
Convergence Rates for Gradient Descent on the Edge of Stability for Overparametrised Least Squares

Lachlan Ewen MacDonald Hancheng Min* Leandro Palma* Salma Tarmoun*
Ziqing Xu* René Vidal

Innovation in Data Engineering and Science (IDEAS)
University of Pennsylvania
Pennsylvania, PA 19104
lemacdonald@protonmail.com

Abstract

Classical optimisation theory guarantees monotonic objective decrease for gradient descent (GD) when employed in a small step size, or “stable”, regime. In contrast, gradient descent on neural networks is frequently performed in a large step size regime called the “edge of stability”, in which the objective decreases non-monotonically with an observed implicit bias towards flat minima. In this paper, we take a step toward quantifying this phenomenon by providing convergence rates for gradient descent with large learning rates in an overparametrised least squares setting. The key insight behind our analysis is that, as a consequence of overparametrisation, the set of global minimisers forms a Riemannian manifold M , which enables the decomposition of the GD dynamics into components parallel and orthogonal to M . The parallel component corresponds to Riemannian gradient descent on the objective sharpness, while the orthogonal component is a bifurcating dynamical system. This insight allows us to derive convergence rates in three regimes characterised by the learning rate size: (a) the subcritical regime, in which transient instability is overcome in finite time before linear convergence to a suboptimally flat global minimum; (b) the critical regime, in which instability persists for all time with a power-law convergence toward the optimally flat global minimum; and (c) the supercritical regime, in which instability persists for all time with linear convergence to an orbit of period two centred on the optimally flat global minimum.

1 Introduction

The well-known gradient descent (GD) iteration

$$\theta \mapsto \text{GD}^\ell(\eta, \theta) := \theta - \eta \nabla \ell(\theta), \quad \theta \in \mathbb{R}^p \quad (1)$$

for minimisation of an objective function $\ell : \mathbb{R}^p \rightarrow \mathbb{R}$, with learning rate $\eta > 0$, is the foundation of most practical algorithms for training deep neural networks (DNNs). Despite the apparent simplicity of the algorithm and *practical* ease with which DNNs can be trained using GD, *theoretically* GD on DNNs remains poorly understood. In addition to the well-known non-convexity of DNN loss functions ℓ , our theoretical understanding is hampered by the fact that in practice DNNs tend to be trained with larger learning rates than permissible according to standard optimisation theory.

Specifically, classical optimisation theory requires that the learning rate η be less than twice the reciprocal of the largest eigenvalue λ (the *sharpness*) of the Hessian $\nabla^2 \ell$. The reason for this is easily

*Equal contribution, alphabetical order

seen by considering the simplest quadratic objective $\ell(\theta) := \frac{1}{2}\lambda\theta^2$ of a single variable θ : if $\eta > 2/\lambda$, then the iterates $\theta_{t+1} = (1 - \eta\lambda)\theta_t$ of $\text{GD}^\ell(\eta, \cdot)$ grow in magnitude like $|1 - \eta\lambda|^t$ leading to rapid divergence. In contrast, DNNs are typically trained in a regime called the *edge of stability* (EOS) [13], in which η is often (pointwise) strictly larger than $2/\lambda$. Miraculously, despite this large learning rate, GD is capable of converging *non-monotonically* to a global minimum of the loss, exhibiting an apparent bias toward less sharp (*flatter*) minima.

Despite significant theoretical progress in the past few years, GD at the edge of stability is still poorly understood when compared with its small learning rate counterpart. In particular, while convergence rates for GD in the small learning rate regime can be obtained for DNNs using standard methods [25], convergence rates for GD at the edge of stability have so far only been obtained in settings wherein monotonic loss decrease can be eventually guaranteed [42, 41, 11, 32].

Our purpose in this paper is to provide convergence rates in a *least squares* setting, wherein sustained monotonic decrease of the loss *may never occur*. To our knowledge, these are the first rates of this kind to be provided in the literature.

Our analysis synthesises *Riemannian geometry* and *dynamical systems theory* to formalise an intuition that has been folklore in the literature for some time [16, 12], namely that GD with a large step size implicitly favours flatter global minima. This formalisation is achieved for *codimension 1, overparametrised least squares problems*, including deep scalar factorisation, in which the solutions (global minima) are guaranteed to form a $p - 1$ -dimensional Riemannian submanifold M of the p -dimensional Euclidean parameter space $\mathbb{R}^p$. Inspired by the “self-stabilisation” idea proposed in [16], we introduce new coordinates $(\theta^\parallel, \theta^\perp)$ for a neighbourhood of M , where $\theta^\parallel$ is a point in M and $\theta^\perp$ coordinatises the direction orthogonal to M , with respect to which the GD map $(\theta^\parallel, \theta^\perp) \mapsto (\text{GD}^\parallel(\theta^\parallel, \theta^\perp), \text{GD}^\perp(\theta^\parallel, \theta^\perp))$ has approximately the following form:

$$\text{GD}^\parallel : (\theta^\parallel, \theta^\perp) \mapsto \text{GD}_M^\lambda \left(\frac{\eta(\theta^\perp)^2}{2c(\eta, \theta^\parallel)}, \theta^\parallel \right) \quad \text{GD}^\perp : (\theta^\parallel, \theta^\perp) \mapsto (1 - \eta\lambda(\theta^\parallel))\theta^\perp + (\theta^\perp)^3. \quad (2)$$

The left map GD_M^λ is Riemannian gradient descent on λ along M (see (4)), with step size $\eta(\theta^\perp)^2/(2c(\eta, \theta^\parallel))$ for some strictly positive number $c(\eta, \theta^\parallel)$; the right is the normal form for a period-doubling bifurcation [27]. Our analysis thereby clarifies the implicit bias of GD toward flat minima:

In a neighbourhood of the solution manifold, GD with a large learning rate implicitly performs *Riemannian GD on the sharpness* along the solution manifold, and oscillates as a *bifurcating dynamical system* in the direction orthogonal to the solution manifold.

Paper contributions. In the context of codimension 1 least squares problems, we:

1. Make the implicit bias of GD toward flat minima explicit: overparametrisation enables the decomposition of GD into an explicitly sharpness-minimising component along the solution manifold, controlled by an oscillating component orthogonal to the solution manifold.
2. Prove that a class of non-convex scalar factorisation problems fit into our framework. In particular, we prove that despite the non-convexity of the landscape itself, the sharpness in such examples is *geodesically strongly convex* along a geodesic ball in the solution manifold. To our knowledge, this observation has not previously been made.
3. Identify and prove convergence and implicit bias results for large step size GD in three regimes:
 - (a) the *subcritical* regime, in which the iterates can be guaranteed to (eventually) decrease monotonically at a linear rate until convergence to a suboptimally flat minimum;
 - (b) the *critical* regime, in which the iterates converge non-monotonically with a power law rate to the sharpness-minimising global minimum;
 - (c) the *supercritical* regime, in which the iterates converge linearly to a stable period-two orbit orthogonal to the solution manifold at the point of minimum sharpness.

Although the former and the latter have already been observed in some examples, to our knowledge the middle regime has not yet been identified. Moreover, our work is the first to be able to theoretically quantify the rate of convergence and implicit bias in all three regimes.

4. Verify our theory against numerical experiments.

2 Related work

Overparametrisation and convergence guarantees for DNNs. Convergence analyses of GD for non-convex loss landscapes such as DNNs are by now well-established [19, 18, 3, 35, 34, 36, 10]. Although none of these works are able to account for non-monotonic convergence of GD using a large step size, they do point to *overparametrisation* as a key feature of DNN loss landscapes which enables a (local) Polyak-Łojasiewicz inequality and thus convergence guarantees in the absence of convexity. Specifically, overparametrisation enables the map f sending parameters to network outputs to be *submersive*, in the sense that the derivative matrix Df of f is pointwise full-rank. Much of the analysis involved in the aforementioned papers is aimed at quantifying this submersivity by lower-bounding the smallest eigenvalue of the “neural tangent kernel” $Df Df^T$ [24, 10]. Overparametrisation in this sense plays a key role in our work too, since it guarantees that the global minima form a smooth manifold with a well-defined normal direction, which is the foundation for Equation (2) and all the analysis that follows.

GD with large learning rates. Recent work in convex optimisation has demonstrated surprising benefits of large step sizes [21, 22, 23, 4, 5], however the mechanism in these works appears to be distinct from the EOS phenomenon in deep learning. Concerning deep learning, although local stability analysis had already hinted at a relation between step size and sharpness prior to 2020 [43], the seminal work of [13] in 2021 was the first to conduct a systematic empirical study of large step size GD. In particular, [13] exposed the fact that GD on DNNs typically uses a larger step size than admissible by classical theory. Moreover, [13] made the empirical observation that this larger step size implicitly regularises sharpness during the late stages training, and coined the term “edge of stability” for this phase. An explosion of empirical and theoretical work has since been produced in an attempt to account for these empirical observations [2, 7, 40, 38, 39, 16, 30, 44, 1, 12, 26, 42, 41, 11, 17, 32, 20]. At a high level, work up to this point on large step size GD can be classified into one of two categories: *general* analysis attempting to outline the essence of the mechanism of stability in this regime while abstracting from specific architectures [16, 14]; and *specific* analysis focusing on precise architectural details in an attempt to derive more fine-grained results on convergence [42, 41, 11, 32] or implicit bias [38, 39, 12, 17]. The advantage of the former approach is generality and clarity of insight, with the disadvantage of less facility in the proof of quantitative results. The advantage of the latter is the wealth of tools for the proof of quantitative results, with the disadvantage of obfuscating essential mathematical structures with inessential architectural detail.

Our work attempts to achieve the strengths of both approaches with a middle ground of abstraction. On the one hand, in common with and inspired by the seminal work [16], our formulation identifies and hinges on minimal mathematical structures which underlie the dynamics of GD with large step size, independent of architectural details; on the other hand, in common with the latter works, we retain enough mathematical structure in our assumptions to cover some of the specific examples already studied in prior work and obtain quantitative convergence rates and implicit bias characterisations. In particular, we are able to provide convergence rates outside of the “eventually monotonically stable” regimes considered in prior work [42, 41, 11, 32], as well as rigorously quantify oscillatory implicit biases of large step size GD observed empirically and partially accounted for theoretically in [12, 17].

3 Theoretical setting

3.1 Notation

Given a smooth (i.e., infinitely differentiable, C^∞) manifold M without boundary, and a point $\theta \in M$, we will use $T_\theta M$ to denote the tangent space to M at θ , and $TM := \bigsqcup_{\theta \in M} T_\theta M$ to denote its tangent bundle. Given a smooth map $g : M \rightarrow N$ of smooth manifolds, we will denote by $D_M g : TM \rightarrow TN$ its derivative, acting at each point $\theta \in M$ as a linear map $D_M g(\theta) : T_\theta M \rightarrow T_{g(\theta)} N$ between tangent spaces.

A *Riemannian metric* on M is a smoothly-varying inner product $\langle \cdot, \cdot \rangle_\theta$ on each $T_\theta M$; we will usually drop the subscript and denote the metric by simply $\langle \cdot, \cdot \rangle$. Given Riemannian metrics on M and N , any higher derivative $D_M^k g : TM^{\otimes k} \rightarrow TN$ is also defined, and at each point $\theta \in M$ acts as a symmetric, multilinear map $D_M^k g(\theta) : T_\theta M^{\otimes k} \rightarrow T_{g(\theta)} N$ whose value on vectors $v_1, \dots, v_k \in T_\theta M$ we denote $D_M^k g(\theta)[v_1, \dots, v_k]$ (respectively, $D_M^k g(\theta)[v^{\otimes k}]$ if $v_i = v \ \forall i$). This $D_M^k g$ may also act on vector fields $V_1, \dots, V_k : M \rightarrow TM$ to give a map $D_M^k g[V_1, \dots, V_k] : M \rightarrow TN$ defined by the formula

$D_M^k g[V_1, \dots, V_k](\theta) := D_M^k g(\theta)[V_1(\theta), \dots, V_k(\theta)]$ (resp. $D_M^k g[V^{\otimes k}](\theta) := D_M^k g(\theta)[V(\theta)^{\otimes k}]$ if $V_i = V \forall i$) for $\theta \in M$.

