AdaGrad avoids saddle points

Kimon Antonakopoulos, Panayotis Mertikopoulos, Georgios Piliouras, Xiao Wang

ADAGRAD Avoids Saddle Points

Kimon Antonakopoulos [1 2] **Panayotis Mertikopoulos** [3 4] **Georgios Piliouras** [5] **Xiao Wang** [6]

Abstract

Adaptive first-order methods in optimization are prominent in machine learning and data science owing to their ability to automatically adapt to the landscape of the function being optimized. However, their convergence guarantees are typically stated in terms of vanishing gradient norms, which leaves open the issue of converging to undesirable saddle points (or even local maximizers). In this paper, we focus on the ADAGRAD family of algorithms – with scalar, diagonal or full-matrix preconditioning – and we examine the question of whether the method's trajectories avoid saddle points. A major challenge that arises here is that ADAGRAD's step-size (or, more accurately, the method's preconditioner) evolves over time in a filtration-dependent way, i.e., as a function of all gradients observed in earlier iterations; as a result, avoidance results for methods with a constant or vanishing step-size do not apply. We resolve this challenge by combining a series of step-size stabilization arguments with a recursive representation of the ADAGRAD preconditioner that allows us to employ stable manifold techniques and ultimately show that the induced trajectories avoid saddle points from almost any initial condition.

1 Introduction

Deep learning architectures have brought forth a revolution in numerous application areas, from computer vision and recommender systems, to speech recognition and natural language processing [6, 19]. Although gradient descent (and its variants) is the mainstay training tool for such models, it comes with a significant inherent drawback: the gradient steps taken at each iteration are essentially "Markovian", in the sense that information gained about the model's loss landscape over time is not taken into account when performing an update. For this reason, adaptive gradient algorithms have emerged as an essential ingredient of contemporary machine learning models and architectures: by incorporating data and knowledge from gradients observed in earlier iterations, adaptive methods perform more informed gradient steps in later iterations, and they are able to adapt efficiently to the landscape of the function being optimized.

The blueprint for most adaptive first-order methods – including ADAM [17], ADAMNC [32], AMSGRAD [32], and RMSPROP – is the ADAGRAD family of algorithms that was introduced concurrently by Duchi et al. [10] and McMahan & Streeter [24]. In the unconstrained case (which is the playground of choice for most adaptive methods of this type), the ADAGRAD algorithm proceeds as a gradient descent method with a matrix-valued step-size – typically referred to as a *preconditioner* – which is defined recursively by taking the square root of the sum of squares of past gradients (possibly tensored, depending on the specific variant of the method). Owing to this clever preconditioning mechanism, ADAGRAD excels in solving convex-structured problems with sparse gradients, while remaining competitive in environments with full (dense) gradients.

Specifically, in convex problems with Lipschitz continuous objectives, ADAGRAD attains an $\mathcal{O}(1/\sqrt{T})$ value convergence rate after T queries to a first-order oracle (stochastic or deterministic). This rate improves to $\mathcal{O}(\log T/T)$ if the problem's objective is strongly convex [10],[1] and to $\mathcal{O}(1/T)$ if the problem's objective is Lipschitz smooth [1, 2, 22].[2] In this regard, ADAGRAD is not order-optimal because it does not attain the iconic $\mathcal{O}(1/T^2)$ accelerated convergence rate of Nesterov [26]; however, other adaptive methods based on the ADAGRAD template *do* achieve this – like the ACCELEGRAD proposal of [22], or the more recent UNIXGRAD and UNDERGRAD algorithms by [16] and [4, 34] respectively.

In the non-convex world (which is of greater interest to

Authors in alphabetical order. [1]Laboratory for Information and Inference Systems, IEM, STI, EPFL, 1015 Lausanne, Switzerland. [2]This work was done when KA was with Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France. [3]Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France [4]Criteo AI Lab [5]Singapore University of Technology and Design [6]Shanghai University of Finance and Economics. Correspondence to: Xiao Wang <wangxiao@sufe.edu.cn>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

[1]Or if the noise in the method's gradient oracle is a relative percentage of the gradient norm, cf. the recent paper [3].

[2]Importantly, the first of these rates continues to hold in the stochastic case, but the second does not.

deep learning applications), ADAGRAD has been shown to attain an $\mathcal{O}(1/T)$ convergence rate when the optimizer has access to a perfect, deterministic gradient oracle, and an $\mathcal{O}(1/\sqrt{T})$ convergence rate when only stochastic gradients are available [23, 35]. Importantly, the merit function in both cases is no longer the value of the objective function, but the sum of gradient norms squared. As a result, the above guarantees translate to a convergence rate for the method's "best iterate", i.e., the queried point with the least (true) gradient norm.[3] In this regard, ADAGRAD is an order-optimal method in the non-convex case – which, coupled with its simplicity and the capability of exploiting sparse gradients – makes it an ideal choice for many problems with moderate-to-high dimensionality and a sparse solution structure.

At the same time, it should be noted that the only guarantee provided in non-convex settings is that of a *vanishing gradient*, not a value minimization certificate (local or global). This leaves open not only the question of global versus local optimality, but an even more fundamental one:

Do the trajectories of ADAGRAD *avoid saddle points?*

This question is the core of our paper; to put our contributions in the proper context, we begin by discussing the related work on the topic.

Related work. The literature on saddle-point avoidance is quite extensive, so some general remarks are in order. First, to the best of our knowledge, existing results concern almost exclusively *non-adaptive* methods (we discuss the single exception that we are aware of below). These results can be subsequently classified into results for deterministic and stochastic gradient descent (depending on the oracle in question); perhaps surprisingly, these two branches of the literature *do not intersect* and, for reasons that we explain below, the insights and techniques cannot be ported from one regime to the other.

Historically, the first avoidance results were obtained in the stochastic setting in the early 90's, with Pemantle [30] and Brandière & Duflo [7] being the first to establish the avoidance of *hyperbolic* saddle points for stochastic gradient descent methods.[4] Specifically, they showed that the trajectories of any vanishing step-size stochastic approximation of a gradient flow avoid hyperbolic saddle points with probability 1, from any initial condition. More recently, and under a somewhat different set of assumptions for the problem's objective function, Ge et al. [12] showed that stochastic gradient descent (SGD) escapes *strict* saddle points (i.e., stationary points where the Hessian of the objective has at least one negative eigenvalue, but could also have zero eigenvalues), and produces iterates close to second-order optimal stationary points with high probability. Daneshmand et al. [8] further refined this result by obtaining positive probability results for *second-order* stationary points, while Mertikopoulos et al. [25] and Hsieh et al. [14] showed that strict saddles are avoided *with probability* 1, not only by SGD, but by any Robbins-Monro approximation of a gradient flow (including stochastic extra-gradient, optimistic gradient, and several other non-adaptive first-order methods). Finally, Staib et al. [33] examined a variant of SGD with *adaptive preconditioning* (including AMSGRAD and RMSPROP) and obtained a strict saddle escape result in the spirit of Ge et al. [12]; to the best of our knowledge, this is the only escape result for adaptive, stochastic methods, and we discuss it in detail later in the paper.

On the deterministic side of the literature, Lee et al. [20] showed that the trajectories of deterministic gradient descent avoid strict saddle points from any initial condition. In a concurrent paper, Panageas & Piliouras [27] established a more general version of this result concerning non-isolated (i.e., continua of) saddle points – including ridges, talwegs, or other manifolds of non-minimizing stationary points. These two approaches where shortly afterwards unified into a generalized methodology that was able to address numerous distinct first-order optimization methods including gradient descent, block coordinate descent, mirror descent and variants thereof [21]. Since these first works, several extensions and refinements have appeared, regarding the rate of convergence to second-order optimal points [9, 15], zeroth-order methods [11], constrained distributed optimization [29] and, finally, gradient descent with a vanishing step-size [28].

The key difference between the two regimes – deterministic and stochastic – is that stochastic results invariably rely on a "positive excitation" assumption for the noise: not necessarily that it is isotropic, but that it excites all directions in space in a uniformly positive amount (for a precise definition, see [5, 30]). In this regard, the noise in the optimizer's gradient queries can be viewed as providing a "boost" to escaping saddle points (though the extent of this boost can be relatively small in high-dimensional problems, a question that has been examined at depth in the relevant literature). By contrast, this "stochastic boost" is completely absent in the analysis of deterministic gradient descent schemes; as a result, the techniques employed in "stochastic escape" papers are likewise completely different to the techniques employed in the "deterministic avoidance" literature. More to the point, because the noise profile is always assumed to be persistent in the stochastic literature, results about

[3] This creates an issue in the stochastic case because it is not possible to identify the point with the least gradient norm when only a single run of random gradient observations is available. In our paper, we only treat the deterministic case, so this issue does not arise.

[4] A saddle point is hyperbolic if the Hessian of the objective function is invertible and has at least one negative eigenvalue at said point.

stochastic methods do not imply anything for the deterministic case – and, likewise, of course for the converse.

Comparison of techniques. We outline below the technical challenges of our approach and the techniques we employed relative to other works in the literature. All previous papers in the deterministic regime with the single exception of [28], apply standard machinery from the dynamical systems literature and specifically standard variants of the celebrated stable manifold theorem [31]. This is a standard tool in analyzing the behavior of a dynamical system in the neighborhood of a stationary point; however, in its standard formulation, this theorem applies only to autonomous (i.e., time-independent) smooth maps or flows. As such, this framework turns out to be too restrictive for many interesting machine learning applications where the *map* producing the dynamics *is also evolving itself over time*.

Such non-autonomous dynamical systems require particular care and case-by-case analysis as they are intractable in their full generality. As a first step in this direction, gradient descent with vanishing step-sizes was shown to provably avoid saddle points via a novel tailored version of the stable manifold theorem [28]. The dynamical system arising from ADAGRAD (this paper) poses unique and novel challenges that puts us well outside any prior approach. These differences are as follows:

- The update rule of ADAGRAD is a "filtration-dependent" function of all its history depending on *time* as well as *all previous states*.
- The step-size matrix of ADAGRAD could be either vanishing or non-vanishing, depending on the landscape encountered. non-vanishing, the norm converges to a non-zero constant for each initial condition $\mathbf{x}_0$.

In view o the above, there is no a priori reason to expect that ADAGRAD should avoid saddle points in the same way that gradient descent does. In more detail, the vanishing step-size regime requires a completely different center-stable theorem than the constant step-size case, and since ADAGRAD interpolates between the two, neither analysis or technique is sufficient for this. In addition, the fact that we have a *matrix-valued* step-size – the preconditioning matrix – it is crucial to establish sufficient control over its spectrum (and, in particular, the minimum and maximum eigenvalues thereof). We achieve this by proving the existence of a *strictly* positive-definite limit for the method's preconditioner which, combined with a "Markovian" reframing of the ADAGRAD update sequence, allows us to decompose the underlying dynamics into a stable system plus a "small" residual term which does not affect the method's convergence properties. Then, leveraging a Lipschitz-type condition for this remainder term allows us to derive an ADAGRAD-tailored stable manifold theorem that precludes the algorithm's convergence to strict saddle points.

2 Setup and preliminaries

2.1 Setup

In this paper, we will focus on non-convex optimization problems of the form

$$\begin{aligned} &\text{minimize} \;\; f(\mathbf{x}), \\ &\text{subject to} \;\; \mathbf{x} \in \mathbb{R}^d. \end{aligned} \tag{Opt}$$

In the above, $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is assumed to be lower bounded and continuously differentiable, i.e., the following holds:

1. $\inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) > -\infty$.
2. There exists some positive constant $L > 0$ such that

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \tag{LS}$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

An important property which follows directly from (LS) is the so-called *descent inequality*:

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 \tag{Descent}$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Heuristically, (Descent) allows us to upper-bound our objective by a quadratic majorant. This property will play a crucial role in what follows.

Another condition for f is concerns its critical points. In particular, we say that $\mathbf{x}^* \in \mathbb{R}^d$ is a *strict saddle point* if $\mathbf{x}^*$ is a critical point and the Hessian of f at $\mathbf{x}^*$ has at least one negative eigenvalue, i.e.,

$$\|\nabla f(\mathbf{x}^*)\| = 0 \;\; \text{and} \;\; \lambda_{\min}(\nabla^2 f(\mathbf{x}^*)) < 0.$$

Throughout this paper we use saddle point for short when there is no ambiguity.

2.2 Algorithm and Problem

In this paper, we will investigate gradient descent with an adaptive step-size in the spirit of the ADAGRAD algorithm of Duchi et al. [10]. Formally, the algorithm is described by the following recursive formula:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \Gamma_t \nabla f(\mathbf{x}_t), \tag{ADAGRAD}$$

where the step-size Γ_t has following three variants:

- ADAGRAD with squared norm adaptation

$$\Gamma_t = \frac{1}{\sqrt{\delta_0^2 + \sum_{s=0}^{t} \|\nabla f(\mathbf{x}_s)\|^2}} \tag{ADANORM}$$

- ADAGRAD with diagonal step-size scaling

$$\Gamma_t = G_t^{-\frac{1}{2}} \qquad \text{(ADADIAG)}$$

where

$$G_t = \delta_0^2 I + \operatorname{diag}\left(\sum_{s=0}^{t} \nabla f(\mathbf{x}_s)\nabla f(\mathbf{x}_s)^\top\right).$$

- ADAGRAD with full matrix preconditioning

$$\Gamma_t = G_t^{-\frac{1}{2}} \qquad \text{(ADAFULL)}$$

where

$$G_t = \delta_0^2 I + \sum_{s=0}^{t} \nabla f(\mathbf{x}_s)\nabla f(\mathbf{x}_s)^\top.$$

We clarify that throughout this paper, all the step-size policies and gradients are deterministic. The problem we address in this article is: *Do* ADAGRAD *algorithms provably avoid saddle points?*

This saddle avoidance type question stimulates the study of non-convex optimization, machine learning and dynamical systems in recent years. It is an essential part leading us to a better understanding of the power of deterministic ADAGRAD algorithms.

2.3 Technical Preliminaries

For posterity, we list some fundamental concepts and results that will be frequently referred to in our analysis and proofs. Standard references for matrix calculus and Banach fixed point arguments can be found in [13] and [18] respectively.

Dynamical systems. Let $\mathbb{T} = \mathbb{R}$ or $\mathbb{Z}$. A smooth dynamical system on $\mathbb{R}^d$ is a continuous differentiable mapping $\phi : \mathbb{T} \times \mathbb{R}^d \to \mathbb{R}^d$, where $\phi(t, \mathbf{x}) = \phi_t(\mathbf{x})$ satisfies

- $\phi_0 : \mathbb{R}^d \to \mathbb{R}^d$ is the identity mapping.
- The composition $\phi_t \circ \phi_s = \phi_{t+s}$ for each $t, s \in \mathbb{T}$.

In our setting, the T compositions of mappings defined by the update rule can be seen as a mapping $\phi(T, \mathbf{x}_0)$ from the initial point $\mathbf{x}_0$ to $\phi_T(\mathbf{x}_0)$.

Diffeomorphism. A differentiable mapping $\phi : \mathbb{R}^d \to \mathbb{R}^d$ is a local diffeomorphism at $\mathbf{x}$ if the Jacobian matrix $D\phi(\mathbf{x})$ is invertible.

Remark 1. A typical example of diffeomorphism is gradient descent with small constant step-size, whose update mapping is $\phi(\mathbf{x}) = \mathbf{x} - \eta\nabla f(\mathbf{x})$. In our setting, the step-size policy is time-dependent and the mapping $\phi(t, \mathbf{x}) = \mathbf{x} - \Gamma_t \nabla f(\mathbf{x})$ is expected to be a diffeomorphism on $\mathbb{R}^d$ for each $t \in \mathbb{N}$.

Matrix preliminaries. For Hermitian matrices G, H we write $G \preccurlyeq H$ or $H \succcurlyeq G$ to mean $H - G$ is positive semidefinite. In particular, $H \succcurlyeq 0$ indicates that H is positive semidefinite. This order is known as the *Löwner partial order*. If H is positive definite, i.e., positive semidefinite and invertible, we write $H \succ 0$. The following two results are frequently applied in stabilization analysis of the adaptive step-size matrix Γ_t.

- Löwner-Heinz inequality: If $A \succcurlyeq B \succcurlyeq 0$ and $0 \leq r \leq 1$ then $A^r \succcurlyeq B^r$.
- Weyl's monotonicity lemma: If H is positive, and the eigenvalues of $A + H$ and A are ordered as

$$|\lambda_1(A + H)| \geq \cdots \geq |\lambda_d(A + H)|$$

and

$$|\lambda_1(A)| \geq \cdots \geq |\lambda_d(A)|.$$

Then

$$\lambda_i(A + H) \geq \lambda_i(A) \text{ for all } i = 1, \ldots, d.$$

Banach fixed point theorem. Let (X, d) be a complete metric space, then each contraction map $T : X \to X$ has unique fixed point.

Remark 2. The matrix preliminaries are used in the stabilization analysis of the step-size policy Γ_t. The Banach fixed point theorem plays a crucial role in the proving Theorem 1, where the complete metric space consists of sequences, and the operator T is a discrete analogy of the integral operator in the continuous time dynamical system. Uniqueness means that the sequence (generated by ADAGRAD) converging to the saddle point under discussion is unique.

3 Analysis and results

In this section we provide the results that (ADANORM), (ADADIAG) and (ADAFULL) avoid saddle points from almost every initial condition. Our approach consists of three parts:

- The sequence of adaptive matrices Γ_t, $t = 1, 2, \ldots$, converges to a (strictly) positive-definite matrix.
- The analysis of the local structure of the ADAGRAD iterative dynamics based on the stabilization of the method's preconditioner.
- Prove the local stable manifold theorem for the ADAGRAD dynamics and then extend the result to global.

We will elaborate on this in the following subsections.

3.1 Stabilization of the preconditioner

Proposition 1. *Let Γ_t be one of adaptive step-size policies of* (AdaNorm),(AdaDiag) *or* (AdaFull)*. Then the following statements hold:*

1. *For each initial point $\mathbf{x}_0$, the eigenvalues of Γ_t converges to strictly positive numbers, i.e.,*
$$\lim_{t\to\infty} \lambda_i(\Gamma_t) > 0 \text{ for all } i = 1, \dots, d.$$
2. *For each sequence $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ generated by* AdaGrad*, the sum of square of gradient norms is finite, i.e.,*
$$\sum_{t=0}^{\infty} \|\nabla f(\mathbf{x}_t)\|^2 < \infty.$$
3. *The limit of $\{\Gamma_t\}_{t\in\mathbb{N}}$ exists and in particular, the limit is positive definite, i.e.,*
$$\lim_{t\to\infty} \Gamma_t = \Gamma \text{ with } \Gamma \succ 0.$$

Remark 3. An immediate consequence of the above stablization result is that the AdaGrad algorithms under study have a gradient decay rate of $\mathcal{O}(1/T)$. In particular, by Part 2 of Proposition 1 and the construction of the AdaGrad preconditioning matrix (see also Propositions A.1–A.3 in Appendix A), we immediately get that

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(X_t)\|^2 \leq \frac{L}{2}\frac{f(X_1) - \inf f}{T} = \mathcal{O}\left(\frac{1}{T}\right)$$

In turn, this yields the rate $\min_{t=1,\dots,T} \|\nabla f(X_t)\|^2 = O(1/T)$ or, if we pick $\bar{X}_T$ uniformly at random from $X_1, \dots, X_T$, we get $\mathbb{E}[\|\nabla f(\bar{X}_T)\|^2] = \mathcal{O}(1/T)$.

In the first statement of the above proposition, we briefly regard Γ_t as the product of the adaptive scalar and the identity matrix. In the proof of the proposition, we show the stabilization of Γ_t and summability of $\|\nabla f(\mathbf{x})\|^2$ independently for three adaptive policies. To provide some intuition, we will below sketch the main idea of the proof for (AdaNorm), the proof of the other two AdaGrad methods following the same strategy with additional techniques from theory of matrix analysis.

In what follows, we use the notation in the Appendix, i.e., we denote

$$\gamma_t = \frac{1}{\sqrt{\delta_0^2 + \sum_{s=0}^{t} \|\nabla f(\mathbf{x}_s)\|^2}},$$

to emphasize that γ_t is the *scalar* step-size in (AdaNorm).

We begin by noting that, since γ_t is decreasing and bounded from below, its limit exists, i.e.,

$$\lim_{t\to\infty} \gamma_t = \inf_{t\in\mathbb{N}} \gamma_t = \gamma_\infty \geq 0.$$

The proof is completed by contradiction. Assume that $\gamma_\infty = 0$. Then by the fact that f is smooth, we have

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2.$$

By the update rule of (AdaNorm), we further have

$$\mathbf{x}_{t+1} - \mathbf{x}_t = -\gamma_t \nabla f(\mathbf{x}_t). \tag{1}$$

Thus, by rearraning the descent inequality, we obtain the uper bound:

$$f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) + \frac{\gamma_t}{2}[L\gamma_t - 1]\|\nabla f(\mathbf{x}_t)\|^2.$$

Summing over all the terms for $t = 1, \dots, T$, and using the fact that $f(\mathbf{x}_T) \geq \inf f(\mathbf{x}_t)$, we have

$$\frac{1}{2}\sum_{t=0}^{T} \gamma_t \|\nabla f(\mathbf{x}_t)\|^2 \leq f(\mathbf{x}_1) - \inf_t f(\mathbf{x}_t) + \sum_{t=0}^{T} \frac{\gamma_t}{2}[L\gamma_t - 1]\|\nabla f(\mathbf{x}_t)\|^2.$$

Since we assume that $\gamma_t \to 0$ as $t \to \infty$, there must exist some t_0 such that $L\gamma_t - 1 < 0$ for all $t > t_0$. Therefore, the right hand side of the above inequality is finite because $\sum_{t=0}^{T} \frac{\gamma_t}{2}[L\gamma_t - 1]\|\nabla f(\mathbf{x}_t)\|^2$ spikes at $T = t_0$, and thus we have

$$\frac{1}{2}\sum_{t=0}^{\infty} \gamma_t \|f(\mathbf{x}_t)\|^2 < +\infty.$$

However, we can also have the following lower bound inequality by direct calculation in Appendix, i.e.,

$$\frac{1}{2\gamma_T} - \frac{\delta_0}{2} \leq \frac{1}{2}\sum_{t=0}^{T} \gamma \|\nabla f(\mathbf{x}_t)\|^2.$$

Again by our assumption that $\lim_{t\to\infty} \gamma_t = 0$, letting $t \to \infty$, we conclude that,

$$\infty \leq \frac{\delta_0}{2} + \sum_{t=0}^{\infty} \gamma_t \|\nabla f(\mathbf{x}_t)\|^2 < \infty,$$

a contradiction which yields the desired result.

