphi}_i(\psi(\boldsymbol{u}))$. To build up better understanding, we summarize this part into the following technical lemma:

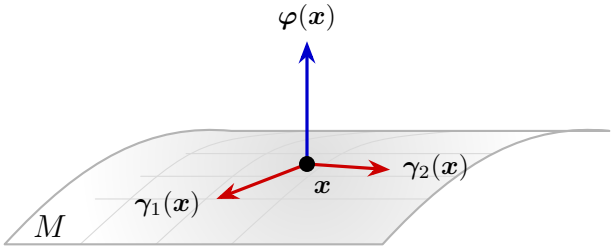


Figure 8: Local coordinate system visualization with $n = 3$ and $m = 1$. At a minimum $\boldsymbol{x} \in M$, we show the tangent vectors $\boldsymbol{\gamma}_1(\boldsymbol{x})$ and $\boldsymbol{\gamma}_2(\boldsymbol{x})$, and the normal vector $\boldsymbol{\varphi}(\boldsymbol{x})$. The coordinate transformation G decouples the GD dynamics into motion along these directions.

Lemma 8.3. *Let M be the regular surface of minima, and $H(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x})$ be the Hessian matrix. The eigenvalues and their corresponding unit eigenvectors are established to be smooth. Specifically, by the implicit function theorem, for each $1 \le i \le m$, the unit eigenvector $\boldsymbol{\varphi}_i(\boldsymbol{x})$ associated with the strictly positive eigenvalue $\lambda_i(\boldsymbol{x})$ depends smoothly on $\boldsymbol{x} \in M$. Furthermore, for $\boldsymbol{u} \in U$, the zero eigenvalue of $H(\psi(\boldsymbol{u}))$ has a multiplicity of $n - m$, and its corresponding independent eigenvectors are exactly the columns of $D\psi(\boldsymbol{u})$, denoted as $\boldsymbol{\gamma}_j(\boldsymbol{u})$. It is verified that*

$$H(\psi(\boldsymbol{u}))\boldsymbol{\gamma}_j(\boldsymbol{u}) = \mathbf{0}. \tag{138}$$

Now, similarly to the two-dimensional example introduced in the previous section, we define the following change of coordinates:

$$G : U \times \mathbb{R}^m \mapsto \mathbb{R}^n, \tag{139}$$

with

$$G(\boldsymbol{u}, \boldsymbol{v}) = \psi(\boldsymbol{u}) + \sum_{i=1}^{m} v_i \boldsymbol{\varphi}_i(\psi(\boldsymbol{u})). \tag{140}$$

Specifically,

$$G(\mathbf{0}, \mathbf{0}) = \mathbf{0}. \tag{141}$$

To build the intuition of the analysis, the constructed local coordinate system adapted to the the minimum set is shown in Figure 8. We know that

$$DG = \left(D\psi + \sum_{i=1}^{m} v_i D(\boldsymbol{\varphi}_i \circ \psi), \boldsymbol{\varphi}_1, \cdots, \boldsymbol{\varphi}_m \right). \tag{142}$$

Then, at $(\mathbf{0}, \mathbf{0})$, since

$$DG = (D\psi, \boldsymbol{\varphi}_1, \cdots, \boldsymbol{\varphi}_m) \tag{143}$$

has rank n, $\det DG(\mathbf{0}, \mathbf{0}) \neq 0$, and G is a diffeomorphism on a neighborhood of $(\mathbf{0}, \mathbf{0})$. For clarity, without loss of generality, in the following context, we still use the open set U defined in equation 98, i.e., we assume that

$$G : U \times V \mapsto G(U \times V) \tag{144}$$

is a diffeomorphism, where

$$V = B_m(\mathbf{0}, \varepsilon_0) \tag{145}$$

is a m-dimensional open ball centered at $\mathbf{0}$ with radius ε_0. Moreover, let C_0 be the constant such that

$$\|DG\|, \|DG^{-1}\| \le C_0 \tag{146}$$

hold on $U \times V$ and $G(U \times V)$ respectively.

In the following sections, we prove the two theorems in Section 8.1.

8.3 EIGENVALUES AND LEARNING RATE

Having established the local coordinate system, we now turn to characterizing the contractive behavior of the GD update. The objective of this section is to derive precise estimates for the decay rates of the components normal to the minimum set M. By bounding the eigenvalues of the Hessian and controlling the higher-order terms in the Taylor expansion, we establish the inequalities necessary to prove the forward-invariance property.

Firstly, by equation 102, there exists a constant $C_1 < 1$ such that for all $\boldsymbol{u} \in U$,

$$\max_{1\le i\le m} |1-\eta\lambda_i(\psi(\boldsymbol{u}))| < C_1. \tag{147}$$

Next, define

$$a(\boldsymbol{u}) = \max_{1\le i\le m} |1-\eta\lambda_i(\psi(\boldsymbol{u}))| \tag{148}$$

and

$$b(\boldsymbol{u}) = \begin{cases} \max\limits_{2\le i\le m} & |1-\eta\lambda_i(\psi(\boldsymbol{u}))|, \text{ if } \eta < \frac{2}{\lambda_1(\psi(\mathbf{0}))+\lambda_m(\psi(\mathbf{0}))} \\ \max\limits_{1\le i\le m-1} & |1-\eta\lambda_i(\psi(\boldsymbol{u}))|, \text{ if } \frac{2}{\lambda_1(\psi(\mathbf{0}))+\lambda_m(\psi(\mathbf{0}))} < \eta < \frac{2}{\lambda_m(\psi(\mathbf{0}))}. \end{cases} \tag{149}$$

Moreover, we define

$$a_+ = \sup_{\boldsymbol{u}\in U} a(\boldsymbol{u}), \quad a_- = \inf_{\boldsymbol{u}\in U} a(\boldsymbol{u}), \quad \text{and} \quad b = \sup_{\boldsymbol{u}\in U} b(\boldsymbol{u}). \tag{150}$$

As a remark, we need to split into two cases to continue our formulation. If

$$\eta < \frac{2}{\lambda_1(\psi(\mathbf{0})) + \lambda_m(\psi(\mathbf{0}))}, \tag{151}$$

there exists $\varepsilon'' > 0$, such that for all $\boldsymbol{u} \in B_{n-m}(\mathbf{0}, \varepsilon'')$, it holds that

$$\eta < \frac{2}{\lambda_1(\psi(\boldsymbol{u})) + \lambda_m(\psi(\boldsymbol{u}))}. \tag{152}$$

Without loss of generality, we assume that equation 152 holds for all $\boldsymbol{u} \in U$. In this case, we have

$$a(\mathbf{0}) = 1 - \eta\lambda_1(\mathbf{0}) \tag{153}$$

and

$$b(\mathbf{0}) = \max\left\{|1-\eta\lambda_2(\mathbf{0})|, |1-\eta\lambda_m(\mathbf{0})|\right\}. \tag{154}$$

Clearly, we have

$$1 > C_1 > a(\mathbf{0}) > b(\mathbf{0}). \tag{155}$$

In this way, by the continuity of $a(\boldsymbol{u})$ and $b(\boldsymbol{u})$, there exists an open set $U^* \subset U$ with $\mathbf{0} \in U^*$, such that

$$\inf_{\boldsymbol{u}\in U^*} a(\boldsymbol{u}) > \sup_{\boldsymbol{u}\in U^*} b(\boldsymbol{u}). \tag{156}$$

and

$$\inf_{\boldsymbol{u}\in U^*} a(\boldsymbol{u}) > \left(\sup_{\boldsymbol{u}\in U^*} a(\boldsymbol{u})\right)^2. \tag{157}$$

Without loss of generality, we also take $U^* = U$ for clarity. In this way, the above defined constants satisfy

$$\max\{b, a_+^2\} < a_- \le a_+ < 1. \tag{158}$$

Similarly, if

$$\frac{2}{\lambda_1(\psi(\mathbf{0})) + \lambda_m(\psi(\mathbf{0}))} < \eta < \frac{2}{\lambda_m(\psi(\mathbf{0}))}, \tag{159}$$

there exists $\varepsilon''' > 0$, such that for all $\boldsymbol{u} \in B_{n-m}(\mathbf{0}, \varepsilon''')$, it holds that

$$\frac{2}{\lambda_1(\psi(\boldsymbol{u})) + \lambda_m(\psi(\boldsymbol{u}))} < \eta < \frac{2}{\lambda_m(\psi(\boldsymbol{u}))}. \tag{160}$$

Therefore, without loss of generality, we assume that equation 160 holds for all $\boldsymbol{u} \in U$. In this case, we have

$$a(\mathbf{0}) = \eta\lambda_m(\mathbf{0}) - 1 \tag{161}$$

and

$$b(\mathbf{0}) = \max\left\{|1 - \eta\lambda_1(\mathbf{0})|, |1 - \eta\lambda_{m-1}(\mathbf{0})|\right\}. \tag{162}$$

Clearly, we also have equation 155-equation 158. Therefore, in both cases, equation 158 holds. For better understanding, we summarize this part into the following lemma:

Lemma 8.4. *Let $\lambda = \sup_{\boldsymbol{x}\in G(U\times V)} \lambda_m(\boldsymbol{x})$ and assume the learning rate is bounded by $\eta < \frac{2}{\lambda}$. For all tangential coordinate $\boldsymbol{u} \in U$,*

$$a(\boldsymbol{u}) = \max_{1\le i\le m} |1 - \eta\lambda_i(\psi(\boldsymbol{u}))|. \tag{163}$$

Furthermore, $b(\boldsymbol{u})$ is determined conditionally based on the chosen learning rate regime:

$$b(\boldsymbol{u}) = \begin{cases} \max\limits_{2\le i\le m} |1 - \eta\lambda_i(\psi(\boldsymbol{u}))|, & \text{if } \eta < \frac{2}{\lambda_1(\psi(\mathbf{0}))+\lambda_m(\psi(\mathbf{0}))}, \\ \max\limits_{1\le i\le m-1} |1 - \eta\lambda_i(\psi(\boldsymbol{u}))|, & \text{if } \frac{2}{\lambda_1(\psi(\mathbf{0}))+\lambda_m(\psi(\mathbf{0}))} < \eta < \frac{2}{\lambda_m(\psi(\mathbf{0}))}. \end{cases} \tag{164}$$

Then for all $\boldsymbol{u} \in U$, $a_+ = \sup_{\boldsymbol{u}\in U} a(\boldsymbol{u})$, $a_- = \inf_{\boldsymbol{u}\in U} a(\boldsymbol{u})$, and $b = \sup_{\boldsymbol{u}\in U} b(\boldsymbol{u})$, it is established that the following inequalities hold:

$$\max\left\{b, a_+^2\right\} < a_- \le a_+ < 1. \tag{165}$$

8.4 Important Estimations

With the definitions above, we proceed to derive the key inequalities required for our main theorems. This section presents important estimations that govern the local behavior of the objective function. Our primary objective is to bound the higher-order terms appearing in the coordinate transformation G and the GD mapping F. These estimations will allow us to decouple the dynamics of the error terms, a crucial step in proving the existence of the convergence direction.