If $g : M \rightarrow \mathbb{R}$ is scalar-valued, then $\nabla_M^k g : TM^{\otimes k-1} \rightarrow TM$ will be used to denote the map obtained by dualising one of the inputs of $D_M^k g$ using the Riemannian metric. Specifically, given any $\theta \in M$ and tangent vectors $v_2, \dots, v_k \in T_\theta M$, $\nabla_M^k g(\theta)[v_2, \dots, v_k]$ is the unique element of $T_\theta M$ satisfying

$$\langle \nabla_M^k g(\theta)[v_2, \dots, v_k], v_1 \rangle = D_M^k g(\theta)[v_1, v_2, \dots, v_k], \quad \forall v_1 \in T_\theta M, \quad (3)$$

where $\langle \cdot, \cdot \rangle$ is the Riemannian metric in $T_\theta M$. The objects $\nabla_M g$ and $\nabla_M^2 g$ are the *Riemannian gradient* and *Riemannian Hessian* respectively. In particular, $D_{\mathbb{R}^p}^k g$ and $\nabla_{\mathbb{R}^p}^k g$ will be denoted simply $D^k g$ and $\nabla^k g$; ∇g and $\nabla^2 g$ then are the ordinary Euclidean gradient and Hessian respectively.

A *geodesic* is a locally length-minimising curve in M (for instance, geodesics in Euclidean space are just straight lines). We will use $d_M(\theta, \theta')$ to denote the geodesic distance between $\theta, \theta' \in M$, which is the length of the shortest geodesic connecting θ to θ' ; d_M makes M into a metric space. For any $\theta \in M$ and any sufficiently small $v \in T_\theta M$, there is a unique geodesic $\gamma : [0, 1] \rightarrow M$ such that $\gamma(0) = \theta$ and $\dot{\gamma}(0) = v$; the *exponential map* on (θ, v) is then by definition $\exp_\theta(v) := \gamma(1)$; in Euclidean space $\exp_\theta(v) = \theta + v$. See Appendix A for more details. Any smooth function $g : M \rightarrow \mathbb{R}$ is associated to a *Riemannian gradient descent* map $(\eta, \theta) \mapsto \text{GD}_M^g(\eta, \theta) \in M$ defined in an open neighbourhood of $\{0\} \times M \subset \mathbb{R}_{\geq 0} \times M$ by

$$\text{GD}_M^g(\eta, \theta) := \exp_\theta(-\eta \nabla_M g(\theta)). \quad (4)$$

This reduces to the familiar $(\eta, \theta) \mapsto \theta - \eta \nabla g(\theta)$ in the Euclidean setting, wherein it is denoted simply GD^g with the subscript dropped.

Recall that a *regular value* of a function $f : \mathbb{R}^p \rightarrow \mathbb{R}^d$ is a point $y \in \mathbb{R}^d$ such that $f^{-1}\{y\}$ is nonempty and f is C^∞ in a neighbourhood of $f^{-1}\{y\}$ with derivative $Df(\theta) : T_\theta \mathbb{R}^p \rightarrow T_{f(\theta)} \mathbb{R}^d$ surjective for all $\theta \in f^{-1}\{y\}$. Moreover, if y is a regular value of f , then $M := f^{-1}\{y\}$ is a smooth submanifold of $\mathbb{R}^p$, and if f is also analytic in a neighbourhood of M , then M is an analytic manifold.

3.2 Problem setting

Given a natural number $p > 1$, function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ and target value $y \in \mathbb{R}$, we aim to solve the *codimension 1 least squares problem*

$$\min_\theta \ell(\theta) := \frac{1}{2}(f(\theta) - y)^2, \quad (5)$$

using gradient descent with constant step size $\eta > 0$:

$$\theta \mapsto \text{GD}^\ell(\eta, \theta) := \theta - \eta \nabla \ell(\theta). \quad (6)$$

We will make the following assumptions on f and y .

Assumption 3.1 (Regularity and analyticity). The point $y \in \mathbb{R}$ is a regular value of f , and f is analytic in a neighbourhood of the pre-image $M := f^{-1}\{y\}$.

The assumption that y is a regular value of f is a strong version of *overparametrisation*: it is equivalent to assuming that the ‘‘neural tangent kernel’’ $Df Df^T = \|Df\|^2$ is non-vanishing along M , as is often insisted upon in overparametrisation theory for deep learning [24, 10]. Analyticity of f in a neighbourhood of $M := f^{-1}\{y\}$ will be used to invoke powerful results from holomorphic dynamics [6] in our convergence theorems.

The key consequence of Assumption 3.1 is that, by the regular value theorem, the solution set M of (5) is a $(p-1)$ -dimensional analytic submanifold of $\mathbb{R}^p$. Points in M will be denoted $\theta^\parallel$ in what follows. For any $\theta^\parallel \in M$, the tangent space $T_{\theta^\parallel} M$ to M at $\theta^\parallel$ is the kernel of $Df(\theta^\parallel)$, and any singular value decomposition of $Df(\theta^\parallel)$ admits precisely one nonzero singular value corresponding to a singular vector orthogonal to $T_{\theta^\parallel} M$. This singular vector is the *normal vector* $n(\theta^\parallel) := \nabla f(\theta^\parallel) / \|\nabla f(\theta^\parallel)\|$.

It follows from the chain rule and the fact that $f \equiv y$ along M that the Hessian $\nabla^2 \ell$ of ℓ satisfies the identity

$$\nabla^2 \ell|_M \equiv (\nabla f \nabla f^T + (f - y) \nabla^2 f)|_M \equiv \nabla f \nabla f^T|_M \quad (7)$$

along M . We denote by λ the largest eigenvalue of $\nabla^2 \ell$, which is equal to $\|\nabla f\|^2$. We assume M to be equipped with the Riemannian metric inherited from its embedding into the Euclidean space $\mathbb{R}^p$.

Our new coordinate representation (2) for gradient descent is essentially a consequence of coordinatising a tubular neighbourhood N of M by a $\parallel$ -coordinate *along* M , and a $\perp$ -coordinate *orthogonal* to M (Figure 1). However, the rigorous form of (2) requires two additional assumptions on the solution manifold M . The first of these provides an invariant line segment about which GD can oscillate, and enables the correct decay rate of the error terms in the $\parallel$ -component of GD.

Assumption 3.2 (Orthogonal stability). There is $\theta_*^\parallel \in M$ and a line segment $\mathcal{L}$ through $\theta_*^\parallel$ orthogonal to M which is invariant under gradient descent on ℓ (see Figure 1).

The second assumption necessary for the rigorous form of (2), concerning the $\perp$ -equation, is a standard assumption from bifurcation theory enabling the realisation of the $\perp$ -equation in a well-known normal form which is easily analysed [27, Theorem 4.3].

Assumption 3.3 (Genericity). Recall the normal vector field $n := \nabla f / \|\nabla f\|$. At each point $\theta^\parallel \in M$, for all η in a neighbourhood of $2/\lambda(\theta^\parallel)$, one has

$$c(\eta, \theta^\parallel) := \left(\frac{\eta}{2} D^3 \ell[n^{\otimes 3}](\theta^\parallel) \right)^2 - \frac{\eta}{6} D^4 \ell[n^{\otimes 4}](\theta^\parallel) > 0. \quad (8)$$

Assumptions 3.1, 3.2 and 3.3 are sufficient to obtain the rigorous form of (2) (Theorem 4.1). As is evident from Equation (2), GD with a large step size behaves like gradient descent on the sharpness along M . To utilise this fact in our convergence theorems, we will make the following additional assumptions on M and the sharpness λ .

Assumption 3.4 (Strong convexity of λ). There is a ball $B_M(\theta_*^\parallel, r) \subset M$, centred on $\theta_*^\parallel$ and of radius $r > 0$ with respect to the geodesic distance, which is *geodesically convex* in the sense that any two points in $B_M(\theta_*^\parallel, r)$ are connected by a unique distance-minimising geodesic, over which the sharpness $\lambda : M \rightarrow \mathbb{R}$ is μ -geodesically strongly convex and has L -geodesically Lipschitz gradients, in the sense that the Riemannian Hessian $\nabla_M^2 \lambda$ of λ satisfies

$$\mu I_{TM} \preceq \nabla_M^2 \lambda \preceq L I_{TM} \quad (9)$$

uniformly over $B_M(\theta_*^\parallel, r)$, where I_{TM} is the identity

map on TM . We assume moreover that λ achieves its minimum value $\lambda(\theta_*^\parallel)$ in $B_M(\theta_*^\parallel, r)$ uniquely at $\theta_*^\parallel$, where one moreover has

$$\nabla_M^2 \lambda(\theta_*^\parallel) = \nu I_{TM}(\theta_*^\parallel) \quad (10)$$

for some $\nu > 0$, where $I_{TM}(\theta_*^\parallel)$ is the identity map on the tangent space $T_{\theta_*^\parallel} M$ of M at $\theta_*^\parallel$.

These assumptions need not be satisfied for general functions f . The geodesic strong convexity of λ in particular seems at first to be an alarmingly strong assumption, since it is well-known that DNN loss landscapes are *non-convex*. Surprisingly, these non-trivial assumptions *can be guaranteed to hold* for the following class of toy examples of deep learning non-convex loss landscapes.

Multilayer scalar factorisation. The map $f : \mathbb{R}^p \rightarrow \mathbb{R}$ defined by $f(\theta^1, \dots, \theta^p) := \theta^p \dots \theta^1$ corresponds to a linear network of depth p and width 1 on a single input datum. Any nonzero target value $y \neq 0$ is a regular value for f , and $f^{-1}\{y\}$ is then a union of hypersurfaces (Proposition B.1). Assuming without loss of generality that $y > 0$ and setting $\theta_*^\parallel := y^{\frac{1}{p}} 1_p$, one takes M to be the connected component of $\theta_*^\parallel$. The line $\mathcal{L}$ is given by the span of the ones-vector 1_p (Proposition B.2). Assumption 3.3 holds by Proposition B.3, while Proposition B.6 demonstrates that λ is geodesically strongly convex on a geodesic ball centred at $\theta_*^\parallel$, with

$$\nabla_M^2 \lambda(\theta_*^\parallel) = 4y^{2-4/p} I_{TM}(\theta_*^\parallel) \quad (11)$$

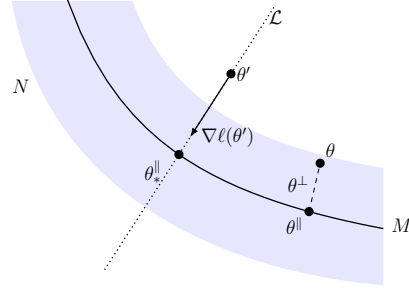


Figure 1: $M = \{(x, y) : xy = 1\}$ with tubular neighbourhood N (shaded) and line $\mathcal{L}$ (dotted). Inside N , any point θ is closest to a unique point $\theta^\parallel$ on M , with $\theta - \theta^\parallel = \theta^\perp n(\theta^\parallel)$ orthogonal to M for some $\theta^\perp \in \mathbb{R}$. Assumption 3.2 says that $\nabla \ell(\theta')$ is parallel to $\mathcal{L}$ at any point $\theta' \in \mathcal{L}$.

so that Assumption 3.4 holds. One may also take $f(\theta^1, \dots, \theta^p) := \theta^p \phi(\theta^{p-1} \phi(\dots (\phi(\theta^1)) \dots))$, with ϕ any nonlinearity that is the identity on a neighbourhood of $y^{1/p}$ and still satisfy all assumptions.