Remark 4. The proof of Proposition 1 for the matrix-based variants of AdaGrad follows the same template. Specifically, the first step is to show that the method's preconditioning matrix is non-increasing in the Löwner order, i.e., $\Gamma_t \succcurlyeq \Gamma_{t+1}$ for all t. Then, by Weyl's monotonicity theorem, this result subsequently translates to the eigenvalues of Γ_t, i.e., $\lambda_i(\Gamma_t) \geq \lambda_i(\Gamma_{t+1})$ for all $i = 1, \dots, d$ (where $\lambda_i(\cdot)$ denotes the i-th eigenvalue of the matrix in question). In view of this, by invoking a similar series of steps based on the descent inequality and an eigenvalue-per-eigenvalue decomposition, it can be shown that the limit of each $\lambda_i(\Gamma_t)$

is strictly positive, which in turn can be used to show that Γ_t converges to a (strictly) positive-definite matrix and the sum of the gradient norms of f is finite. For the precise statements and proofs, we refer the reader to Propositions A.2 and A.3 in Appendix A.

3.2 Local structure of the ADAGRAD dynamics

The first essential technique in proving saddle avoidance of ADAGRAD algorithms is the following. Since the increment of each iterate is the product of a preconditioned matrix and gradient vector, all three ADAGRAD algorithms can be decomposed into a stabilized matrix and a residual term.

More precisely, all of the above algorithms have the following form:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \Gamma_t \nabla f(\mathbf{x}_t) \tag{2}$$

where Γ_t is a sequence of positive definite matrices that converge to a symmetric positive definite matrix Γ, and we assume that

$$\|\Gamma\|_2 \leq \|\Gamma_0\|_2 \leq \frac{1}{\delta_0} \leq \frac{1}{L}.$$

Since we have assumed that Γ_t has the limit matrix Γ, and then we can write

$$\Gamma_t = \Gamma + \Gamma_t - \Gamma.$$

Thus, the algorithm (2) can be written as

$$\begin{aligned}\mathbf{x}_{t+1} &= \mathbf{x}_t - \Gamma_t \nabla f(\mathbf{x}_t) \\ &= \mathbf{x}_t - (\Gamma + \Gamma_t - \Gamma)\nabla f(\mathbf{x}_t) \\ &= \mathbf{x}_t - \Gamma \nabla f(\mathbf{x}_t) - (\Gamma_t - \Gamma)\nabla f(\mathbf{x}_t)\end{aligned}$$

Without loss of generality, we assume that $\mathbf{0}$ is a strict saddle point of f, i.e., $\nabla f(\mathbf{0}) = \mathbf{0}$, then the Taylor expansion of f at $\mathbf{0}$ is the following

$$\begin{aligned}\nabla f(\mathbf{x}) &= \nabla f(\mathbf{0}) + \nabla^2 f(\mathbf{0})\mathbf{x} + \theta(\mathbf{x}) \\ &= \nabla^2 f(\mathbf{0})\mathbf{x} + \theta(\mathbf{x}).\end{aligned}$$

With the Taylor expansion of $\nabla f(\mathbf{x})$ in a neighborhood of $\mathbf{0}$, we replace the first $\nabla f(\mathbf{x}_t)$ with

$$\nabla f(\mathbf{x}_t) = \nabla^2 f(\mathbf{0})\mathbf{x}_t + \theta(\mathbf{x}_t),$$

provided that $\mathbf{x}_t$ is taken from a small neighborhood of $\mathbf{0}$ where the Taylor expansion is performed. We have an equivalent expression of the dynamical system (2) through the following calculation.

$$\begin{aligned}\mathbf{x}_{t+1} &= \mathbf{x}_t - \Gamma_t \nabla f(\mathbf{x}_t) \\ &= \mathbf{x}_t - \Gamma \left(\nabla^2 f(\mathbf{0})\mathbf{x}_t + \theta(\mathbf{x}_t)\right) - (\Gamma_t - \Gamma)\nabla f(\mathbf{x}_t) \\ &= \mathbf{x}_t - \Gamma \nabla^2 f(\mathbf{0})\mathbf{x}_t - \Gamma\theta(\mathbf{x}_t) - (\Gamma_t - \Gamma)\nabla f(\mathbf{x}_t) \\ &= \left(I - \Gamma \nabla^2 f(\mathbf{0})\right)\mathbf{x}_t - \Gamma\theta(\mathbf{x}_t) - (\Gamma_t - \Gamma)\nabla f(\mathbf{x}_t)\end{aligned}$$

We denote the non-linear part of the above dynamical system by $\eta(t, \mathbf{x})$, i.e.,

$$\eta(t, \mathbf{x}) = -\Gamma\theta(\mathbf{x}) - (\Gamma_t - \Gamma)\nabla f(\mathbf{x}). \tag{3}$$

In Lyapunov-Perron method, the Lipschitz type condition of the whole remainder is an crucial property that enable us to prove the existence of local stable manifold. By a Taylor expansion, we trivially get $\theta(\mathbf{x}) = \nabla f(\mathbf{x}) - \nabla^2 f(\mathbf{0})\mathbf{x}$ so the differential of $\theta(\mathbf{x})$ becomes

$$D\theta(\mathbf{x}) = D(\nabla f(\mathbf{x}) - \nabla^2 f(\mathbf{0})\mathbf{x}) = \nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{0}),$$

with the Lipschitzness of $\theta(\mathbf{x})$ being a consequence of the boundedness of $D\theta(\mathbf{x})$. Sice the gradient $\nabla f(\mathbf{x})$ is Lipschitz by assumption, the other part of $\eta(t, \mathbf{x})$ satisfies the Lipschitz type condition as long as t is large enough since the norm of $\Gamma_t - \Gamma$ becomes arbitrarily small as t goes to infinity. The formal statement is provided in the following proposition and we defer the detailed proof to Appendix B.

Proposition 2. *By the definition of $\eta(t, \mathbf{x})$ in (3), we have that for any $\epsilon > 0$, there exist a neighborhood $\mathbb{B}$ of $\mathbf{0}$ and some $t_0 \in \mathbb{N}$, such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{B}$ and $t \geq t_0$, we have*

$$\|\eta(t, \mathbf{x}) - \eta(t, \mathbf{y})\| \leq \epsilon \|\mathbf{x} - \mathbf{y}\|$$

3.3 ADAGRAD avoids saddle points

Given the properties of the remainder $\eta(t, \mathbf{x})$, we are now ready to state the local stable-manifold theorem corresponding to the ADAGRAD family of algorithms.

Theorem 1. *Suppose $1 \geq \lambda_1 \geq \ldots \geq \lambda_s > 0 > \lambda_{s+1} \geq \ldots \geq \lambda_d$, H is a diagonal matrix of the form:*

$$H = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix}.$$

Suppose that for any $\epsilon > 0$, there exists a ball centering at $\mathbf{0}$ with radius δ, such that the mapping $\eta(t, \mathbf{x})$ satisfes

- $\eta(t, \mathbf{0}) = \mathbf{0}$,
- $\|\eta(t, \mathbf{x}) - \eta(t, \mathbf{y})\| \leq \epsilon \|\mathbf{x} - \mathbf{y}\|$ *for all* $\mathbf{x}, \mathbf{y} \in \mathbb{B}(\delta)$.

Suppose further that $\gamma(t, \mathbf{x})$ is a function satisfying

$$0 < c \leq \gamma(t, \mathbf{x}) \leq \frac{1}{\lambda_1}.$$

Then the dynamical system

$$\mathbf{x}_{t+1} = (I - \gamma(t, \mathbf{x}_0)H)\mathbf{x}_t + \eta(t, \mathbf{x}_t) \tag{4}$$

has an s-dimensional stable manifold at $\mathbf{0}$. In particular, stable manifold exists for the case when $\gamma(t, \mathbf{x}) = 1$, i.e., the dynamical system is

$$\mathbf{x}_{t+1} = (I - H)\mathbf{x}_t + \eta(t, \mathbf{x}_t).$$

By an s-dimensional manifold, we mean a set that can be represented as a graph of a function. The scheme of proving existence of stable manifold for a dynamical system is called Lyapunov-Perron method, which originated from the study of structure stability of dynamical system defined by ordinary differential equations. The goal of all effort is to show that if the dynamical system converges to a unstable fixed point (in our settings, it means a saddle point) with an initial point $\mathbf{x}_0$, then this initial point $\mathbf{x}_0$ lies on the graph of some function from the stable space to unstable subspace with respect to the eigenspace decomposition of the linearization of the dynamical system. We give a quick review of the Lyapunov-Perron method for continuous time dynamical system. For a detailed study of Lyapunov-Perron method, we recommend [31] as a reference. Consider the dynamical system defined by

$$\frac{d\mathbf{x}}{dt} = A(t)\mathbf{x} + R(t, \mathbf{x}) \tag{5}$$

where $A(t)$ is a time-dependent matrix. If the solution $u(t, \mathbf{x}_0)$ generated by the dynamical system with some initial condition $\mathbf{x}_0$ converges to an unstable fixed point, then it must hold that for the integral operator T:

$$Tu(t, \mathbf{x}_0) = U(t)\mathbf{x}_0 + \int_0^t U(t-s)R(s, u(s, \mathbf{x}_0))ds - \int_t^\infty V(t-s)R(s, u(s, \mathbf{x}_0))ds,$$

$u(t, \mathbf{x}_0)$ is a fixed point of T. We will skip the discussion on $U(t)$ and $V(t)$ since they play no role other than giving intuition on the form of the discrete version of T. The Banach space consists of curves converging to the fixed point and having initial points whose stable components are all equal. Banach fixed point theorem ensures the existence and uniqueness of local stable manifold of the dynamical system (5). The main challenge of proving Theorem 1 is to formulate the discrete and adaptive version of operator T and then to show that T is a contraction map on the space of sequences converging to saddle points.

The reason for which one cannot treat the dynamical system arising from ADAGRAD algorithms as straightforwardly as the previous works is because the step-size is a matrix that involves all the historic iterations. Our main claim is the following: suppose the sequence $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ is generated by dynamical system of ADAGRAD algorithm and this sequence converges to a saddle point, WLOG, $\mathbf{0}$. Then we only have to focus on the sequence of step-size matrices $\{\Gamma_t\}_{t\in\mathbb{N}}$ which is generated by the iterations. We will treat $\{\Gamma_t\}_{t\in\mathbb{N}}$ as a fixed sequence that is pre-generated, furthermore, this sequence actually depends only on the initial point $\mathbf{x}_0$, so whenever necessary, we use $\Gamma_t(\mathbf{x}_0)$ to emphasize this property. Theorem 1 implies a more general result: for any sequence $\{\Gamma_t(\mathbf{x}_0)\}_{t\in\mathbb{N}}$, if the initial condition $\mathbf{x}_0$ makes the sequence $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ converge to a saddle point, then the initial condition $\mathbf{x}_0$ lies on an s-dimensional manifold which is a graph of a function from the stable space to unstable space.

To obtain a full discrete analogy of the solution $u(t, \mathbf{x}_0)$ in the continuous case, we regard the sequence $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ a sequence of functions $\{\mathbf{x}_t(\mathbf{x}_0)\}_{t\in\mathbb{N}}$ of the initial condition $\mathbf{x}_0$. This notion is essential in the proof of Theorem 1. The discrete version of T acting on space of sequences is as follows:

$$(T\mathbf{x})_{t+1} = B(t, 0)\mathbf{x}_0^+ + \sum_{i=0}^{t} B(t, i+1)\eta^+(i, \mathbf{x}_0, \mathbf{x}_i) - \sum_{i=0}^{\infty} C(t+1+i, t+1)^{-1}\eta^-(t+1+i, \mathbf{x}_0, \mathbf{x}_{t+1+i})$$

where the definition of $B(t, 0)$, $C(t+1+i, t+1)$ will be elaborated in Appendix. T transform a sequence $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ to a new sequence $\{(T\mathbf{x})_t\}_{t\in\mathbb{N}}$ and locally we will focus on the space of sequences converging to the saddle point. The main ingredients of showing the existence of local stable manifold consist of the following:

- The transformed sequence $\{(T\mathbf{x})_t\}_{t\in\mathbb{N}}$ converges to the fixed point $\mathbf{0}$ as long as $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ does, (Proposition C.1);
- The sequences $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ converging to $\mathbf{0}$ whose initial points have the same stable component, say $\mathbf{x}_0^+ = \mathbf{a}$, form a complete metric space, denoted as $X(\mathbf{a}, \mathbf{0})$, (Lemma C.4);
- The operator T is a contraction mapping acting on $X(\mathbf{a}, \mathbf{0})$, (Lemma C.6).

Therefore, applying Banach Fixed Point Theorem, we can conclude that there exists a unique sequence converging to $\mathbf{0}$ for each fixed stable component of the initial condition $\mathbf{x}_0$, and this sequence gives the unique correspondence between the stable-unstable components of $\mathbf{x}_0$, and this can be written as $\mathbf{x}_0^- = \phi(\mathbf{x}_0^+)$.

Note that the dynamical system of ADAGRAD algorithms is conjugate to the dynamical system above. By assumption, the Hessian of $f(\mathbf{x})$ at isolated strict saddle point $\mathbf{0}$ is symmetric and diagonalizable. For a symmetric positive definite matrix Γ, the matrix

$$\Gamma^{-\frac{1}{2}}\Gamma\nabla^2 f(\mathbf{0})\Gamma^{\frac{1}{2}} = \Gamma^{\frac{1}{2}}\nabla^2 f(\mathbf{0})\Gamma^{\frac{1}{2}}$$

has the same eigenvalues as $\Gamma\nabla^2 f(\mathbf{0})$, while $\Gamma^{\frac{1}{2}}\nabla^2 f(\mathbf{0})\Gamma^{\frac{1}{2}}$ has the same number of positive and negative eigenvalues as $\nabla^2 f(\mathbf{0})$ does. Thus $\Gamma^{\frac{1}{2}}\nabla^2 f(\mathbf{0})\Gamma^{\frac{1}{2}}$ is diagonalizable and this implies that $\Gamma\nabla^2 f(\mathbf{0})$ is also diagonalizable. Moreover, the number of positive and negative eigenvalues of $\Gamma\nabla^2 f(\mathbf{0})$ agree with that of $\nabla^2 f(\mathbf{0})$. Suppose that the diagonalization of $\Gamma\nabla^2 f(\mathbf{0})$ can be completed by the linear transformation matrix Q, i.e., H is a diagonal matrix and $\Gamma\nabla^2 f(\mathbf{0}) = Q^{-1}HQ$, then we have

$$\mathbf{x}_{t+1} = (I - Q^{-1}HQ)\mathbf{x}_t + \eta(t, \mathbf{x}_t) \tag{6}$$
$$= Q^{-1}(I - H)Q\mathbf{x}_t + \eta(t, \mathbf{x}_t) \tag{7}$$

which is equavalent to

$$Q\mathbf{x}_{t+1} = (I - H)Q\mathbf{x}_t + Q\eta(t, \mathbf{x}_t).$$

If we denote $\mathbf{y}_t = Q\mathbf{x}_t$, we have a dynamical system interms of $\mathbf{y}_t$ as follows,

$$\mathbf{y}_{t+1} = (I - H)\mathbf{y}_t + Q\eta(t, Q^{-1}\mathbf{y}_t). \tag{8}$$

We leave the complete argument showing that $Q\eta(t, Q^{-1}\mathbf{y}_t)$ satisfies the Lipschitzness condition of Theorem 1 and the existence of local stable manifold of the dynamical system of (8) and that of $\mathbf{x}_t$ to Appendix (Lemma C.7).

Note that the existence of local stable manifold of ADAGRAD algorithm implies that the set of initial points converging to saddle point is of measure zero, but to extend the result to the whole Euclidean space, we need the following proposition assuring that the mapping defined by ADAGRAD algorithms are diffeomorphism.

Proposition 3. *There exists a positive number $\delta_0 \geq L$ for Lipschitz number L, such that* (ADANORM)*,* (ADADIAG) *and* (ADAFULL) *are local diffeomorphisms on $\mathbb{R}^d$.*

The main difficulty in understanding that ADAGRAD algorithms are diffeomorphisms comes from the fact that these algorithms depend on all iterations. However, it suffices to show that for each iterate, the map defined by the algorithm is a diffeomorphism, i.e., for the iteration

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \Gamma_t\nabla f(\mathbf{x}_t),$$

we need to show that map

$$\varphi(\mathbf{x}) \stackrel{\text{def}}{=} \mathbf{x} - \Gamma_t\nabla f(\mathbf{x}),$$

where Γ_t is the preconditioned step matrix from (ADAFULL), (ADANORM), or (ADADIAG), is a diffeomorphism. Take (ADAFULL) for example,

$$\Gamma_t = \left(\delta_0^2 I + \sum_{s=0}^{t}\nabla f(\mathbf{x}_s)\nabla f(\mathbf{x}_s)^T\right)^{-\frac{1}{2}},$$

to show that $\varphi(\mathbf{x})$ is a diffeomorphism on the t'th iterate, we split Γ_t as follows:

$$\Gamma_t = \left(\delta_0^2 I + S + \nabla f(\mathbf{x}_t)\nabla f(\mathbf{x}_t)^T\right)^{-\frac{1}{2}}$$

where

$$S = \sum_{s=0}^{t-1}\nabla f(\mathbf{x}_s)\nabla f(\mathbf{x}_s)^T.$$

Note that only $\nabla f(\mathbf{x}_t)\nabla f(\mathbf{x}_t)^T$ depends on $\mathbf{x}_t$, and thus the rest terms ξI and S can be treated as constants in proving that the t'th iterate is a diffeomorphism. To be precise, the mapping

$$\varphi(\mathbf{x}) = \mathbf{x} - \left(\delta_0^2 I + S + \nabla f(\mathbf{x})\nabla f(\mathbf{x})^T\right)^{-\frac{1}{2}}\nabla f(\mathbf{x})$$

is expected to be a diffeomorphism as long as δ_0 is properly tuned. The standard approach of showing $\varphi(\mathbf{x})$ to be a diffeomorphism is to compute the Jacobian matrix of $\varphi(\mathbf{x})$, and the diffeomorphism follows if the determinant of $D\varphi(\mathbf{x})$ is positive by Inverse Function Theorem.

Now, regarding the calculation of the Jacobian matrix, (ADANORM) is fundamentally different from the other two variants. The reason for this is that in (ADANORM), the preconditioned matrix Γ_t is essentially a scalar function, i.e.,

$$\Gamma_t = \frac{I}{\sqrt{\delta_0^2 + \sum_{s=0}^{t}\|\nabla f(\mathbf{x}_s)\|^2}},$$

so the Jacobian matrix of $\varphi(\mathbf{x})$ can be computed explicitly if we write it as

$$\varphi(\mathbf{x}) = \mathbf{x} - \frac{1}{\sqrt{\delta_0^2 + S + \|\nabla f(\mathbf{x})\|^2}}\nabla f(\mathbf{x})$$

where $S = \sum_{s=0}^{t-1}\|\nabla f(\mathbf{x}_s)\|^2$.

By contrast, the diffeomorphism arguments for (ADAFULL) and (ADADIAG) require a different set of arguments. Especially for (ADAFULL), it is more convenient to analyze the determinant of Jacobian according to the condition Γ_t satisfies. Since the essential strategy is to show that the determinant of $D\varphi(\mathbf{x})$ is positive, we have the intuition that once ξ is taken to be large, the determinant of $D\varphi(\mathbf{x})$ can be arbitrarily close to 1. The assumptions on the boundedness of (higher-order) partial derivatives of f guarantee that the determinant of $D\varphi(\mathbf{x})$ is close to 1 as long as ξ is large. The detailed analysis is provided in Appendices A and B.

Combining the new Stable Manifold Theorem and Proposition 3, we finally obtain:

Theorem 2. (ADANORM)*,* (ADADIAG) *and* (ADAFULL) *avoid strict saddle points from almost any initialization.*

It is straightforward to conclude the saddle avoidance if the saddle points are assumed to be isolated. The proof of the

last theorem extends the saddle avoidance guarantee from the countable saddle points to uncountable. We leave the full proof to the appendix.

3.4 Comparison with related results

We conclude this section with a discussion of a series of related results in the literature. To put things in context, it is important to first describe in detail the type of statements obtained in the "escape" literature – to which the work of Staib et al. [33] belongs.

To begin with, the typical "*escape result*" – and, in particular, the work of Staib et al. [33] – follows the template below:

1. Fix some probability threshold $\delta > 0$.
2. Run a stochastic gradient-based algorithm for a predetermined number of iterations T with a fixed step-size η (both η and T given as a function of δ).
3. Then, with probability at least $1 - \delta$, *at least one* of the iterates produced by the algorithm under study will be close to a second-order stationary point (and hence away from all saddle points). [The existence of a second-order stationary point is assumed by default].

In view of this, the first important difference with results of this type is that escape results only guarantee that *some* iterate of the algorithm under study will be close to a second-order stationary point. In particular, the work of Staib et al. [33] on adaptive algorithms leaves open the possibility that the algorithm may revisit a saddle point infinitely often, and it cannot exclude the event that the limit points of the algorithm may contain strict saddles. By contrast, our paper rules out exactly this behavior, so it is complementary to the analysis of Staib et al. [33] in this regard.