For a given $\boldsymbol{x} \in G(U \times V)$, take

$$(\boldsymbol{u}, \boldsymbol{v}) = G^{-1}(\boldsymbol{x}) \in U \times V, \tag{166}$$

by which $\nabla f(\boldsymbol{x})$ can be considered as a function of $\boldsymbol{v}$. Applying Taylor expansion at $\psi(\boldsymbol{u})$, by equation 140, as $\nabla f(\psi(\boldsymbol{u})) = \mathbf{0}$, we have

$$\begin{aligned} \nabla f(\boldsymbol{x}) =& \nabla f(\boldsymbol{x}) - \nabla f(\psi(\boldsymbol{u})) \\ =& H(\psi(\boldsymbol{u}))(\boldsymbol{x} - \psi(\boldsymbol{u})) + O(\|\boldsymbol{x} - \psi(\boldsymbol{u})\|^2) \\ =& \sum_{i=1}^{m} v_i H(\psi(\boldsymbol{u}))\boldsymbol{\varphi}_i(\psi(\boldsymbol{u})) + O(\|\boldsymbol{x} - \psi(\boldsymbol{u})\|^2) \\ =& \sum_{i=1}^{m} v_i \lambda_i(\psi(\boldsymbol{u}))\boldsymbol{\varphi}_i(\psi(\boldsymbol{u})) + O(\|\boldsymbol{x} - \psi(\boldsymbol{u})\|^2), \end{aligned} \tag{167}$$

where the remainder term $O(\|\boldsymbol{x} - \psi(\boldsymbol{u})\|^2)$ is the following integral

$$\int_0^1 (1-t)\nabla^3 f((1-t)\psi(\boldsymbol{u}) + t\boldsymbol{x})[\boldsymbol{x} - \psi(\boldsymbol{u}), \boldsymbol{x} - \psi(\boldsymbol{u})]\mathrm{d}t. \tag{168}$$

That is to say, for each $\boldsymbol{x} \in G(U \times V)$,

$$\left\| \nabla f(\boldsymbol{x}) - \sum_{i=1}^{m} v_i \lambda_i(\psi(\boldsymbol{u}))\boldsymbol{\varphi}_i(\psi(\boldsymbol{u})) \right\| \le O(\|\boldsymbol{v}\|^2). \tag{169}$$

Therefore, there exists a constant $C_2 > 0$ such that for each $\boldsymbol{x} = G(\boldsymbol{u}, \boldsymbol{v}) \in G(U \times V)$,

$$\|\nabla f(\boldsymbol{x})\| \le \lambda\|\boldsymbol{v}\| + C_2\|\boldsymbol{v}\|^2, \tag{170}$$

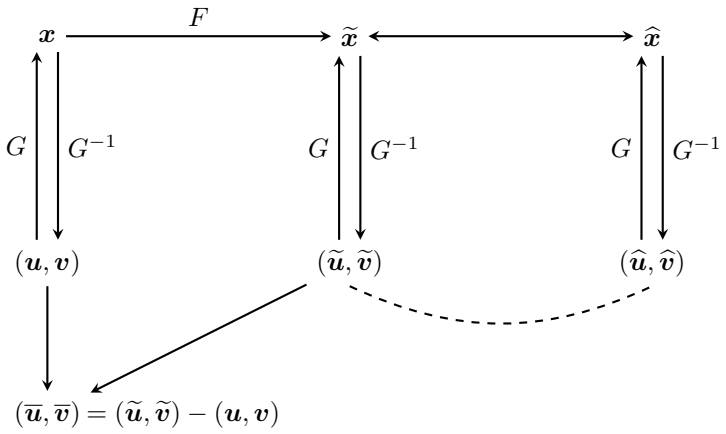


Figure 9: Relations between mappings and coordinates.

where λ is defined in equation 107. Next, for simplicity, we take

$$\widetilde{\boldsymbol{x}} = F(\boldsymbol{x}) = \boldsymbol{x} - \eta \nabla f(\boldsymbol{x}). \tag{171}$$

By equation 140 and equation 167, we have

$$\widetilde{\boldsymbol{x}} = \psi(\boldsymbol{u}) + \sum_{i=1}^{m} (1 - \eta \lambda_i(\psi(\boldsymbol{u}))) v_i \boldsymbol{\varphi}_i(\psi(\boldsymbol{u})) - \eta O(\|\boldsymbol{x} - \psi(\boldsymbol{u})\|^2), \tag{172}$$

where the remainder term $O(\|\boldsymbol{x} - \psi(\boldsymbol{u})\|^2)$ is the integral defined in equation 168. In this way, by taking

$$\widehat{\boldsymbol{v}} = \Lambda(\boldsymbol{u}) \boldsymbol{v}, \tag{173}$$

with a $m \times m$ matrix $\Lambda(\boldsymbol{u}) = \mathrm{diag}((1 - \eta \lambda_1(\psi(\boldsymbol{u}))), \cdots, (1 - \eta \lambda_m(\psi(\boldsymbol{u}))))$, and

$$\widehat{\boldsymbol{x}} = G(\boldsymbol{u}, \widehat{\boldsymbol{v}}), \tag{174}$$

we have

$$\|\widetilde{\boldsymbol{x}} - \widehat{\boldsymbol{x}}\| \leq \eta C_2 \|\boldsymbol{v}\|^2. \tag{175}$$

For better understanding the quantities defined above, one is able to refer Figure 9. Moreover, unless otherwise specified, the notations $\boldsymbol{u}, \boldsymbol{v}, \widehat{\boldsymbol{v}}, \widetilde{\boldsymbol{x}}, \widehat{\boldsymbol{x}}$ are used in the following context. Moreover, we also take $\widehat{\boldsymbol{x}} = G(\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{v}})$ with $\widehat{\boldsymbol{u}} = \boldsymbol{u}$ in equation 174 for convenience.

At the end of this section, we summarize the important estimates in the following lemma.

Lemma 8.5. *For a given $\boldsymbol{x} \in G(U \times V)$, let*

$$(\boldsymbol{u}, \boldsymbol{v}) = G^{-1}(\boldsymbol{x}) \in U \times V, \tag{176}$$

then there exists a constant $C_2 > 0$ such that for all $\boldsymbol{x} \in G(U \times V)$,

$$\|\nabla f(\boldsymbol{x})\| \leq \lambda \|\boldsymbol{v}\| + C_2 \|\boldsymbol{v}\|^2, \tag{177}$$

where λ is defined in equation 107, and

$$\|\widetilde{\boldsymbol{x}} - \widehat{\boldsymbol{x}}\| \leq \eta C_2 \|\boldsymbol{v}\|^2, \tag{178}$$

where $\widetilde{\boldsymbol{x}}$ and $\widehat{\boldsymbol{x}}$ are defined in equation 171 and equation 174, respectively.

8.5 Existence of Forward-invariance Set

In this section, we introduce the following lemma, which supports the existence of a forward-invariant set. We also present the proof of Theorem 8.1 in this section.

Lemma 8.6. *There exists a constant $\varepsilon_1 > 0$, such that $G(B_n(\mathbf{0}, \varepsilon_1))$ is a forward-invariance set. i.e., for all $\boldsymbol{x} \in G(B_n(\mathbf{0}, \varepsilon_1))$, $F(\boldsymbol{x}) \in G(B_n(\mathbf{0}, \varepsilon_1))$.*

Proof. First, take an open set $B_n(\mathbf{0}, \delta) \subset G(U \times V)$ with a constant $\widetilde{\varepsilon} > 0$ satisfying

$$\widetilde{\varepsilon} < \min\left\{\delta/(3\eta\lambda), \sqrt{\delta/(3\eta C_2)}, \varepsilon_0\right\}, \tag{179}$$

where λ, ε_0 and C_2 are defined in equation 107, equation 145 and equation 169 respectively. Then for each $\boldsymbol{x} \in \widetilde{\Omega}$, where

$$\widetilde{\Omega} = G(U \times B_m(\mathbf{0}, \widetilde{\varepsilon})) \cap B_n(\mathbf{0}, \delta/3), \tag{180}$$

by equation 170 and equation 171, we have

$$\begin{aligned}
\|\widetilde{\boldsymbol{x}}\| \leq & \|\boldsymbol{x}\| + \eta\|\nabla f(\boldsymbol{x})\| \\
\leq & \delta/3 + \eta\lambda\|\boldsymbol{v}\| + \eta C_2\|\boldsymbol{v}\|^2 \\
\leq & \delta/3 + \eta\lambda\widetilde{\varepsilon} + \eta C_2\widetilde{\varepsilon}^2 \\
\leq & \delta/3 + \delta/3 + \delta/3 \\
= & \delta.
\end{aligned} \tag{181}$$

Then we know that, for the open set $\widetilde{\Omega}$ containing $\mathbf{0}$, $\boldsymbol{x} \in \widetilde{\Omega}$ implies that $F(\boldsymbol{x}) \in B_n(\mathbf{0}, \delta) \subset G(U \times V)$. Now, taking a constant $\varepsilon_1 > 0$ such that

$$G(B_n(\mathbf{0}, \varepsilon_1)) \subset \widetilde{\Omega}, \tag{182}$$

and

$$\varepsilon_1 < \min\left\{1, \frac{1 - a_+^2}{4K + 2K^2}\right\}, \tag{183}$$

where a_+ is defined in equation 150, and K is a constant defined by

$$K = \eta C_0 C_2. \tag{184}$$

Then, for each $\boldsymbol{x} \in G(B_n(\mathbf{0}, \varepsilon_1))$, we know that $(\boldsymbol{u}, \boldsymbol{v}) \in B_n(\mathbf{0}, \varepsilon_1)$ and $\widetilde{\boldsymbol{x}} = F(\boldsymbol{x}) \in U \times V$. Here, $G(B_n(\mathbf{0}, \varepsilon_1))$ is the forward-invariant set we need. i.e. We are going to prove that for all $\boldsymbol{x} \in G(B_n(\mathbf{0}, \varepsilon_1))$, $F(x) \in G(B_n(\mathbf{0}, \varepsilon_1))$. Next, consider

$$(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{v}}) = G^{-1}(\widetilde{\boldsymbol{x}}) \tag{185}$$

and

$$(\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{v}}) = G^{-1}(\widehat{\boldsymbol{x}}), \tag{186}$$

by equation 146, equation 175 and the mean value theorem, it yields that

$$\|(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{v}}) - (\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{v}})\| \leq C_0\|\widetilde{\boldsymbol{x}} - \widehat{\boldsymbol{x}}\| \leq K\|\boldsymbol{v}\|^2. \tag{187}$$

Thus, as $\widehat{\boldsymbol{u}} = \boldsymbol{u}$, we have

$$\|\widetilde{\boldsymbol{u}} - \boldsymbol{u}\| \leq K\|\boldsymbol{v}\|^2 \tag{188}$$

and

$$\|\widetilde{\boldsymbol{v}} - \widehat{\boldsymbol{v}}\| \leq K\|\boldsymbol{v}\|^2. \tag{189}$$

Meanwhile, by equation 173 and equation 150, we have

$$\|\widehat{\boldsymbol{v}}\| \leq a_+\|\boldsymbol{v}\|. \tag{190}$$

Then, by equation 188, as $(\boldsymbol{u}, \boldsymbol{v}) \in B_n(\mathbf{0}, \varepsilon_1)$,

$$\begin{aligned}
\|\widetilde{\boldsymbol{u}}\|^2 \leq & (\|\boldsymbol{u}\| + \|\widetilde{\boldsymbol{u}} - \boldsymbol{u}\|)^2 \\
= & \|\boldsymbol{u}\|^2 + 2\|\boldsymbol{u}\|\|\widetilde{\boldsymbol{u}} - \boldsymbol{u}\| + \|\widetilde{\boldsymbol{u}} - \boldsymbol{u}\|^2 \\
\leq & \|\boldsymbol{u}\|^2 + 2K\|\boldsymbol{u}\|\|\boldsymbol{v}\|^2 + K^2\|\boldsymbol{v}\|^4 \\
\leq & \|\boldsymbol{u}\|^2 + 2K\varepsilon_1\|\boldsymbol{v}\|^2 + K^2\varepsilon_1^2\|\boldsymbol{v}\|^2 \\
\leq & \|\boldsymbol{u}\|^2 + (2K\varepsilon_1 + K^2\varepsilon_1)\|\boldsymbol{v}\|^2.
\end{aligned} \tag{191}$$