Some thought reveals that these assumptions can be relaxed in more-or-less straightforward ways to deep linear networks of sufficient constant width on multi-point datasets. However, the scalar case we consider in this paper is already sufficiently instructive and non-trivial that we leave a more general elaboration of this framework to future work.

4 A normal form for gradient descent about the solution manifold

In this section we state a rigorous form (Theorem 4.1) of the approximate update equations alluded to in Equation (2), and give an idea of its proof. This section makes use only of Assumptions 3.1, 3.2 and 3.3. The geodesic convexity of Assumption 3.4 is *not* needed to derive the normal form of GD about M , and is necessary only for the convergence theorems to be given in the next section.

Our analysis of the dynamics of GD is inspired by that of [16]; it proceeds from a Taylor expansion². However, while [16] Taylor expands around a set of points with sharpness $\leq 2/\eta$, we Taylor expand around the solution manifold M . Our derivation is novel and provides a number of advantages: the assumptions are more transparent and easily checked, and we are guaranteed that M is a Riemannian manifold, while the set of points with sharpness $\leq 2/\eta$ considered in [16] need not be.

Specifically, every point θ sufficiently close to M admits a unique nearest point $\theta^\parallel$ on M . The difference $\theta - \theta^\parallel$ is then a scalar multiple $\theta^\perp n(\theta^\parallel)$ of the normal vector $n(\theta^\parallel) = \nabla f(\theta^\parallel) / \|\nabla f(\theta^\parallel)\|$ at $\theta^\parallel$. Thus, pairs $(\theta^\parallel, \theta^\perp)$ with $\theta^\parallel \in M$ and sufficiently small $\theta^\perp \in \mathbb{R}$ suffice to completely coordinatise a neighbourhood of M in $\mathbb{R}^p$ known as a *tubular neighbourhood* N [28, p. 147]. Our first result is then the following characterisation of the GD dynamics in terms of $\theta^\parallel, \theta^\perp$, the sharpness λ and its Riemannian gradient descent map GD_M^λ .

Theorem 4.1 (Normal form for GD about M). *There is an analytic change of coordinates $(\eta, \theta^\parallel, \theta^\perp) \mapsto (\eta, \theta)$ for a tubular neighbourhood N of M in which the gradient descent map $\text{GD}^\ell : (\eta, \theta) \mapsto (\eta, \theta - \eta \nabla \ell(\theta^\parallel))$ takes the form $(\eta, \theta^\parallel, \theta^\perp) \mapsto (\eta, \text{GD}^\parallel(\eta, \theta^\parallel, \theta^\perp), \text{GD}^\perp(\eta, \theta^\parallel, \theta^\perp))$, where*

$$\text{GD}^\parallel(\eta, \theta^\parallel, \theta^\perp) = \text{GD}_M^\lambda \left(\frac{\eta(\theta^\perp)^2}{2c(\eta, \theta^\parallel)}, \theta^\parallel \right) + O((\theta^\perp)^3 \min\{1, d_M(\theta^\parallel, \theta_*^\parallel)\}), \quad (12)$$

$$\text{GD}^\perp(\eta, \theta^\parallel, \theta^\perp) = (1 - \eta\lambda(\theta^\parallel))\theta^\perp + (\theta^\perp)^3 + O((\theta^\perp)^4), \quad (13)$$

with $c(\eta, \theta^\parallel)$ defined in Assumption 3.3.

Outline of proof. Recalling the normal vector field $n := \nabla f / \|\nabla f\|$, we approximate the gradient $\nabla \ell(\theta)$ at $\theta \in N$ by the Taylor series

$$\nabla \ell(\theta) = \nabla \ell(\theta^\parallel) + \theta^\perp \nabla^2 \ell(\theta^\parallel)[n(\theta^\parallel)] + \frac{(\theta^\perp)^2}{2} \nabla^3 \ell(\theta^\parallel)[n(\theta^\parallel), n(\theta^\parallel)] + O((\theta^\perp)^3). \quad (14)$$

Since M consists of global minima for ℓ , $\nabla \ell(\theta^\parallel) = 0$. Additionally, since $n(\theta^\parallel)$ is the sole eigenvector along which $\nabla^2 \ell(\theta^\parallel) = \nabla f(\theta^\parallel) \nabla f(\theta^\parallel)^T$ is nontrivial, with eigenvalue $\lambda(\theta^\parallel)$, one has $\nabla^2 \ell(\theta^\parallel)[n(\theta^\parallel)] = \lambda(\theta^\parallel)n(\theta^\parallel)$. One thus has:

$$\nabla \ell(\theta) = \lambda(\theta^\parallel)\theta^\perp n(\theta^\parallel) + \frac{|\theta^\perp|^2}{2} \nabla^3 \ell(\theta^\parallel)[n(\theta^\parallel), n(\theta^\parallel)] + O((\theta^\perp)^3). \quad (15)$$

Plugging this into the gradient descent update formula, applying the orthogonal projection P_{TM} of $T\mathbb{R}^p$ onto TM and n^T of $T\mathbb{R}^p$ onto the normal direction respectively, and noting that

$$P_{TM} \nabla^3 \ell[n, n] = \nabla_M (\nabla^2 \ell[n, n]) = \nabla_M \lambda \quad (16)$$

²The idea of higher order terms inducing implicit bias is also present in literature on Sharpness-Aware Minimisation [8].

since $\nabla(n^T n) = \nabla(1) = 0$, yields the following formulae for the GD update in $(\theta^\parallel, \theta^\perp)$:

$$\theta^\parallel \mapsto \theta^\parallel - \frac{\eta(\theta^\perp)^2}{2} \nabla_M \lambda(\theta^\parallel) + O((\theta^\perp)^3) \quad (17)$$

$$\theta^\perp \mapsto (1 - \eta\lambda(\theta^\parallel))\theta^\perp - \frac{\eta(\theta^\perp)^2}{2} D^3 \ell(\theta^\parallel) [n(\theta^\parallel)^{\otimes 3}] + O((\theta^\perp)^3). \quad (18)$$

The $O((\theta^\perp)^3)$ error in the $\parallel$ -update is tightened to $O((\theta^\perp)^3 \min\{1, d_M(\theta^\parallel, \theta_*^\parallel)\})$ using Assumption 3.2, and $\theta^\parallel - \frac{\eta(\theta^\perp)^2}{2} \nabla_M \lambda(\theta^\parallel)$ is the first order approximation of $\exp_{\theta^\parallel} \left(-\frac{\eta(\theta^\perp)^2}{2} \nabla_M \lambda(\theta^\parallel) \right)$ by the Taylor expansion of the exponential map [33]. The final form is obtained by invoking Assumption 3.3 and applying a standard transformation from bifurcation theory [27, Theorem 4.3]. See Theorem C.6 for a rigorous proof. $\square$

We can immediately make the following qualitative observations of the dynamics in Theorem 4.1, with which we will be able to give an idea of how our convergence theorems work. Let $(\theta_t^\parallel, \theta_t^\perp)$ denote the t^{th} iterate of gradient descent in the $(\theta^\parallel, \theta^\perp)$ coordinates of Theorem 4.1.

The $\parallel$ -component of the GD dynamics is a perturbed version of *Riemannian gradient descent on the sharpness* along the solution manifold M , with time-varying step size determined by the magnitude of the $\perp$ -iterates $\theta_t^\perp$. Recalling that we assume λ to be geodesically strongly convex as in Assumption 3.4, we can be assured of descent guarantees for λ provided $|\theta_t^\perp|$ can be controlled.

The behaviour of $|\theta_t^\perp|$ is governed by the map (13), which is a perturbed, time-varying version of the simpler “flip bifurcation”

$$x \mapsto (1 - \eta\lambda')x + x^3 \quad (19)$$

from bifurcation theory [27]. The iterates x_t of (19) admit the following dynamics.

1. When $\eta < 2/\lambda'$, 0 is a *hyperbolic* attractor, and the iterates $|x_t|$ go to zero exponentially fast. Were this to hold also for the iterates $|\theta_t^\perp|$ of (13), $\theta_t^\parallel$ would evolve like gradient descent on λ with exponentially decaying step size, hence would converge exponentially fast to a suboptimally flat point.
2. When $\eta = 2/\lambda'$, 0 is a *parabolic* attractor, and the iterates $|x_t|$ go to zero like $\Theta(t^{-1/2})$. Were this to hold also for the iterates $|\theta_t^\perp|$ of (13), $\theta_t^\parallel$ would evolve like gradient descent on λ with power-law-decaying stepsize, hence would converge with a power-law rate to the optimally flat point.
3. When $\eta > 2/\lambda'$, there is a hyperbolically attracting orbit of period two with amplitude $\approx \sqrt{\eta\lambda' - 2}$ to which the iterates x_t converge exponentially fast. Were the same to be true for the iterates $\theta_t^\perp$ of (13), $\theta_t^\parallel$ would evolve like gradient descent on λ with step size $\Theta((\eta\lambda' - 2))$, hence would converge exponentially to the optimally flat point.

In the next section, we provide theorems showing that these intuitions gathered from the unperturbed, time-invariant (19) can be rigorously carried over to the perturbed, time-varying system defined by Theorem 4.1.

5 Convergence theorems

In this section, we state our convergence theorems and provide numerical demonstrations of them on a multilayer scalar factorisation problem. In addition to the Assumptions 3.1, 3.2 and 3.3 required for the normal form (Theorem 4.1), we now also invoke Assumption 3.4 to provide rates of convergence. Iterates will be denoted θ_t (respectively $(\theta_t^\parallel, \theta_t^\perp)$) for $t \in \mathbb{N} \cup \{0\}$. For convenience, our $|\theta_t^\perp|$ plots concern the orthogonal coordinate of the intermediate Proposition C.5 rather than that of the final Theorem 4.1; since the transformation going between them is analytic (Theorem C.6), this causes no difference in the convergence rate.

5.1 Subcritical regime

The subcritical regime is when $2/\lambda(\theta_*^\parallel) > \eta > 2/\lambda(\theta_0^\parallel)$. In this regime, considered also in certain examples in prior works [26, 32], the orthogonal component $\theta_t^\perp$ of the iterates exhibits initial, transient oscillation driven by (13) of Theorem 4.1; during this time, the sharpness values $\lambda(\theta_t^\parallel)$ are monotonically decreasing according to Theorem 4.1 to the point of stability $\eta < 2/\lambda(\theta_t^\parallel)$ in a finite number of steps. Following the achievement of stability, $\theta_t^\perp$ decays exponentially fast, resulting in exponentially less aggressive steps in $\theta_t^\parallel$ to decrease $\lambda(\theta_t^\parallel)$, ultimately resulting in convergence to a suboptimally-flat global minimum of ℓ . See Figure 2. The theorem below appears as Theorem D.4 in the appendix.