The second fundamental difference with the escape literature is that it is typically assumed therein that the algorithm under study is subject to persistent noise lower-bounded along any direction (see for example Definition 4.2 in Staib et al. [33] and the discussion right after). More concretely, this means that the iterates of the algorithms studied in this literature are subject to continual random shocks, a fact which greatly facilitates the "escape" from saddle points (in the sense described above). This is readily seen in the bounds for T given by Staib et al. [33], which are of the form $T = O(1/\nu^4)$, with $\nu > 0$ denoting the minimum noise level along any direction. [In particular, it is not possible to get deterministic results by setting the variance of the noise to 0 in Staib et al. [33].]

These differences are also reflected in the divergent tools and techniques required to establish avoidance results compared to the escape literature. Specifically, thanks to the persistent noise in the setup of standard escape results, the analysis does not require any delicate center stable manifold arguments. However, these arguments are indispensable for excluding strict saddles as limit points of the underlying dynamics, a fact which serves to explain the gulf in techniques between Staib et al. [33] and our paper. As far as we are aware, there is no comparable stable manifold theorem for adaptive algorithms in the literature.

In regards to the stochastic setting, to the best of our knowledge, the only avoidance results for stochastic algorithms are [7, 14, 25, 30]. These stochastic avoidance results concern exclusively *non-adaptive* algorithms with persistent noise. Obtaining avoidance results for stochastic adaptive methods would be a very fruitful direction for future research, but not one which can be attacked at this stage.

4 Concluding remarks

In this paper we examined the saddle-point avoidance properties of the ADAGRAD family of methods (ranging from scalar to full-matrix preconditioning), and we showed that all policies under study avoid saddle points from almost any initial condition. A major challenge in our analysis is that the dynamical system arising from ADAGRAD is not only time-dependent, but filtration-dependent. Nonetheless, after an extensive stabilization analysis for the method's preconditioner, the induced dynamical system reduces to a form that enables us to apply the Lyapunov-Perron method to prove a new stable manifold theorem for ADAGRAD. These techniques and results not only advance our understanding of adaptive gradient methods, but they also initiate the study of saddle-point avoidance results for other methods like adaptive mirror descent; we leave this to future work.

Acknowledgments

KA is grateful for financial support by the Swiss National Science Foundation (SNSF) under grant number 200021_205011. PM is grateful for financial support by the French National Research Agency (ANR) in the framework of the "Investissements d'avenir" program (ANR-15-IDEX-02), the LabEx PERSYVAL (ANR-11-LABX-0025-01), MIAI@Grenoble Alpes (ANR-19-P3IA-0003), and the bilateral ANR-NRF grant ALIAS (ANR-19-CE48-0018-01). GP acknowledges that this research/project is supported in part by the National Research Foundation, Singapore under its AI Singapore Program (AISG Award No: AISG2-RP-2020-016), NRF 2018 Fellowship NRF-NRFF2018-07, NRF2019-NRF-ANR095 ALIAS grant, grant PIE-SGP-AI-2020-01, AME Programmatic Fund (Grant No. A20H6b0151) from the Agency for Science, Technology and Research (A*STAR) and Provost's Chair Professorship grant RGEPPV2101. XW acknowledges Grant 202110458 from SUFE and support from the Shanghai Research Center for Data Science and Decision Technology.

References

[1] Antonakopoulos, K. and Mertikopoulos, P. Adaptive first-order methods revisited: Convex optimization without Lipschitz requirements. In *NeurIPS '21: Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021.

[2] Antonakopoulos, K., Belmega, E. V., and Mertikopoulos, P. Adaptive extra-gradient methods for min-max optimization and games. In *ICLR '21: Proceedings of the 2021 International Conference on Learning Representations*, 2021.

[3] Antonakopoulos, K., Pethick, T., Kavis, A., Mertikopoulos, P., and Cevher, V. Sifting through the noise: Universal first-order methods for stochastic variational inequalities. In *NeurIPS '21: Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021.

[4] Antonakopoulos, K., Vu, D. Q., Cevher, V., Levy, K. Y., and Mertikopoulos, P. UnderGrad: A universal black-box optimization method with almost dimension-free convergence rate guarantees. In *ICML '22: Proceedings of the 39th International Conference on Machine Learning*, 2022.

[5] Benaïm, M. and Hirsch, M. W. Asymptotic pseudotrajectories and chain recurrent flows, with applications. *Journal of Dynamics and Differential Equations*, 8(1):141–176, 1996.

[6] Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[7] Brandière, O. and Duflo, M. Les algorithmes stochastiques contournent-ils les pièges ? *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 32(3):395–427, 1996.

[8] Daneshmand, H., Kohler, J., Lucchi, A., and Hofmann, T. Escaping saddles with stochastic gradients. In *ICML '18: Proceedings of the 35th International Conference on Machine Learning*, 2018.

[9] Du, S. S., Jin, C., Lee, J. D., Jordan, M. I., Póczos, B., and Singh, A. Gradient descent can take exponential time to escape saddle points. In *NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

[10] Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

[11] Flokas, L., Vlatakis-Gkaragkounis, E. V., and Piliouras, G. Efficiently avoiding saddle points with zero order methods: No gradients required. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[12] Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle points – Online stochastic gradient for tensor decomposition. In *COLT '15: Proceedings of the 28th Annual Conference on Learning Theory*, 2015.

[13] Horn, R. A. and Johnson, C. R. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.

[14] Hsieh, Y.-P., Mertikopoulos, P., and Cevher, V. The limits of min-max optimization algorithms: Convergence to spurious non-critical sets. In *ICML '21: Proceedings of the 38th International Conference on Machine Learning*, 2021.

[15] Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. How to escape saddle points efficiently. In *ICML '17: Proceedings of the 34th International Conference on Machine Learning*, 2017. URL http://proceedings.mlr.press/v70/jin17a.html.

[16] Kavis, A., Levy, K. Y., Bach, F., and Cevher, V. UnixGrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[17] Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. https://arxiv.org/abs/1412.6980, 2014.

[18] Latif, A. *Topics in Fixed Point Theory*. Springer, 2014.

[19] LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436–444, 2015.

[20] Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. Gradient descent only converges to minimizers. In *COLT '16: Proceedings of the 29th Annual Conference on Learning Theory*, 2016.

[21] Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I., and Recht, B. First-order methods almost always avoid strict saddle points. *Mathematical Programming*, 176 (1):311–337, February 2019.

[22] Levy, K. Y., Yurtsever, A., and Cevher, V. Online adaptive methods, universality and acceleration. In *NeurIPS '18: Proceedings of the 32nd International Conference of Neural Information Processing Systems*, 2018.

[23] Li, X. and Orabona, F. On the convergence of stochastic gradient descent with adaptive stepsizes. In *AISTATS '19: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

[24] McMahan, H. B. and Streeter, M. Adaptive bound optimization for online convex optimization. In *COLT '10: Proceedings of the 23rd Annual Conference on Learning Theory*, 2010.

[25] Mertikopoulos, P., Hallak, N., Kavis, A., and Cevher, V. On the almost sure convergence of stochastic gradient descent in non-convex problems. In *NeurIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[26] Nesterov, Y. A method for unconstrained convex minimization problem with the rate of convergence $\mathcal{O}(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269(543-547), 1983.

[27] Panageas, I. and Piliouras, G. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. In *ITCS '17: Proceedings of the 8th Conference on Innovations in Theoretical Computer Science*, 2017.

[28] Panageas, I., Piliouras, G., and Wang, X. First-order methods almost always avoid saddle points: The case of vanishing step-sizes. In *NeurIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[29] Panageas, I., Piliouras, G., and Wang, X. Multiplicative weights updates as a distributed constrained optimization algorithm: Convergence to second-order stationary points almost always. In *International Conference on Machine Learning*, pp. 4961–4969. PMLR, 2019.

[30] Pemantle, R. Nonconvergence to unstable points in urn models and stochastic aproximations. *Annals of Probability*, 18(2):698–712, April 1990.

[31] Perko, L. *Differential Equations and Dynamical Systems*. Springer, 2001.

[32] Reddi, S. J., Kale, S., and Kumar, S. On the convergence of Adam and beyond. In *ICLR '18: Proceedings of the 2018 International Conference on Learning Representations*, 2018.

[33] Staib, M., Reddi, S., Kale, S., Kumar, S., and Sra, S. Escaping saddle points with adaptive gradient methods. In *ICML '19: Proceedings of the 36th International Conference on Machine Learning*, 2019.

[34] Vu, D. Q., Antonakopoulos, K., and Mertikopoulos, P. Fast routing under uncertainty: Adaptive learning in congestion games with exponential weights. In *NeurIPS '21: Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021.

[35] Ward, R., Wu, X., and Bottou, L. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes, from any initialization. In *ICML '19: Proceedings of the 36th International Conference on Machine Learning*, 2019.

A Stabilization of the ADAGRAD preconditioners

In this section we shall investigate the asymptotic behaviour of the classical gradient descent algorithmic scheme run with different types of step-sizes. More precisely, in what follows we shall investigate the case of a scalar, diagonal and full matrix step-sizes.

For the sake of convenience we develop their respective analysis individually.

A.1 AdaGrad with scalar step-size

We start by investigating the particular case of the gradient descent,

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t \nabla f(\mathbf{x}_t) \tag{GD}$$

run with the adaptive scalar step-size policy:

$$\gamma_t = \frac{1}{\sqrt{\delta_0^2 + \sum_{s=0}^{t} \|\nabla f(\mathbf{x}_s)\|_*^2}} \tag{AdaNorm}$$

with $\delta_0^2 > 0$. The first result concerns the asymptotic stabilization of the adaptive step-size around a (strictly) positive value $\gamma_\infty > 0$.

Proposition A.1. *Assume that $f : \mathbb{R}^d \to \mathbb{R}$ is smooth and $\mathbf{x}_t$ are the iterates of (GD) run with the adaptive step-size policy (AdaNorm). Then, the following hold:*

1. *The (AdaNorm) step-size γ_t converges to some strictly positive value γ_∞, i.e.,*

$$\gamma_t \to \inf_{t \in \mathbb{N}} \gamma_t = \gamma_\infty > 0 \tag{A.1}$$

2. *The sequence $\{\|\nabla f(\mathbf{x}_t)\|_*^2\}_{t\in\mathbb{N}}$ is summable, i.e.,*

$$\sum_{t=0}^{+\infty} \|\nabla f(\mathbf{x}_t)\|_*^2 < +\infty \tag{A.2}$$

Proof. We shall start with the first property. Since γ_t is decreasing and bounded from below by 0, we have. that its limit exists and more precisely:

$$\lim_{t \to +\infty} \gamma_t = \inf_{t \in \mathbb{N}} \gamma_t = \gamma_\infty \geq 0$$

Assume that $\gamma_\infty = 0$. Then, by the fact that f is smooth, we can choose positive number $\beta > L$, and we have:

$$\begin{aligned}
f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{\beta}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
&= f(\mathbf{x}_t) - \gamma_t \|\nabla f(\mathbf{x}_t)\|_*^2 + \frac{\beta \gamma_t^2}{2} \|\nabla f(\mathbf{x}_t)\|_*^2 \\
&= f(\mathbf{x}_t) - \frac{1}{2}\gamma_t \|\nabla f(\mathbf{x}_t)\|_*^2 - \frac{1}{2}\gamma_t \|\nabla f(\mathbf{x}_t)\|_*^2 + \frac{\beta \gamma_t^2}{2} \|\nabla f(\mathbf{x}_t)\|_*^2 \\
&= f(\mathbf{x}_t) - \frac{1}{2}\gamma_t \|\nabla f(\mathbf{x}_t)\|_*^2 + \frac{\gamma_t}{2} \left[\beta \gamma_t - 1\right] \|\nabla f(\mathbf{x}_t)\|_*^2
\end{aligned}$$

which yields after rearranging,

$$\frac{1}{2}\gamma_t \|\nabla f(\mathbf{x}_t)\|_*^2 \leq f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) + \frac{\gamma_t}{2} \left[\beta \gamma_t - 1\right] \|\nabla f(\mathbf{x}_t)\|_*^2 \tag{A.3}$$

One the other hand, one may directly verify that the quantity $\frac{\gamma_t}{2} \left[\beta \gamma_t - 1\right] \|\nabla f(\mathbf{x}_t)\|_*^2$ becomes non-positive whenever $\gamma_t \leq 1/\beta$. Now, by the assumption that $\gamma_\infty = 0$ there exists some $t_0 \in \mathbb{N}$ such that:

$$\gamma_t \leq \frac{1}{\beta} \quad \text{for all} \quad t > t_0$$

So, after telescoping (A.3), we have:

$$\begin{aligned}\frac{1}{2}\sum_{t=0}^{T}\gamma_t\|\nabla f(\mathbf{x}_t)\|_*^2 &\leq f(\mathbf{x}_1) - \inf_t f(\mathbf{x}_t) + \sum_{t=0}^{T}\frac{\gamma_t}{2}\left[\beta\gamma_t - 1\right]\|\nabla f(\mathbf{x}_t)\|_*^2 \\ &= f(\mathbf{x}_1) - \inf_t f(\mathbf{x}_t) + \sum_{t=0}^{t_0}\frac{\gamma_t}{2}\left[\beta\gamma_t - 1\right]\|\nabla f(\mathbf{x}_t)\|_*^2 + \sum_{t=t_0+1}^{T}\frac{\gamma_t}{2}\left[\beta\gamma_t - 1\right]\|\nabla f(\mathbf{x}_t)\|_*^2 \\ &\leq f(\mathbf{x}_1) - \inf_t f(\mathbf{x}_t) + \sum_{t=0}^{t_0}\frac{\gamma_t}{2}\left[\beta\gamma_t - 1\right]\|\nabla f(\mathbf{x}_t)\|_*^2\end{aligned}$$

with the last inequality being obtained by the definition of t_0. We proceed by bounding the quantity $\frac{1}{2}\sum_{t=0}^{T}\gamma_t\|\nabla f(\mathbf{x}_t)\|_*^2$ from below. More precisely, we have:

$$\begin{aligned}\frac{1}{2}\sum_{t=0}^{T}\gamma_t\|\nabla f(\mathbf{x}_t)\|_*^2 &= \frac{1}{2}\sum_{t=0}^{T}\frac{\|\nabla f(\mathbf{x}_t)\|_*^2}{\sqrt{\delta_0^2 + \sum_{s=0}^{t}\|\nabla f(\mathbf{x}_s)\|_*^2}} \\ &\geq \frac{1}{2\sqrt{\delta_0^2 + \sum_{t=0}^{T}\|\nabla f(\mathbf{x}_t)\|_*^2}}\sum_{t=0}^{T}\|\nabla f(\mathbf{x}_t)\|_*^2 \\ &= \frac{1}{2\sqrt{\delta_0^2 + \sum_{t=0}^{T}\|\nabla f(\mathbf{x}_t)\|_*^2}}\left[\delta_0^2 + \sum_{t=1}^{T}\|\nabla f(\mathbf{x}_t)\|_*^2 - \delta_0^2\right] \\ &= \frac{\gamma_T}{2}\left[\frac{1}{\gamma_T^2} - \delta_0^2\right] \\ &= \frac{1}{2\gamma_T} - \frac{\gamma_T\delta_0^2}{2} \\ &\geq \frac{1}{2\gamma_T} - \frac{\delta_0}{2}\end{aligned}$$

with the last inequality being obtained by the fact that $\gamma_t \leq 1/\delta_0$ for all $t = 1, 2, \ldots$ So, summarizing the above estimations we get:

$$\frac{1}{2\gamma_T} \leq f(\mathbf{x}_1) - \inf_t f(\mathbf{x}_t) + \frac{\delta_0}{2} + \sum_{t=0}^{t_0}\frac{\gamma_t}{2}\left[\beta\gamma_t - 1\right]\|\nabla f(\mathbf{x}_t)\|_*^2 \tag{A.4}$$

Now, by letting $T \to +\infty$ we have that $\frac{1}{2\gamma_T} \to +\infty$, since we assumed that $\gamma_t \to 0$ and hence we get that:

$$+\infty \leq f(\mathbf{x}_1) - \inf_t f(\mathbf{x}_t) + \frac{\delta_0}{2} + \sum_{t=0}^{t_0}\frac{\gamma_t}{2}\left[\beta\gamma_t - 1\right]\|\nabla f(\mathbf{x}_t)\|_*^2 < +\infty \tag{A.5}$$

which is a contradiction. Hence, we readily get that $\gamma_\infty > 0$ and the result follows.

For the second claim, we have that:

$$\begin{aligned}\sum_{t=0}^{+\infty}\|\nabla f(\mathbf{x}_t)\|_*^2 &= \lim_{T\to+\infty}\sum_{t=0}^{T}\|\nabla f(\mathbf{x}_t)\|_*^2 \\ &= \lim_{T\to+\infty}\left[\frac{1}{\gamma_T^2} - \delta_0^2\right] \\ &= \frac{1}{\gamma_\infty} - \delta_0^2 \\ &< +\infty\end{aligned}$$

with the last strict inequality being obtained by the fact that $\gamma_\infty > 0$ being invoking our first claim and hence the result follows. ■

A.2 AdaGrad with diagonal adaptation

We next investigate the diagonal AdaGrad method. This is given by the following formula:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \Gamma_t \nabla f(\mathbf{x}_t) \tag{A.6}$$

where $\Gamma_t \in \mathbb{R}^{d\times d}$ is a sequence of matrices defined as the inverse square root of G_t:

$$\Gamma_t = G_t^{-\frac{1}{2}} = \left[\delta_0^2 I + \operatorname{diag}\left(\sum_{s=0}^{t} \nabla f(\mathbf{x}_s)\nabla f(\mathbf{x}_s)^\top\right)\right]^{-\frac{1}{2}} \tag{AdaDiag}$$

Lemma A.1. *Assume that $\mathbf{x}_t$ aer the iterates of AdaGrad algorithm with step-size policy* (AdaDiag)*. Then the following hold:*

1. *The sequence of matrices $\{\Gamma_t\}_{t\in\mathbb{N}}$ is non-increasing in the Löwner sense, i.e.,*

$$\Gamma_t \succcurlyeq \Gamma_{t+1} \text{ for all } t = 1, 2, ...$$

2. *The sequence of eigenvalues is non-increasing, i.e.,*

$$\lambda_i(\Gamma_t) \geq \lambda_i(\Gamma_{t+1})$$

Proof. We show the first claim. By definition of G_t in the (AdaDiag), we have that

$$G_{t+1} - G_t = \operatorname{diag}\left(\nabla f(\mathbf{x}_{t+1})\nabla f(\mathbf{x}_{t+1})^\top\right) \succcurlyeq 0$$

which means that G_t is a Löwner non-decreasing sequence, i.e.,

$$G_{t+1} \succcurlyeq G_t.$$

Therefore, by applying Löwner-Heinz inequality, we can have

$$G_{t+1}^{\frac{1}{2}} \succcurlyeq G_t^{\frac{1}{2}} \text{ for all } t = 1, 2, ...$$

and then

$$\Gamma_t \succcurlyeq \Gamma_{t+1}.$$

The second claim is straightforward since G_t is diagonal, thus Γ_t is diagonal, and the eigenvalues are the diagonal entries,i.e.,we have that

$$\lambda_i(\Gamma_t) \geq \lambda_i(\Gamma_{t+1}).$$

■

Proposition A.2. *Assume that $f : \mathbb{R}^d \to \mathbb{R}$ is smooth and $\mathbf{x}_t$ are the iterates run with with adaptive step-size policy* (AdaDiag)*. Then, the following hold:*

1. *The sequence of the eigenvalues $\{\lambda_i(\Gamma_t)\}_{t\in\mathbb{N}}$ converges to strictly positive value λ_i^∞ for all $i = 1, 2, ...$ i.e.,*

$$\lim_{t\to\infty} \lambda_i(\Gamma_t) = \inf_{t\in\mathbb{N}} \lambda_i(\Gamma_t) = \lambda_i^\infty > 0 \text{ for all } i = 1, 2, ...d$$

2. *The sequence $\{\|\nabla f(\mathbf{x}_t)\|_*^2\}_{t\in\mathbb{N}}$ is summable, i.e.,*

$$\sum_{t=0}^{\infty} \|\nabla f(\mathbf{x}_t)\|_*^2 < \infty.$$

Proof. Since $\{\Gamma_t\}_{t\in\mathbb{N}}$ is a decreasing sequence of diagonal positive definite matrices, the eigenvalues $\lambda_i(\Gamma_t)$ is bounded below by 0, i.e.,

$$\lim_{t\to\infty}\lambda_i(\Gamma_t)=\inf_{t\in\mathbb{N}}\lambda_i(\Gamma_t)=\lambda_i^\infty\geq 0$$

for some λ_i^∞. We will show that λ_i^∞ is actually strictly positive.