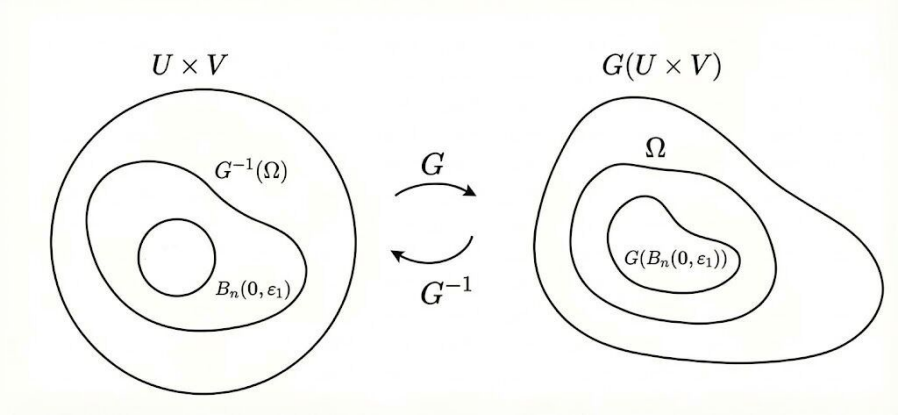


Figure 10: Relationship between important sets in Lemma 8.6

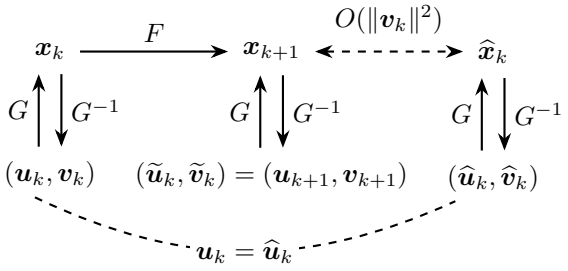


Figure 11: The relationships in GD trajectory.

At the same time, by equation 189 and equation 190,

$$\begin{aligned}
\|\widetilde{\boldsymbol{v}}\|^2 \leq & (\|\widehat{\boldsymbol{v}}\| + \|\widetilde{\boldsymbol{v}} - \widehat{\boldsymbol{v}}\|)^2 \\
= & \|\widehat{\boldsymbol{v}}\|^2 + 2\|\widehat{\boldsymbol{v}}\|\|\widetilde{\boldsymbol{v}} - \widehat{\boldsymbol{v}}\| + \|\widetilde{\boldsymbol{v}} - \widehat{\boldsymbol{v}}\|^2 \\
\leq & a_+^2\|\boldsymbol{v}\|^2 + 2a_+\|\boldsymbol{v}\| \cdot K\|\boldsymbol{v}\|^2 + K\|\boldsymbol{v}\|^2 \cdot K\|\boldsymbol{v}\|^2 \\
\leq & a_+^2\|\boldsymbol{v}\|^2 + 2a_+\varepsilon_1 K\|\boldsymbol{v}\|^2 + K^2\varepsilon_1^2\|\boldsymbol{v}\|^2 \\
\leq & (a_+^2 + 2K\varepsilon_1 + K^2\varepsilon_1)\|\boldsymbol{v}\|^2.
\end{aligned} \tag{192}$$

As a result of equation 191 and equation 192, and by the definition of ε_1 in equation 183,

$$\|\widetilde{\boldsymbol{u}}\|^2 + \|\widetilde{\boldsymbol{v}}\|^2 \leq \|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2. \tag{193}$$

Therefore, $(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{v}}) \in B_n(\mathbf{0}, \varepsilon_1)$ as well, and then $\widetilde{\boldsymbol{x}} = G(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{v}}) \in G\left(B_n(\mathbf{0}, \varepsilon_1)\right)$. This proves the lemma by taking $\Omega = G(B_n(\mathbf{0}, \varepsilon_1))$. □

Here we stress that, for all $\boldsymbol{x} \in G(B_n(\mathbf{0}, \varepsilon_1))$, equation 188, equation 189, equation 190 and equation 193 hold and are frequently applied in the subsequent proof. For a better understanding the relations between the sets and points in Lemma 8.6, one can refer Figure 9. Again, for convenience, in the following context we use $(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{v}})$ defined in equation 185 unless otherwise specified.

8.6 Proof of Theorem 8.1

In this section, we prove the forward-invariance property for GD. Having constructed the locally invariant neighborhood $G(B_n(\mathbf{0}, \varepsilon_1))$ in Lemma 8.6, we now demonstrate the forward-invariance set of GD dynamics by the previous established coordinate G. Specifically, we show that the normal component of the trajectory contracts geometrically to zero, while the tangential component forms a Cauchy sequence, thereby guaranteeing convergence to a unique point on the minimum set.

By Lemma 8.6, we know that $\boldsymbol{x}_{k+1} \in G(B_n(\mathbf{0}, \varepsilon_1))$ if $\boldsymbol{x}_k \in G(B_n(\mathbf{0}, \varepsilon_1))$. Then, $G(B_n(\mathbf{0}, \varepsilon_1))$ is a forward-invariance set for GD trajectory. For convenience, we redefine and list all important quantities and estimates above. Frist, recall that the GD trajectory is defined inductively as, for $k \geq 0$,

$$\boldsymbol{x}_{k+1} = F(\boldsymbol{x}_k). \tag{194}$$

Then, define

$$(\boldsymbol{u}_k, \boldsymbol{v}_k) = G^{-1}(\boldsymbol{x}_k) \tag{195}$$

and

$$(\widetilde{\boldsymbol{u}}_k, \widetilde{\boldsymbol{v}}_k) = (\boldsymbol{u}_{k+1}, \boldsymbol{v}_{k+1}). \tag{196}$$

Moreover, we take $\widehat{\boldsymbol{u}}_k = \boldsymbol{u}_k$ and define

$$\widehat{\boldsymbol{v}}_k = \Lambda(\boldsymbol{u}_k)\boldsymbol{v}_k, \tag{197}$$

then we take

$$\widehat{\boldsymbol{x}}_k = G(\widehat{\boldsymbol{u}}_k, \widehat{\boldsymbol{v}}_k). \tag{198}$$

For clarity, see Figure 11 for more details about the relation between the notations defined above. With these notations, by equation 188, for all $k \geq 0$, we have

$$\|\boldsymbol{u}_{k+1} - \boldsymbol{u}_k\| \leq K\|\boldsymbol{v}_k\|^2. \tag{199}$$

By equation 189 and equation 190, for all $k \geq 0$, we have

$$\|\boldsymbol{v}_{k+1} - \widehat{\boldsymbol{v}}_k\| \leq K\|\boldsymbol{v}_k\|^2. \tag{200}$$

Here K is defined in equation 184. Moreover, by equation 193, we have

$$\|\boldsymbol{u}_{k+1}\|^2 + \|\boldsymbol{v}_{k+1}\|^2 \leq \|\boldsymbol{u}_k\|^2 + \|\boldsymbol{v}_k\|^2. \tag{201}$$

These estimations on $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$ are also necessary for the later on proof. Next, we start to prove the convergence property of Theorem 8.1. As $(\boldsymbol{u}_k, \boldsymbol{v}_k) \in B_n(\mathbf{0}, \varepsilon_1)$ and by equation 190

$$\|\boldsymbol{v}_{k+1}\| \leq \|\widehat{\boldsymbol{v}}_k\| + K\|\boldsymbol{v}_k\|^2 \leq (a_+ + K\varepsilon_1)\,\|\boldsymbol{v}_k\|. \tag{202}$$

Moreover, by the definition of ε_1 in equation 183,

$$\varepsilon_1 < \frac{1 - a_+^2}{4K} = \frac{(1 - a_+)}{K}\frac{1 + a_+}{4} < \frac{1 - a_+}{K}. \tag{203}$$

It means that

$$a_+ + K\varepsilon_1 < 1. \tag{204}$$

As a result, $\{\boldsymbol{v}_k\}_{k\geq 0}$ is contractive and

$$\lim_{k\to\infty} \boldsymbol{v}_k = \boldsymbol{v}^*, \tag{205}$$

and clearly,

$$\boldsymbol{v}^* = \mathbf{0}. \tag{206}$$

Hence, for all $p \geq 0$, by equation 199 and equation 202,

$$\begin{aligned} \|\boldsymbol{u}_{k+p} - \boldsymbol{u}_k\| &\leq \sum_{i=0}^{p-1} \|\boldsymbol{u}_{k+i+1} - \boldsymbol{u}_{k+i}\| \leq K\sum_{i=0}^{p-1} \|\boldsymbol{v}_{k+i}\|^2 \\ &\leq K\sum_{i=0}^{p-1} (a_+ + K\varepsilon_1)^{2i}\,\|\boldsymbol{v}_k\|^2 \leq \frac{K}{1 - (a_+ + K\varepsilon_1)^2}\,\|\boldsymbol{v}_k\|^2. \end{aligned} \tag{207}$$

Therefore, $\{\boldsymbol{u}_k\}_{k\geq 0}$ is a Cauchy sequence, thus $\{\boldsymbol{u}_k\}_{k\geq 0}$ is convergent and we take

$$\lim_{k\to\infty} \boldsymbol{u}_k = \boldsymbol{u}^*. \tag{208}$$

By equation 201, we know that

$$(\boldsymbol{u}^*, \boldsymbol{v}^*) \in B_n(\mathbf{0}, \varepsilon_1). \tag{209}$$

Finally,

$$\lim_{k\to\infty} \boldsymbol{x}_k = \lim_{k\to\infty} G(\boldsymbol{u}_k, \boldsymbol{v}_k) = G(\boldsymbol{u}^*, \boldsymbol{v}^*) = \psi(\boldsymbol{u}^*). \tag{210}$$

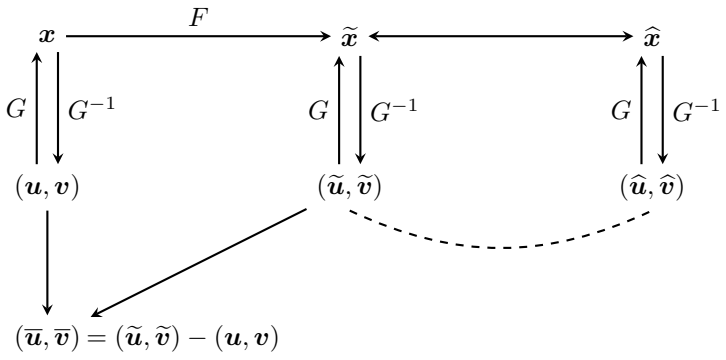


Figure 12: Relations between mappings and coordinates.

Here the convergence property is proved. Lastly, apply Taylor expansion, it follows that

$$\begin{aligned} & f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k) \\ =& -\eta \|\nabla f(\boldsymbol{x}_k)\|^2 + \eta^2 \int_0^1 (1-t)\, \nabla f(\boldsymbol{x}_k)^\top \nabla^2 f\left(\boldsymbol{x}_k - t\eta \nabla f(\boldsymbol{x}_k)\right) \nabla f(\boldsymbol{x}_k)\, \mathrm{d}t \\ \leq& -\eta \|\nabla f(\boldsymbol{x}_k)\|^2 + \eta^2 \left(\int_0^1 (1-t)\, \lambda_m \left(\boldsymbol{x}_k - t\eta \nabla f(\boldsymbol{x}_k)\right)\, \mathrm{d}t \right) \|\nabla f(\boldsymbol{x}_k)\|^2 \\ \leq& \left(-\eta + \frac{\eta^2}{2} \lambda \right) \|\nabla f(\boldsymbol{x}_k)\|^2. \end{aligned} \tag{211}$$

Here we complete the proof of theorem 8.1 by selecting a constant $C = \eta - \frac{\eta^2}{2}\lambda > 0$.