Theorem 5.1. *Assume that $\eta < 2/\lambda(\theta_*^\parallel)$. Then there is a constant $\gamma > 0$ such that for all $\theta_0^\perp$ sufficiently close to zero and all $\theta_0^\parallel \neq \theta_*^\parallel \in B_M(\theta_*^\parallel, r)$, if $\eta > 2/\lambda(\theta_0^\parallel)$ is sufficiently small, then there is*

$$\tau \leq O\left((\theta_0^\perp)^{-2} \left(\frac{\lambda(\theta_0^\parallel) - \lambda(\theta_*^\parallel)}{2/\eta - \lambda(\theta_*^\parallel)}\right)^\gamma\right) \quad (20)$$

such that $\eta < 2/\lambda(\theta_t^\parallel)$ for all $t \geq \tau$ and $\eta \geq 2/\lambda(\theta_t^\parallel)$ for all $t < \tau$. Consequently, setting $\beta := 1 - (2 - \eta\lambda(\theta_\tau^\parallel)) < 1$, the iterates $(\theta_t^\parallel, \theta_t^\perp)$ converge to a global minimum $(\theta_\infty^\parallel, 0)$ of ℓ in M with suboptimal sharpness

$$\lambda(\theta_\infty^\parallel) - \lambda(\theta_*^\parallel) \geq \exp\left(-O((\theta_\tau^\perp)^2(1 - \beta^2)^{-1})\right)(\lambda(\theta_\tau^\parallel) - \lambda_*). \quad (21)$$

at a rate $O(\beta^t)$.

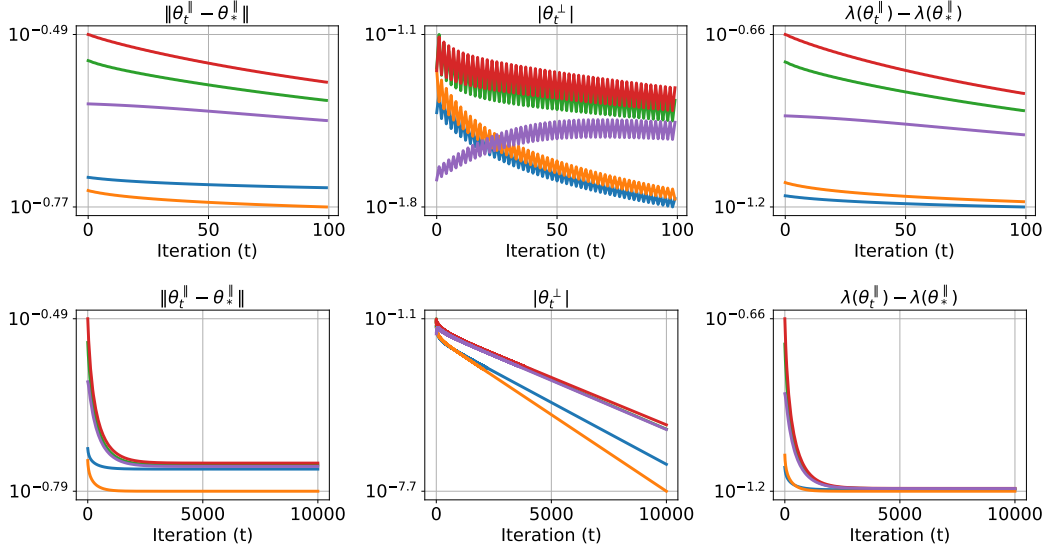


Figure 2: Log-scale plots of distance from $\theta_t^\parallel$ to $\theta_*^\parallel$ (left), magnitude of $\theta_t^\perp$ (centre) and sharpness suboptimality gap (right) for gradient descent on depth 5 scalar factorisation in the **subcritical regime**. Trajectories from five different initialisations shown. Initial instability in $|\theta_t^\perp|$ (top) is overcome in finite time with rapid convergence to a suboptimally flat global minimum (bottom).

5.2 Critical regime

The critical regime happens when $\eta = 2/\lambda(\theta_*^\parallel)$. To our knowledge, this regime has not been observed previously in the literature. In this regime, the dynamics of GD are non-hyperbolic, with $\theta_t^\perp$ exhibiting 2-periodic decay to zero with a rate $\Theta(t^{-\frac{1}{2}})$. Consequently, the dynamics of $\theta_t^\parallel$ are essentially those of gradient descent on λ with a step size decaying according to a power-law. Ultimately then, θ_t does converge to the optimally flat global minimum of ℓ , but does so at a slow rate. See Figure 3. The theorem below appears as Theorem D.9 in the appendix.

Theorem 5.2. Assume that $\eta = 2/\lambda(\theta_*^\parallel)$ and that $\nu/(c(2/\lambda_*, \theta_*^\parallel)\lambda_*) < 1$, where ν and $c(\cdot, \cdot)$ are defined as in Assumptions 3.3 and 3.4 respectively. Then for all $\theta_0^\parallel$ sufficiently close to $\theta_*^\parallel$ and all $\theta_0^\perp \neq 0$ sufficiently small, one has $d_M(\theta_t^\parallel, \theta_*^\parallel) = \Theta(t^{-1/2})$ and $|\theta_t^\perp| = \Theta(t^{-1/2})$.

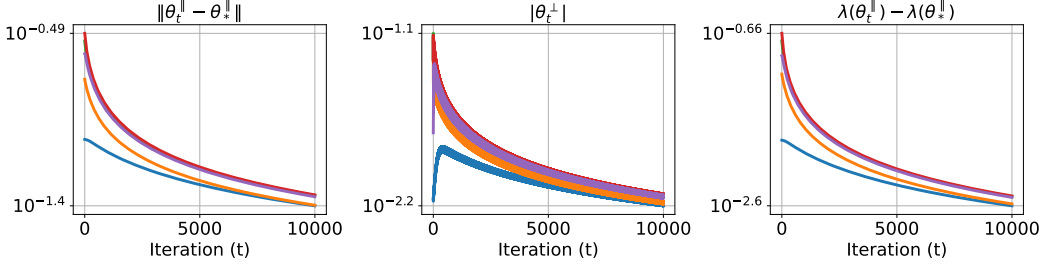


Figure 3: Log- y -scale plots of distance from $\theta_t^\parallel$ to $\theta_*^\parallel$ (left), magnitude of $\theta_t^\perp$ (centre) and sharpness suboptimality gap (right) on depth 5 scalar factorisation in the **critical regime**. Trajectories from five different initialisations shown. The iterates $|\theta_t^\perp|$ may or may not initially increase; in both cases, asymptotic, power law decrease of $|\theta_t^\perp|$ and $\|\theta_t^\parallel - \theta_*^\parallel\|$ to zero indicates power law convergence to the optimally flat global minimum.

Some further remarks about the critical regime are necessary. First, the additional hypothesis that $\nu/(c(2/\lambda_*, \theta_*^\parallel)\lambda_*) < 1$ (which is satisfied for multilayer scalar factorisation, see Proposition B.7) is not strictly necessary for convergence; power-law convergence still occurs to the same minimum without this assumption, but is in that case faster for $d_M(\theta_t^\parallel, \theta_*^\parallel)$ than the $\Theta(t^{-1/2})$ rate given in the theorem statement. However, we are unaware of examples having this faster rate. Second, the $|\theta_t^\perp|$ iterates may initially increase substantially as in Figure 3 depending on the initialisation; our theorem accommodates this behaviour explicitly (see Theorem D.8).

5.3 Supercritical regime

The supercritical regime is when $\eta > 2/\lambda(\theta_*^\parallel)$. In this case, when $\eta > 2/\lambda(\theta_*^\parallel)$ is sufficiently small, Equation (13) exhibits linear convergence to a stable orbit of period two with amplitude $\approx (\eta\lambda(\theta_*^\parallel) - 2)^{1/2}$. It follows that the step sizes in $\theta_t^\parallel$ in its descent on λ are of *approximately constant* size, so that $\theta_t^\parallel$ will converge *linearly* to the optimally flat $\theta_*^\parallel$. However, the complete iterates θ_t of GD do not converge to a global minimum, but asymptotically oscillate orthogonally to M about the point $\theta_*^\parallel$ along the line $\mathcal{L}$ of Assumption 3.2. See Figure 4. Although this regime has been observed in prior works [12, 17, 20], no general quantitative convergence theorem in this regime has yet been proved. The theorem below appears as Theorem D.10 in the appendix.

Theorem 5.3. Assume that $\eta > 2/\lambda(\theta_*^\parallel)$ is sufficiently small. Then there are positive constants C_1, C_2, C_3 and a stable orbit of period two of the form $(\theta_*^\parallel, \pm(\eta\lambda(\theta_*^\parallel) - 2)^{1/2} + O(\eta\lambda(\theta_*^\parallel) - 2))$ to which the iterates $(\theta_t^\parallel, \theta_t^\perp)$ starting from any $(\theta_0^\parallel, \theta_0^\perp)$ satisfying $0 < |\theta_0^\perp| \leq C_1(\eta\lambda(\theta_*^\parallel) - 2)^{1/2}$ and $d_M(\theta_0^\parallel, \theta_*^\parallel) \leq C_2(\eta\lambda(\theta_*^\parallel) - 2)^{1/2}$ converge with rate $O((1 - C_3(\eta\lambda(\theta_*^\parallel) - 2))^t)$.

6 Limitations, discussion and conclusion

Our work has a number of limitations, all of which point toward avenues for further exploration along the lines we have developed.

Higher order orthogonal dynamics. Prior work has indicated that beyond the 2-periodic behaviour we studied in this work, higher-order periodicity and chaos emerge as the learning rate increases [17]. Our Theorem 4.1 provides new insight into this behaviour as a manifestation of a bifurcating system oscillating orthogonally to the solution manifold. Our analysis only handles the simplest (namely 2-periodic) case of this supercritical behaviour. We expect our framework to be able to admit convergence proofs of higher oscillatory and chaotic behaviour also, using ideas from dynamical systems theory, however we do not expect this to be easy and have not attempted it in this paper.

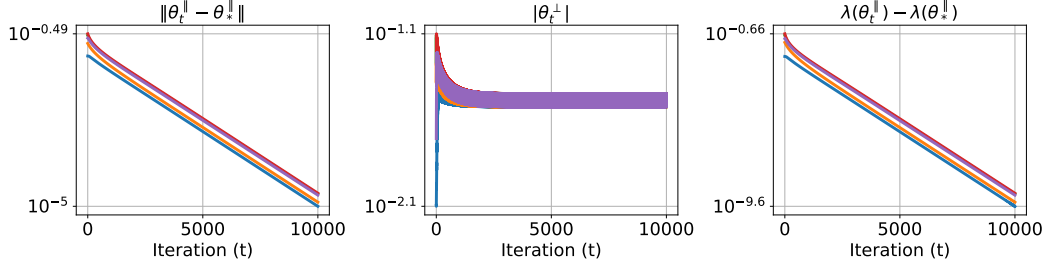


Figure 4: Log-scale plots of distance from $\theta_t^{\parallel}$ to $\theta_*^{\parallel}$ (left), magnitude of $\theta_t^{\perp}$ (centre) and sharpness suboptimality gap (right) on depth 5 scalar factorisation in the **supercritical regime**. Trajectories from five different initialisations shown. The iterates $|\theta_t^{\perp}|$ converge to a stable, period-two orbit driving linear convergence of $\theta_t^{\parallel}$ to the optimally flat global minimum $\theta_*^{\parallel}$.

Higher codimension least squares. Our theory only allows us to treat codimension 1 problems. This is clearly a severe limitation, ruling out for instance overparametrised regression on multiple datapoints or multiple outputs. It is not difficult to extend our definitions to higher codimension, however proving anything in this setting seems very challenging, as it would require an understanding of higher-dimensional bifurcating dynamical systems. We leave this question to future work.

Geodesic convexity of sharpness for more general solution manifolds. One of the most surprising results of our work is that the sharpness of the loss is geodesically strongly convex over a geodesic ball in the solution manifold for a number of overparametrised non-convex problems. Should this prove to be a more general fact, which we suspect may be the case, it would go a long way to explaining the still-mysterious ease of optimisation and implicit bias of GD on DNNs. We leave a more thorough investigation of this question to future work.