By the fact that f is smooth, we can choose $\beta > L$ and have:

$$\begin{aligned}
f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t)+\langle\nabla f(\mathbf{x}_t),\mathbf{x}_{t+1}-\mathbf{x}_t\rangle+\frac{\beta}{2}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2\\
&= f(\mathbf{x}_t)-\nabla f(\mathbf{x}_t)^\top\Gamma_t\nabla f(\mathbf{x}_t)+\frac{\beta}{2}\|\Gamma_t\nabla f(\mathbf{x}_t)\|^2\\
&= f(\mathbf{x}_t)-\frac{1}{2}\nabla f(\mathbf{x}_t)^\top\Gamma_t\nabla f(\mathbf{x}_t)-\frac{1}{2}\nabla f(\mathbf{x}_t)^\top[I-\beta\Gamma_t]\Gamma_t\nabla f(\mathbf{x}_t)\\
&= f(\mathbf{x}_t)-\frac{1}{2}\nabla f(\mathbf{x}_t)^\top\Gamma_t\nabla f(\mathbf{x}_t)+\frac{1}{2}\sum_{i=1}^d\left[\frac{\partial f}{\partial x_i}(\mathbf{x}_t)\right]^2(\beta\lambda_i^2(\Gamma_t)-\lambda_i(\Gamma_t)).
\end{aligned}$$

By rearranging, we have

$$\frac{1}{2}\nabla f(\mathbf{x}_t)^\top\Gamma_t\nabla f(\mathbf{x}_t)\leq f(\mathbf{x}_t)-f(\mathbf{x}_{t+1})+\frac{1}{2}\sum_{i=1}^d\left[\frac{\partial f}{\partial x_i}(\mathbf{x}_t)\right]^2(\beta\lambda_i^2(\Gamma_t)-\lambda_i(\Gamma_t)) \tag{A.7}$$

and then

$$\frac{1}{2}\sum_{t=0}^T\nabla f(\mathbf{x}_t)^\top\Gamma_t\nabla f(\mathbf{x}_t)\leq f(\mathbf{x}_1)-\inf_t f(\mathbf{x}_t)+\frac{1}{2}\sum_{t=0}^T\sum_{i=1}^d\left[\frac{\partial f}{\partial x_i}(\mathbf{x}_t)\right]^2(\beta\lambda_i^2(\Gamma_t)-\lambda_i(\Gamma_t)). \tag{A.8}$$

Since $\lambda_i(\Gamma_t)$ is assumed to approach 0 as $t\to\infty$, there must exist some t_0, such that for all $t > t_0$,

$$\beta\lambda_i^2(\Gamma_t)-\lambda_i(\Gamma_t)<0,$$

and this implies

$$\frac{1}{2}\sum_{t=0}^T\nabla f(\mathbf{x}_t)^\top\Gamma_t\nabla f(\mathbf{x}_t)\leq f(\mathbf{x}_1)-\inf_t f(\mathbf{x}_t)+\frac{1}{2}\sum_{t=0}^{t_0}\sum_{i=1}^d\left[\frac{\partial f}{\partial x_i}(\mathbf{x}_t)\right]^2(\beta\lambda_i^2(\Gamma_t)-\lambda_i(\Gamma_t))$$

for $T > t_0$.

On the other hand, the lower bound can be estimated as follows:

$$\begin{aligned}
\frac{1}{2}\sum_{t=0}^T\nabla f(\mathbf{x}_t)^\top\Gamma_t\nabla f(\mathbf{x}_t) &\geq \frac{1}{2}\sum_{t=0}^T\nabla f(\mathbf{x}_t)^\top\Gamma_T\nabla f(\mathbf{x}_t)\\
&= \frac{1}{2}\sum_{t=0}^T\sum_{i=1}^d\frac{\left[\frac{\partial f}{\partial x_i}(\mathbf{x}_t)\right]^2}{\sqrt{\delta_0^2+\sum_{s=1}^T\left[\frac{\partial f}{\partial x_i}(\mathbf{x}_s)\right]^2}}\\
&\geq \frac{1}{2\sqrt{\delta_0^2+\sum_{s=0}^T\left[\frac{\partial f}{\partial x_i}(\mathbf{x}_s)\right]^2}}\sum_{t=0}^T\left[\frac{\partial f}{\partial x_i}(\mathbf{x}_t)\right]^2
\end{aligned}$$

where i is the one with respect to $\lambda_i(\Gamma_t)$ that goes to 0 as $t\to\infty$.

Note that

$$\frac{1}{2}\sum_{t=0}^T\nabla f(\mathbf{x}_t)^\top\Gamma_t\nabla f(\mathbf{x}_t)\geq \frac{1}{2\sqrt{\delta_0^2+\sum_{s=0}^T\left[\frac{\partial f}{\partial x_i}(\mathbf{x}_s)\right]^2}}\sum_{t=0}^T\left[\frac{\partial f}{\partial x_i}(\mathbf{x}_t)\right]^2$$

$$
\begin{aligned}
&= \frac{1}{2\sqrt{\delta_0^2 + \sum_{s=0}^{T}\left[\frac{\partial f}{\partial x_i}(\mathbf{x}_s)\right]^2}}\left(\delta_0^2 + \sum_{t=0}^{T}\left[\frac{\partial f}{\partial x_i}(\mathbf{x}_t)\right]^2 - \delta_0^2\right)\\
&= \frac{\gamma_T}{2}\left(\frac{1}{\gamma_T^2} - \delta_0^2\right)\\
&= \frac{1}{2\gamma_T} - \frac{\gamma_T\delta_0^2}{2}\\
&\geq \frac{1}{\gamma_T} - \frac{\delta_0}{2}
\end{aligned}
$$

where we denote

$$
\gamma_T = \frac{1}{\sqrt{\delta_0^2 + \sum_{s=0}^{T}\left[\frac{\partial f}{\partial x_i}(\mathbf{x}_s)\right]^2}},
$$

and the last inequality comes from the fact that we can choose T such that $\gamma_T < \frac{1}{\delta_0}$.

By rearranging and combining previous estimate, we have:

$$
\begin{aligned}
\frac{1}{\gamma_T} &\leq \frac{1}{2}\sum_{t=0}^{T}\nabla f(\mathbf{x}_t)^\top \Gamma_t \nabla f(\mathbf{x}_t) + \frac{\delta_0}{2}\\
&\leq \frac{\delta_0}{2} + f(\mathbf{x}_1) - \inf_t f(\mathbf{x}_t) + \frac{1}{2}\sum_{t=0}^{t_0}\sum_{i=1}^{d}\left[\frac{\partial f}{\partial x_i}(\mathbf{x}_t)\right]^2\left(\beta\lambda_i^2(\Gamma_t) - \lambda_i(\Gamma_t)\right)\\
&< +\infty,
\end{aligned}
$$

and letting $T \to \infty$, we have $\gamma_T \to 0$, and therefore

$$
+\infty \leq \frac{1}{2}\sum_{t=0}^{\infty}\nabla f(\mathbf{x}_t)^\top \Gamma_t \nabla f(\mathbf{x}_t) + \frac{\delta_0}{2} < +\infty
$$

which is a contradiction.

To show that the sequence of gradient norm is summable, it sufficies to show that for each component i, $\sum_{t=0}^{\infty}\left[\frac{\partial f}{\partial x_i}(\mathbf{x}_t)\right]^2 < \infty$. This is true since we have

$$
\begin{aligned}
\sum_{t=0}^{\infty}\left[\frac{\partial f}{\partial x_i}(\mathbf{x}_t)\right]^2 &= \lim_{T\to\infty}\sum_{t=0}^{T}\left[\frac{\partial f}{\partial x_i}(\mathbf{x}_t)\right]^2\\
&= \lim_{T\to\infty}\left[\frac{1}{\gamma_T^2} - \delta_0^2\right]\\
&= \lim_{T\to\infty}\frac{1}{\gamma_T^2} - \delta_0^2\\
&< +\infty
\end{aligned}
$$

where γ_T is from the argument in proving the first claim, and the last inequality holds also because of the first claim, since for all i, the limit $\lim_{T\to\infty}\frac{1}{\gamma_T^2}$ exists and is finite. This proves our assertion and completes our proof. ■

A.3 AdaGrad with full matrix adaptation

Finally, we examine the full matrix version of the AdaGrad-type methods. In particular, this is given by the following recursive formula:

$$
\mathbf{x}_{t+1} = \mathbf{x}_t - \Gamma_t \nabla f(\mathbf{x}_t) \tag{A.9}
$$

where $\Gamma_t \in \mathbb{R}^{d\times d}$ is a sequence of symmetric (full dimensional) matrices defined as the inverse square root of G_t:

$$\Gamma_t \equiv G_t^{-\frac{1}{2}} = \left[\delta_0^2 I + \sum_{s=0}^{t} \nabla f(\mathbf{x}_s)\nabla f(\mathbf{x}_s)^\top\right]^{-\frac{1}{2}} \tag{AdaFull}$$

Lemma A.2. *Assume that $\mathbf{x}_t$ are the iterates of (A.9) run with the adaptive step-size policy (AdaFull). Then, the following hold:*

1. *The sequence of matices $\{\Gamma_t\}_{t\in\mathbb{N}}$ is non-increasing in the Löwner sense, i.e.,*

$$\Gamma_t \succcurlyeq \Gamma_{t+1} \;\; \textit{for all} \;\; t = 1, 2, \dots \tag{A.10}$$

2. *The sequence of eigenvalues $\lambda_i(\Gamma_t)$ is non-increasing, i.e.,*

$$\lambda_i(\Gamma_t) \geq \lambda_i(\Gamma_{t+1}) \tag{A.11}$$

Proof. We begin with the first claim. By definition of G_t we have:

$$G_{t+1} - G_t = \nabla f(\mathbf{x}_{t+1})\nabla f(\mathbf{x}_{t+1})^\top \succcurlyeq 0 \tag{A.12}$$

which in turn yields that G_t is a Löwner non-decreasing sequence, i.e.,:

$$G_{t+1} \succcurlyeq G_t \;\; \text{for all} \;\; t = 1, 2, \dots \tag{A.13}$$

Hence, by applying Löwner-Heinz inequality we readily get:

$$G_{t+1}^{\frac{1}{2}} \succcurlyeq G_t^{\frac{1}{2}} \;\; \text{for all} \;\; t = 1, 2, \dots \tag{A.14}$$

and so by applying Weyl's monotonicity thoerem for $\Gamma_t = G_t^{-\frac{1}{2}}$ we have:

$$\Gamma_t \succcurlyeq \Gamma_{t+1}$$

and

$$\lambda_i(\Gamma_t) \geq \lambda_i(\Gamma_{t+1}).$$

■

We proceed by providing the following proposition

Proposition A.3. *Assume that $f : \mathbb{R}^d \to \mathbb{R}$ is smooth and $\mathbf{x}_t$ are the iterates of (A.9) run with the adaptive step-size policy (AdaFull). Then, the following hold:*

1. *The sequence of the eigenvalues $\{\lambda_i(\Gamma_t)\}_{t\in\mathbb{N}}$ converges to a strictly positive value λ_i^∞ for all $i = 1, 2, \dots, d$, i.e.,*

$$\lim_{t\to+\infty} \lambda_i(\Gamma_t) = \inf_{t\in\mathbb{N}} \lambda_i(\Gamma_t) = \lambda_i^\infty > 0 \;\; \textit{for all} \;\; i = 1, 2, \dots d \tag{A.15}$$

2. *The sequence $\{\|\nabla f(\mathbf{x}_t)\|_*^2\}_{t\in\mathbb{N}}$ is summable, i.e.,*

$$\sum_{t=0}^{+\infty} \|\nabla f(\mathbf{x}_t)\|_*^2 < +\infty \tag{A.16}$$

Proof. We start with the proof of our first claim. By combining the fact that $\{\Gamma_t\}_{t\in\mathbb{N}}$ is a decreasing sequence of matrices and Weyl's monotonicity inequality we have that for every $i = 1, 2, \dots d$ the respective sequence of eigenvalues $\lambda_i(\Gamma_t)$ is non-increasing. Moreover, since Γ_t is positive definite for all $t = 1, 2, \dots$ we readily get for all $i = 1, 2, \dots d$ the sequence $\{\lambda_i(\Gamma_t)\}_{t\in\mathbb{N}}$ is bounded from below by 0. Therefore, for all $i = 1, 2, \dots d$ the limit of the sequence $\{\lambda_i(\Gamma_t)\}_{t\in\mathbb{N}}$ exists and more precisely:

$$\lim_{t\to+\infty} \lambda_i(\Gamma_t) = \inf_{t\in\mathbb{N}} \lambda_i(\Gamma_t) = \lambda_i^\infty \geq 0 \;\; \text{for all} \;\; i = 1, 2, \dots d \tag{A.17}$$

We assume now that there exists some $i_0 \in \{1, 2, \dots, d\}$ such that:

$$\lim_{t \to +\infty} \lambda_{i_0}(\Gamma_t) = 0 \tag{A.18}$$

Then, by the fact that f is smooth, we can choose $\beta > L$ and have:

$$\begin{aligned}
f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{\beta}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
&= f(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)^\top \Gamma_t \nabla f(\mathbf{x}_t) + \frac{\beta}{2} \|\Gamma_t \nabla f(\mathbf{x}_t)\|^2 \\
&= f(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)^\top \Gamma_t \nabla f(\mathbf{x}_t) + \frac{\beta}{2} \nabla f(\mathbf{x}_t)^\top \Gamma_t^\top \nabla f(\mathbf{x}_t) \Gamma_t \\
&= f(\mathbf{x}_t) - \frac{1}{2} \nabla f(\mathbf{x}_t)^\top \Gamma_t \nabla f(\mathbf{x}_t) - \frac{1}{2} \nabla f(\mathbf{x}_t)^\top \left[I - \beta \Gamma_t^\top \right] \Gamma_t \nabla f(\mathbf{x}_t)
\end{aligned}$$

and by rearranging we have:

$$\begin{aligned}
\frac{1}{2} \nabla f(\mathbf{x}_t)^\top \Gamma_t \nabla f(\mathbf{x}_t) &\leq f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) - \frac{1}{2} \nabla f(\mathbf{x}_t)^\top \left[I - \beta \Gamma_t^\top \right] \Gamma_t \nabla f(\mathbf{x}_t) \\
&= f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) - \frac{1}{2} \nabla f(\mathbf{x}_t)^\top \left[I - \beta \Gamma_t \right]^\top \Gamma_t \nabla f(\mathbf{x}_t)
\end{aligned}$$

Now, since Γ_t is decreasing we have:

$$\frac{1}{\delta_0} I = \Gamma_0 \succcurlyeq \Gamma_t \tag{A.19}$$

where under the assumption $\delta_0 > \beta$ yields:

$$I - \beta \Gamma_t \succ 0 \ \text{ for all } \ t = 1, 2, \dots \tag{A.20}$$

and hence by telescoping we have:

$$\frac{1}{2} \sum_{t=0}^{T} \nabla f(\mathbf{x}_t)^\top \Gamma_t \nabla f(\mathbf{x}_t) \leq f(\mathbf{x}_1) - \inf_{t \in \mathbb{N}} f(\mathbf{x}_t) \tag{A.21}$$

On the other hand, we bound the quantity $\frac{1}{2} \sum_{t=0}^{T} \nabla f(\mathbf{x}_t)^\top \Gamma_t \nabla f(\mathbf{x}_t)$ from below as follows:

$$\begin{aligned}
\frac{1}{2} \sum_{t=0}^{T} \nabla f(\mathbf{x}_t)^\top \Gamma_t \nabla f(\mathbf{x}_t) &\geq \sum_{t=0}^{T} \nabla f(\mathbf{x}_t)^\top \Gamma_T \nabla f(\mathbf{x}_t) \\
&= \sum_{t=0}^{T} \operatorname{tr}(\Gamma_T \nabla f(\mathbf{x}_t) \nabla f(\mathbf{x}_t)^\top) \\
&= \operatorname{tr}(\Gamma_T \sum_{t=0}^{T} \nabla f(\mathbf{x}_t) \nabla f(\mathbf{x}_t)^\top) \\
&= \operatorname{tr}(\Gamma_T \left[\delta_0^2 I + \sum_{t=0}^{T} \nabla f(\mathbf{x}_t) \nabla f(\mathbf{x}_t)^\top - \delta_0^2 I \right]) \\
&= \operatorname{tr}(\Gamma_T G_T) - \delta_0^2 \operatorname{tr}(\Gamma_T)
\end{aligned}$$

Now since by definition $\Gamma_t = G_t^{-\frac{1}{2}}$ we have that $\Gamma_T G_T = G_T^{\frac{1}{2}} = \Gamma_T^{-1}$. Furthermore by the monotonicity of the trace operator (cf. +++) we have that:

$$\delta_0^2 \operatorname{tr}(\Gamma_T) \geq \delta_0^2 \operatorname{tr}(\Gamma_0) \tag{A.22}$$

and by the fact $\Gamma_0 = \frac{1}{\delta_0} I$ we have:

$$\delta_0^2 \operatorname{tr}(\Gamma_T) \geq \delta_0 d \tag{A.23}$$

and hence we have:

$$\begin{aligned}\frac{1}{2}\sum_{t=0}^{T}\nabla f(\mathbf{x}_t)^\top\Gamma_t\nabla f(\mathbf{x}_t) &\geq \mathrm{tr}(\Gamma_T^{-1}) - \delta_0 d\\ &\geq \sum_{i=1}^{d}\frac{1}{\lambda_i(\Gamma_T)} - \delta_0 d\\ &\geq \frac{1}{\lambda_{i_0}(\Gamma_T)} - \delta_0 d\end{aligned}$$

Therefore, summarizing the above estimations we get:

$$\frac{1}{\lambda_{i_0}(\Gamma_T)} \leq f(\mathbf{x}_1) - \inf_{t\in\mathbb{N}} f(\mathbf{x}_t) + \delta_0 d \tag{A.24}$$

Now, by letting $T \to +\infty$ we get that $\frac{1}{\lambda_{i_0}(\Gamma_T)} \to +\infty$ since we assumed that $\lambda_{i_0}(\Gamma_T) \to 0$. So, summarizing the above estimations:

$$+\infty \leq f(\mathbf{x}_1) - \inf_{t\in\mathbb{N}} f(\mathbf{x}_t) + \delta_0 d < +\infty \tag{A.25}$$

which is a contradiction. Hence, for all $i = 1, 2, \dots d$ the sequence of the eigenvalues $\lambda_i(\Gamma_t)$ converges to a strictly positive value $\lambda_i^\infty > 0$ and therefore the result follows.

Now, we turn our attention towards the proof of our second claim. In particular, by the above we have:

$$\sum_{t=0}^{T}\nabla f(\mathbf{x}_t)^\top\Gamma_t\nabla f(\mathbf{x}_t) \leq f(\mathbf{x}_1) - \inf_{t\in\mathbb{N}} f(\mathbf{x}_t) \tag{A.26}$$

Moreover, we have:

$$\begin{aligned}\sum_{t=0}^{T}\nabla f(\mathbf{x}_t)^\top\Gamma_t\nabla f(\mathbf{x}_t) &\geq \sum_{t=0}^{T}\nabla f(\mathbf{x}_t)^\top\Gamma_T\nabla f(\mathbf{x}_t)\\ &\geq \lambda_{\min}(\Gamma_T)\sum_{t=0}^{T}\nabla f(\mathbf{x}_t)^\top\nabla f(\mathbf{x}_t)\\ &\geq \lambda_{\min}(\Gamma_T)\sum_{t=0}^{T}\|\nabla f(\mathbf{x}_t)\|_*^2\end{aligned}$$

So, we have:

$$\lambda_{\min}(\Gamma_T)\sum_{t=1}^{T}\|\nabla f(\mathbf{x}_t)\|_*^2 \leq f(\mathbf{x}_1) - \inf_{t\in} f(\mathbf{x}_t) \tag{A.27}$$

Now, by letting $T \to +\infty$ and recalling from our previous claim that:

$$\lim_{T\to+\infty}\lambda_{\min}(\Gamma_T) = \lambda_{\min}^\infty > 0 \tag{A.28}$$

we get:

$$\sum_{t=0}^{T}\|\nabla f(\mathbf{x}_t)\|_*^2 \leq \frac{1}{\lambda_{\min}^\infty}\left[f(\mathbf{x}_1) - \inf_{t\in\mathbb{N}} f(\mathbf{x}_t)\right] < +\infty \tag{A.29}$$

and the result follows. ■

We proceed by showing that matrix sequence $\{G_t\}_{t\in\mathbb{N}}$ itself stabilizes asymptotically around a positive definite matix G^∞. Formally, we have the following proposition.

Proposition A.4. *Assume that $f : \mathbb{R}^d \to \mathbb{R}$ is smooth and $\mathbf{x}_t$ are the iterates of (A.9) run with the adaptive step-size policy (ADAFULL). Then, the limit of $\{G_t\}_{t\in\mathbb{N}}$ exists and in particular we have:*

$$\lim_{t\to+\infty} G_t = G^\infty \;\; \textit{with} \;\; G^\infty \succ 0 \tag{A.30}$$

Proof. We start by showing the limit existence of G_t. Fix $i \neq j \in \{1, 2, \dots d\}$. We then have:

- For the diagonal terms of G_t: $[G_t]_{i,i}$:

$$\begin{aligned}
[G_t]_{i,i} &= \left[\delta_0^2 I + \sum_{s=0}^{t} \nabla f(\mathbf{x}_s)\nabla f(\mathbf{x}_s)^\top\right]_{i,i} \\
&= \delta_0^2 + \sum_{s=0}^{t} \left[\nabla f(\mathbf{x}_s)\nabla f(\mathbf{x}_s)^\top\right]_{i,i} \\
&= \delta_0^2 + \sum_{s=0}^{t} (\nabla f^i(\mathbf{x}_s))^2 \\
&\leq \delta_0^2 + \sum_{t=0}^{+\infty} \nabla f(\mathbf{x}_s)^\top \nabla f(\mathbf{x}_s) \\
&< +\infty
\end{aligned}$$

  with the last inequality being obtained by Proposition A.3. So, since the sequence $\{[G_t]_{i,i}\}_{t\in\mathbb{N}}$ is a non-decreasing and upper-bounded its limit exists.

- For the terms $[G_t]_{i,j}$ we have:

$$\begin{aligned}
[G_t]_{i,j} &= \left[\delta_0^2 I + \sum_{s=0}^{t} \nabla f(\mathbf{x}_s)\nabla f(\mathbf{x}_s)^\top\right]_{i,j} \\
&= \sum_{s=0}^{t} \left[\nabla f(\mathbf{x}_s)\nabla f(\mathbf{x}_s)^\top\right]_{i,j} \\
&= \sum_{s=0}^{t} \nabla f^i(\mathbf{x}_s)\nabla f^j(\mathbf{x}_s)
\end{aligned}$$

  Now, by applying Cauchy-Schwartz inequality we have:

$$\begin{aligned}
\sum_{s=0}^{t} |\nabla f^i(\mathbf{x}_s)\nabla f^j(\mathbf{x}_s)| &\leq \sqrt{\sum_{s=0}^{t} (\nabla f^i(\mathbf{x}_s))^2}\sqrt{\sum_{s=0}^{t} (\nabla f^i(\mathbf{x}_s))^2} \\
&\leq \left(\sum_{s=1}^{+\infty} \nabla f(\mathbf{x}_t)^\top \nabla f(\mathbf{x}_t)\right)^2 \\
&< +\infty
\end{aligned}$$

  with the last strict inequality being obtained by Proposition A.3. So, we get that $\sum_{s=1}^{t} \nabla f^i(\mathbf{x}_s)\nabla f^j(\mathbf{x}_s)$ converges absolutely, which yields that it converges. This in turn yields that the limit of $[G_t]_{i,j}$ exists.