8.7 Proof of Theorem 8.2, preparation

We now turn our attention to the second main result, Theorem 8.2, which characterizes the asymptotic behavior of the GD trajectory. To establish the existence of a convergence direction or an alternating convergence direction, we must refine our analysis to account for the interplay between the linear contraction rates and the nonlinear perturbations inherent in the coordinate transformation.

Next, consider

$$(\overline{\boldsymbol{u}}, \overline{\boldsymbol{v}}) = (\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{v}}) - (\boldsymbol{u}, \widehat{\boldsymbol{v}}) \tag{212}$$

as functions of $(\boldsymbol{u}, \boldsymbol{v})$ on $B_n(\mathbf{0}, \varepsilon_1)$. More specifically, $(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{v}}) = G^{-1} \circ F \circ G(\boldsymbol{u}, \boldsymbol{v}) - (\boldsymbol{u}, \Lambda(\boldsymbol{u})\boldsymbol{v})$, where $\Lambda(\boldsymbol{u})$ is defined in equation 173.

When GD takes a step, we measure exactly where it lands relative to the valley floor. The proof introduces two different ways to evaluate this next step. $(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{v}})$ represents the exact next position the algorithm takes in this coordinate space. $(\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{v}}) = (\boldsymbol{u}, \widehat{\boldsymbol{v}})$ represents the approximate next position. We first consider the error between these two quantities. One can refer figure 12 for clarity. Here we begin with one important lemma including the estimation of $(\overline{\boldsymbol{u}}, \overline{\boldsymbol{v}})$.

Lemma 8.7. *Consider $(\overline{\boldsymbol{u}}, \overline{\boldsymbol{v}})$ defined in equation 212 as functions of $(\boldsymbol{u}, \boldsymbol{v})$, then there exists a constant $C_3 > 0$ such that for all $(\boldsymbol{u}, \boldsymbol{v}) \in B_n(\mathbf{0}, \varepsilon_1)$, the following inequalities hold:*

$$\|(\overline{\boldsymbol{u}}, \overline{\boldsymbol{v}})\| \leq C_3 \|\boldsymbol{v}\|^2, \tag{213}$$

$$\left\| \frac{\partial (\overline{\boldsymbol{u}}, \overline{\boldsymbol{v}})}{\partial \boldsymbol{u}} \right\| \leq C_3 \|\boldsymbol{v}\|^2, \tag{214}$$

$$\left\| \frac{\partial (\overline{\boldsymbol{u}}, \overline{\boldsymbol{v}})}{\partial \boldsymbol{v}} \right\| \leq C_3 \|\boldsymbol{v}\|. \tag{215}$$

Proof. Since $G(\boldsymbol{u}, \mathbf{0}) = \psi(\boldsymbol{u}) \in M$ and $F(\psi(\boldsymbol{u})) = \psi(\boldsymbol{u})$, it follows that

$$\widetilde{\boldsymbol{u}}(\boldsymbol{u}, \mathbf{0}) = \boldsymbol{u} \text{ and } \widetilde{\boldsymbol{v}}(\boldsymbol{u}, \mathbf{0}) = \mathbf{0} \tag{216}$$

and

$$\overline{\boldsymbol{u}}(\boldsymbol{u},\mathbf{0})=\mathbf{0} \text{ and } \overline{\boldsymbol{v}}(\boldsymbol{u},\mathbf{0})=\mathbf{0} \tag{217}$$

Moreover, by equation 140, one knows

$$DG(\boldsymbol{u},\mathbf{0})=(D\psi,\boldsymbol{\varphi}_1,\cdots,\boldsymbol{\varphi}_m). \tag{218}$$

Then, by equation 171,

$$DF=I-\eta H(\boldsymbol{x}), \tag{219}$$

and by equation 136,

$$DFDG(\boldsymbol{u},\mathbf{0})=(D\psi,(1-\eta\lambda_1(\boldsymbol{u}))\boldsymbol{\varphi}_1,\cdots,(1-\eta\lambda_m(\boldsymbol{u}))\boldsymbol{\varphi}_m). \tag{220}$$

Therefore, we have

$$DG^{-1}DFDG(\boldsymbol{u},\mathbf{0})=\begin{pmatrix} I_{n-m} & \mathbf{0} \\ \mathbf{0} & \Lambda(\boldsymbol{u}) \end{pmatrix}. \tag{221}$$

Moreover, by equation 136, equation 140, and equation 173, we obtain that

$$\frac{\partial(\widetilde{\boldsymbol{u}},\widetilde{\boldsymbol{v}})}{\partial(\boldsymbol{u},\boldsymbol{v})}(\boldsymbol{u},\mathbf{0})=\begin{pmatrix} I_{n-m} & \mathbf{0} \\ \mathbf{0} & \Lambda(\boldsymbol{u}) \end{pmatrix}. \tag{222}$$

Hence, for all $\boldsymbol{u}\in B_{n-m}(\mathbf{0},\varepsilon_1)$,

$$\frac{\partial(\overline{\boldsymbol{u}},\overline{\boldsymbol{v}})}{\partial(\boldsymbol{u},\boldsymbol{v})}(\boldsymbol{u},\mathbf{0})=\mathbf{0}. \tag{223}$$

Specifically, for $\boldsymbol{u}\in B_{n-m}(\mathbf{0},\varepsilon_1)$,

$$\frac{\partial(\overline{\boldsymbol{u}},\overline{\boldsymbol{v}})}{\partial\boldsymbol{v}}(\boldsymbol{u},\mathbf{0})=\mathbf{0}. \tag{224}$$

It is easy to obtain that

$$\frac{\partial^2(\overline{\boldsymbol{u}},\overline{\boldsymbol{v}})}{\partial\boldsymbol{u}\partial\boldsymbol{v}}(\boldsymbol{u},\mathbf{0})=\mathbf{0}. \tag{225}$$

Next, take $(\overline{\boldsymbol{u}},\overline{\boldsymbol{v}})$ as functions of $\boldsymbol{v}$. Using Taylor expansion at $\boldsymbol{v}=\mathbf{0}$, by equation 217 and equation 224, it yields that:

$$(\overline{\boldsymbol{u}},\overline{\boldsymbol{v}})=\int_0^1(1-t)\frac{\partial^2(\overline{\boldsymbol{u}},\overline{\boldsymbol{v}})}{\partial\boldsymbol{v}^2}(\boldsymbol{u},t\boldsymbol{v})[\boldsymbol{v},\boldsymbol{v}]\,dt. \tag{226}$$

Similarly, take $\frac{\partial(\overline{\boldsymbol{u}},\overline{\boldsymbol{v}})}{\partial\boldsymbol{u}}$ and $\frac{\partial(\overline{\boldsymbol{u}},\overline{\boldsymbol{v}})}{\partial\boldsymbol{v}}$ as functions of $(\boldsymbol{u},\boldsymbol{v})$. By Taylor expansion at $\boldsymbol{v}=\mathbf{0}$, equation 223, equation 224 and equation 225, we obtain that

$$\frac{\partial(\overline{\boldsymbol{u}},\overline{\boldsymbol{v}})}{\partial\boldsymbol{u}}=\int_0^1(1-t)\frac{\partial^3(\overline{\boldsymbol{u}},\overline{\boldsymbol{v}})}{\partial\boldsymbol{u}\partial\boldsymbol{v}^2}(\boldsymbol{u},t\boldsymbol{v})[\boldsymbol{v},\boldsymbol{v}]\,dt, \tag{227}$$

and

$$\frac{\partial(\overline{\boldsymbol{u}},\overline{\boldsymbol{v}})}{\partial\boldsymbol{v}}=\int_0^1\frac{\partial^2(\overline{\boldsymbol{u}},\overline{\boldsymbol{v}})}{\partial\boldsymbol{v}^2}(\boldsymbol{u},t\boldsymbol{v})[\boldsymbol{v}]\,dt. \tag{228}$$

Combining equation 226, equation 227, and equation 228, we finish the proof of this lemma.

□

At the end of this subsection, we state the following important estimates that are frequently used in the following proof.

Denote $\overline{u}_{k,i}$ and $\overline{v}_{k,i}$ as the i-th component of $\overline{\boldsymbol{u}}_k$ and $\overline{\boldsymbol{v}}_k$ respectively, where $\overline{\boldsymbol{u}}_k$ and $\overline{\boldsymbol{v}}_k$ are defined as

$$\overline{\boldsymbol{u}}_k=\boldsymbol{u}_{k+1}-\boldsymbol{u}_k, \tag{229}$$

and

$$\overline{\boldsymbol{v}}_k=\boldsymbol{v}_{k+1}-\widehat{\boldsymbol{v}}_k. \tag{230}$$

By lemma 8.7, we immediately obtain the following lemma containing estimates that will apply in the later proof:

Lemma 8.8. *For all $(\boldsymbol{u}_0, \boldsymbol{v}_0) \in B_n(\mathbf{0}, \varepsilon_1)$ and $k \geq 0$, it follows that*

$$\|(\overline{\boldsymbol{u}}_k, \overline{\boldsymbol{v}}_k)\| \leq C_3 \|\boldsymbol{v}_k\|^2, \tag{231}$$

$$\left\| \frac{\partial(\overline{\boldsymbol{u}}_k, \overline{\boldsymbol{v}}_k)}{\partial \boldsymbol{u}_k} \right\| \leq C_3 \|\boldsymbol{v}_k\|^2, \tag{232}$$

$$\left\| \frac{\partial(\overline{\boldsymbol{u}}_k, \overline{\boldsymbol{v}}_k)}{\partial \boldsymbol{v}_k} \right\| \leq C_3 \|\boldsymbol{v}_k\|. \tag{233}$$

Having established the estimation on $(\overline{\boldsymbol{u}}, \overline{\boldsymbol{v}})$ and its corresponding partial derivatives, these local bounds must now be extended to analyze the accumulation of errors along the GD trajectory. To achieve this, the partial derivatives of the k-th iterates $(\boldsymbol{u}_k, \boldsymbol{v}_k)$ with respect to the initial coordinates are investigated.

8.8 Proof of Theorem 8.2 under equation 109, Part I

In this section, we prove Theorem 8.2 under equation 109.