Finding the tubular neighbourhood. Our theory is premised on the *assumption* that GD is initialised in a tubular neighbourhood of the solution manifold. While there is no reason to think this is the case with standard initialisation schemes used in practice, we conjecture that GD does converge rapidly to a tubular neighbourhood from a standard initialisation during the phase known as progressive sharpening [13], after which the familiar dynamics described here would apply. We leave exploration of this question to future work.

Relation to stochasticity. Prior works [9, 15, 31] studying *stochastic* gradient descent have demonstrated an implicit bias toward flat minimisers arising from stochasticity. All of these works consider a smaller learning rate than those considered in our work; consequently, the mechanism behind the implicit bias considered in [9, 15, 31] is fundamentally different than that considered in our work. Studying how these distinct implicit biases interact is an important direction for future research.

Acknowledgments and Disclosure of Funding

This research was supported by NSF 2031985 and Simons Foundation 814201 (Theorinet), ONR MURI 503405-78051 and University of Pennsylvania Startup Funds.

References

- [1] Atish Agarwala, Fabian Pedregosa, and Jeffrey Pennington. Second-order regression models exhibit progressive sharpening to the edge of stability. In *ICML*, 2023.
- [2] K. Ahn, J. Zhang, and S. Sra. Understanding the unstable convergence of gradient descent. In *ICML*, 2022.
- [3] Z. Allen-Zhu, Y. Li, and Z. Song. A Convergence Theory for Deep Learning via Over-Parameterization. In *ICML*, pages 242–252, 2019.
- [4] J. M. Altschuler and P. Parrilo. Acceleration by Stepsize Hedging: Multi-Step Descent and the Silver Stepsize Schedule. *Journal of the ACM*, 2023.

- [5] J. M. Altschuler and P. Parrilo. Acceleration by stepsize hedging: Silver Stepsize Schedule for smooth convex optimization. *Mathematical Programming*, pages 1–14, 2024.
- [6] M. Arizzi and J. Raissy. On Écale-hakim’s theorems on holomorphic dynamics. *Frontiers in Complex Dynamics*, 2014.
- [7] S. Arora, Z. Li, and A. Panigrahi. Understanding Gradient Descent on Edge of Stability in Deep Learning. In *ICML*, 2022.
- [8] P. L. Bartlett, P. M. Long, and O. Bousquet. The Dynamics of Sharpness-Aware Minimization: Bouncing Across Ravines and Drifting Towards Wide Minima. *Journal of Machine Learning Research*, 24:1–36, 2023.
- [9] G. Blanc, N. Gupta, G. Valiant, and P. Valiant. Implicit regularization for deep neural networks driven by an Ornstein-Uhlenbeck like process. In *COLT*, 2020.
- [10] S. Bombari, M. H. Amani, and M. Mondelli. Memorization and Optimization in Deep Neural Networks with Minimum Over-parameterization. In *NeurIPS*, 2022.
- [11] Y. Cai, J. Wu, S. Mei, M. Lindsey, and P. L. Bartlett. Large Stepsize Gradient Descent for Non-Homogeneous Two-Layer Networks: Margin Improvement and Fast Optimization. In *NeurIPS*, 2024.
- [12] Lei Chen and Joan Bruna. Beyond the Edge of Stability via Two-step Gradient Updates. In *ICML*, 2023.
- [13] J. Cohen, S. Kaur, Y. Li, J. Zico Kolter, and A. Talwalkar. Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability. In *ICLR*, 2021.
- [14] Jeremy Cohen, Alex Damian, Ameet Talwalkar, J Zico Kolter, and Jason D. Lee. Understanding Optimization in Deep Learning with Central Flows. In *ICLR*, 2025.
- [15] A. Damian, T. Ma, and J. Lee. Label Noise SGD Provably Prefers Flat Global Minimizers. In *NeurIPS*, 2021.
- [16] A. Damian, E. Nichani, and J. Lee. Self-Stabilization: The Implicit Bias of Gradient Descent at the Edge of Stability. In *ICLR*, 2023.
- [17] Dayal Singh Kalra and Tianyu He and Maissam Barkeshli. Universal Sharpness Dynamics in Neural Network Training: Fixed Point Analysis, Edge of Stability, and Route to Chaos. In *ICLR*, 2025.
- [18] S. S. Du, J. Lee, H. Li, L. Wang, and X. Zhai. Gradient Descent Finds Global Minima of Deep Neural Networks. In *ICML*, pages 1675–1685, 2019.
- [19] S. S. Du, X. Zhai, B. Póczos, and A. Singh. Gradient Descent Provably Optimizes Over-parameterized Neural Networks. In *ICLR*, 2019.
- [20] Avrajit Ghosh, Soo Min Kwon, Rongrong Wang, Saiprasad Ravishankar, and Qing Qu. Learning dynamics of deep matrix factorization beyond the edge of stability. In *ICLR*, 2025.
- [21] B. Grimmer. Provably faster gradient descent via long steps. *SIAM Journal on Optimization*, 34:2588–2608, 2024.
- [22] B. Grimmer, K. Shu, and A. L. Wang. Accelerated gradient descent via long steps. arXiv:2309.09961, 2023.
- [23] B. Grimmer, K. Shu, and A. L. Wang. Accelerated objective gap and gradient norm convergence for gradient descent via long steps. *INFORMS Journal on Optimization*, 7:156–169, 2025.
- [24] A. Jacot, F. Gabriel, and C. Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *NeurIPS*, pages 8571–8580, 2018.
- [25] H. Karimi, J. Nutini, and M. Schmidt. Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition. In *ECML PKDD*, pages 795–811, 2016.

- [26] Itai Kreisler, Mor Shpigel Nacson, Daniel Soudry, and Yair Carmon. Gradient descent monotonically decreases the sharpness of gradient flow solutions in scalar networks and beyond. In *ICML*, 2023.
- [27] Y. A. Kuznetsov. *Elements of Applied Bifurcation Theory, 4th Edition*. Springer, 2023.
- [28] J. M. Lee. *Introduction to Smooth Manifolds, Second Edition*. Springer, 2013.
- [29] J. M. Lee. *Introduction to Riemannian Manifolds, Second Edition*. Springer, 2018.
- [30] S. Lee and C. Jang. A new characterization of the edge of stability based on a sharpness measure aware of batch gradient distribution. In *ICLR*, 2023.
- [31] Z. Li, T. Wang, and S. Arora. What Happens after SGD Reaches Zero Loss? –A Mathematical Framework. In *ICLR*, 2022.
- [32] Liming Liu, Zixuan Zhang, Simon Du, and Tuo Zhao. A minimalist example of edge-of-stability and progressive sharpening, 2025.
- [33] Maria G Monera, A Montesinos-Amilibia, and Esther Sanabria-Codesal. The taylor expansion of the exponential map and geometric applications. *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas*, 108:881–906, 2014.
- [34] Q. Nguyen. On the Proof of Global Convergence of Gradient Descent for Deep ReLU Networks with Linear Widths. In *NeurIPS*, 2021.
- [35] Q. Nguyen and M. Mondelli. Global Convergence of Deep Networks with One Wide Layer Followed by Pyramidal Topology. In *NeurIPS*, 2020.
- [36] Q. Nguyen, M. Mondelli, and G. Montufar. Tight Bounds on the Smallest Eigenvalue of the Neural Tangent Kernel for Deep ReLU Networks. In *ICML*, 2021.
- [37] P. Petersen. *Riemannian Geometry*. Springer, 2016.
- [38] Yuqing Wang, Minshuo Chen, Tuo Zhao, and Molei Tao. Large Learning Rate Tames Homogeneity: Convergence and Balancing Effect. In *ICLR*, 2022.
- [39] Yuqing Wang, Zhenghao Xu, Tuo Zhao, and Molei Tao. Good regularity creates large learning rate implicit biases: edge of stability, balancing, and catapult. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023.
- [40] Z. Wang, Z. Li, and J. Li. Analyzing Sharpness along GD Trajectory: Progressive Sharpening and Edge of Stability. In *NeurIPS*, 2022.
- [41] J. Wu, P. L. Bartlett, M. Telgarsky, and B. Yu. Large Stepsize Gradient Descent for Logistic Loss: Non-Monotonicity of the Loss Improves Optimization Efficiency. In *COLT*, 2024.
- [42] J. Wu, V. Braverman, and J. Lee. Implicit Bias of Gradient Descent for Logistic Regression at the Edge of Stability. In *NeurIPS*, 2023.
- [43] L. Wu, C. Ma, and W. E. How SGD Selects the Global Minima in Over-parameterized Learning: A Dynamical Stability Perspective. In *NeurIPS*, 2018.
- [44] X. Zhu, Z. Wang, X. Wang, M. Zhou, and R. Ge. Understanding Edge-of-Stability Training Dynamics with a Minimalist Example. In *ICLR*, 2023.

A Differential geometry background

The purpose of this section is to give a brief overview of essential notions from Riemannian geometry. A good source for the first subsection is [29]. We have not been able to locate a good source for the latter subsection; in particular, to our knowledge no attempt has yet been made to consider the Riemannian Hessian of the square gradient norm of a function defining a hypersurface, however the calculation is elementary.

A.1 Riemannian metrics, connections and geodesics

Recall that a *Riemannian metric* on a manifold M is a smoothly-varying family of positive-definite inner products $\langle \cdot, \cdot \rangle_\theta$ on the tangent spaces $T_\theta M$ of M , with associated norm $\| \cdot \|_\theta$. Any submanifold M of a Riemannian manifold (N, g_N) inherits a Riemannian metric from N ; precisely, given tangent vectors $v, w \in T_\theta M \subset T_\theta N$, one defines

$$g_M(v, w) := g_N(v, w). \quad (22)$$

Since we work with submanifolds of Euclidean space, whose tangent spaces are all themselves subspaces of Euclidean space equipped with the Euclidean inner product, we will often use the notation $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ without further decoration when considering the metric of a submanifold of Euclidean space.

Denote by T^*M the cotangent bundle of M , whose fibre over $\theta \in M$ is the space of linear functionals $T_\theta^*M \rightarrow \mathbb{R}$. Given $(k, l) \in \mathbb{N} \cup \{0\}$ a (k, l) -tensor at the point $\theta \in M$ is an element of $T_\theta M^{\otimes k} \otimes T_\theta^*M^{\otimes l}$, and a (k, l) -tensor field is a smooth section of the bundle $TM^{\otimes k} \otimes T^*M^{\otimes l}$, that is, a smooth assignment of a tensor $T(\theta) \in T_\theta M^{\otimes k} \otimes T_\theta^*M^{\otimes l}$ to each $\theta \in M$. The space of such sections will be denoted $\Gamma(TM^{\otimes k} \otimes T^*M^{\otimes l})$.