So, summarizing we get that the limit of G_t exists, i.e.,

$$\lim_{t\to+\infty} G_t = G^\infty \tag{A.31}$$

On the other hand, concerning the positive definiteness of G^∞ we have for all $i = 1, 2, \dots d$:

$$\lambda_i(G^\infty) = \lim_{t\to+\infty} \lambda_i(G_t)$$

$$
\begin{aligned}
&= \lim_{t\to+\infty} \lambda_i(\Gamma_t^{-2}) \\
&= \lim_{t\to+\infty} \frac{1}{\lambda_i(\Gamma_t)^2} \\
&= \frac{1}{(\lambda_i^\infty)^2}
\end{aligned}
$$

which by invoking Proposition A.3 yields that:

$$\lambda_i(G^\infty) > 0 \;\text{ for all }\; i = 1, 2, \dots d \tag{A.32}$$

which concludes the proof. ■

B Proof of Proposition 2

Proof. Recall that in the context before proposition 2, it is denoted that

$$\eta(t, \mathbf{x}) = -\Gamma\theta(\mathbf{x}) - (\Gamma_t - \Gamma)\nabla f(\mathbf{x}).$$

The proof consists of two parts.

Part 1: Recall that $\theta(\mathbf{x})$ is the remainder of the Taylor expansion of $\nabla f(\mathbf{x})$ at $\mathbf{0}$, i.e.

$$\nabla f(\mathbf{x}) = \nabla f(\mathbf{0}) + \nabla^2 f(\mathbf{0})\mathbf{x} + \theta(\mathbf{x}) \tag{B.1}$$
$$= \nabla^2 f(\mathbf{0})\mathbf{x} + \theta(\mathbf{x}) \tag{B.2}$$

and then

$$\theta(\mathbf{x}) = \nabla f(\mathbf{x}) - \nabla^2 f(\mathbf{0})\mathbf{x}.$$

The differential of $\theta(\mathbf{x})$ is

$$D\theta(\mathbf{x}) = D(\nabla f(\mathbf{x}) - \nabla^2(\mathbf{0})\mathbf{x}) = \nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{0}).$$

From the Fundamental Theorem of Calculus and chain rule, we have that

$$\theta(\mathbf{x}) - \theta(\mathbf{y}) = \int_0^1 \frac{d}{dt}\theta(t\mathbf{x} + (1-t)\mathbf{y})dt \tag{B.3}$$
$$= \int_0^1 \left(D\theta(\mathbf{z})|_{\mathbf{z}=t\mathbf{x}+(1-t)\mathbf{y}}\right) \cdot \frac{d}{dt}(t\mathbf{x} + (1-t)\mathbf{y})dt \tag{B.4}$$
$$= \int_0^1 \left(D\theta(\mathbf{z})|_{\mathbf{z}=t\mathbf{x}+(1-t)\mathbf{y}}\right) \cdot (\mathbf{x} - \mathbf{y})dt. \tag{B.5}$$

Thus, the norm of the difference $\theta(\mathbf{x}) - \theta(\mathbf{y})$ can be estimated as

$$\|\theta(\mathbf{x}) - \theta(\mathbf{y})\| = \|\int_0^1 \left(D\theta(\mathbf{z})|_{\mathbf{z}=t\mathbf{x}+(1-t)\mathbf{y}}\right) \cdot (\mathbf{x} - \mathbf{y})dt\| \tag{B.6}$$
$$\leq \int_0^1 \|\left(D\theta(\mathbf{z})|_{\mathbf{z}=t\mathbf{x}+(1-t)\mathbf{y}}\right) \cdot (\mathbf{x} - \mathbf{y})\|dt \tag{B.7}$$
$$\leq \int_0^1 \|D\theta(\mathbf{z})|_{\mathbf{z}=t\mathbf{x}+(1-t)\mathbf{y}}\| \cdot \|\mathbf{x} - \mathbf{y}\|dt \tag{B.8}$$
$$\leq \left(\int_0^1 \|D\theta(\mathbf{z})|_{\mathbf{z}=t\mathbf{x}+(1-t)\mathbf{y}}\|dt\right) \|\mathbf{x} - \mathbf{y}\|, \tag{B.9}$$

where

$$D\theta(\mathbf{z})|_{\mathbf{z}=t\mathbf{x}+(1-t)\mathbf{y}} = \left(\nabla^2 f(\mathbf{z}) - \nabla^2 f(\mathbf{0})\right)|_{\mathbf{z}=t\mathbf{x}+(1-t)\mathbf{y}}$$

Since $\nabla^2 f(\mathbf{x})$ is assumed to be Lipschitz, for any $\epsilon > 0$, there exists a δ-ball $\mathbb{B}$ of $\mathbf{0}$, such that for any $\mathbf{z} \in \mathbb{B}$, it holds that

$$\|\nabla^2 f(\mathbf{z}) - \nabla^2 f(\mathbf{0})\| \leq \epsilon,$$

and this completes the proof of part 1.

Part 2: Since the limit of Γ_t is Γ and then given any $\epsilon > 0$, there exists t_0, such that for all $t > t_0$, $\|\Gamma_t - \Gamma\| < \epsilon$. On the other hand, the gradient $\nabla f(\mathbf{x})$ is assumed to be Lipschitz, thus for $\mathbf{x}, \mathbf{y} \in \mathbb{B}$, and $t > t_0$, it holds that

$$\|\Gamma_t - \Gamma\| \cdot \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq \epsilon L \|\mathbf{x} - \mathbf{y}\|.$$

Therefore, the norm of difference $\eta(t, \mathbf{x}) - \eta(t, \mathbf{y})$ is estimated as

$$\begin{aligned}
\|\eta(t, \mathbf{x}) - \eta(t, \mathbf{y})\| &\leq \|-\Gamma\theta(\mathbf{x}) - (\Gamma_t - \Gamma)\nabla f(\mathbf{x}) + \Gamma\theta(\mathbf{y}) + (\Gamma_t - \Gamma)\nabla f(\mathbf{y})\| && \text{(B.10)}\\
&\leq \|\Gamma\theta(\mathbf{x}) - \Gamma\theta(\mathbf{y})\| + \|(\Gamma_t - \Gamma)\nabla f(\mathbf{x}) - (\Gamma_t - \Gamma)\nabla f(\mathbf{y})\| && \text{(B.11)}\\
&\leq \|\Gamma\| \cdot \|\theta(\mathbf{x}) - \theta(\mathbf{y})\| + \|\Gamma_t - \Gamma\| \cdot \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| && \text{(B.12)}
\end{aligned}$$

For a given $\epsilon > 0$, find a neighborhood $\mathbb{B}$ of $\mathbf{0}$ such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{B}$,

$$\|\theta(\mathbf{x}) - \theta(\mathbf{y})\| \leq \frac{\epsilon}{\|\Gamma\|}\|\mathbf{x} - \mathbf{y}\|,$$

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|,$$

and

$$\|\Gamma_t - \Gamma\| \leq \frac{\epsilon}{L} \text{ for } t \geq t_0.$$

Thus, we have

$$\|\eta(t, \mathbf{x}) - \eta(t, \mathbf{y})\| \leq \epsilon\|\mathbf{x} - \mathbf{y}\|,$$

and the proof completes. ■

C Proof of Theorem 1

In the proof of Theorem 1, we will denote $\xi = \delta_0^2$ for convenience.

Proof. Let

$$\begin{aligned}
A(m, n) &= (I - \gamma(m, \mathbf{x}_0)H)...(I - \gamma(n, \mathbf{x}_0)H) && \text{(C.1)}\\
&= \begin{bmatrix} B(m, n) & \\ & C(m, n) \end{bmatrix} && \text{(C.2)}
\end{aligned}$$

for $m \geq n$ and I for $m < n$, where

$$B(m, n) = (I - \gamma(m, \mathbf{x}_0)H^+)...(I - \gamma(n, \mathbf{x}_0)H^+)$$

and

$$C(m, n) = (I - \gamma(m, \mathbf{x}_0)H^-)...(I - \gamma(n, \mathbf{x}_0)H^-).$$

Let $\mathbf{v}$ be a vector, we denote $\mathbf{v}^+$ the stable component and $\mathbf{v}^-$ the unstable component of $\mathbf{v}$, i.e., $\mathbf{v}^+ \in E^s$ and $\mathbf{v}^- \in E^u$, $E^+ \oplus E^- = T_{\mathbf{0}}\mathbb{R}^d = \mathbb{R}^d$. Then the solution $\mathbf{x}_{t+1} = (\mathbf{x}_{t+1}^+, \mathbf{x}_{t+1}^-)$ can be written in the form of stable-unstable:

$$\mathbf{x}_{t+1}^+ = B(t, 0)\mathbf{x}_0^+ + \sum_{i=0}^{t} B(t, i+1)\eta^+(i, \mathbf{x}_i)$$

and

$$\mathbf{x}_{t+1}^- = C(t, 0)\mathbf{x}_0^- + \sum_{i=0}^{t} C(t, i+1)\eta^-(i, \mathbf{x}_i).$$

Then $\mathbf{x}_0^-$ can be written as

$$\mathbf{x}_0^- = C(t, 0)^{-1}\mathbf{x}_{t+1}^- - C(t, 0)^{-1}\sum_{i=0}^{t} C(t, i+1)\eta^-(i, \mathbf{x}_i). \quad \text{(C.3)}$$

Simplifying the notation by denoting

$$C_i = I - \gamma(i, \mathbf{x}_0)H^-,$$

we have

$$C(t,0) = C_t ... C_0, \quad \text{and} \quad C(t,0)^{-1} = C_0^{-1} ... C_t^{-1}. \tag{C.4}$$

Since H^- is the diagonal matrix of all negative eigenvalues, the inverse of C_i is the following:

$$C_i^{-1} = \begin{bmatrix} \frac{1}{1-\gamma(i,\mathbf{x}_0)\lambda_{s+1}} & & \\ & \ddots & \\ & & \frac{1}{1-\gamma(i,\mathbf{x}_0)\lambda_d} \end{bmatrix}$$

and the entries satisfy

$$1 > \frac{1}{1-\gamma(i,\mathbf{x}_0)\lambda_{s+1}} \geq \frac{1}{1-\gamma(i,\mathbf{x}_0)\lambda_d} > 0.$$

Since $\gamma(i, \mathbf{x}_0) > c$ is uniformly bounded above 0 in a neighborhood of $\mathbf{0}$, then the norm $\|C_i^{-1}\|$ satisfies

$$\|C_i^{-1}\| \leq \frac{1}{1-c\lambda_{s+1}} < 1,$$

and then

$$C(t,0)^{-1} \to 0, \quad \text{as} \quad t \to \infty.$$

Using the expression of (C.4), (C.3) can be written as

$$\mathbf{x}_0^- = C(t,0)^{-1}\mathbf{x}_{t+1}^- - C(t,0)^{-1}\sum_{i=0}^{t} C(t,i+1)\eta^-(i,\mathbf{x}_0,\mathbf{x}_i) \tag{C.5}$$

$$= C_0^{-1} ... C_t^{-1}\mathbf{x}_{t+1}^- - \left(C_0^{-1}\eta^-(0,\mathbf{x}_0,\mathbf{x}_0) + ... + C_0^{-1} ... C_t^{-1}\eta^-(t,\mathbf{x}_0,\mathbf{x}_t)\right). \tag{C.6}$$

If $\mathbf{x}_t \to 0$, then the sequence is bounded in a neighborhood of $\mathbf{x}_0$. So assuming $\mathbf{x}_t$ is bounded, we can push t to ∞ (since the identity holds for any t) and we have

$$\mathbf{x}_0^- = -\sum_{i=1}^{\infty} C(i-1,0)^{-1}\eta^-(i-1,\mathbf{x}_0,\mathbf{x}_{i-1}). \tag{C.7}$$

So far the above derivation is heuristic and no existence or uniqueness is guaranteed. But we can go one step further to see where the "Stable Manifold" comes from. We say that the initial condition $\mathbf{x}_0$ lies on a manifold is equivalent to saying that $\mathbf{x}_0$ lies on a graph of some mapping from stable space to unstable space, i.e.,

$$(\mathbf{x}_0^+, \mathbf{x}_0^-) \text{ satisfies that } \mathbf{x}_0^- = \varphi(\mathbf{x}_0^+) \text{ for some } \varphi : E^s \to E^u.$$

The right hand side of C.7 contains $C(i-1,0)^{-1}$ that involves $\mathbf{x}_0 = (\mathbf{x}_0^+, \mathbf{x}_0^-)$ and the sequence $\mathbf{x}_{i-1}$ is also determined by the initial condition $\mathbf{x}_0$, so we can also write $\mathbf{x}_{i-1}$ as a function of $\mathbf{x}_0$ in the following

$$\mathbf{x}_{i-1} = \mathbf{x}_{i-1}(\mathbf{x}_0^+, \mathbf{x}_0^-).$$

Therefore the equation C.7 can be rewritten as

$$\mathbf{x}_0^- = -\sum_{i=1}^{\infty} C(i-1,0,(\mathbf{x}_0^+,\mathbf{x}_0^-))^{-1}\eta^-(i-1,(\mathbf{x}_0^+,\mathbf{x}_0^-),\mathbf{x}_{i-1}(\mathbf{x}_0^+,\mathbf{x}_0^-)),$$

where the right hand side is a function of $(\mathbf{x}_0^+, \mathbf{x}_0^-)$ and we can denote the right hand side as $\Phi(\mathbf{x}_0^+, \mathbf{x}_0^-)$. Then the stable manifold $\mathbf{x}_0^- = \varphi(\mathbf{x}_0^+)$ (as an implicit function) is expected to be solved from the equation

$$\mathbf{x}_0^- = \Phi(\mathbf{x}_0^+, \mathbf{x}_0^-).$$

It suffices to show that for the sequence generated by the dynamical system of AdaGrad algorithm with initial condition $\mathbf{x}_0 = \mathbf{a} \oplus \mathbf{x}_0^-$ (when $\mathbf{a}$ satisfies certain condition), the unstable component $\mathbf{x}_0^-$ in uniquely determined by $\mathbf{a}$. This comes from that the operator T is a contraction mapping on the space of convergent sequences (with $\mathbf{0}$ the limit) whose 0'th terms have the same stable component (the rest of the paper). Since the sequence that converges to $\mathbf{0}$ while generated by the dynamical system is the unique fixed point of T, the initial condition $\mathbf{x}_0$ is also unique. But the stable component $\mathbf{x}_0^+$ is fixed, so the uniqueness of the unstable component $\mathbf{x}_0^-$ implies the existence of some function $\varphi : E^s \to E^u$ so that

$$\mathbf{x}_0^- = \varphi(\mathbf{x}_0^+).$$

■

Lemma C.1. *Suppose*

$$\mathbf{x}_0^- = -\sum_{i=1}^{\infty} C(i-1,0)^{-1}\eta^-(i-1,\mathbf{x}_0,\mathbf{x}_{i-1}).$$

Then

$$\mathbf{x}_{t+1} = \left(B(t,0)\mathbf{x}_0^+ + \sum_{i=0}^{t} B(t,i+1)\eta^+(i,\mathbf{x}_0,\mathbf{x}_i)\right) \oplus \left(-\sum_{i=0}^{\infty} C(t+1+i,t+1)^{-1}\eta^-(t+1+i,\mathbf{x}_0,\mathbf{x}_{t+1+i})\right) \tag{C.8}$$

Proof. It suffices to show that

$$\mathbf{x}_{t+1}^- = -\sum_{i=0}^{\infty} C(t+1+i,t+1)^{-1}\eta^-(t+1+i,\mathbf{x}_0,\mathbf{x}_{t+1+i}). \tag{C.9}$$

It has been shown that

$$\mathbf{x}_{t+1}^- = C(t,0)\mathbf{x}_0^- + \sum_{i=0}^{t} C(t,i+1)\eta^-(i,\mathbf{x}_0,\mathbf{x}_i) \tag{C.10}$$

and assumed

$$\mathbf{x}_0^- = -\sum_{i=1}^{\infty} C(i-1,0)^{-1}\eta^-(i-1,\mathbf{x}_0,\mathbf{x}_{i-1}). \tag{C.11}$$

Plug C.11 into C.10:

$$\mathbf{x}_{t+1}^- = \underbrace{C(t,0)\left(-\sum_{i=1}^{\infty} C(i-1,0)^{-1}\eta^-(i-1,\mathbf{x}_0,\mathbf{x}_{i-1})\right)}_{\mathrm{I}} + \sum_{i=0}^{t} C(t,i+1)\eta^-(i,\mathbf{x}_0,\mathbf{x}_i). \tag{C.12}$$

Splitting $-\sum_{i=1}^{\infty} C(i-1,0)^{-1}\eta^-(i-1,\mathbf{x}_0,\mathbf{x}_{i-1})$ into

$$-\sum_{i=1}^{t+1} C(i-1,0)^{-1}\eta^-(i-1,\mathbf{x}_0,\mathbf{x}_{i-1}) - \sum_{i=t+2}^{\infty} C(i-1,0)^{-1}\eta^-(i-1,\mathbf{x}_0,\mathbf{x}_{i-1}),$$

and we have

$$\mathrm{I} = C(t,0)\left(-\sum_{i=1}^{t+1} C(i-1,0)^{-1}\eta^-(i-1,\mathbf{x}_0,\mathbf{x}_{i-1}) - \sum_{i=t+2}^{\infty} C(i-1,0)^{-1}\eta^-(i-1,\mathbf{x}_0,\mathbf{x}_{i-1})\right) \tag{C.13}$$

$$= -C(t,0)\sum_{i=1}^{t+1} C(i-1,0)\eta^-(i-1,\mathbf{x}_0,\mathbf{x}_{i-1}) - C(t,0)\sum_{i=k+2}^{\infty} C(i-1,0)^{-1}\eta^-(i-1,\mathbf{x}_{i-1}) \tag{C.14}$$

$$= -\sum_{i=0}^{t} C(t, i+1)\eta^{-}(i, \mathbf{x}_0, \mathbf{x}_i) - \sum_{i=0}^{\infty} C(t+1+i, t+1)^{-1}\eta^{-}(t+1+i, \mathbf{x}_0, \mathbf{x}_{t+1+i}). \tag{C.15}$$

Therefore, by putting I back into C.12, we have

$$\mathbf{x}_{t+1}^{-} = -\sum_{i=0}^{\infty} C(t+1+i, t+1)^{-1}\eta^{-}(t+1+i, \mathbf{x}_0, \mathbf{x}_{t+1+i})$$

which is exactly same as C.9. ■

The idea of Lyapunov-Perron method is to consider the right hand side of C.8 as an operator acting on a sequence $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$, note that this sequence $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ is arbitrary and not necessarily generated by gradient descent or any other algorithm. Specifically we call the action T on a sequence $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ as

$$(T\mathbf{x})_{t+1} = \left(B(t,0)\mathbf{x}_0^{+} + \sum_{i=0}^{t} B(t, i+1)\eta^{+}(i, \mathbf{x}_0, \mathbf{x}_i)\right) \oplus \left(-\sum_{i=0}^{\infty} C(t+1+i, t+1)^{-1}\eta^{-}(t+1+i, \mathbf{x}_0, \mathbf{x}_{t+1+i})\right)$$

where $t \geq 0$. Moreover, by the definition of $B(m, n) = I$ if $m < n$, we have formally that the 0'th term of the transformed sequence $\{(T\mathbf{x})\}_{t\in\mathbb{N}}$ is

$$(T\mathbf{x})_0 = \mathbf{x}_0^{+} \oplus \left(-\sum_{i=0}^{\infty} C(i, 1)^{-1}\eta^{-}(i, \mathbf{x}_0, \mathbf{x}_i)\right).$$

This means that the action of the operator T preserves the stable component of the 0'th term of any sequence on which T acts.

Another important property of T is this: Consider the definition of $B(m, n)$, $C(m, n)$ and $\eta(t, \mathbf{x}_0, \mathbf{x}_t)$, we can conclude that if the sequence $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ is generated by the algorithm (adaptive gradient descent), then such sequence is a "fixed point" of the operator T, i.e.,

$$\mathbf{x}_{t+1} = (I - \gamma(t, \mathbf{x}_0)H)\,\mathbf{x}_t + \eta(t, \mathbf{x}_0, \mathbf{x}_t) \Longrightarrow T\mathbf{x} = \mathbf{x}$$

where $\mathbf{x} = \{\mathbf{x}_t\}_{t\in\mathbb{N}}$ and $T\mathbf{x} = \{(T\mathbf{x})_t\}_{t\in\mathbb{N}}$.

Proposition C.1. $\lim_{t\to\infty}(T\mathbf{x})_t = \mathbf{0}$ *if* $\mathbf{x}_t \to \mathbf{0}$.