Firstly, without loss of generality, since λ_i, ψ, $\overline{\boldsymbol{u}}$, and $\overline{\boldsymbol{v}}$ are smooth for all $1 \leq i \leq m$, then for all $(\boldsymbol{u}, \boldsymbol{v}) \in B_n(\mathbf{0}, \varepsilon_1)$, we can also obtain that

$$\left\| \frac{\partial(\lambda_i \circ \psi)}{\partial \boldsymbol{u}} \right\| \leq C_3. \tag{234}$$

Next, pick $\widetilde{a}$ such that

$$a_+ < \widetilde{a} < \sqrt{a_-} < 1. \tag{235}$$

Take $\varepsilon_2 > 0$ such that

$$0 < \varepsilon_2 < \min\left\{1, \varepsilon_1, \frac{\widetilde{a} - a_+}{K}, \frac{\widetilde{a} - a_+}{mC_3}, \frac{a_- - b}{2m^2 K}\right\}, \tag{236}$$

$$\varepsilon_2 \left(1 + (1+\eta)\exp\left(\frac{\widetilde{C_3}\varepsilon_2}{1-\widetilde{a}}\right)\right) < \frac{\widetilde{a} - a_+}{\widetilde{C_3}} \tag{237}$$

$$\frac{(1+\eta)n^2 C_3^2 \varepsilon_2^2}{(1-\widetilde{a}^2)\big(\widetilde{a} - a_+ - \widetilde{C}_3\varepsilon_2\big)} \exp\left(\frac{\widetilde{C}_3\varepsilon_2^2}{1-\widetilde{a}^2}\right) < 1, \tag{238}$$

and

$$\varepsilon_2 \exp\left(\frac{\widetilde{C_3}\varepsilon_2}{1-\widetilde{a}}\right) < \frac{(a_- - \widetilde{a}^2)^2}{2C_3 + K\eta} \tag{239}$$

where $\widetilde{C}_3 = nC_3$.

With these preparations, we have the following lemma.

Lemma 8.9. *Assume that equation 109 holds. For $1 \leq j \leq m$, define*

$$\rho_{k,j} := \sum_{i=1}^{n-m} \left|\frac{\partial u_{k,i}}{\partial v_{0,j}}\right|, \qquad \mu_{k,j} := \sum_{i=1}^{m} \left|\frac{\partial v_{k,i}}{\partial v_{0,j}}\right|, \tag{240}$$

where $u_{k,i}$ and $v_{k,i}$ are the i-th component of $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$ respectively. Then, for all $(\boldsymbol{u}_0, \boldsymbol{v}_0) \in B_n(\mathbf{0}, \varepsilon_2)$ and for all $k \geq 0$,

$$\rho_{k,j} \leq \exp\left(\frac{\widetilde{C}_3\varepsilon_2}{1-\widetilde{a}}\right) \qquad \mu_{k,j} \leq \widetilde{a}^{\,k}. \tag{241}$$

Proof. First, we prove a new estimation for $\boldsymbol{v}_k$. By equation 189, equation 190, and equation 236, one has

$$\|\boldsymbol{v}_{k+1}\| \leq \|\widehat{\boldsymbol{v}}_k\| + \|\overline{\boldsymbol{v}}_k\| \leq a_+\|\boldsymbol{v}_k\| + K\|\boldsymbol{v}_k\|^2 \leq (a_+ + K\varepsilon_2)\|\boldsymbol{v}_k\| < \widetilde{a}\,\|\boldsymbol{v}_k\|. \tag{242}$$

Therefore, inductively for all $k \geq 0$,

$$\|\boldsymbol{v}_k\| \leq \varepsilon_2 \widetilde{a}^k. \tag{243}$$

Next, by equation 229 and equation 230, we have

$$\boldsymbol{u}_{k+1} = \boldsymbol{u}_k + \overline{\boldsymbol{u}}_k, \qquad \boldsymbol{v}_{k+1} = \Lambda(\boldsymbol{u}_k)\boldsymbol{v}_k + \overline{\boldsymbol{v}}_k. \tag{244}$$

Differentiating equation 244 with respect to $v_{0,j}$, one gets, for $1 \leq i \leq n-m$,

$$\frac{\partial u_{k+1,i}}{\partial v_{0,j}} = \frac{\partial u_{k,i}}{\partial v_{0,j}} + \sum_{l=1}^{n-m} \frac{\partial \overline{u}_{k,i}}{\partial u_{k,l}} \frac{\partial u_{k,l}}{\partial v_{0,j}} + \sum_{s=1}^{m} \frac{\partial \overline{u}_{k,i}}{\partial v_{k,s}} \frac{\partial v_{k,s}}{\partial v_{0,j}}, \tag{245}$$

and, for $1 \leq i \leq m$,

$$\begin{aligned} \frac{\partial v_{k+1,i}}{\partial v_{0,j}} = &\left(1 - \eta\lambda_i(\psi(\boldsymbol{u}_k))\right) \frac{\partial v_{k,i}}{\partial v_{0,j}} - \eta \sum_{l=1}^{n-m} \frac{\partial(\lambda_i \circ \psi)}{\partial u_l}(\boldsymbol{u}_k) \frac{\partial u_{k,l}}{\partial v_{0,j}} \, v_{k,i} \\ &+ \sum_{l=1}^{n-m} \frac{\partial \overline{v}_{k,i}}{\partial u_l}(\boldsymbol{u}_k, \boldsymbol{v}_k) \frac{\partial u_{k,l}}{\partial v_{0,j}} + \sum_{s=1}^{m} \frac{\partial \overline{v}_{k,i}}{\partial v_{k,s}} \frac{\partial v_{k,s}}{\partial v_{0,j}}. \end{aligned} \tag{246}$$

Summing the absolute values in equation 245, and using equation 231, equation 232, equation 233, and equation 234, it follows that

$$\begin{aligned} \rho_{k+1,j} &\leq \left(1 + (n-m)C_3\|\boldsymbol{v}_k\|^2\right) \rho_{k,j} + (n-m)C_3\|\boldsymbol{v}_k\| \, \mu_{k,j} \\ &\leq \left(1 + nC_3\|\boldsymbol{v}_k\|^2\right) \rho_{k,j} + nC_3\|\boldsymbol{v}_k\| \, \mu_{k,j} \end{aligned} \tag{247}$$

and

$$\begin{aligned} \mu_{k+1,j} &\leq (a_+ + mC_3\|\boldsymbol{v}_k\|) \, \mu_{k,j} + \eta m C_3\|\boldsymbol{v}_k\| \, \rho_{k,j} + mC_3\|\boldsymbol{v}_k\|^2 \rho_{k,j} \\ &\leq (a_+ + nC_3\|\boldsymbol{v}_k\|) \, \mu_{k,j} + (1+\eta) n C_3\|\boldsymbol{v}_k\| \, \rho_{k,j}. \end{aligned} \tag{248}$$

Since $\varepsilon_2 < 1$, combining with equation 242, we have

$$\rho_{k+1,j} \leq (1 + \widetilde{C}_3\varepsilon_2\widetilde{a}^{2k})\rho_{k,j} + \widetilde{C}_3\varepsilon_2\widetilde{a}^k\mu_{k,j} \tag{249}$$

and

$$\mu_{k+1,j} \leq (a_+ + \widetilde{C}_3\varepsilon_2)\mu_{k,j} + (1+\eta)\widetilde{C}_3\varepsilon_2\widetilde{a}^k\rho_{k,j}. \tag{250}$$

Also,

$$\rho_{0,j} = 0, \qquad \mu_{0,j} = 1. \tag{251}$$

Next we use induction to prove that $\mu_{k,j} \leq \widetilde{a}^k$. For $k = 0$, from equation 251 it obviously holds. Then assume that $\mu_{l,j} \leq \widetilde{a}^{\,l}$ holds for all $0 \leq l \leq k$. By equation 249 and equation 251, one has

$$\rho_{k+1,j} + 1 \leq (1 + \widetilde{C}_3\varepsilon_2\widetilde{a}^k)(\rho_{k,j} + 1) \leq \exp(\widetilde{C}_3\varepsilon_2\widetilde{a}^k)(\rho_{k,j} + 1) \leq \exp\left(\widetilde{C}_3\varepsilon_2 \sum_{l=0}^{k} \widetilde{a}^l\right). \tag{252}$$

Therefore,

$$\rho_{k+1,j} \leq \exp\left(\frac{\widetilde{C}_3\varepsilon_2}{1-\widetilde{a}}\right) - 1 \leq \exp\left(\frac{\widetilde{C}_3\varepsilon_2}{1-\widetilde{a}}\right). \tag{253}$$

Thus, for all $k \geq 0$,

$$\rho_{k,j} \leq \exp\left(\frac{\widetilde{C}_3\varepsilon_2}{1-\widetilde{a}}\right). \tag{254}$$

Next, by equation 248, equation 242, equation 254, and equation 237,

$$\mu_{k+1,j} \leq (a_+ + \widetilde{C}_3\varepsilon_2)\mu_{k,j} + (1+\eta)\widetilde{C}_3\varepsilon_2\widetilde{a}^k \exp\left(\frac{\widetilde{C}_3\varepsilon_2}{1-\widetilde{a}}\right) \tag{255}$$

$$\leq (a_+ + \widetilde{C}_3\varepsilon_2)\widetilde{a}^k + (1+\eta)\widetilde{C}_3\varepsilon_2\widetilde{a}^k \exp\left(\frac{\widetilde{C}_3\varepsilon_2}{1-\widetilde{a}}\right) \tag{256}$$

$$\leq \left(a_+ + \widetilde{C}_3\varepsilon_2 + (1+\eta)\widetilde{C}_3\varepsilon_2 \exp\left(\frac{\widetilde{C}_3\varepsilon_2}{1-\widetilde{a}}\right)\right)\widetilde{a}^k \leq \widetilde{a}^{\,k+1}. \tag{257}$$

Thus equation 241 follows by induction. Here we finish the proof of the lemma. □

With these important estimates on the partial derivatives, we also have the following technical lemma.

Lemma 8.10. *Define*

$$a_k := 1 - \eta\lambda_1(\psi(\boldsymbol{u}_k)), \qquad A_k := \prod_{l=0}^{k} a_l. \tag{258}$$

Then the series

$$\sum_{k=0}^{\infty} \frac{\overline{v}_{k,1}}{A_k}, \qquad \sum_{k=0}^{\infty} \frac{\partial}{\partial v_{0,j}} \left(\frac{\overline{v}_{k,1}}{A_k} \right), \tag{259}$$

are absolutely and uniformly convergent on $B_n(\mathbf{0}, \varepsilon_2)$. In particular,

$$\left| \sum_{k=0}^{\infty} \frac{\partial}{\partial v_{0,j}} \left(\frac{\overline{v}_{k,1}}{A_k} \right) \right| < 1, \tag{260}$$

Proof. Since $a_k \geq a_-$ for all $k \geq 0$, one has

$$A_k \geq a_-^k > a_-^{k+1}. \tag{261}$$

Also, by equation 189 and equation 242,

$$|\overline{v}_{k,1}| \leq K\|\boldsymbol{v}_k\|^2 \leq K\varepsilon_2^2 \widetilde{a}^{\,2k}. \tag{262}$$

Therefore,

$$\sum_{k=0}^{\infty} \left| \frac{\overline{v}_{k,1}}{A_k} \right| \leq K\varepsilon_2^2 \sum_{k=0}^{\infty} \frac{\widetilde{a}^{\,2k}}{a_-^{k+1}} = \frac{K\varepsilon_2^2}{a_- - \widetilde{a}^2}, \tag{263}$$

so the series $\sum_{k=0}^{\infty} \overline{v}_{k,1}/A_k$ is convergent on $B_n(\mathbf{0}, \varepsilon_2)$.