Whether or not M carries a Riemannian metric, the derivative of a map $f : M \rightarrow \mathbb{R}$ makes sense as a map $D_M f : TM \rightarrow T\mathbb{R}$ which is fibrewise-linear in the sense that its restriction $D_M f(\theta) : T_\theta M \rightarrow T_{f(\theta)}\mathbb{R}$ to each fibre is a linear map. The derivative $D_M f$ can in this sense be thought of as a $(0, 1)$ -tensor field. Higher derivatives $D_M^k f$ of f are thus derivatives of tensor fields, and can be defined using a Riemannian metric. Specifically, assuming that M has a Riemannian metric, there is a distinguished derivative operator $D_M : \Gamma(TM^{\otimes k} \otimes T^*M^{\otimes l}) \rightarrow \Gamma(TM^{\otimes k} \otimes T^*M^{\otimes l+1})$ defined for all $(k, l) \in \mathbb{N} \cup \{0\}$ called the *Levi-Civita connection*. The additional T^*M -slot gained by applying D_M to a tensor T is to be thought of as a *direction* in which T is differentiated by D_M . In particular, given $T \in \Gamma(TM^{\otimes k} \otimes T^*M^{\otimes l})$ and $X \in \Gamma(TM)$, the notation $D_M T[X] \in \Gamma(TM^{\otimes k} \otimes T^*M^{\otimes l})$ is the *directional derivative of T in the direction X* . Note that when $M = \mathbb{R}^p$, $D := D_{\mathbb{R}^p}$ is the usual multivariate derivative, and the higher derivatives $D^k := D_{\mathbb{R}^p}^k$ coincide with the usual higher-order derivatives in Euclidean space. If $g : M \rightarrow \mathbb{R}$ is a smooth, scalar-valued function, then $\nabla_M^k g \in \Gamma(TM \otimes T^*M^{\otimes k-1})$ will be used to denote $D_M^k g$ with one of the T^*M -slots dualised to a TM -slot as in (3). In particular, $\nabla_M g$ and $\nabla_M^2 g$ are the *Riemannian gradient* and *Riemannian Hessian* respectively; they coincide with the usual gradient ∇g and Hessian $\nabla^2 g$ respectively when $M = \mathbb{R}^p$.

A special case of the Levi-Civita connection is that inherited by a submanifold. If M is a submanifold of a Riemannian manifold N , then the Levi-Civita connection associated to the induced metric (22) is $D_M := P_{TM} D_N$, where P_{TM} is the orthogonal projection $TN \rightarrow TM$ induced by the metric on N .

Given $(\theta, v) \in TM$, classical ordinary differential equation (ODE) theory guarantees the existence of a unique M -valued solution γ to the TM -valued second order ODE

$$(D_M \dot{\gamma})[\dot{\gamma}] = 0, \quad \gamma(0) = \theta, \quad \dot{\gamma}(0) = v \quad (23)$$

on some interval $[0, \epsilon]$. This unique solution is called the *geodesic* through (θ, v) . In particular, for all $v \in T_\theta M$ sufficiently small, $\gamma(1)$ makes sense and is the value of the *exponential map* $\exp_\theta(v)$. The exponential map $\exp_\theta : B_{T_\theta M}(0, R) \rightarrow M$ is invertible onto its image $\exp_\theta(B_{T_\theta M}(0, R)) \subset M$ for all $R > 0$ sufficiently small, and its inverse, denoted $\log_\theta : \exp_\theta(B_{T_\theta M}(0, R)) \rightarrow B_{T_\theta M}(0, R)$, is called the *logarithm*. When $M = \mathbb{R}^p$ with the Euclidean metric, $\exp_\theta(v)$ is defined for all $(\theta, v) \in T\mathbb{R}^p$ and is equal to $\theta + v$. The *length* of a geodesic γ on $[0, b]$ is given by $\int_0^b \|\dot{\gamma}(t)\| dt$, and the *geodesic distance* $d_M(\theta, \theta')$ between two points $\theta, \theta' \in M$ is the length of the shortest

geodesic connecting θ to θ' . Associated to any geodesic $\gamma : [0, b] \rightarrow M$ and any $t \in [0, b]$ is a unique orthogonal map $\Pi(\gamma)_t : T_{\gamma(0)}M \rightarrow T_{\gamma(t)}M$ such that $\Pi(\gamma) : t \mapsto \Pi(\gamma)t$ satisfies the ODE

$$(D_M \Pi(\gamma))[\dot{\gamma}] = 0, \quad \Pi(\gamma)_0 = I_{T_{\gamma(0)}M}. \quad (24)$$

This map $\Pi(\gamma)_t$ is called the *parallel transport* morphism, and extends to a map $T_{\gamma(0)}M^{\otimes k} \otimes T_{\gamma(0)}^*M^{\otimes l} \rightarrow T_{\gamma(t)}M^{\otimes k} \otimes T_{\gamma(t)}^*M^{\otimes l}$ of arbitrary tensors. In particular, if $\theta, \theta' \in M$ admit a unique geodesic $\gamma : [0, b] \rightarrow M$ with $\gamma(0) = \theta$ and $\gamma(b) = \theta'$, $\Pi(\gamma)_b : T_\theta M \rightarrow T_{\theta'}M$ will be denoted simply $\Pi_{\theta \rightarrow \theta'}$.

Tensor fields admit *covariant Taylor expansions* on M defined as follows. Fix $\theta \in M$ and suppose that $\theta' \in M$ is sufficiently close to θ that there is a unique geodesic $\gamma : [0, 1] \rightarrow M$ such that $\gamma(0) = \theta$ and $\gamma(1) = \theta'$. Consider then the function $\tilde{T}(t) := \Pi(\gamma)_t^{-1}T(\gamma(t)) \in T_\theta M^{\otimes k} \otimes T_\theta^*M^{\otimes l}$. Taylor's theorem in a single variable implies that

$$\tilde{T}(1) = \tilde{T}(0) + D\tilde{T}(0) + \frac{1}{2}D^2\tilde{T}(0) + \dots \quad (25)$$

Since $(D_M \Pi(\gamma))[\dot{\gamma}] = 0$ and $(D_M \dot{\gamma})[\dot{\gamma}] = 0$, one has:

$$D^k \tilde{T}(0) = D^k(\Pi(\gamma)^{-1} \circ T \circ \gamma)(t)|_{t=0} \quad (26)$$

$$= (\Pi(\gamma)_t^{-1} D_M^k T(\gamma(t))[\dot{\gamma}(t)^{\otimes k}])|_{t=0} \quad (27)$$

$$= D_M^k T(\theta)[\log_\theta(\theta')^{\otimes k}]. \quad (28)$$

Substituting these formulae into (25) yields the *covariant Taylor expansion*

$$T(\theta') = \Pi_{\theta \rightarrow \theta'} \left(T(\theta) + D_M T(\theta)[\log_\theta(\theta')] + \frac{1}{2} D_M^2 T(\theta)[\log_\theta(\theta')^{\otimes 2}] + O(\|\log_\theta(\theta')\|^3) \right). \quad (29)$$

A.2 Riemannian geometry of submanifolds and hypersurfaces in Euclidean space

When M is a hypersurface Euclidean space $\mathbb{R}^p$, with normal vector field n , the Hessian of a smooth function takes the following form.

Lemma A.1. *Let M be a smooth hypersurface in $\mathbb{R}^p$ with normal vector field n . The Riemannian Hessian $\nabla_M^2 \lambda$ of a smooth map $\lambda : \mathbb{R}^p \rightarrow \mathbb{R}$ along M is the $p \times p$ matrix-valued function*

$$P_{TM}(\nabla^2 \lambda - \langle n, \nabla \lambda \rangle Dn)P_{TM}, \quad (30)$$

on M .

Proof. When M is an embedded submanifold of $\mathbb{R}^p$, the action of the Levi-Civita connection on a function λ (i.e., the Riemannian gradient operator) is given by $P_{TM} \nabla \lambda$, where ∇ is the ordinary Euclidean gradient. The action of the Levi-Civita connection on a vector field $X : \mathbb{R}^p \rightarrow \mathbb{R}^p$ is given by $P_{TM} DX P_{TM}$. Thus:

$$\nabla_M^2 \lambda = P_{TM} D(P_{TM} \nabla \lambda) P_{TM} = P_{TM} (\nabla^2 \lambda - Dn \langle n, \nabla \lambda \rangle - n((Dn)^T \nabla \lambda)^T - nn^T \nabla^2 \lambda) P_{TM} \quad (31)$$

$$= P_{TM} (P_{TM} \nabla^2 \lambda - \langle n, \nabla \lambda \rangle Dn - n((Dn)^T \nabla \lambda)^T) P_{TM} \quad (32)$$

$$= P_{TM} (\nabla^2 \lambda - \langle n, \nabla \lambda \rangle Dn) P_{TM}, \quad (33)$$

where the final line follows from

$$P_{TM}(nv^T) = (I - nn^T)(nv^T) = nv^T - nv^T = 0 \quad (34)$$

for any vector v . $\square$

We will be particularly interested in the case of a hypersurface $M = f^{-1}\{y\} \subset \mathbb{R}^p$, where $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is a smooth submersion, with metric g inherited from the ambient Euclidean space. Specifically, when n is the unit normal vector field defined by

$$n(\theta) := \frac{\nabla f(\theta)}{\|\nabla f(\theta)\|}, \quad \theta \in M, \quad (35)$$

the Riemannian metric g on M is defined by

$$g(\theta) := P_{TM}(\theta) := I - n(\theta)n(\theta)^T, \quad \theta \in M \quad (36)$$

where we use $P_{TM}(\theta)$ to denote the projection onto $T_\theta M$. We will be particularly interested in computing the Riemannian Hessian of the function $\lambda := \|\nabla f\|^2$, which admits the following formula entirely in terms of f and its derivatives.

Lemma A.2. *For a hypersurface $M := f^{-1}\{y\}$ of Euclidean space defined by a smooth function $f : \mathbb{R}^p \rightarrow \mathbb{R}$, the Riemannian Hessian of $\lambda := \|\nabla f\|^2$ is given by*

$$\nabla_M^2 \lambda = 2P_{TM}((\nabla^3 f \nabla f + (\nabla^2 f)^2) - \langle n, \nabla^2 f n \rangle \nabla^2 f) P_{TM} \quad (37)$$

where $\nabla^3 f$ is regarded as a map $\mathbb{R}^p \rightarrow \mathbb{R}^{p \times p}$, or as the map sending a direction vector v to the directional derivative of the Hessian matrix $\nabla^2 f$ in the direction v .

Proof. By Lemma A.1, we have that

$$\nabla_M^2 \lambda = P_{TM}(\nabla^2 \lambda - \langle n, \nabla \lambda \rangle Dn) P_{TM}, \quad (38)$$

so it remains only to compute the various quantities. The following identities are elementary to verify:

$$\nabla \lambda = 2\nabla^2 f \nabla f, \quad (39)$$

$$\nabla^2 \lambda = 2(\nabla^3 f \nabla f + (\nabla^2 f)^2), \quad (40)$$

and

$$Dn = P_{TM} \frac{\nabla^2 f}{\|\nabla f\|}. \quad (41)$$

The result now follows. $\square$

Our next lemma allows us to lower-bound the *convexity radius* $\text{conv}(\theta)$ at $\theta \in M$ which is the largest $r > 0$ such that the geodesic ball $B_M(\theta, r)$ is geodesically convex. This will be used in verifying Assumption 3.4 for multilayer scalar factorisation. If M is a hypersurface in $\mathbb{R}^p$ with normal vector field n , then the *second fundamental form* of M is the map $h : TM \otimes TM \rightarrow \mathbb{R}$ defined by

$$h(\theta)[v, w] := -\langle Dn(\theta)[v], w \rangle \quad (42)$$

for all $\theta \in M$ and $v, w \in T_\theta M$.