Proof. The matrix $B(t, 0)$ is in the form of:

$$(I - \gamma(t, \mathbf{x}_0)H^{+}) \cdots (I - \gamma(0, \mathbf{x}_0)H^{+})$$

where H^{+} is diagonal of positive eigenvalues, so $B(t, 0) \to 0$ as $t \to \infty$. Combining with Lemma C.2 and C.3, we can conclude the result. ■

Lemma C.2. *Suppose that* $\lim_{t\to\infty} \mathbf{x}_t = \mathbf{0}$. *Then*

$$\lim_{t\to\infty} \sum_{i=0}^{t} B(t, i+1)\eta^{-}(i, \mathbf{x}_0, \mathbf{x}_i) = \mathbf{0}.$$

Proof. Denote $\gamma_t = \gamma(t, \mathbf{x}_0)$ for short, and we do the following estimation.

$$\begin{aligned}
\|\sum_{i=0}^{t} B(t, i+1)\eta^{-}(i, \mathbf{x}_0, \mathbf{x}_i)\| &\leq \sum_{i=0}^{t} \|B(t, i+1)\eta^{-}(i, \mathbf{x}_0, \mathbf{x}_i)\| \\
&\leq \sum_{i=0}^{t} \|B(t, i+1)\| \cdot \|\eta^{-}(i, \mathbf{x}_0, \mathbf{x}_i)\| \\
&\leq \sum_{i=0}^{t} \|B(t, i+1)\| \cdot \|\eta(i, \mathbf{x}_0, \mathbf{x}_i)\|
\end{aligned}$$

$$\leq \sum_{i=0}^{t} \|B(t,i+1)\| \cdot \gamma_i \epsilon \|\mathbf{x}_i\|$$
$$= \epsilon \sum_{i=0}^{t} \|B(t,i+1)\| \cdot \gamma_i \|\mathbf{x}_i\|$$
$$= \epsilon \sum_{i=0}^{t} (1-\gamma_t\lambda_s)\cdots(1-\gamma_{i+1}\lambda_s)\gamma_i\|\mathbf{x}_i\|.$$

The last equality from the above, i.e.,

$$\|B(t,i+1)\| = (1-\gamma_t\lambda_s)\cdots(1-\gamma_{i+1}\lambda_s)$$

comes from the notation of $B(t,i+1)$ which is defined as:

$$B(t,i+1) = \begin{bmatrix} 1-\gamma_t\lambda_1 & & \\ & \ddots & \\ & & 1-\gamma_t\lambda_s \end{bmatrix} \cdots \begin{bmatrix} 1-\gamma_{i+1}\lambda_1 & & \\ & \ddots & \\ & & 1-\gamma_{i+1}\lambda_s \end{bmatrix}$$

where $\lambda_1 \geq ... \geq \lambda_s > 0$.

Denote

$$S_t = \sum_{i=0}^{t} (1-\gamma_t\lambda_s)\cdots(1-\gamma_{i+1}\lambda_s)\gamma_i\|\mathbf{x}_i\|$$
$$= (1-\gamma_t\lambda_s)\cdots(1-\gamma_1\lambda_s)\gamma_0\|\mathbf{x}_0\| + \cdots + \gamma_t\|\mathbf{x}_t\|$$

and then

$$\begin{aligned} S_{t+1} &= (1-\gamma_{t+1}\lambda_s)...(1-\gamma_1\lambda_s)\gamma_0\|\mathbf{x}_0\| + \cdots + (1-\gamma_{t+1}\lambda_s)\gamma_t\|\mathbf{x}_t\| + \gamma_{t+1}\|\mathbf{x}_{t+1}\| \\ &= (1-\gamma_{t+1}\lambda_s)\left((1-\gamma_t\lambda_s)\cdots(1-\gamma_1\lambda_s)\gamma_0\|\mathbf{x}_0\| + \cdots + \gamma_t\|\mathbf{x}_t\|\right) + \gamma_{t+1}\|\mathbf{x}_{t+1}\| \\ &= (1-\gamma_{t+1}\lambda_s)S_t + \gamma_{t+1}\|\mathbf{x}_{t+1}\| \\ &= S_t - \gamma_{t+1}\lambda_s S_t + \gamma_{t+1}\|\mathbf{x}_{t+1}\| \\ &= S_t + \gamma_{t+1}(\|\mathbf{x}_{t+1}\| - \lambda_s S_t). \end{aligned}$$

Subtracting S_t, we have that

$$S_{t+1} - S_t = \gamma_{t+1}(\|\mathbf{x}_{t+1}\| - \lambda_s S_t),$$

and the following observation for the sequence $\{S_t\}_{t\in\mathbb{N}}$ is immediate:

- Case 1: If $S_{t+1} - S_t > 0$, then $\|\mathbf{x}_{t+1}\| - \lambda_s S_t > 0$;
- Case 2: If $S_{t+1} - S_t < 0$, then $\|\mathbf{x}_{t+1}\| - \lambda_s S_t < 0$;
- Case 3: If $S_{t+1} - S_t = 0$, then $\|\mathbf{x}_{t+1}\| - \lambda_s S_t = 0$.

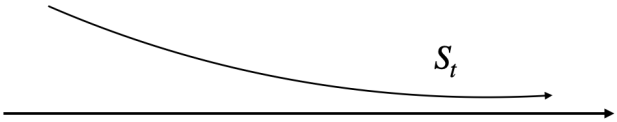
Note that Case 3 trivially implies that $S_t \to 0$ by the assumption in the lemma that $\mathbf{x}_t \to \mathbf{0}$. Moreover, Case 1 can be interpreted as follows: If the sequence S_t is successively increasing over an interval of integers, i.e., $t \in [n, ...n+m]$, then the terms of S_t over this interval are dominated by terms of the convergent sequence $\{\frac{\|\mathbf{x}_t\|}{\lambda_s}\}_{t\in\mathbb{N}}$ because

$$\|\mathbf{x}_{t+1}\| - \lambda_s S_t > 0 \iff \frac{\|\mathbf{x}_{t+1}\|}{\lambda_s} > S_t.$$

Therefore, the interval of integers over which S_t is increasing successively cannot have infinite length, and thus, Case 2 has to occur infinitely often. Then the sequence S_t must have one of the following two patterns:

- Pattern 1: S_t has infinitely many intervals of integers on which S_t is increasing successively,

- Pattern 2: S_t is decreasing after certain $t^* \in \mathbb{N}$.



If S_t is of Pattern 1, then we have

$$\limsup_t S_t \leq \lim_{t\to\infty} \|\mathbf{x}_t\| = 0.$$

If S_t is of Pattern 2, then S_t is convergent and we can denote its limit by $S^* \in [0, \infty)$. Taking limit for $t \leftarrow \infty$ in the following relation

$$S_{t+1} - S_t = \gamma_{t+1}(\|\mathbf{x}_t\| - \lambda_s S_t)$$

we have that

$$0 = S^* - S^* = c(0 - \lambda_s S^*).$$

Since c is positive (recall that $\gamma_t = \gamma(t, \mathbf{x}_0)$ is uniformly bounded away from 0 if $\mathbf{x}_0$ is taken from a ball centering at $\mathbf{0}$), it must hold that

$$S^* = 0.$$

So by the estimation in the beginning of the proof, we have

$$\lim_{t\to\infty} \|\sum_{i=0}^{t} B(t, i+1)\eta^-(i, \mathbf{x}_0, \mathbf{x}_i)\| \leq \epsilon \lim_{t\to\infty} S_t = 0$$

provided $\lim_{t\to\infty} \mathbf{x}_t = \mathbf{0}$. The proof completes. ■

Lemma C.3. *Suppose that* $\lim_{t\to\infty} \mathbf{x}_t = \mathbf{0}$. *Then*

$$\lim_{t\to\infty} -\sum_{i=0}^{\infty} C(t+1+i, t+1)^{-1}\eta^-(t+1+i, \mathbf{x}_0, \mathbf{x}_{t+1+i}) = \mathbf{0}.$$

Proof. The estimate gives

$$\begin{aligned}
&\|-\sum_{i=0}^{\infty} C(t+1+i, t+1)^{-1}\eta^-(t+1+i, \mathbf{x}_0, \mathbf{x}_{t+1+i})\| \\
&\leq \sum_{i=0}^{\infty} \|C(t+1+i, t+1)^{-1}\eta^-(t+1+i, \mathbf{x}_0, \mathbf{x}_{t+1+i})\| \\
&\leq \sum_{i=0}^{\infty} \|C(t+1+i, t+1)^{-1}\| \cdot \|\eta^-(t+1+i, \mathbf{x}_0, \mathbf{x}_{t+1+i})\| \\
&\leq \sum_{i=0}^{\infty} \|C(t+1+i, t+1)^{-1}\| \cdot \|\eta(t+1+i, \mathbf{x}_0, \mathbf{x}_{t+1+i})\|
\end{aligned}$$

$$\leq \sum_{i=0}^{\infty} \|C(t+1+i,t+1)^{-1}\| \cdot \gamma_{t+1+i}\epsilon \|\mathbf{x}_{t+1+i}\|$$
$$= \epsilon \sum_{i=0}^{\infty} \|C(t+1+i,t+1)^{-1}\| \cdot \gamma_{t+1+i} \|\mathbf{x}_{t+1+i}\|.$$

Since $\Gamma(\mathbf{x}_0) = \lim_t \gamma(t, \mathbf{x}_0)$ is a continuous function of $\mathbf{x}_0$, so if $\mathbf{x}_0$ is taken from a compact ball centering at $\mathbf{0}$, $\lim_t \gamma(t, \mathbf{x}_0)$ is bounded above uniformly. Thus there exists a constant $K > 0$ such that

$$\gamma(t, \mathbf{x}_0) \leq K \text{ for all } t \in \mathbb{N}, \ \mathbf{x}_0 \in \mathbb{B}(\mathbf{0}), \mathbf{x}_0^+ = \mathbf{a}.$$

On the other hand, the norm of $\mathbf{x}_{t+1+i}$ for $i \geq 0$ satisfies

$$\|\mathbf{x}_{t+1+i}\| \leq \sup_{k>t} \|\mathbf{x}_k\|,$$

where $\sup_{k>t}\|\mathbf{x}_k\|$ only depends on t. So

$$\epsilon \sum_{i=0}^{\infty} \|C(t+1+i,t+1)^{-1}\| \cdot \gamma_{t+1+i} \|\mathbf{x}_{t+1+i}\| \leq \epsilon K \sup_{k>t} \|\mathbf{x}_k\| \sum_{i=0}^{\infty} \|C(t+1+i,t+1)^{-1}\|. \tag{C.16}$$

The definition of $C(m,n)$ gives the following

$$\begin{aligned}
&C(t+1+i,t+1)^{-1} \\
&= (I - \gamma_{t+1}H^-)^{-1} \cdots (I - \gamma_{t+1+i}H^-)^{-1} \\
&= \begin{bmatrix} 1-\gamma_{t+1}\lambda_{s+1} & & \\ & \ddots & \\ & & 1-\gamma_{t+1}\lambda_d \end{bmatrix}^{-1} \cdots \begin{bmatrix} 1-\gamma_{t+1+i}\lambda_{s+1} & & \\ & \ddots & \\ & & 1-\gamma_{t+1+i}\lambda_d \end{bmatrix}^{-1} \\
&= \begin{bmatrix} \frac{1}{1-\gamma_{t+1}\lambda_{s+1}} & & \\ & \ddots & \\ & & \frac{1}{1-\gamma t+1\lambda_d} \end{bmatrix} \cdots \begin{bmatrix} \frac{1}{1-\gamma_{t+1+i}\lambda_{s+1}} & & \\ & \ddots & \\ & & \frac{1}{1-\gamma_{t+1+i}\lambda_d} \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{(1-\gamma_{t+1}\lambda_{s+1})\cdots(1-\gamma_{t+1+i}\lambda_{s+1})} & & \\ & \ddots & \\ & & \frac{1}{(1-\gamma_{t+1}\lambda_d)\cdots(1-\gamma_{t+1+i}\lambda_d)} \end{bmatrix}.
\end{aligned}$$

Recall that

$$0 > \lambda_{s+1} \geq ... \geq \lambda_d$$

so we have that

$$\frac{1}{(1-\gamma_{t+1}\lambda_{s+1})\cdots(1-\gamma_{t+1+i}\lambda_{s+1})} \geq ... \geq \frac{1}{(1-\gamma_{t+1}\lambda_d)\cdots(1-\gamma_{t+1+i}\lambda_d)},$$

and then the operator norm of $C(t+1+i,t+1)^{-1}$ is

$$\|C(t+1+i,t+1)^{-1}\| = \frac{1}{(1-\gamma_{t+1}\lambda_{s+1})\cdots(1-\gamma_{t+1+i}\lambda_{s+1})}.$$

Furthermore, $\gamma_t = \gamma(t, \mathbf{x}_0) \geq c$, i.e., is uniformly bounded away from 0 if $\mathbf{x}_0$ is taken from a ball centering at $\mathbf{0}$, and then the following inequality holds:

$$\frac{1}{1-\gamma_t\lambda_{s+1}} \leq \frac{1}{1-c\lambda_{s+1}}.$$

Thus the norm satisfies

$$\|C(t+1+i,t+1)^{-1}\| = \frac{1}{(1-\gamma_{t+1}\lambda_{s+1})\cdots(1-\gamma_{t+1+i}\lambda_{s+1})}$$

$$\leq \left(\frac{1}{1-c\lambda_{s+1}}\right)^{i+1}.$$

Since the ratio $\frac{1}{1-c\lambda_{s+1}}$ is less than 1, the series on the right hand side of the following inequality,

$$\sum_{i=0}^{\infty}\|C(t+1+i,t+1)^{-1}\| \leq \sum_{i=0}^{\infty}\left(\frac{1}{1-c\lambda_{s+1}}\right)^{i+1},$$

converges to the finite number $C^* = -\frac{1}{c\lambda_{s+1}}$. Combining with the estimation in the beginning and C.16, we can conclude that

$$\|-\sum_{i=0}^{\infty} C(t+1+i,t+1)^{-1}\eta^{-}(t+1+i,\mathbf{x}_0,\mathbf{x}_{t+1+i})\| \tag{C.17}$$

$$\leq \epsilon\sum_{i=0}^{\infty}\|C(t+1+i,t+1)^{-1}\|\cdot\gamma_{t+1+i}\|\mathbf{x}_{t+1+i}\| \tag{C.18}$$

$$\leq \epsilon K\sup_{k>t}\|\mathbf{x}_k\|\sum_{i=0}^{\infty}\|C(t+1+i,t+1)^{-1}\| \tag{C.19}$$

$$\leq \epsilon K\sup_{k>t}\|\mathbf{x}_k\|\sum_{i=0}^{\infty}\left(\frac{1}{1-c\lambda_{s+1}}\right)^{i+1} \tag{C.20}$$

$$= \epsilon KC^*\sup_{k>t}\|\mathbf{x}_k\| \tag{C.21}$$

$$= \epsilon K\left(-\frac{1}{c\lambda_{s+1}}\right)\sup_{k>t}\|\mathbf{x}_k\|. \tag{C.22}$$

By the assumption that $\lim_{t\to\infty}\mathbf{x}_t = \mathbf{0}$, we have that $\sup_{k>t}\|\mathbf{x}_k\|$ as $t\to\infty$, so

$$\|-\sum_{i=0}^{\infty} C(t+1+i,t+1)^{-1}\eta^{-}(t+1+i,\mathbf{x}_0,\mathbf{x}_{t+1+i})\| \longrightarrow 0 \text{ as } t\to\infty,$$

and the proof completes. ■

We summerize two important properties of the integral operator T: Suppose $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ is sequence in a ball centering at $\mathbf{0}$ such that $\lim_{t\to\infty}\mathbf{x}_t = \mathbf{0}$ and $\mathbf{x}_0^+ = \mathbf{a}$. Then the transformed sequence $\{(T\mathbf{x})_t\}_{t\in\mathbb{N}}$ satisfies

1. $(T\mathbf{x})_0^+ = \mathbf{a}$;
2. $\lim_{t\to\infty}(T\mathbf{x})_t = \mathbf{0}$.

In another word, the operator T transforms a sequence converging to $\mathbf{0}$ whose stable component of the 0'th term is $\mathbf{a}$ to another sequence converging to $\mathbf{0}$ whose stable component of the 0'th term is $\mathbf{a}$.

Let $X(\mathbf{a},\mathbf{0}) = \{\{\mathbf{x}_t\}_{t\in\mathbb{N}} : \lim_{t\to\infty}\mathbf{x}_t = \mathbf{0}, \mathbf{x}_0^+ = \mathbf{a}\}$. Define the metric on $X(\mathbf{a},\mathbf{0})$ as follows:

Definition 1. Let $\mathbf{u} = \{\mathbf{u}_t\}_{t\in\mathbb{N}}$ and $\mathbf{v} = \{\mathbf{v}_t\}_{t\in\mathbb{N}}$ be two sequences in $X(\mathbf{a},\mathbf{0})$, then the distance $d(\mathbf{u},\mathbf{v})$ is defined as

$$d(\mathbf{u},\mathbf{v}) = \sup_{n\geq 0}\{\|\mathbf{u}_n - \mathbf{v}_n\|\}.$$

The following lemma shows that the sequences satisfying above two properties form a complete metric space.

Lemma C.4. *$X(\mathbf{a},\mathbf{0})$ is a complete metric space.*

Proof. Let $\mathbf{u}_1 = \{\mathbf{u}_{1,j}\}_{j\in\mathbb{N}}$, $\mathbf{u}_2 = \{\mathbf{u}_{2,j}\}_{j\in\mathbb{N}}$,..., $\{\mathbf{u}_{i,j}\}_{j\in\mathbb{N}}$... be a sequence of sequences that converges to $\mathbf{0}$ and $\mathbf{u}_{i,0}^+ = \mathbf{a}$, i.e., the 0'th term of $\mathbf{u}_i$ has stable component equal to $\mathbf{a}$. Suppose that $\{\mathbf{u}_i\}_{i\in\mathbb{N}}$ is Cauchy, i.e., given any $\epsilon > 0$, there exists an integer $L > 0$, such that

$$d(\mathbf{u}_n,\mathbf{u}_m) = \sup_{j\geq 0}\{\|\mathbf{u}_{n,j} - \mathbf{u}_{m,j}\|\} < \epsilon$$

for all $n, m > L$. This implies that for each j, $\|\mathbf{u}_{n,j} - \mathbf{u}_{m,j}\| < \epsilon$ for all $n, m > L$. Furthermore, this is equivalent to say that each sequence $\{\mathbf{u}_{k,j}\}_{k\in\mathbb{N}}$ with fixed j, is Cauchy. Therefore, for each j, there exists a point $\mathbf{u}_{*,j}$ such that

$$\lim_{k\to\infty} \mathbf{u}_{k,j} = \mathbf{u}_{*,j}$$

We denote the limit sequence

$$\mathbf{u}_* = \{\mathbf{u}_{*,j}\}_{j\in\mathbb{N}}.$$

Letting $m \to \infty$, we have that

$$\|\mathbf{u}_{n,j} - \mathbf{u}_{*,j}\| < \epsilon$$

for all $n > L$. Since $\lim_{j\to\infty} \mathbf{u}_{n,j} = \mathbf{0}$, we can conclude that

$$\lim_{j\to\infty} \mathbf{u}_{*,j} = \mathbf{0},$$

and this means that the limit sequence $\mathbf{u}_*$ belongs to $X(\mathbf{a}, \mathbf{0})$. ■

Denote $X(\mathbf{0}, \mathbf{a}, \delta)$ the space of sequences $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ satisfying

- $\lim_{t\to\infty} \mathbf{x}_t = \mathbf{0}$;
- $\mathbf{x}_0^+ = \mathbf{a}$;
- $\|\mathbf{x}_t\| \le \delta$ for all $t \in \mathbb{N}$.