For the $v_{0,j}$-derivatives, equation 231, equation 233, and equation 241 imply that

$$\left| \frac{\partial \overline{v}_{k,1}}{\partial v_{0,j}} \right| \leq C_3\|\boldsymbol{v}_k\|^2\rho_{k,j} + C_3\|\boldsymbol{v}_k\|\mu_{k,j} \leq C_3\varepsilon_2^2\widetilde{a}^{\,2k} \exp\left(\frac{\widetilde{C}_3\varepsilon_2}{1-\widetilde{a}} \right) + C_3\varepsilon_2\widetilde{a}^{\,k}\widetilde{a}^{\,k}. \tag{264}$$

Hence, we have

$$\left| \frac{\partial \overline{v}_{k,1}}{\partial v_{0,j}} \right| \leq 2C_3\varepsilon_2\widetilde{a}^{2k} \exp\left(\frac{\widetilde{C}_3\varepsilon_2}{1-\widetilde{a}} \right) \tag{265}$$

Then by equation 261 we know that

$$\begin{aligned} \sum_{k=0}^{\infty} \frac{1}{A_k} \left| \frac{\partial \overline{v}_{k,1}}{\partial v_{0,j}} \right| &\leq 2C_3\varepsilon_2 \exp\left(\frac{\widetilde{C}_3\varepsilon_2}{1-\widetilde{a}} \right) \sum_{k=0}^{\infty} \frac{\widetilde{a}^{2k}}{a_-^k} \leq 2C_3\varepsilon_2 \exp\left(\frac{\widetilde{C}_3\varepsilon_2}{1-\widetilde{a}} \right) \frac{a_-}{a_- - \widetilde{a}^2} \\ &\leq 2C_3\varepsilon_2 \exp\left(\frac{\widetilde{C}_3\varepsilon_2}{1-\widetilde{a}} \right) \frac{1}{a_- - \widetilde{a}^2} \leq 2C_3\varepsilon_2 \exp\left(\frac{\widetilde{C}_3\varepsilon_2}{1-\widetilde{a}} \right) \frac{1}{(a_- - \widetilde{a}^2)^2} \end{aligned} \tag{266}$$

Then, by equation 234, we have

$$\left| \frac{\partial a_k}{\partial v_{0,j}} \right| = \eta \left| \frac{\partial(\lambda_1 \circ \psi)(\boldsymbol{u}_k)}{\partial v_{0,j}} \right| \leq \eta C_3\rho_{k,j} \leq \eta C_3 \exp\left(\frac{\widetilde{C}_3\varepsilon_2}{1-\widetilde{a}} \right). \tag{267}$$

Hence

$$\left| \frac{\partial}{\partial v_{0,j}} \left(\frac{1}{A_k} \right) \right| \leq \frac{1}{A_k} \sum_{l=0}^{k} \left| \frac{1}{a_l} \frac{\partial a_l}{\partial v_{0,j}} \right| \leq \eta C_3 \exp\left(\frac{\widetilde{C}_3\varepsilon_2}{1-\widetilde{a}} \right) \frac{k+1}{a_-^{k+2}}. \tag{268}$$

Then we have

$$\begin{aligned} \sum_{k=0}^{\infty} |\overline{v}_{k,1}| \left| \frac{\partial}{\partial v_{0,j}} \left(\frac{1}{A_k} \right) \right| &\leq \sum_{k=0}^{\infty} K\varepsilon_2^2\widetilde{a}^{2k}\eta C_3 \exp\left(\frac{\widetilde{C}_3\varepsilon_2}{1-\widetilde{a}} \right) \frac{k+1}{a_-^{k+2}} \\ &\leq K\varepsilon_2^2\eta C_3 \exp\left(\frac{\widetilde{C}_3\varepsilon_2}{1-\widetilde{a}} \right) \frac{1}{(a_- - \widetilde{a}^2)^2} \end{aligned} \tag{269}$$

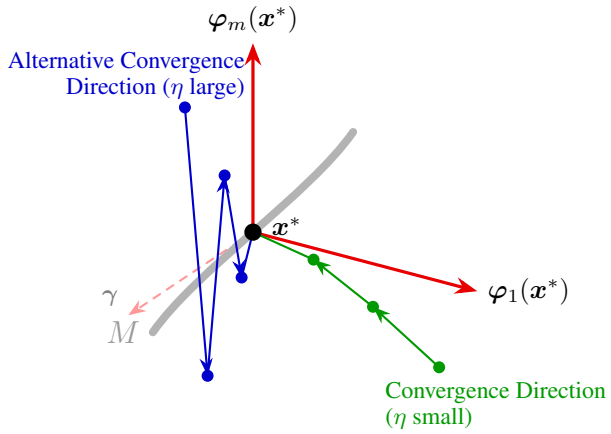


Figure 13: Visualization of the asymptotic convergence direction. The minimum set M (gray curve) is locally characterized by the tangent direction $\boldsymbol{\gamma}$ and the normal eigenvectors $\boldsymbol{\varphi}_1(\boldsymbol{x}^*)$ and $\boldsymbol{\varphi}_m(\boldsymbol{x}^*)$, corresponding to the smallest and largest positive eigenvalues, respectively. The **green trajectory** (η small) illustrates the convergence direction, where the iterates approach $\boldsymbol{x}^*$ aligned with $\boldsymbol{\varphi}_1(\boldsymbol{x}^*)$. The **blue trajectory** (η large) illustrates the alternative convergence direction, where the iterates zig-zag along $\boldsymbol{\varphi}_m(\boldsymbol{x}^*)$.

Combining equation 264, equation 268, equation 262, equation 261, equation 266 and equation 269, one gets

$$\sum_{k=0}^{\infty}\left|\frac{\partial}{\partial v_{0,j}}\left(\frac{\overline{v}_{k,1}}{A_k}\right)\right| \leq \sum_{k=0}^{\infty}\frac{1}{A_k}\left|\frac{\partial \overline{v}_{k,1}}{\partial v_{0,j}}\right| + \sum_{k=0}^{\infty}|\overline{v}_{k,1}|\left|\frac{\partial}{\partial v_{0,j}}\left(\frac{1}{A_k}\right)\right| \tag{270}$$

$$\leq (2C_3 + K\eta)\varepsilon_2 \exp\left(\frac{\widetilde{C_3}\varepsilon_2}{1-\widetilde{a}}\right)\frac{1}{(a_- - \widetilde{a}^2)^2} < 1. \tag{271}$$

□

With the convergence results of Lemma 8.10 established, the necessary preparations for the proof of Theorem 8.2 are complete. The bounds on $\overline{\boldsymbol{u}}$, $\overline{\boldsymbol{v}}$, and their derivatives provide sufficient conditions to control the nonlinear distortion of the coordinate system. In the subsequent section, we use these estimates to construct a global diffeomorphism on the local neighborhood. This construction allows us to invoke Sard's theorem, thereby proving that the set of initializations leading to bad behavior constitutes a set of Lebesgue measure zero.

8.9 Proof of Theorem 8.2 under equation 109, Part II

In this section, we rigorously eliminate the theoretical possibility of bad convergence behavior. To establish that the convergence direction $\boldsymbol{d}_k$ defined in equation 110 is well-defined for almost all initializations, we must prove that the set of starting points $\widetilde{S}$ leading to such degenerate behaviors has Lebesgue measure zero. We achieve this by leveraging the estimates from Lemma 8.10 to construct a mapping whose critical values capture these problematic trajectories, subsequently invoking Sard's theorem.

We again provide the same illustration in Figure 13 for the convergence direction, to aid understanding and build intuition.

First, in this section, we have

$$\eta < \frac{2}{\lambda_1(\mathbf{0}) + \lambda_m(\mathbf{0})}. \tag{272}$$

Recall that equation 152 means that, for all $\boldsymbol{u} \in U$,

$$\eta < \frac{2}{\lambda_1(\psi(\boldsymbol{u})) + \lambda_m(\psi(\boldsymbol{u}))}. \tag{273}$$

Therefore, for all $\boldsymbol{u} \in U$, we have

$$a(\boldsymbol{u}) = 1 - \eta\lambda_1(\psi(\boldsymbol{u})). \tag{274}$$

Define

$$\widetilde{S} = \left\{ \boldsymbol{x}_0 \in G(B_n(\mathbf{0}, \varepsilon_2)) : \forall k \geq 0, |v_{k,1}| < \sum_{i=2}^{m} |v_{k,i}| \right\}, \tag{275}$$

where ε_2 defined in Lemma 8.10. We now show that $\widetilde{S}$ is a measure-zero set.

For clarity, we take, for $k \geq 0$,

$$b_k = b(\boldsymbol{u}_k), \tag{276}$$

where the function $b(\boldsymbol{u})$ is defined in equation 149. Then, by equation 186, we have for all $k \geq 0$,

$$|\widehat{v}_{k,1}| \leq a_k |v_{k,1}| \leq a_+ |v_{k,1}| \tag{277}$$

and for each $2 \leq i \leq m$,

$$|\widehat{v}_{k,i}| \leq b_k |v_{k,i}| \leq b |v_{k,i}|, \tag{278}$$

where a_k, a_+, b_k, b are defined in equation 148, equation 258, equation 149 and equation 276 respectively.

Now we state the following two lemmas.

Lemma 8.11. *If $\boldsymbol{x}_0 \in \widetilde{S}$, then for all $k \geq 0$,*

$$\sum_{i=2}^{m} |v_{k,i}| \leq m\varepsilon_2 \left(b + 2m^2 K \varepsilon_2\right)^k \tag{279}$$

Proof. Firstly, by equation 173, equation 278 and equation 200, we have

$$\begin{aligned} \sum_{i=2}^{m} |v_{k+1,i}| \leq & \sum_{i=2}^{m} |\widehat{v}_{k,i}| + \sum_{i=2}^{m} |v_{k+1,i} - \widehat{v}_{k,i}| \\ \leq & b \sum_{i=2}^{m} |v_{k,i}| + mK \|\boldsymbol{v}_k\|^2. \end{aligned} \tag{280}$$

Moreover,

$$\begin{aligned} \|\boldsymbol{v}_k\|^2 = & v_{k,1}^2 + \sum_{i=2}^{m} v_{k,i}^2 \leq \left(\sum_{i=2}^{m} |v_{k,i}| \right)^2 + \left(\sum_{i=2}^{m} |v_{k,i}| \right)^2 = 2 \left(\sum_{i=2}^{m} |v_{k,i}| \right)^2 \\ \leq & 2(m-1)\varepsilon_2 \sum_{i=2}^{m} |v_{k,i}| \leq 2m\varepsilon_2 \sum_{i=2}^{m} |v_{k,i}|. \end{aligned} \tag{281}$$

Thus, substituting equation 281 back into equation 280 yields,

$$\sum_{i=2}^{m} |v_{k+1,i}| \leq \left(b + 2m^2 K \varepsilon_2\right) \sum_{i=2}^{m} |v_{k,i}|, \tag{282}$$

which completes the proof of this lemma by induction. □

Lemma 8.12. *If $\boldsymbol{x}_0 \in \widetilde{S}$, then we have*

$$\sum_{k=0}^{\infty} \frac{\overline{v}_{k,1}}{A_k} + v_{0,1} = 0, \tag{283}$$

where $A_k = \prod_{i=0}^{k} a_i$, and a_i is defined in equation 276.