Lemma A.3. *Let M be a smooth hypersurface in $\mathbb{R}^p$ with normal vector field n , and fix $\theta_* \in M$. Assume that for all $R > 0$, there exists $C_R > 0$ such that*

$$|h(\theta)[u, v]| \leq C_R \quad (43)$$

for all $\theta \in B_M(\theta_*, 2R)$ and all orthonormal vectors $u, v \in T_\theta M$. Then for all $R > 0$, one has

$$\text{conv}(\theta_*) \geq \frac{1}{2} \min \left\{ \frac{\pi}{C_R}, R \right\}. \quad (44)$$

Proof. The bound follows from a well-known relationship between the convexity radius and the *injectivity radius*, which, at a point $\theta \in M$, is the largest $r > 0$ such that $\exp_\theta : B_{T_\theta M}(0, r) \rightarrow M$ is injective. We thus first bound the injectivity radius. Fix $R > 0$. By the Gauss equation and the hypothesis, at any point $\theta \in B_M(\theta_*, 2R)$, the sectional curvature $K(\theta)$ of M satisfies

$$|K(\theta)[u, v]| = |h(\theta)[u, u]h(\theta)[v, v] - h(\theta)[u, v]^2| \leq C_R^2 \quad (45)$$

for all unit, orthogonal pairs $u, v \in T_\theta M$. Consequently, by Klingenberg's formula [37, Lemma 6.4.7], the injectivity radius $\text{inj}(\theta)$ at θ , namely the largest value of r such that $\exp_\theta : B_{T_\theta M}(0, r) \rightarrow M$ is injective, satisfies

$$\text{inj}(\theta) \geq \min \left\{ \frac{\pi}{C_R}, \frac{1}{2}L(\theta) \right\}, \quad (46)$$

where $L(\theta)$ is the length of the shortest geodesic loop based at θ . If $L(\theta) \geq 2R$, then this simply yields $\text{inj} \geq \min\{\pi/C_R, R\}$. If $L(\theta) < 2R$, then the shortest geodesic loop $\gamma : [0, L(\gamma)] \rightarrow M$ at θ is contained entirely in $B(\theta_*, 2R)$. Fenchel's theorem then combines with the hypothesis to give

$$C_R \geq \int_0^{L(\theta)} |h(\gamma(t))[\dot{\gamma}(t), \dot{\gamma}(t)]| dt \geq \frac{2\pi}{L(\gamma)}, \quad (47)$$

so that $(1/L(\theta)) \geq \pi/C_R$. Consequently, one has the bound

$$\inf_{\theta \in B(\theta_*, R)} \text{inj}(\theta) \geq \left\{ \frac{\pi}{C_R}, R \right\} \quad (48)$$

in general.

Now, setting

$$r_R := \frac{1}{2} \min \left\{ \frac{\pi}{C_R}, R \right\}, \quad (49)$$

one sees that the hypotheses of [37, Theorem 6.4.8] hold, implying that $B(\theta_*, r_R)$ is geodesically convex. $\square$

B Case study: deep scalar factorisation

In this section, we verify that all of our assumptions are satisfied for deep scalar factorisation, where $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is defined by

$$f(\theta_1, \dots, \theta_p) := \theta_1 \cdots \theta_p, \quad \theta := (\theta_1, \dots, \theta_p) \in \mathbb{R}^p. \quad (50)$$

For $\theta \in \mathbb{R}^p$ having all entries nonzero, it will be convenient to denote

$$v(\theta) := (\theta_1^{-1}, \dots, \theta_p^{-1}). \quad (51)$$

We first demonstrate that Assumption 3.1 is satisfied.

Proposition B.1. *Any $y \neq 0$ is a regular value of f . Consequently, Assumption 3.1 is satisfied.*

Proof. Since $y \neq 0$, any point $\theta^\parallel \in f^{-1}\{y\}$ has all coordinates being nonzero. Hence $Df(\theta^\parallel) = y v(\theta^\parallel) \neq 0$ making y a regular value of f . $\square$

The pre-image manifold $f^{-1}\{y\}$ of any $y \neq 0$ has several connected components. Without loss of generality, we will assume from hereon that $y > 0$ and denote by M the component of $f^{-1}\{y\}$ contained in the positive orthant. We next demonstrate that Assumption 3.2 is satisfied.

Proposition B.2. *Set $\theta_*^\parallel := y^{1/p} 1_p$, where 1_p is the vector of 1s. Then the line spanned by the normal vector $n(\theta_*^\parallel)$ is invariant under gradient descent on $\ell = (1/2)(f - y)^2$ for any η , so that Assumption 3.2 is satisfied.*

Proof. At $\theta_*^\parallel := y^{1/p} 1_p$, one has $n(\theta_*^\parallel) = p^{-1/2} 1_p$. For any $\alpha \in \mathbb{R}$, one sees that

$$\nabla \ell(\alpha n(\theta_*^\parallel)) = (f(\alpha p^{-1/2} 1_p) - y) \nabla f(\alpha p^{-1/2} 1_p) \quad (52)$$

$$= (f(\alpha p^{-1/2} 1_p) - y) \alpha^{p-1} p^{-(p-1)/2} 1_p \quad (53)$$

is a multiple of $n(\theta_*^\parallel)$, implying that the span of $n(\theta_*^\parallel)$ is invariant under gradient descent on ℓ . $\square$

We will continue to denote $\theta_*^\parallel := y^{1/p} 1_p$, which is often referred to in the literature as the ‘‘balanced’’ solution. We next come to Assumption 3.3.

Proposition B.3. *Assumption 3.3 holds along M . In particular, at $\theta_*^\parallel$,*

$$c(2/\lambda(\theta_*^\parallel), \theta_*^\parallel) = \left(\frac{32}{p^3} \binom{p}{2}^2 - \frac{8}{p^2} \binom{p}{3} \right) y^{-2/p}, \quad (54)$$

where we use the convention $\binom{p}{3} = 0$ if $p < 3$.

Proof. Given symmetric, multilinear functionals A, B taking k and l inputs respectively, let $A \odot B$ denote their symmetric product:

$$(A \odot B)[v_1, \dots, v_{k+l}] = \frac{1}{(k+l)!} \sum_{\sigma \in \text{Perm}(k+l)} A[v_{\sigma(1)}, \dots, v_{\sigma(k)}] B[v_{\sigma(k+1)}, \dots, v_{\sigma(k+l)}], \quad (55)$$

where $\text{Perm}(k+l)$ is the group of permutations of $k+l$ elements. Since $\ell \equiv (1/2)(f-y)^2$, one has

$$D\ell = (f-y)Df \Rightarrow D\ell|_M = 0, \quad (56)$$

$$D^2\ell = Df \odot Df + (f-y)D^2f \Rightarrow D^2\ell|_M = Df \odot Df \quad (57)$$

$$D^3\ell = 3D^2f \odot Df + (f-y)D^3f \Rightarrow D^3\ell|_M = 3D^2f \odot Df \quad (58)$$

$$D^4\ell = 4D^3f \odot Df + 3D^2f \odot D^2f + (f-y)D^4f \Rightarrow D^4\ell|_M = 4D^3f \odot Df + 3D^2f \odot D^2f. \quad (59)$$

One then sees that

$$c(\eta, \theta^\parallel) = \left(\frac{\eta}{2} D^3\ell[n^{\otimes 3}](\theta^\parallel) \right)^2 - \frac{\eta}{6} D^4\ell[n^{\otimes 4}](\theta^\parallel) \quad (60)$$

$$= \left(\frac{3\eta}{2} D^2f[n^{\otimes 2}]Df[n] \right)^2 (\theta^\parallel) - \frac{\eta}{6} \left(4D^3f[n^{\otimes 3}]Df[n] + 3(D^2f[n^{\otimes 2}])^2 \right) (\theta^\parallel) \quad (61)$$

$$= \left(\frac{9\eta^2}{4} \|Df(\theta^\parallel)\|^2 - \frac{\eta}{2} \right) (D^2f(\theta^\parallel)[n(\theta^\parallel)^{\otimes 2}])^2 - \frac{4\eta}{6} \|Df(\theta^\parallel)\| D^3f(\theta^\parallel)[n(\theta^\parallel)^{\otimes 3}]. \quad (62)$$

By continuity, it suffices to prove the claim at $\eta = 2/\lambda(\theta^\parallel) = 2/\|Df(\theta^\parallel)\|^2$; substituting this into the above yields

$$c(2/\lambda(\theta^\parallel), \theta^\parallel) = \frac{8}{\|Df(\theta^\parallel)\|^2} D^2f(\theta^\parallel)[n(\theta^\parallel)^{\otimes 2}]^2 - \frac{8}{6\|Df(\theta^\parallel)\|} D^3f(\theta^\parallel)[n(\theta^\parallel)^{\otimes 3}]. \quad (63)$$

One computes

$$Df_l(\theta^\parallel) = \frac{y}{\theta_l}, \quad (64)$$

$$D^2f_{l_1 l_2}(\theta^\parallel) = (1 - \delta_{l_1 l_2}) \frac{y}{\theta_{l_1} \theta_{l_2}}, \quad (65)$$

$$D^3f_{l_1 l_2 l_3}(\theta^\parallel) = \begin{cases} (1 - \delta_{l_1 l_2})(1 - \delta_{l_1 l_3})(1 - \delta_{l_2 l_3}) \frac{y}{\theta_{l_1} \theta_{l_2} \theta_{l_3}} & \text{if } L \geq 3 \\ 0 & \text{if } L = 2 \end{cases}. \quad (66)$$

Consequently,

$$D^2f(\theta^\parallel)[n(\theta^\parallel)^{\otimes 2}] = \frac{1}{\|Df(\theta^\parallel)\|^2} D^2f(\theta^\parallel)[Df(\theta^\parallel)^{\otimes 2}] \quad (67)$$

$$= \frac{1}{\|Df(\theta^\parallel)\|^2} \sum_{l_1, l_2} (1 - \delta_{l_1 l_2}) \frac{y}{\theta_{l_1} \theta_{l_2}} \frac{y}{\theta_{l_1}} \frac{y}{\theta_{l_2}} \quad (68)$$

$$= \frac{2}{\|Df(\theta^\parallel)\|^2} \sum_{l_1 < l_2} \frac{y^3}{\theta_{l_1}^2 \theta_{l_2}^2} \neq 0. \quad (69)$$

If $p = 2$, then, since $D^3f \equiv 0$, Assumption 3.3 trivially holds for all $\theta^\parallel \in M$ and all η close to $2/\lambda(\theta^\parallel)$. Let us assume then that $p \geq 3$. One has:

$$D^3f(\theta^\parallel)[n(\theta^\parallel)^{\otimes 3}] = \frac{1}{\|Df(\theta^\parallel)\|^3} D^3f(\theta^\parallel)[Df(\theta^\parallel)^{\otimes 3}] \quad (70)$$

$$= \frac{1}{\|Df(\theta^\parallel)\|^3} \sum_{l_1, l_2, l_3} (1 - \delta_{l_1 l_2})(1 - \delta_{l_1 l_3})(1 - \delta_{l_2 l_3}) \frac{y}{\theta_{l_1} \theta_{l_2} \theta_{l_3}} \frac{y}{\theta_{l_1}} \frac{y}{\theta_{l_2}} \frac{y}{\theta_{l_3}} \quad (71)$$

$$= \frac{6}{\|Df(\theta^\parallel)\|^3} \sum_{l_1 < l_2 < l_3} \frac{y^4}{\theta_{l_1}^2 \theta_{l_2}^2 \theta_{l_3}^2}. \quad (72)$$

Thus one has

$$c(2/\lambda(\theta^\parallel), \theta^\parallel) = \frac{32y^6}{\|Df(\theta^\parallel)\|^6} \left(\sum_{l_1 < l_2} \frac{1}{\theta_{l_1}^2 \theta_{l_2}^2} \right)^2 - \frac{8y^4}{\|Df(\theta^\parallel)\|^4} \sum_{l_1 < l_2 < l_3} \frac{1}{\theta_{l_1}^2 \theta_{l_2}^2 \theta_{l_3}^2}. \quad (73)$$

Substituting

$$\|Df(\theta^\parallel)\|^2 = \sum_{l=1}^p \frac{y^2}{\theta_l^2} \quad (74)$$

then yields

$$c(2/\lambda(\theta^\parallel), \theta^\parallel) = 32 \left(\sum_{l_1 < l_2} \theta_{l_1}^{-2} \theta_{l_2}^{-2} \right)^2 \left(\sum_l \theta_l^{-2} \right)^{-3} - 8 \left(\sum_{l_1 < l_2 < l_3} \theta_{l_1}^{-2} \theta_{l_2}^{-2} \theta_{l_3}^{-2} \right) \left(\sum_l \theta_l^{-2} \right)^{-2} \quad (75)$$

$$= \left(\sum_l \theta_l^{-2} \right)^{-3} \left(32 \left(\sum_{l_1 < l_2} \theta_{l_1}^{-2} \theta_{l_2}^{-2} \right)^2 - 8 \left(\sum_{l_1 < l_2 < l_3} \theta_{l_1}^{-2} \theta_{l_2}^{-2} \theta_{l_3}^{-2} \right) \left(\sum_l \theta_l^{-2} \right) \right) \quad (76)$$

$$= (pS_1(\theta^\parallel))^{-3} \left(32 \binom{p}{2}^2 S_2(\theta^\parallel)^2 - 8p \binom{p}{3} S_1(\theta^\parallel) S_3(\theta^\parallel) \right), \quad (77)$$

where $S_i(\theta^\parallel)$ are the elementary symmetric means in the variables θ_l^{-2} . Newton's inequality for the elementary symmetric means gives $S_2(\theta^\parallel)^2 \geq S_1(\theta^\parallel) S_3(\theta^\parallel)$, so that $c(2/\lambda(\theta^\parallel), \theta^\parallel) > 0$ since $\binom{p}{2}^2 = p^2(p-1)^2/4 > p^2(p-1)(p-2)/6 = p \binom{p}{3}$.