Lemma C.5. *Suppose (ϵ, δ) is a pair of positive constants such that the Lipschitz condition is satisfied. Then ϵ can be adjusted so that for a sequence $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ with $\mathbf{x}_0^+ = \mathbf{a}$ ($\|\mathbf{a}\| \le \delta$), $\|\mathbf{x}_t^+\| \le \delta$, $\|\mathbf{x}_t^-\| \le \delta$ for all $t \ge 0$, the transformed sequence $T\mathbf{x}$ satisfies $(T\mathbf{x})_0^+ = \mathbf{a}$, $\|(T\mathbf{x})_t^+\| \le \delta$, and $\|(T\mathbf{x})_t^-\| \le \delta$ for all $t \ge 0$.*

Proof. When $t = 0$, we have

$$(T\mathbf{x})_0 = \mathbf{a} \oplus \left(-\sum_{i=0}^{\infty} C(i,0)^{-1} \eta^-(i, \mathbf{x}_0, \mathbf{x}_i) \right),$$

and then the estimate gives the following

$$\|(T\mathbf{x})_0^-\| = \|-\sum_{i=0}^{\infty} C(i,0)^{-1} \eta^-(i, \mathbf{x}_0, \mathbf{x}_i)\| \tag{C.23}$$

$$\le \sum_{i=0}^{\infty} \|C(i,0)^{-1} \eta^-(i, \mathbf{x}_0, \mathbf{x}_i)\| \tag{C.24}$$

$$\le \sum_{i=0}^{\infty} \|C(i,0)^{-1}\| \cdot \|\eta^-(i, \mathbf{x}_0, \mathbf{x}_i)\| \tag{C.25}$$

Since

$$\|\eta^-(i, \mathbf{x}_0, \mathbf{x})\| \le \|\eta(i, \mathbf{x}_0, \mathbf{x})\| = \gamma(i, \mathbf{x}_0)\|\theta(\mathbf{x}_i)\|,$$

with the Lipschitz condition that

$$\|\theta(\mathbf{x})\| \le \epsilon\|\mathbf{x}\|$$

we have

$$\|\eta^-(i, \mathbf{x}_0, \mathbf{x})\| \le \gamma(i, \mathbf{x}_0)\|\theta(\mathbf{x}_i)\| \tag{C.26}$$

$$\le \gamma(i, \mathbf{x}_0)\epsilon\|\mathbf{x}_i\| \tag{C.27}$$

$$= \gamma(i, \mathbf{x}_0)\epsilon\sqrt{\|\mathbf{x}_i^+\|^2 + \|\mathbf{x}_i^-\|^2} \tag{C.28}$$

$$\le \gamma(i, \mathbf{x}_0)\epsilon\sqrt{\delta^2 + \delta^2} \tag{C.29}$$

$$= \gamma(i, \mathbf{x}_0)\epsilon\sqrt{2}\delta. \tag{C.30}$$

Then the estimate of $\|(T\mathbf{x})_0^-\|$ can be done as follows.

$$\|(T\mathbf{x})_0^-\| \leq \sum_{i=0}^{\infty} \|C(i,0)^{-1}\| \cdot \|\eta^-(i, \mathbf{x}_0, \mathbf{x}_i)\| \tag{C.31}$$

$$\leq \epsilon\sqrt{2}\delta \cdot \left(\sum_{i=0}^{\infty} \|C(i,0)^{-1}\| \cdot \gamma(i, \mathbf{x}_0) \right) \tag{C.32}$$

$$\leq \gamma(0, \mathbf{x}_0)\epsilon\sqrt{2}\delta \cdot \left(\sum_{i=0}^{\infty} \|C(i,0)^{-1}\| \right) \tag{C.33}$$

$$\leq \gamma(0, \mathbf{x}_0)\epsilon\sqrt{2}\delta \left(-\frac{1}{c\lambda_{s+1}} \right) \tag{C.34}$$

$$\leq \frac{1}{\sqrt{\xi}}\epsilon\sqrt{2}\delta \left(-\frac{1}{c\lambda_{s+1}} \right) \tag{C.35}$$

The last inequality is due to that $\gamma(0, \mathbf{x}_0)$ is uniformly bounded above by $\frac{1}{\sqrt{\xi}}$. As long as ϵ is chosen according to

$$\frac{1}{\sqrt{\xi}}\sqrt{2}\epsilon \left(-\frac{1}{c\lambda_{s+1}} \right) \leq 1 \iff \epsilon \leq -c\lambda_{s+1}\sqrt{\frac{\xi}{2}}$$

we have that

$$\|(T\mathbf{x})_0^-\| \leq \delta.$$

To estimate the norm $\|(T\mathbf{x})_t^+\|$ and $\|(T\mathbf{x})_t^-\|$ for $t \geq 1$, it is equivalent to estimate $\|(T\mathbf{x})_{t+1}\|$ for $t \geq 0$. We have that

$$\|(T\mathbf{x})_{t+1}^+\| \leq \|B(t,0)\mathbf{a}\| + \sum_{i=0}^{t} \|B(t, i+1)\eta^+(i, \mathbf{x}_0, \mathbf{x}_i)\| \tag{C.36}$$

$$\leq \|B(t,0)\mathbf{a}\| + \sum_{i=0}^{t} \|B(t, i+1)\| \cdot \|\eta^+(i, \mathbf{x}_0, \mathbf{x}_i)\| \tag{C.37}$$

where

$$\|\eta^+(i, \mathbf{x}_0, \mathbf{x}_i)\| \leq \|\eta(i, \mathbf{x}_0, \mathbf{x}_i)\| \tag{C.38}$$

$$= \gamma(i, \mathbf{x}_0)\|\theta(\mathbf{x}_i)\| \tag{C.39}$$

$$\leq \gamma(i, \mathbf{x}_0)\epsilon\|\mathbf{x}_i\| \tag{C.40}$$

$$\leq \gamma(i, \mathbf{x}_0)\epsilon\sqrt{2}\delta. \tag{C.41}$$

Furthermore, recall that

$$B(t, i+1) = (I - \gamma(t, \mathbf{x}_0)H^+)...(I - \gamma(i+1, \mathbf{x}_0)H^+),$$

and $\gamma(i, \mathbf{x}_0) \geq c$, the norm satisfies

$$\|B(t, i+1)\| = (1 - \gamma(t, \mathbf{x}_0)\lambda_s)...(1 - \gamma(i+1, \mathbf{x}_0)\lambda_s) \tag{C.42}$$

$$\leq \underbrace{(1 - c\lambda_s)...(1 - c\lambda_s)}_{t - i \text{ copies}} \tag{C.43}$$

$$= (1 - c\lambda_s)^{t-i}. \tag{C.44}$$

So the sum

$$\sum_{i=0}^{t} \|B(t, i+1)\| = \sum_{i=0}^{t} (1 - c\lambda_s)^{t-i} \tag{C.45}$$

$$= \frac{1-(1-c\lambda_s)^{t+1}}{1-(1-c\lambda_s)} \tag{C.46}$$

$$= \frac{1-(1-c\lambda_s)^{t+1}}{c\lambda_s} \tag{C.47}$$

$$\leq \frac{1}{c\lambda_s} \tag{C.48}$$

And then

$$\|(T\mathbf{x})_{t+1}^{+}\| \leq \|B(t,0)\mathbf{a}\| + \sum_{i=0}^{t}\|B(t,i+1)\| \cdot \gamma(i,\mathbf{x}_0)\epsilon\sqrt{2}\delta \tag{C.49}$$

$$\leq \|B(t,0)\mathbf{a}\| + \gamma(0,\mathbf{x}_0)\epsilon\sqrt{2}\delta\left(\sum_{i=0}^{t}\|B(t,i+1)\|\right) \tag{C.50}$$

$$\leq \|B(t,0)\|\delta + \gamma(0,\mathbf{x}_0)\epsilon\sqrt{2}\delta\left(\frac{1}{c\lambda_s}\right) \tag{C.51}$$

$$\leq (1-c\lambda_s)^{t+1}\delta + \frac{1}{\sqrt{\xi}}\epsilon\sqrt{2}\delta\left(\frac{1}{c\lambda_s}\right) \tag{C.52}$$

$$\leq (1-c\lambda_s)\delta + \frac{1}{\sqrt{\xi}}\epsilon\sqrt{2}\delta\left(\frac{1}{c\lambda_s}\right). \tag{C.53}$$

The last inequality gives

$$\|(T\mathbf{x})_{t+1}^{+}\| \leq \left((1-c\lambda_s) + \sqrt{\frac{2}{\xi}}\epsilon\left(\frac{1}{c\lambda_s}\right)\right)\delta,$$

and the ϵ should be chosen according to

$$(1-c\lambda_s) + \sqrt{\frac{2}{\xi}}\epsilon\left(\frac{1}{c\lambda_s}\right) < 1 \Longleftrightarrow \epsilon < c^2\lambda_s^2\sqrt{\frac{\xi}{2}}$$

so that

$$\|(T\mathbf{x})_{t+1}^{+}\| < \delta$$

for all $t \geq 0$.

To estimate $\|(T\mathbf{x})_{t+1}^{-}\|$, we firstly recall that

$$(T\mathbf{x})_{t+1}^{-} = -\sum_{i=0}^{\infty} C(t+1+i,t+1)^{-1}\eta^{-}(t+1+i,\mathbf{x}_0,\mathbf{x}_{t+1+i}).$$

Therefore

$$\|(T\mathbf{x})_{t+1}^{-}\| = \|-\sum_{i=0}^{\infty} C(t+1+i,t+1)^{-1}\eta^{-}(t+1+i,\mathbf{x}_0,\mathbf{x}_i)\| \tag{C.54}$$

$$\leq \sum_{i=0}^{\infty}\|C(t+1+i,t+1)^{-1}\eta^{-}(t+1+i,\mathbf{x}_0,\mathbf{x}_{t+1+i})\| \tag{C.55}$$

$$\leq \sum_{i=0}^{\infty}\|C(t+1+i,t+1)^{-1}\| \cdot \|\eta^{-}(t+1+i,\mathbf{x}_0,\mathbf{x}_{t+1+i})\| \tag{C.56}$$

where

$$\|\eta^{-}(t+1+i,\mathbf{x}_0,\mathbf{x}_{t+1+i})\| \leq \|\eta(t+1+i,\mathbf{x}_0,\mathbf{x}_{t+1+i})\| \tag{C.57}$$

$$= \gamma(t+1+i,\mathbf{x}_0)\|\theta(\mathbf{x}_{t+1+i})\| \tag{C.58}$$

$$\leq \gamma(t+1+i, \mathbf{x}_0)\epsilon\|\mathbf{x}_{t+1+i}\| \tag{C.59}$$

$$\leq \gamma(t+1+i, \mathbf{x}_0)\epsilon\sqrt{2}\delta \tag{C.60}$$

$$\leq \frac{1}{\sqrt{\xi}}\epsilon\sqrt{2}\delta. \tag{C.61}$$

And then we have

$$\|(T\mathbf{x})^-_{t+1}\| \leq \sqrt{\frac{2}{\xi}}\epsilon\delta\left(\sum_{i=0}^{\infty}\|C(t+1+i, t+1)^{-1}\|\right) \tag{C.62}$$

$$\leq \sqrt{\frac{2}{\xi}}\epsilon\delta\left(\sum_{i=0}^{\infty}\left(\frac{1}{1-c\lambda_{s+1}}\right)^{i+1}\right) \tag{C.63}$$

$$\leq \sqrt{\frac{2}{\xi}}\epsilon\delta\left(-\frac{1}{c\lambda_{s+1}}\right). \tag{C.64}$$

From the above estimate we need to chose ϵ according to

$$\sqrt{\frac{2}{\xi}}\epsilon\left(-\frac{1}{c\lambda_{s+1}}\right) \leq 1 \Longleftrightarrow \epsilon \leq -c\lambda_{s+1}\sqrt{\frac{\xi}{2}}$$

and then

$$\|(T\mathbf{x})^-_{t+1}\| \leq \delta.$$

In summery, ϵ needs to be chosen according to

$$\epsilon \leq \min\left\{-c\lambda_{s+1}\sqrt{\frac{\xi}{2}}, c^2\lambda_s^2\sqrt{\frac{\xi}{2}}\right\} \tag{C.65}$$

$$= \sqrt{\frac{\xi}{2}}\cdot\min\{-c\lambda_{s+1}, c^2\lambda_s^2\}, \tag{C.66}$$

so that the operator T maps the sequences in the way described in the lemma. ■

Lemma C.6. *Suppose that the function $\eta(t, \mathbf{z}, \mathbf{x})$ has the form of*

$$\eta(t, \mathbf{z}, \mathbf{x}) = \gamma(t, \mathbf{z})\theta(\mathbf{x})$$

and $\theta(\mathbf{x})$ satisfies that for any $\epsilon > 0$, there exists $\delta > 0$, such that

$$\|\theta(\mathbf{x}) - \theta(\mathbf{y})\| \leq \epsilon\|\mathbf{x} - \mathbf{y}\|$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{B}(\delta)$. Then T is a contraction mapping on $X(\mathbf{0}, \mathbf{a}, \delta)$, i.e., there exists a positive constant $\kappa < 1$ such that

$$d(T\mathbf{u}, T\mathbf{v}) \leq \kappa d(\mathbf{u}, \mathbf{v})$$

for all $\mathbf{u}, \mathbf{v} \in X(\mathbf{0}, \mathbf{a}, \delta)$.

Remark 5. In this lemma we use $\gamma(t, \mathbf{z})$ for any sequences ignoring that different sequences have different initial conditions. This is because for any sequence $\gamma(t, \mathbf{z})$, we can associate with a dynamical system, e.g., the initial one we are interested in, and we are actually looking for the stable manifold of this dynamical system at $\mathbf{0}$, so the sequence $\gamma(t, \mathbf{z})$ can be regarded independent of the sequences $\{\mathbf{x}_t\}_{t\in\mathbb{N}}$ fed into T.

Proof. Let $\mathbf{u} = \{\mathbf{u}_t\}_{t\in\mathbb{N}}$ and $\mathbf{v} = \{\mathbf{v}_t\}_{t\in\mathbb{N}}$ be two sequences in $X(\mathbf{0}, \mathbf{a}, \delta)$. Then the 0'th term of the difference is

$$(T\mathbf{u})_0 - (T\mathbf{v})_0 = \mathbf{u}_0^+ \oplus \left(-\sum_{i=0}^{\infty}C(i,0)^{-1}\eta^-(i, \mathbf{u}_0, \mathbf{u}_i)\right) - \mathbf{v}_0^+ \oplus \left(-\sum_{i=0}^{\infty}C(i,0)^{-1}\eta^-(i, \mathbf{v}_0, \mathbf{v}_i)\right) \tag{C.67}$$

$$= (\mathbf{u}_0^+ - \mathbf{v}_0^+) \oplus \left(-\sum_{i=0}^{\infty} C(i,0)^{-1}(\eta^-(i,\mathbf{u}_0,\mathbf{u}_i) - \eta^-(i,\mathbf{v}_0,\mathbf{v}_i)) \right) \tag{C.68}$$

$$= (\mathbf{a} - \mathbf{a}) \oplus \left(-\sum_{i=0}^{\infty} C(i,0)^{-1}(\eta^-(i,\mathbf{u}_0,\mathbf{u}_i) - \eta^-(i,\mathbf{v}_0,\mathbf{v}_i)) \right) \tag{C.69}$$

$$= -\sum_{i=0}^{\infty} C(i,0)^{-1}(\eta^-(i,\mathbf{u}_0,\mathbf{u}_i) - \eta^-(i,\mathbf{v}_0,\mathbf{v}_i)). \tag{C.70}$$

So the norm is estimated as follows:

$$\|(T\mathbf{u})_0 - (T\mathbf{v})_0\| = \|(T\mathbf{u})_0^- - (T\mathbf{v})_0^-\| \tag{C.71}$$

$$\leq \|\sum_{i=0}^{\infty} C(i,0)^{-1}(\eta^-(i,\mathbf{z},\mathbf{u}_i) - \eta^-(i,\mathbf{z},\mathbf{v}_i))\| \tag{C.72}$$

$$\leq \sum_{i=0}^{\infty} \|C(i,0)^{-1}\| \cdot \|\eta^-(i,\mathbf{z},\mathbf{u}_i) - \eta^-(i,\mathbf{z},\mathbf{v}_i)\| \tag{C.73}$$

$$\leq \sum_{i=0}^{\infty} \|C(i,0)^{-1}\| \cdot \|\eta(i,\mathbf{z},\mathbf{u}_i) - \eta(i,\mathbf{z},\mathbf{v}_i)\| \tag{C.74}$$

$$\leq \sum_{i=0}^{\infty} \|C(i,0)^{-1}\| \cdot \|\gamma(i,\mathbf{z})\theta(\mathbf{u}_i) - \gamma(i,\mathbf{z})\theta(\mathbf{v}_i)\| \tag{C.75}$$

$$\leq \sum_{i=0}^{\infty} \|C(i,0)^{-1}\| \cdot \gamma(i,\mathbf{z})\|\theta(\mathbf{u}_i) - \theta(\mathbf{v}_i)\|. \tag{C.76}$$

Since

$$\|\theta(\mathbf{u}_i) - \theta(\mathbf{v}_i)\| \leq \epsilon\|\mathbf{u}_i - \mathbf{v}_i\|, \text{ and } \gamma(i,\mathbf{z}) \leq \gamma(0,\mathbf{z}),$$

we have that

$$\|(T\mathbf{u})_0 - (T\mathbf{v})_0\| \leq \sum_{i=0}^{\infty} \|C(i,0)^{-1}\| \cdot \gamma(i,\mathbf{z})\epsilon\|\mathbf{u}_i - \mathbf{v}_i\| \tag{C.77}$$

$$\leq \gamma(0,\mathbf{z})\epsilon \sum_{i=0}^{\infty} \|C(i,0)^{-1}\| \cdot \|\mathbf{u}_i - \mathbf{v}_i\| \tag{C.78}$$

$$\leq \gamma(0,\mathbf{z})\epsilon \sum_{i=0}^{\infty} \|C(i,0)^{-1}\| \cdot \sup_{i\geq 0} \|\mathbf{u}_t - \mathbf{v}_t\| \tag{C.79}$$

$$= \gamma(0,\mathbf{z})\epsilon \sum_{i=0}^{\infty} \|C(i,0)^{-1}\| \cdot d(\mathbf{u},\mathbf{v}) \tag{C.80}$$

$$= \gamma(0,\mathbf{z})\epsilon \left(-\frac{1}{c\lambda_{s+1}} \right) d(\mathbf{u},\mathbf{v}) \tag{C.81}$$

$$\leq \frac{1}{\sqrt{\xi}}\epsilon \left(-\frac{1}{c\lambda_{s+1}} \right) d(\mathbf{u},\mathbf{v}). \tag{C.82}$$

Therefore, the first condition for ϵ so that

$$\frac{1}{\sqrt{\xi}}\epsilon \left(-\frac{1}{c\lambda_{s+1}} \right) < 1$$

which implies that

$$\epsilon < -c\lambda_{s+1}\sqrt{\xi}$$

The rest terms,i.e., $(T\mathbf{x})_{t+1}$ with $t \geq 0$, give the following estimate.

$$(T\mathbf{u} - T\mathbf{v})_{t+1} \tag{C.83}$$

$$= (T\mathbf{u})_{t+1} - (T\mathbf{v})_{t+1} \tag{C.84}$$

$$= \left(B(t,0)\mathbf{u}_0^+ + \sum_{i=0}^{t} B(t,i+1)\eta^+(i,\mathbf{z},\mathbf{u}_i) \right) \tag{C.85}$$

$$\oplus \left(-\sum_{i=0}^{\infty} C(t+1+i,t+1)^{-1}\eta^-(t+1+i,\mathbf{z},\mathbf{u}_{t+1+i}) \right) \tag{C.86}$$

$$- \left(B(t,0)\mathbf{v}_0^+ + \sum_{i=0}^{t} B(t,i+1)\eta^+(i,\mathbf{z},\mathbf{v}_i) \right) \tag{C.87}$$

$$\oplus \left(-\sum_{i=0}^{\infty} C(t+1+i,t+1)^{-1}\eta^-(t+1+i,\mathbf{z},\mathbf{v}_{t+1+i}) \right) \tag{C.88}$$

$$= \left(B(t,0)(\mathbf{u}_0^+ - \mathbf{v}_0^+) + \sum_{i=0}^{t} B(t,i+1)(\eta^+(i,\mathbf{z},\mathbf{u}_i) - \eta^+(i,\mathbf{z},\mathbf{v}_i)) \right) \tag{C.89}$$

$$\oplus \left(-\sum_{i=0}^{\infty} C(t+1+i,t+1)^{-1}(\eta^-(t+1+i,\mathbf{z},\mathbf{u}_{t+1+i}) - \eta^-(t+1+i,\mathbf{z},\mathbf{v}_{t+1+i})) \right) \tag{C.90}$$

$$= \left(\sum_{i=0}^{t} B(t,i+1)(\eta^+(i,\mathbf{z},\mathbf{u}_i) - \eta^+(i,\mathbf{z},\mathbf{v}_i)) \right) \tag{C.91}$$

$$\oplus \left(-\sum_{i=0}^{\infty} C(t+1+i,t+1)^{-1}(\eta^-(t+1+i,\mathbf{z},\mathbf{u}_{t+1+i}) - \eta^-(t+1+i,\mathbf{z},\mathbf{v}_{t+1+i})) \right). \tag{C.92}$$

So the norm of the difference is the following

$$\|(T\mathbf{u})_{t+1} - (T\mathbf{v})_{t+1}\| \leq \sum_{i=0}^{t} \|B(t,i+1)\| \|\eta^+(i,\mathbf{z},\mathbf{u}_i) - \eta^+(i,\mathbf{z},\mathbf{v}_i)\| \tag{C.93}$$

$$+ \sum_{i=0}^{\infty} \|C(t+1+i,t+1)^{-1}\| \|\eta^-(t+1+i,\mathbf{z},\mathbf{u}_{t+1+i}) - \eta^-(t+1+i,\mathbf{z},\mathbf{v}_{t+1+i})\|. \tag{C.94}$$

Since $\eta(t,\mathbf{z},\mathbf{x}) = \gamma(t,\mathbf{z})\theta(\mathbf{x})$, and then by Lipschitz condition on $\theta(\mathbf{x})$, we have that

$$\|\eta^+(i,\mathbf{z},\mathbf{u}_i) - \eta^+(i,\mathbf{z},\mathbf{v}_i)\| \leq \|\eta(i,\mathbf{z},\mathbf{u}_i) - \eta(i,\mathbf{z},\mathbf{v}_i)\| \tag{C.95}$$

$$= \|\gamma(i,\mathbf{z})\theta(\mathbf{u}_i) - \gamma(i,\mathbf{z})\theta(\mathbf{v}_i)\| \tag{C.96}$$

$$= \gamma(i,\mathbf{z})\epsilon\|\theta(\mathbf{u}_i) - \theta(\mathbf{v}_i)\| \tag{C.97}$$

$$\leq \gamma(i,\mathbf{z})\epsilon\|\mathbf{u}_i - \mathbf{v}_i\|. \tag{C.98}$$

Same argument gives estimate on η^-:

$$\|\eta^-(t+1+i,\mathbf{z},\mathbf{u}_{t+1+i}) - \eta^-(t+1+i,\mathbf{z},\mathbf{v}_{t+1+i})\| \leq \gamma(t+1+i,\mathbf{z})\epsilon\|\mathbf{u}_{t+1+i} - \mathbf{v}_{t+1+i}\|.$$

The norm $\|(T\mathbf{u})_{t+1} - (T\mathbf{v})_{t+1}\|$ can be estimated as

$$\|(T\mathbf{u})_{t+1} - (T\mathbf{v})_{t+1}\| \leq \sum_{i=0}^{t} \|B(t,i+1)\| \cdot \gamma(i,\mathbf{z})\epsilon\|\mathbf{u}_i - \mathbf{v}_i\| \tag{C.99}$$

$$+ \sum_{i=0}^{\infty} \|C(t+1+i,t+1)^{-1}\| \cdot \gamma(t+1+i,\mathbf{z})\epsilon\|\mathbf{u}_{t+1+i} - \mathbf{v}_{t+1+i}\| \tag{C.100}$$

$$\leq \sum_{i=0}^{t} \|B(t,i+1)\| \cdot \gamma(i,\mathbf{z})\epsilon \sup_{t\geq 0}\|\mathbf{u}_t - \mathbf{v}_t\| \tag{C.101}$$

$$+\sum_{i=0}^{\infty}\|C(t+1+i,t+1)^{-1}\|\cdot\gamma(t+1+i,\mathbf{z})\epsilon\sup_{t\geq 0}\|\mathbf{u}_t-\mathbf{v}_t\| \tag{C.102}$$

$$=\sum_{i=0}^{t}\|B(t,i+1)\|\cdot\gamma(i,\mathbf{z})\epsilon d(\mathbf{u},\mathbf{v}) \tag{C.103}$$

$$+\sum_{i=0}^{\infty}\|C(t+1+i,t+1)^{-1}\|\cdot\gamma(t+1+i,\mathbf{z})\epsilon d(\mathbf{u},\mathbf{v}) \tag{C.104}$$

Since

$$\gamma(i,\mathbf{z})\leq\gamma(0,\mathbf{z}),$$

the above inequality can be simplified to

$$\|(T\mathbf{u})_{t+1}-(T\mathbf{v})_{t+1}\|\leq\left(\sum_{i=0}^{t}\|B(t,i+1)\|+\sum_{i=0}^{\infty}\|C(t+1+i,t+1)^{-1}\|\right)\gamma(0,\mathbf{z})\epsilon d(\mathbf{u},\mathbf{v}) \tag{C.105}$$

$$\leq\left(\frac{1}{c\lambda_s}-\frac{1}{c\lambda_{s+1}}\right)\gamma(0,\mathbf{z})\epsilon d(\mathbf{u},\mathbf{v}) \tag{C.106}$$

$$\leq\left(\frac{1}{c\lambda_s}-\frac{1}{c\lambda_{s+1}}\right)\frac{1}{\sqrt{\xi}}\epsilon d(\mathbf{u},\mathbf{v}) \tag{C.107}$$

The coefficient

$$\left(\frac{1}{c\lambda_s}-\frac{1}{c\lambda_{s+1}}\right)\frac{1}{\sqrt{\xi}}\epsilon$$

needs to be adjusted so that it is less than 1, and this can be achieved by tuning ϵ so that

$$\epsilon<\frac{c\lambda_s\lambda_{s+1}\sqrt{\xi}}{\lambda_{s+1}-\lambda_s}.$$

Since $-c\lambda_{s+1}\sqrt{\frac{\xi}{2}}<-c\lambda_{s+1}\sqrt{\xi}$, combining with the condition for ϵ in the previous lemma, we have that

$$\epsilon<\min\left\{-c\lambda_{s+1}\sqrt{\frac{\xi}{2}},c^2\lambda_s^2\sqrt{\frac{\xi}{2}},\frac{c\lambda_s\lambda_{s+1}\sqrt{\xi}}{\lambda_{s+1}-\lambda_s}\right\}$$

suffices to make T a contraction mapping, i.e., there exists $\kappa<1$ such that

$$d(T\mathbf{u},T\mathbf{v})\leq\kappa d(\mathbf{u},\mathbf{v}).$$

■

The following lemma guarantees that by a linear transformation, whose matrix comes from diagonalization of Hessian matrix at saddle point, the stable manifold of the diagonalized dynamical system can be carried to the stable manifold of the diagonalizable dynamical system.