Proof. First, since as $a_k \geq a_-$ holds for each $k \geq 0$,

$$\left|\sum_{k=0}^{\infty} \frac{\overline{v}_{k,1}}{A_k}\right| \leq \sum_{k=0}^{\infty} \frac{|\overline{v}_{k,1}|}{a_-^{k+1}}, \tag{284}$$

thus by Lemma 8.10, the series $\sum_{k=0}^{\infty} \overline{v}_{k,1}/A_k$ is convergent. Next, suppose that

$$\sum_{k=0}^{\infty} \frac{\overline{v}_{k,1}}{A_k} + v_{0,1} = \theta. \tag{285}$$

Notice that, by equation 230,

$$\overline{v}_{k,1} = v_{k+1,1} - a_k v_{k,1}. \tag{286}$$

Then, for all $l \geq 0$, by equation 230, one can use induction to prove that

$$\theta = \frac{v_{l+1,1}}{A_l} + \sum_{k=l}^{\infty} \frac{\overline{v}_{k,1}}{A_k}. \tag{287}$$

Consequently, by Lemma 8.11 and $a_k > a_-$, for all $l \geq 0$, we have

$$|\theta| \leq \left(\frac{b + 2m^2 K\varepsilon_2}{a_-}\right)^l m\varepsilon_2 + \left|\sum_{k=l}^{\infty} \frac{\overline{v}_{k,1}}{A_k}\right|. \tag{288}$$

Since $\sum_{k=0}^{\infty} \overline{v}_{k,1}/A_k$ is convergent and by 8.10, it implies that $\theta = 0$ by taking $l \to \infty$. This proves the lemma. □

Next, we define two functions

$$\sigma : B_n(\mathbf{0}, \varepsilon_2) \to \mathbb{R} \text{ and } h : B_n(\mathbf{0}, \varepsilon_2) \to \mathbb{R}^n \tag{289}$$

as

$$\sigma(\boldsymbol{u}, \boldsymbol{v}) = \sum_{k=0}^{\infty} \frac{\overline{v}_{k,1}(\boldsymbol{u}, \boldsymbol{v})}{A_k} \tag{290}$$

and

$$h(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u}, \sigma(\boldsymbol{u}, \boldsymbol{v}) + v_1, v_2, \ldots, v_m). \tag{291}$$

Here, as $\overline{\boldsymbol{v}}_k$ is a function of $(\boldsymbol{u}_0, \boldsymbol{v}_0)$, for clarity, we omit the subscripts "0" and consider $\overline{\boldsymbol{v}}_k$ as a function of $(\boldsymbol{u}, \boldsymbol{v})$. Since $\sum_{k=0}^{\infty} \overline{v}_{k,1}/A_k$ is convergent, σ and h are well-defined.

We first prove that h is injective on $B_n(\mathbf{0}, \varepsilon_2)$. By contradiction, if there exists $(\boldsymbol{u}', \boldsymbol{v}') \neq (\boldsymbol{u}'', \boldsymbol{v}'')$ such that $h(\boldsymbol{u}', \boldsymbol{v}') = h(\boldsymbol{u}'', \boldsymbol{v}'')$, we have $\boldsymbol{u}' = \boldsymbol{u}''$, $v_i' = v_i''$ for $2 \leq i \leq m$ and

$$\frac{\sigma(\boldsymbol{u}', \boldsymbol{v}') - \sigma(\boldsymbol{u}', \boldsymbol{v}'')}{v_1' - v_1''} = -1. \tag{292}$$

Here, v_i' and v_i'' are i-th component of $\boldsymbol{v}'$ and $\boldsymbol{v}''$ respectively. As a result, there exists ξ between v_1' and v_1'' such that

$$\partial_{v_1} \sigma(\boldsymbol{u}', \xi, v_2', \ldots, v_m') = -1. \tag{293}$$

This contradicts equation 271.

Moreover, the Jacobian Dh is given by

$$Dh = \begin{pmatrix} I_{n-m} & \mathbf{0} & \mathbf{0} \\ \nabla_{\boldsymbol{u}} \sigma & 1 + \partial_{v_1} \sigma & \nabla_{\boldsymbol{v}_{2:m}} \sigma \\ \mathbf{0} & \mathbf{0} & I_{m-1} \end{pmatrix}. \tag{294}$$

It follows that

$$\det(Dh) = 1 + \partial_{v_1} \sigma \neq 0. \tag{295}$$

Hence h is a C^1 diffeomorphism from $B_n(\mathbf{0}, \varepsilon_2)$ to $h(B_n(\mathbf{0}, \varepsilon_2))$. Define then

$$W = \left\{(\boldsymbol{u}, \boldsymbol{z}) \in \mathbb{R}^{n-1} : (\boldsymbol{u}, 0, \boldsymbol{z}) \in B_n(\mathbf{0}, \varepsilon_2)\right\}. \tag{296}$$

It is obvious that W is an open set in $\mathbb{R}^{n-1}$. We also define

$$\widetilde{\varphi} : W \to B_n(\mathbf{0}, \varepsilon_2) \tag{297}$$

as

$$\widetilde{\varphi}(\boldsymbol{u}, \boldsymbol{z}) = h^{-1}(\boldsymbol{u}, 0, \boldsymbol{z}). \tag{298}$$

It is clear that $\mathrm{rank}(D\widetilde{\varphi}) \leq n-1$. Therefore, by Sard's theorem, the set $\widetilde{\varphi}(W)$ has measure zero.

Finally, for each $\boldsymbol{x}_0 \in \widetilde{S}$, Lemma 8.12 implies that

$$(\boldsymbol{u}_0, \boldsymbol{v}_0) \in \widetilde{\varphi}(W). \tag{299}$$

Therefore, $\widetilde{S} \subset G(\widetilde{\varphi}(W))$. Since G is a diffeomorphism and maps a measure zero set to a measure zero set, $\widetilde{S}$ has measure zero.

Now define, for each $k \geq 0$,

$$S_k = \{\boldsymbol{x}_0 \in G(B_n(\mathbf{0}, \varepsilon_2)) : \boldsymbol{x}_k \in M\}. \tag{300}$$

Since M is a regular surface of dimension $n-m < n$, M has measure zero. Then, by the assumption on F, $S_k \subset F^{-k}(M)$ has measure zero. As a result, the following set

$$S = \widetilde{S} \cup (\cup_{k \geq 0} S_k) \tag{301}$$

has measure zero as well.

Here we include additional important estimates.

Lemma 8.13. *For all $\boldsymbol{x}_0 \in G(B_n(\mathbf{0}, \varepsilon_2)) \setminus S$, there exists $k_0 \geq 0$, such that for all $k \geq k_0$,*

$$|v_{k,1}| \geq \sum_{i=2}^{m} |v_{k,i}|. \tag{302}$$

Proof. First, since $\boldsymbol{x}_0 \notin \widetilde{S}$, then there exists a constant $k_0 \geq 0$ such that equation 302 hold for $k = k_0$. Then, as for each $k \geq k_0$,

$$|v_{k+1,1}| \geq |\hat{v}_{k,1}| - |v_{k+1,1} - \hat{v}_{k,1}| \geq a_- |v_{k,1}| - K\|\boldsymbol{v}_k\|^2. \tag{303}$$

We now use induction to prove equation 302. Suppose that equation 302 holds for a given $k \geq k_0$. In this case, since

$$\|\boldsymbol{v}_k\|^2 \leq v_{k,1}^2 + \left(\sum_{i=2}^{m} |v_{k,i}|\right)^2 \leq 2|v_{k,1}|^2, \tag{304}$$

then, by equation 303

$$|v_{k+1,1}| \geq a_- |v_{k,1}| - 2K|v_{k,1}|^2 \geq (a_- - 2K\varepsilon_2)|v_{k,1}|. \tag{305}$$

On the other hand, for $i = 2, \cdots, m$:

$$|v_{k+1,i}| \leq b|v_{k,i}| + K\|\boldsymbol{v}_k\|^2 \leq b|v_{k,i}| + 2K|v_{k,1}|^2. \tag{306}$$

Thus,

$$\sum_{i=2}^{m} |v_{k+1,i}| \leq b \sum_{i=2}^{m} |v_{k,i}| + (m-1)K\|\boldsymbol{v}_k\|^2 \leq b \sum_{i=2}^{m} |v_{k,i}| + 2(m-1)K|v_{k,1}|^2. \tag{307}$$

By the definition of ε_2 in equation 236, one knows:

$$|v_{k+1,1}| \geq \sum_{i=2}^{m} |v_{k+1,i}|. \tag{308}$$

Therefore, this lemma can be proved inductively. □

Moreover, by equation 236, $b < a_- - 2K\varepsilon_2$. Therefore, one can pick $1 < \alpha < 2$ such that

$$q = \frac{b}{(a_- - 2K\varepsilon_2)^\alpha} < 1. \tag{309}$$

Therefore,

$$\frac{|v_{k+1,i}|}{|v_{k+1,1}|^\alpha} \leq \frac{b|v_{k,i}|}{(a_- - 2K\varepsilon_2)^\alpha |v_{k,1}|^\alpha} + \frac{2K|v_{k,1}|^{2-\alpha}}{(a_- - 2K\varepsilon_2)^\alpha} \tag{310}$$

Since $|v_{k,1}| < 1$, it reduces to

$$\frac{|v_{k+1,i}|}{|v_{k+1,1}|^\alpha} \leq \frac{b|v_{k,i}|}{(a_- - 2K\varepsilon_2)^\alpha |v_{k,1}|^\alpha} + \frac{2K}{(a_- - 2K\varepsilon_2)^\alpha} \tag{311}$$

By induction, for all $k \geq k_0$ and $i = 2, \cdots, m$,

$$\frac{|v_{k+1,i}|}{|v_{k+1,1}|^\alpha} \leq q^{k-k_0} \frac{|v_{k_0,i}|}{|v_{k_0,1}|^\alpha} + \frac{2K}{(a_- - 2K\varepsilon_2)^\alpha} \sum_{j=0}^{k-k_0} q^j. \tag{312}$$

As $\sum_{j=0}^{k-k_0} q^j < 1/(1-q)$, it yields that for all $k \geq k_0$ and $i = 2, \cdots, m$,

$$\frac{|v_{k+1,i}|}{|v_{k+1,1}|^\alpha} \leq q^{k-k_0} \frac{|v_{k_0,i}|}{|v_{k_0,1}|^\alpha} + \frac{2K}{(1-q)(a_- - 2K\varepsilon_2)^\alpha}. \tag{313}$$

Therefore, there exists a constant C_5, such that for all $k \geq k_0$ and $i = 2, \cdots, m$,

$$\frac{|v_{k,i}|}{|v_{k,1}|^\alpha} \leq C_5. \tag{314}$$

Next, we claim that there exists a $k_1 \geq k_0$, such that for all $k \geq k_1$ and $(\boldsymbol{u}_0, \boldsymbol{v}_0) \in B_n(\mathbf{0}, \varepsilon_2)$, $v_{k,1} \neq 0$. Suppose there exists a $\widetilde{k} \geq k_1$, such that $v_{\widetilde{k},1} = 0$, then by equation 308, it implies for $i = 2, \cdots, m$, $|v_{\widetilde{k},i}| < |v_{\widetilde{k},1}| = 0$. Hence $\boldsymbol{v}_{\widetilde{k}} = \mathbf{0}$. Moreover, by equation 199, one knows $\boldsymbol{u}_{\widetilde{k}} = \boldsymbol{u}^*$. Therefore, $\boldsymbol{x}_{\widetilde{k}} = \boldsymbol{x}^*$. Then this implies $\boldsymbol{x}_0 \in S_{\widetilde{k}}$, which is a contradiction.

Therefore, it is not difficult to see that for all $k \geq k_1$,

$$\|\boldsymbol{v}_k\| = O(|v_{k,1}|). \tag{315}$$

Similarly, by equation 146, equation 188,

$$\|\psi(\boldsymbol{u}_k) - \psi(\boldsymbol{u}^*)\| \leq O(|v_{k,1}|^2). \tag{316}$$

Meanwhile, for all $i = 2, \cdots, m$, since $\boldsymbol{\varphi}_i$ and ψ are continuous,

$$\lim_{k\to\infty} \langle \boldsymbol{\varphi}_i(\psi(\boldsymbol{u}_k)), \boldsymbol{\varphi}_1(\psi(\boldsymbol{u}^*))\rangle = 0 \tag{317}$$

and

$$\lim_{k\to\infty} \langle \boldsymbol{\varphi}_1(\psi(\boldsymbol{u}_k)), \boldsymbol{\varphi}_1(\psi(\boldsymbol{u}^*))\rangle = 1. \tag{318}$$

Next, for all $\boldsymbol{x}_0 \in G(B_n(\mathbf{0}, \varepsilon_2)) \setminus S$, denote $\boldsymbol{x}^* = \lim_{k\to\infty} \boldsymbol{x}_k$, and define

$$\beta_k = \langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\varphi}_1(\psi(\boldsymbol{u}^*))\rangle \tag{319}$$

and

$$\zeta_k = \beta_k / v_{k,1}. \tag{320}$$

We have the following lemma.

Lemma 8.14. *There are at most finitely many $k > 0$ such that $\beta_k = 0$.*

Proof. We prove this lemma by contradiction. First, by equation 140

$$\beta_k = \langle \psi(\boldsymbol{u}_k) - \psi(\boldsymbol{u}^*) + \sum_{i=1}^m v_{k,i} \boldsymbol{\varphi}_i(\psi(\boldsymbol{u}_k)), \boldsymbol{\varphi}_1(\boldsymbol{u}^*)\rangle. \tag{321}$$

By equation 316, equation 317 and equation 318, there exists $k_2 \geq k_1$, such that for all $k \geq k_2$,

$$|\zeta_k| \geq \frac{1}{2}. \tag{322}$$

Therefore, there are infinitely many $\beta_k = 0$ implies that there are infinitely many $v_{k,1} = 0$ for some $k \geq k_2$, which is a contradiction. This completes the proof of the lemma. □

Finally, we begin the last step of the proof. By equation 195 and equation 199, applying the Taylor expansion gives

$$\boldsymbol{x}_k - \boldsymbol{x}^* = D\psi(\boldsymbol{u}^*)(\boldsymbol{u}_k - \boldsymbol{u}^*) + O(\|\boldsymbol{v}_k\|^4) + \sum_{j=1}^{m} \langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\varphi}_i(\psi(\boldsymbol{u}^*))\rangle \boldsymbol{\varphi}_i(\psi(\boldsymbol{u}^*)) + O(\|\boldsymbol{v}_k\|^2), \tag{323}$$

By equation 316, equation 317 equation 318, and equation 322 one has for all $i = 1, \cdots, n-m$,

$$\frac{\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\gamma}_i(\psi(\boldsymbol{u}^*))\rangle}{|v_{k,1}|} \longrightarrow 0, \tag{324}$$

and for all $j = 2, \cdots, m$

$$\frac{\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\varphi}_j(\psi(\boldsymbol{u}^*))\rangle}{|v_{k,1}|} \longrightarrow 0. \tag{325}$$

and for all $k \geq k_2$,

$$\frac{|v_{k,1}|}{\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\varphi}_1(\psi(\boldsymbol{u}^*))\rangle} \leq 2. \tag{326}$$

Hence, $\boldsymbol{x}_k - \boldsymbol{x}^*$ has only minor components in the tangent directions at $\boldsymbol{x}^*$. Therefore, for all $i = 1, \cdots, n-m$,

$$\frac{\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\gamma}_i(\psi(\boldsymbol{u}^*))\rangle}{\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\varphi}_1(\psi(\boldsymbol{u}^*))\rangle} = \frac{\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\gamma}_i(\psi(\boldsymbol{u}^*))\rangle}{|v_{k,1}|} \frac{|v_{k,1}|}{\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\varphi}_1(\psi(\boldsymbol{u}^*))\rangle} \longrightarrow 0. \tag{327}$$

and for all $j = 2, \cdots, m$,

$$\frac{\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\varphi}_j(\psi(\boldsymbol{u}^*))\rangle}{\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\varphi}_1(\psi(\boldsymbol{u}^*))\rangle} = \frac{\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\varphi}_j(\psi(\boldsymbol{u}^*))\rangle}{|v_{k,1}|} \frac{|v_{k,1}|}{\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\varphi}_1(\psi(\boldsymbol{u}^*))\rangle} \longrightarrow 0. \tag{328}$$

Therefore,

$$\lim_{k\to\infty} |\cos\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\varphi}_1(\psi(\boldsymbol{u}^*))\rangle| = 1. \tag{329}$$

The last step is to discuss the sign of this cosine. From equation 303, one knows there exists $k_3 \geq k_2$ such that for all $k \geq k_3$, $v_{k,1}$ will preserve its sign and

$$|\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\varphi}_1(\psi(\boldsymbol{u}^*))\rangle| > \frac{1}{2}|v_{k,1}|. \tag{330}$$

Therefore,

$$\lim_{k\to\infty} \boldsymbol{d}_k = \pm\boldsymbol{\varphi}_1(\psi(\boldsymbol{u}^*)) \tag{331}$$

depending on the sign of $v_{k_3,1}$.

8.10 Proof of Theorem 8.2 Under equation 111

Most of the proof for this case is similar to the previous case. Here we provide a sketch.

Recall equation 160, in this case, we have

$$a(\boldsymbol{u}) = \eta\lambda_m(\psi(\boldsymbol{u})) - 1. \tag{332}$$

Similarly, we define $\widetilde{S}$ as

$$\widetilde{S} = \left\{ \boldsymbol{x}_0 \in B_n(\boldsymbol{0}, \varepsilon_2) : \forall k \geq 0, |v_{k,m}| < \sum_{i=1}^{m-1} |v_{k,i}| \right\}, \tag{333}$$

One can prove that $\widetilde{S}$ is a measure-zero set by exactly the same steps.

Next, using the same set S_k as in equation 300 for each $k \geq 0$, by the assumption on F, it follows that S_k has measure zero. Therefore, the set $S = \widetilde{S} \cup (\cup_{k\geq 0} S_k)$ also has measure zero.

Therefore, for all $\boldsymbol{x}_0 \in G(B_n(\boldsymbol{0}, \varepsilon_2)) \setminus S$, there exists $k_0 \geq 0$, such that for all $k \geq k_0$,

$$|v_{k,m}| \geq \sum_{i=1}^{m-1} |v_{k,i}|. \tag{334}$$

Similarly, there exists a constant C_5, such that for all $k \geq k_0$ and $i = 1, \cdots, m-1$,

$$\frac{|v_{k,i}|}{|v_{k,m}|^\alpha} \leq C_5. \tag{335}$$

In addition, there exists $k_1 \geq k_0$, such that for all $k \geq k_1$,

$$v_{k,m} \neq 0. \tag{336}$$

Thus, similarly, one has

$$\|\boldsymbol{v}_k\| = O(|v_{k,m}|), \tag{337}$$

$$\|\psi(\boldsymbol{u}_k) - \psi(\boldsymbol{u}^*)\| \leq O(|v_{k,m}|^2), \tag{338}$$

for all $i = 1, \cdots, m-1$,

$$\lim_{k\to\infty} \langle \boldsymbol{\varphi}_i(\psi(\boldsymbol{u}_k)), \boldsymbol{\varphi}_m(\psi(\boldsymbol{u}^*))\rangle = 0 \tag{339}$$

and

$$\lim_{k\to\infty} \langle \boldsymbol{\varphi}_m(\psi(\boldsymbol{u}_k)), \boldsymbol{\varphi}_m(\psi(\boldsymbol{u}^*))\rangle = 1. \tag{340}$$

Next, for all $\boldsymbol{x}_0 \in G(B_n(\mathbf{0}, \varepsilon_2)) \setminus S$, take $\boldsymbol{x}^* = \lim_{k\to\infty} \boldsymbol{x}_k$, define

$$\beta_k = \langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\varphi}_m(\psi(\boldsymbol{u}^*))\rangle \tag{341}$$

and

$$\zeta_k = \beta_k / v_{k,m}. \tag{342}$$

Similarly, there were at most finitely many $k > 0$ such that $\beta_k = 0$. Moreover, there exists $k_2 \geq k_1$, such that for all $k \geq k_2$,

$$|\zeta_k| \geq \frac{1}{2}. \tag{343}$$

The last step of the proof is similar to the previous case:

$$\frac{\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\gamma}_i(\psi(\boldsymbol{u}^*))\rangle}{|v_{k,m}|} \longrightarrow 0, \tag{344}$$

and for all $j = 1, \cdots, m-1$

$$\frac{\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\varphi}_j(\psi(\boldsymbol{u}^*))\rangle}{|v_{k,m}|} \longrightarrow 0. \tag{345}$$

and for all $k \geq k_2$,

$$\frac{|v_{k,m}|}{\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\varphi}_m(\psi(\boldsymbol{u}^*))\rangle} \leq 2. \tag{346}$$

Hence, $\boldsymbol{x}_k - \boldsymbol{x}^*$ also has only minor components in the tangent directions at $\boldsymbol{x}^*$ in this case. Therefore, similarly, for all $i = 1, \cdots, n-m$,

$$\frac{\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\gamma}_i(\psi(\boldsymbol{u}^*))\rangle}{\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\varphi}_m(\psi(\boldsymbol{u}^*))\rangle} = \frac{\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\gamma}_i(\psi(\boldsymbol{u}^*))\rangle}{|v_{k,m}|} \frac{|v_{k,m}|}{\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\varphi}_m(\psi(\boldsymbol{u}^*))\rangle} \longrightarrow 0. \tag{347}$$

and for all $j = 1, \cdots, m-1$,

$$\frac{\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\varphi}_j(\psi(\boldsymbol{u}^*))\rangle}{\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\varphi}_m(\psi(\boldsymbol{u}^*))\rangle} = \frac{\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\varphi}_j(\psi(\boldsymbol{u}^*))\rangle}{|v_{k,m}|} \frac{|v_{k,m}|}{\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\varphi}_m(\psi(\boldsymbol{u}^*))\rangle} \longrightarrow 0. \tag{348}$$

Therefore,

$$\lim_{k\to\infty} |\cos\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\varphi}_m(\psi(\boldsymbol{u}^*))\rangle| = 1. \tag{349}$$

The last step is again to discuss the sign of this cosine. From equation 303, one knows there exists $k_3 \geq k_2$, such that for all $k \geq k_3$, $v_{k,m}$ alternates its sign and

$$|\langle \boldsymbol{x}_k - \boldsymbol{x}^*, \boldsymbol{\varphi}_m(\psi(\boldsymbol{u}^*))\rangle| > \frac{1}{2}|v_{k,m}|. \tag{350}$$

Therefore,

$$\lim_{k\to\infty} (-1)^k \boldsymbol{d}_k = \pm\boldsymbol{\varphi}_m(\boldsymbol{u}^*) \tag{351}$$

depending on the sign of $v_{k_3,m}$.

8.11 TECHNICAL THEOREMS

Theorem 8.15 (Cholesky Factorization Theorem, Higham (2002))**.** *Let $A \in \mathbb{R}^{n\times n}$ be positive semidefinite of rank r. Then there exists at least one upper triangular P with nonnegative diagonal elements such that $A = P^T P$.*

Theorem 8.16 (Sard's Theorem, Lee (2012))**.** *Let $F : U \to \mathbb{R}^n$ be a $\mathcal{C}^p$ mapping with $p \geq 1$, where U is an open subset of $\mathbb{R}^m$. Define $A = \{\boldsymbol{x} \in U \mid \boldsymbol{x}$ is a critical point of $F\}$, then the image $F(A)$ has measure zero in $\mathbb{R}^n$. Here, $\boldsymbol{x}$ is a critical point if* $\operatorname{rank}(DF(\boldsymbol{x})) < n$.