Lemma C.7. *Let G be a diagonalizable real matrix. Then the dynamical system*

$$\mathbf{x}_{t+1}=(I-\gamma(t,\mathbf{x}_0)G)\mathbf{x}_t-\gamma(t,\mathbf{x}_0)\theta(\mathbf{x}_t).$$

has a local stable manifold at $\mathbf{0}$.

Proof. In this proof, we denote $\gamma_t=\gamma(t,\mathbf{x}_0)$ for short. Since G is diagonalizable, there exists an invertible matrix Q such that

$$G=Q^{-1}HQ,$$

and hence we have

$$QGQ^{-1} = H,$$

where

$$H = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix}.$$

Consider the linear map

$$\mathbf{z} = \varphi(\mathbf{x}) = Q\mathbf{x},$$

note that φ induces a new dynamical system of $\mathbf{z}$:

$$Q^{-1}\mathbf{z}_{t+1} = (I - \gamma_t G)Q^{-1}\mathbf{z}_t - \gamma_t \theta(Q^{-1}\mathbf{z}_k).$$

Multiplying by Q from the left on both sides, we have

$$\mathbf{z}_{t+1} = Q(I - \gamma_t G)Q^{-1}\mathbf{z}_t - \gamma_t Q\theta(Q^{-1}\mathbf{z}_t) \tag{C.108}$$

$$= (I - \gamma_t H)\mathbf{z}_t - \gamma_t \hat{\theta}(\mathbf{z}_t), \tag{C.109}$$

where

$$\hat{\theta}(\mathbf{z}_t) = Q\theta(Q^{-1}\mathbf{z}_t).$$

We next verify that $\hat{\theta}$ satisfies the Lipschitz condition, i.e., given any $\epsilon > 0$, there exists a $\delta' > 0$, such that

$$\|\hat{\theta}(\mathbf{w}_1) - \hat{\theta}(\mathbf{w}_2)\| = \|Q\theta(Q^{-1}\mathbf{w}_1) - Q\theta(Q^{-1}\mathbf{w}_2)\| \tag{C.110}$$

$$\leq \epsilon \|\mathbf{w}_1 - \mathbf{w}_2\| \tag{C.111}$$

for all $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{B}(\mathbf{0}, \delta')$.

For any given $\epsilon > 0$ and a fixed linear isomorphism Q, with respect to

$$\frac{\epsilon}{\|Q\|\|Q^{-1}\|}$$

there exists a $\delta > 0$, such that

$$\|\theta(\mathbf{u}_1) - \theta(\mathbf{u}_2)\| \leq \frac{\epsilon}{\|Q\|\|Q^{-1}\|}\|\mathbf{u}_1 - \mathbf{u}_2\|$$

for all $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{B}(\mathbf{0}, \delta)$. Denote

$$V := Q\left(\mathbb{B}(\mathbf{0}, \delta)\right),$$

i.e.,

$$V = \{\mathbf{w} \in \mathbb{R}^d : \mathbf{w} = Q(\mathbf{u}) \text{ for some } \mathbf{u} \in \mathbb{B}(\mathbf{0}, \delta)\}.$$

Since $Q\mathbf{u}$ is a linear diffeomorphism (change of basis is of full rank) from the open ball $\mathbb{B}(\mathbf{0}, \delta)$ to $\mathbb{R}^d$, V is an open neighborhood of $\mathbf{0}$. Therefore, there exists an open ball at $\mathbf{0}$ with radius δ', denoted as $\mathbb{B}(\mathbf{0}, \delta')$, such that $\mathbb{B}(\mathbf{0}, \delta') \subset V$. By definition of V, we have that for any $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{B}(\mathbf{0}, \delta') \subset V$, there exist $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{B}(\mathbf{0}, \delta)$, such that

$$\begin{cases} \mathbf{w}_1 = Q\mathbf{u}_1 \\ \mathbf{w}_2 = Q\mathbf{u}_2, \end{cases} \tag{C.112}$$

and the inverse transformation is given by

$$\begin{cases} \mathbf{u}_1 = Q^{-1}\mathbf{w}_1 \\ \mathbf{u_2} = Q^{-1}\mathbf{w}_2. \end{cases} \tag{C.113}$$

And then we have

$$\|\hat{\theta}(\mathbf{w}_1) - \hat{\theta}(\mathbf{w}_2)\| = \|Q\theta(Q^{-1}\mathbf{w}_1) - Q\theta(Q^{-1}\mathbf{w}_2)\| \tag{C.114}$$

$$= \|Q\theta(\mathbf{u}_1) - Q\theta(\mathbf{u}_2)\| \tag{C.115}$$
$$\leq \|Q\|\|\theta(\mathbf{u}_1) - \theta(\mathbf{u}_2)\| \tag{C.116}$$
$$\leq \|Q\| \frac{\epsilon}{\|Q\|\|Q^{-1}\|} \|\mathbf{u}_1 - \mathbf{u}_2\| \tag{C.117}$$
$$= \|Q\| \frac{\epsilon}{\|Q\|\|Q\|^{-1}} \|Q^{-1}\mathbf{w}_1 - Q^{-1}\mathbf{w}_2\| \tag{C.118}$$
$$\leq \|Q\| \frac{\epsilon}{\|Q\|\|Q\|^{-1}} \|\mathbf{w}_1 - \mathbf{w}_2\| \tag{C.119}$$
$$= \epsilon\|\mathbf{w}_1 - \mathbf{w}_2\|. \tag{C.120}$$

The verification completes. Thus the stable manifold (measure zero set of initial condition) of dynamical system with diagonal linear part can be carried to dynamical system with diagonalizable linear part by the linear map Q. ■

D Proof of Proposition 3

We proceed by showing that AdaGrad-Norm, AdaGrad-Diag and FullAdaGrad are diffeomorphism if the parameter δ_0 is chosen properly. Throughout the proof, we denote $\xi = \delta_0^2$ for convenience.

Proof. By diffeomorphism, we mean the last iterate acting on $\mathbf{x}_t$ is a diffeomorphism. Recall that the AdaGrad algorithms have the following form:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \Gamma_t \nabla f(\mathbf{x}_t),$$

where

$$\Gamma_t = \left(\xi I + \sum_{s=1}^{t} \nabla f(\mathbf{x}_s) \nabla f(\mathbf{x}_s)^\top \right)^{-\frac{1}{2}}$$

for the FullAdaGrad, which we consider first. Since the matrix Γ_t can be written as

$$\Gamma_t = \left(\xi I + S + \nabla f(\mathbf{x}_t) \nabla f(\mathbf{x}_t) \right)^{-\frac{1}{2}},$$

where

$$S = \sum_{s=1}^{t-1} \nabla f(\mathbf{x}_s) \nabla f(\mathbf{x}_s),$$

then the update rule considered a mapping acting on $\mathbf{x}_t$ can be reduce to the following form

$$\varphi(\mathbf{x}) = \mathbf{x} - \Omega(\mathbf{x}) \nabla f(\mathbf{x})$$

where

$$\Omega(\mathbf{x}) = \Gamma_t(\mathbf{x}) = \begin{bmatrix} \omega_{11} & \dots & \omega_{1d} \\ \vdots & & \\ \omega_{d1} & \dots & \omega_{dd} \end{bmatrix}$$

and the entries ω_{ij} are functions of $\mathbf{x}$, i.e., $\omega_{ij} = \omega_{ij}(\mathbf{x})$.

The differential of $\varphi(\mathbf{x})$ is

$$D\varphi(\mathbf{x}) = I - D\left(\Omega(\mathbf{x}) \nabla f(\mathbf{x})\right)$$

By matrix differentiation rule, we have

$$D\left(\Omega(\mathbf{x}) \nabla f(\mathbf{x})\right) = \begin{bmatrix} \frac{\partial \Omega_1(\mathbf{x}) \cdot \nabla f(\mathbf{x})}{\partial \mathbf{x}} \\ \vdots \\ \frac{\partial \Omega_d(\mathbf{x}) \cdot \nabla f(\mathbf{x})}{\partial \mathbf{x}} \end{bmatrix} \tag{D.1}$$

$$= \begin{bmatrix} \Omega_1(\mathbf{x}) \frac{\partial \nabla f(\mathbf{x})}{\partial \mathbf{x}} + \nabla f(\mathbf{x})^\top \frac{\partial \Omega_1(\mathbf{x})}{\partial \mathbf{x}} \\ \vdots \\ \Omega_d(\mathbf{x}) \frac{\partial \nabla f(\mathbf{x})}{\partial \mathbf{x}} + \nabla f(\mathbf{x})^\top \frac{\partial \Omega_d(\mathbf{x})}{\partial \mathbf{x}} \end{bmatrix} \tag{D.2}$$

Note that the matrix

$$\begin{bmatrix} \Omega_1(\mathbf{x}) \frac{\partial \nabla f(\mathbf{x})}{\partial \mathbf{x}} \\ \vdots \\ \Omega_d(\mathbf{x}) \frac{\partial \nabla f(\mathbf{x})}{\partial \mathbf{x}} \end{bmatrix}$$

can have bounded norm as small as possible if ξ is taken to be small enough. On the other hand,

$$\frac{\partial \Omega_i(\mathbf{x})}{\partial \mathbf{x}}$$

is of small norm if ξ is taken to be large.

Note that the matrix Ω satisfies the following property

$$\Omega^2(\mathbf{x}) \left(\xi I + S + \nabla f(\mathbf{x}) \nabla f(\mathbf{x})^\top \right) = I,$$

where S is a matrix independent of $\mathbf{x}$, and

$$(\Omega^2(\mathbf{x}))_{ij} = \sum_s \omega_{is} \omega_{sj}.$$

Therefore

$$\sum_k (\Omega^2)_{ik} (\xi \delta_{kj} + S_{kj} + \partial_k f \partial_j f) \tag{D.3}$$

$$= \sum_k \left(\sum_s \omega_{is} \omega_{sk} \right) (\xi \delta_{kj} + S_{kj} + \partial_k f \partial_j f) = \delta_{ij} \tag{D.4}$$

Take partial derivative with respect to any x_α, we have that

$$\sum_k \left[\frac{\partial}{\partial x_\alpha} \left(\sum_s \omega_{is} \omega_{sk} \right) (\xi \delta_{kj} + S_{kj} + \partial_k f \partial_j f) + \left(\sum_s \omega_{is} \omega_{sk} \right) \frac{\partial}{\partial x_\alpha} (\xi \delta_{kj} + S_{kj} + \partial_k f \partial_j f) \right] = 0.$$

By assumption that

$$\frac{\partial}{\partial x_\alpha} (\partial_k f \partial_j f)$$

is bounded and since ξ can be taken as large as possible, the latter term

$$\left(\sum_s \omega_{is} \omega_{sk} \right) \frac{\partial}{\partial x_\alpha} (\xi \delta_{kj} + S_{kj} + \partial_k f \partial_j f)$$

is uniformly bounded. And thus

$$\frac{\partial}{\partial x_\alpha} \left(\sum_s \omega_{is} \omega_{sk} \right) \to 0 \text{ as } \xi \to \infty,$$

which will implies that the norm of

$$\frac{\partial \Omega}{\partial \mathbf{x}}$$

approaches 0 as $\xi \to \infty$. We have completed the proof that when $\xi \to \infty$, $\det D\varphi \to 1$, which means that φ is a diffeomorphism. The above argument automatically applies to AdaGrad-Diag, to see that AdaGrad-Norm is also a diffeomorphism, we take an explicit calculation. The algorithm

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{\sqrt{\sum_{i=0}^t \|\nabla f(\mathbf{x}_i)\|^2 + \xi}} \nabla f(\mathbf{x}_t)$$

can be written as the update rule

$$g(t,\mathbf{x}) = \mathbf{x} - \frac{1}{\sqrt{S_t + \|\nabla f(\mathbf{x})\|^2}}\nabla f(\mathbf{x})$$

where

$$S_t = \sum_{i=0}^{t-1}\|\nabla f(\mathbf{x}_i)\|^2 + \xi.$$

For each fixed $t \in \mathbb{N}$, we can ignore the t and denote

$$\gamma(\mathbf{x}) = \frac{1}{\sqrt{S + \|\nabla f(\mathbf{x})\|^2}}$$

and keep in mind that S can be chosen as large as possible. Then the update rule is written as

$$g(\mathbf{x}) = \mathbf{x} - \gamma(\mathbf{x})\nabla f(\mathbf{x}).$$

The Jacobian of g is computed by

$$Dg(\mathbf{x}) = I - \big(\nabla\gamma(\mathbf{x}) \otimes \nabla f(\mathbf{x}) + \gamma(\mathbf{x})\nabla^2 f(\mathbf{x})\big) \tag{D.5}$$

The term $\gamma(\mathbf{x})\nabla^2 f(\mathbf{x})$ is bounded and can be arbitrarily small if $\gamma(\mathbf{x})$ is small because $\|\nabla^2 f(\mathbf{x})\|$ is assumed to be bounded. The first term

$$\nabla\gamma(\mathbf{x}) \otimes \nabla f(\mathbf{x})$$

is a matrix with entries in the form of

$$\frac{\partial\gamma}{\partial x_i}\frac{\partial f}{\partial x_j}.$$

Since in the beginning we assume the algorithm runs in a compact subset of $\mathbb{R}^d$, $\frac{\partial f}{\partial x_j}$ are uniformly bounded. Moreover

$$\frac{\partial\gamma}{\partial x_i} = \frac{\partial}{\partial x_i}\frac{1}{\sqrt{S + \|\nabla f(\mathbf{x})\|^2}} \tag{D.6}$$

$$= -\frac{1}{2}(S + \|\nabla f(\mathbf{x})\|^2)^{-\frac{3}{2}}\frac{\partial\|\nabla f(\mathbf{x})\|^2}{\partial x_i} \tag{D.7}$$

$$= -\frac{1}{2}(S + \|\nabla f(\mathbf{x})\|^2)^{-\frac{3}{2}}\frac{\partial}{\partial x_i}\left(\sum_{k=1}^{d}\left(\frac{\partial f}{\partial x_k}\right)^2\right) \tag{D.8}$$

$$= -\frac{1}{2}(S + \|\nabla f(\mathbf{x})\|^2)^{-\frac{3}{2}}\left(\sum_{k=1}^{d}2\frac{\partial f}{\partial x_k}\frac{\partial^2 f}{\partial x_i\partial x_k}\right) \tag{D.9}$$

Recall that for a matrix A, the ℓ_1 norm is

$$\|A\|_1 = \max\sum_{i=1}^{n}|a_{ij}|,$$

and the inequality

$$\frac{1}{\sqrt{n}}\|A\|_1 \le \|A\|_2 \le \sqrt{n}\|A\|_1$$

shows the boundedness of $\frac{\partial^2 f}{\partial x_i\partial x_k}$ provided $\|\nabla^2 f(\mathbf{x})\|_2 < L$, even without the compactness assumption.

Thus we have proven that $\frac{\partial\gamma}{\partial x_i}$ can be arbitrarily small as long as S is large enough. And this shows that at each point $\mathbf{x}$, the Jacobian $Dg(\mathbf{x})$ can be arbitrarily close to the identity I. Since the determinant $\det(Dg(\mathbf{x}))$ is a continuous function of the entries of $Dg(\mathbf{x})$, so $\det(Dg(\mathbf{x})) \to 1$, as $\gamma(\mathbf{x}) \to 0$, and there exists a large enough S (so that $\gamma(\mathbf{x})$ is close enough to 0) at each point $\mathbf{x}$, such that $\det(Dg(\mathbf{x}))$ is strictly positive. If $\mathbf{x}$ is in a compact set, there exists a uniform S such that the update rule $g(\mathbf{x})$ is a local diffeomorphism everywhere.

■

E Proof of Theorem 2

We complete the proof of the last theorem by extending local existence of stable manifold to global, and then the measure zero result of initial condition follows from the fact that the manifold is of lower dimension than that of $\mathbb{R}^d$.

Proof. Note that the dynamical system defined by AdaGrad algorithms determines each iterate based on time t and the initial condition $\mathbf{x}_0$, thus the t'th iterate can be thought as the image of a mapping depending on t adn $\mathbf{x}_0$, we denote this mapping by $\psi(t, \mathbf{x}_0)$, i.e.

$$\mathbf{x}_{t+1} = \psi(t, \mathbf{x}_0).$$

We define

$$\tilde{\psi}(m, n, \mathbf{x}) = \psi(m, ..., \psi(n+1, \psi(n, \mathbf{x}))...) \text{ for } m > n.$$

The stable set of a set of fixed point $\mathcal{A}^*$, denoted by $W^s(\mathcal{A}^*)$, of the dynamical system defined by $\psi(t, \mathbf{x})$ is

$$W^s(\mathcal{A}^*) = \{\mathbf{x}_0 : \lim_{k\to\infty} \tilde{\psi}(k, 0, \mathbf{x}_0) \in \mathcal{A}^*\}.$$

Fix a point $\mathbf{x}_0 \in W^s(\mathcal{A}^*)$. Since

$$\tilde{\psi}(k, 0, \mathbf{x}_0) \to \mathbf{x}^* \in \mathcal{A}^*,$$

there exists some non-negative integer T and all $t \geq T$, such that

$$\tilde{\psi}(t, 0, \mathbf{x}_0) \in \bigcup_{\mathbf{x}^* \in \mathcal{A}^*} U_{\mathbf{x}^*} = \bigcup_{i=1}^{\infty} U_{\mathbf{x}_i^*}.$$

So $\tilde{\psi}(t, 0, \mathbf{x}_0) \in U_{\mathbf{x}_i^*}$ for some $\mathbf{x}_i^* \in \mathcal{A}^*$ and all $t \geq T$. This is equivalent to

$$\tilde{\psi}(T + k, T, \tilde{\psi}(T, 0, \mathbf{x}_0)) \in U_{\mathbf{x}_i^*}$$

for all $k \geq 0$, and this implies that

$$\tilde{\psi}(T, 0, \mathbf{x}_0) \in \tilde{\psi}^{-1}(T + k, T, U_{\mathbf{x}_i^*})$$

for all $k \geq 0$. And then we have

$$\tilde{\psi}(T, 0, \mathbf{x}_0) \in \bigcap_{k=0}^{\infty} \tilde{\psi}^{-1}(T + k, T, U_{\mathbf{x}_i^*}).$$

Denote $S_{i,T} := \bigcap_{k=0}^{\infty} \tilde{\psi}^{-1}(T + k, T, U_{\mathbf{x}_i^*})$ and the above relation is equivalent to $\mathbf{x}_0 \in \tilde{\psi}^{-1}(T, 0, S_{i,T})$. Take the union for all nonnegative integers T, we have

$$\mathbf{x}_0 \in \bigcup_{T=0}^{\infty} \tilde{\psi}^{-1}(T, 0, S_{i,T}).$$

And union for all i we obtain that

$$\mathbf{x}_0 \in \bigcup_{i=1}^{\infty} \bigcup_{T=0}^{\infty} \tilde{\psi}^{-1}(T, 0, S_{i,T})$$

implying that

$$W^s(\mathcal{A}^*) \subset \bigcup_{i=1}^{\infty} \bigcup_{T=0}^{\infty} \tilde{\psi}^{-1}(T, 0, S_{i,T}).$$

Since $S_{i,T} \subset W_n(\mathbf{x}^*)$, and $W_n(\mathbf{x}^*)$ has codimension at least 1. This implies that $S_{i,T}$ has measure 0 with respect to the volume measure from the Riemannian metric on M. Since the image of set of measure zero under diffeomorphism is of measure zero, and countable union of zero measure sets is still measure zero, we obtain that $W^s(\mathcal{A}^*)$ is of measure zero. ■