To complete the proof, we evaluate at $\theta_*^\parallel = y^{1/p} 1_p$. One has:

$$\sum_l \theta_l^{-2} = py^{-2/p}, \quad \sum_{l_1 < l_2} \theta_{l_1}^{-2} \theta_{l_2}^{-2} = \binom{p}{2} y^{-4/p}, \quad \sum_{l_1 < l_2 < l_3} \theta_{l_1}^{-2} \theta_{l_2}^{-2} \theta_{l_3}^{-2} = \binom{p}{3} y^{-6/p}, \quad (78)$$

from which it follows that

$$c(2/\lambda(\theta_*^\parallel), \theta_*^\parallel) = \frac{32}{p^3} \binom{p}{2}^2 y^{-2/p} - \frac{8}{p^2} \binom{p}{3} y^{-2/p}, \quad (79)$$

thus yielding the claimed formula. $\square$

We finally come to verifying Assumption 3.4. This requires us both to demonstrate the existence of a geodesically convex ball containing $\theta_*^\parallel$, and to demonstrate that λ is geodesically strongly convex in this ball.

We first give an explicit ball about $\theta_*^\parallel$ which is geodesically convex using Lemma A.3.

Lemma B.4. *The geodesic ball of radius*

$$r := \begin{cases} \infty & \text{if } p = 2 \\ y^{1/p} \left(\frac{4\pi+1-\sqrt{16\pi+1}}{16\pi-8} \right) & \text{if } p \geq 3 \end{cases} \quad (80)$$

centred on $\theta_^\parallel$ is geodesically convex.*

Proof. That r can be taken to be ∞ if $p = 2$ follows from the fact that M in this case is a copy of the real line.

Suppose that $p \geq 3$. Observe that for any $\theta^\parallel = (\theta_1, \dots, \theta_p) \in M$ and any $u, v \in T_{\theta^\parallel} M$, one has

$$h(\theta^\parallel)[u, v] = -\langle Dn(\theta^\parallel)[u], v \rangle \quad (81)$$

$$= -\|\nabla f(\theta^\parallel)\|^{-1} \langle u, \nabla^2 f(\theta^\parallel) \rangle \quad (82)$$

$$= -\left(\sum_i \theta_i^{-2}\right)^{-\frac{1}{2}} \langle u, (n(\theta^\parallel)n(\theta^\parallel)^T - \text{diag}((\theta_i^{-2})_{i=1}^p))v \rangle \quad (83)$$

$$= -\left(\sum_i \theta_i^{-2}\right)^{-\frac{1}{2}} \left(\sum_i u_i v_i \theta_i^{-2}\right), \quad (84)$$

since $n(\theta^\parallel)^T w = 0$ for any vector $w \in T_{\theta^\parallel} M$. Consequently, for any $R \in [0, (1/2)y^{1/p}]$, since $B_M(\theta^\parallel, 2R) \subset B_{\mathbb{R}^p}(\theta^\parallel, 2R)$ by the distance minimising property of geodesics, if $\theta^\parallel \in B_M(\theta^\parallel, 2R)$ then $\theta_i \in [y^{1/p} - 2R, y^{1/p} + 2R]$ for all i , so that

$$|h(\theta^\parallel)[u, v]| \leq (y^{1/p} - 2R)^{-2} (y^{1/p} + 2R) =: C_R. \quad (85)$$

By Lemma A.3, for all $R \in [0, (1/2)y^{1/p}]$ the ball of radius

$$r_R := \frac{1}{2} \min \left\{ \frac{\pi}{C_R}, R \right\} \quad (86)$$

is geodesically convex; it thus remains only to optimise this r_R . Since $R \mapsto R$ is monotonically increasing and $R \mapsto \pi/C_R$ is monotonically decreasing, it suffices to find the first smallest value of R for which $\pi/C_R = R$. This equality gives a quadratic whose solution is

$$R = y^{1/p} \left(\frac{4\pi + 1 - \sqrt{16\pi + 1}}{8\pi - 4} \right), \quad (87)$$

from which the result follows. $\square$

We next use Lemma A.2 to compute the Riemannian Hessian of λ along M .

Lemma B.5. *For $\theta^\parallel = (\theta_1, \dots, \theta_p) \in M$, define $s_1(\theta^\parallel) := \sum_i \theta_i^{-2}$ and $s_2(\theta^\parallel) := \sum_i \theta_i^{-4}$. Then the function λ has Riemannian Hessian*

$$\nabla_M^2 \lambda = \begin{cases} 2y^2 P_{TM} \left(3 \text{diag}(v^{\odot 4}) - \frac{s_2}{s_1} \text{diag}(v^{\odot 2}) \right) P_{TM} & \text{if } p > 2 \\ 2y^2 P_{TM} \left(\text{diag}(v^{\odot 4}) + \left(s_1 - \frac{s_2}{s_1} \right) \text{diag}(v^{\odot 2}) \right) P_{TM} & \text{if } p = 2 \end{cases}. \quad (88)$$

In particular, at $\theta_*^\parallel \in U$,

$$\nabla_M^2 \lambda(\theta_*^\parallel) = 4y^{2-4/p} I_{TM}(\theta_*^\parallel). \quad (89)$$

Moreover, if $p = 2$, then λ is 2-geodesically strongly convex over all of M ; if $p > 2$, then for any $\delta \in [0, 2 - \sqrt{3}]$, λ is $2y^{2-4/p}(1+\delta)^{-4}(1-\delta)^{-2}(3(1-\delta)^2 - (1+\delta)^2)$ -geodesically strongly convex over the closed geodesic-distance ball $B_M(\theta_*^\parallel, \delta y^{1/p})$ of radius $\delta y^{1/p}$ about $\theta_*^\parallel$ in M .

Proof. For notational convenience, set $V_1 := \text{diag}(v^{\odot 2})$ and $V_2 := \text{diag}(v^{\odot 4})$, where $\odot$ denotes the Hadamard (componentwise) product and v is the vector field defined in (51). From the identities (64), (65) and (66), one sees that along M one has

$$\nabla f \equiv y v, \quad \nabla^2 f \equiv y(vv^T - V_1), \quad n \equiv s_1^{-1/2} v. \quad (90)$$

The pointwise-bilinear map $\nabla^3 f$ is zero if $p = 2$, and if $p > 2$ it acts on a vector z to give the matrix with components

$$(\nabla^3 f[z])_{ij} = y(1 - \delta_{ij}) \sum_k (1 - \delta_{jk})(1 - \delta_{ik}) v_i v_j v_k z_k \quad (91)$$

$$= y(1 - \delta_{ij}) \left(\sum_k (v_i v_j v_k z_k - \delta_{jk} v_i v_j v_k z_k - \delta_{ik} v_i v_j v_k z_k + \delta_{jk} \delta_{ik} v_i v_j v_k z_k) \right) \quad (92)$$

$$= y(1 - \delta_{ij}) \left((v^T z) v v^T - v(V_1 z)^T - (V_1 z) v^T + \text{diag}(v^{\odot 3} \odot z) \right)_{ij} \quad (93)$$

$$= y \left((v^T z) v v^T - v(V_1 z)^T - (V_1 z) v^T + \text{diag}(v^{\odot 3} \odot z) + \right. \quad (94)$$

$$\left. - (v^T z) V_1 + \text{diag}(v^{\odot 3} \odot z) + \text{diag}(v^{\odot 3} \odot z) - \text{diag}(v^{\odot 3} \odot z) \right)_{ij} \quad (95)$$

$$= y \left((v^T z) (v v^T - V_1) - v(V_1 z)^T - (V_1 z) v^T + 2 \text{diag}(v^{\odot 3} \odot z) \right)_{ij}, \quad (96)$$

where the fourth line follows from the fact that $(1 - \delta_{ij}) A_{ij} = (A - \text{diag}(A))_{ij}$ for any matrix A and the fact that $\text{diag}(v(V_1 z)^T) = \text{diag}((V_1 z) v^T) = \text{diag}(v^{\odot 3} \odot z)$. Substituting $z = \nabla f = y v$ one sees that

$$\nabla^3 f \nabla f = \begin{cases} y^2 (s_1 (v v^T - V_1) - v v^T V_1 - V_1 v v^T + 2V_2) & \text{if } p > 2 \\ 0 & \text{if } p = 2 \end{cases}. \quad (97)$$

$$(\nabla^2 f)^2 = y^2 (s_1 v v^T - v v^T V_1 - V_1 v v^T + V_2) \quad (98)$$

and

$$\langle n, \nabla^2 f n \rangle \nabla^2 f = \frac{y^2}{s_1} \langle v, v v^T v - V_1 v \rangle (v v^T - V_1) = y^2 (s_1 - s_2/s_1) (v v^T - V_1), \quad (99)$$

so that, by Lemma A.2,

$$\nabla_M^2 \lambda = 2P_{TM} (\nabla^3 f \nabla f + (\nabla^2 f)^2 - \langle n, \nabla^2 f \rangle \nabla^2 f) P_{TM} \quad (100)$$

$$= 2y^2 P_{TM} (3V_2 - (s_2/s_1) V_1 + (s_1 + (s_2/s_1)) v v^T - 2v v^T V_1 - 2V_1 v v^T) P_{TM} \quad (101)$$

if $p > 2$ and

$$\nabla_M^2 \lambda = 2y^2 P_{TM} (V_2 + (s_1 - (s_2/s_1)) V_1 - v v^T V_1 - V_1 v v^T + (s_2/s_1) v v^T) P_{TM} \quad (102)$$

if $p = 2$. Now, let τ be any tangent vector field to M . Since v is normal to M one has $v^T \tau = 0$, so that

$$((s_1 + (s_2/s_1)) v v^T - 2v v^T V_1 - 2V_1 v v^T) \tau = -2v v^T V_1 \tau \quad (103)$$

and

$$(-v v^T V_1 - V_1 v v^T + (s_2/s_1) v v^T) \tau = -v v^T V_1 \tau \quad (104)$$

are both multiples of the normal vector v ; it follows that

$$\nabla_M^2 \lambda = \begin{cases} 2y^2 P_{TM} (3V_2 - (s_2/s_1) V_1) P_{TM} & \text{if } p > 2 \\ 2y^2 P_{TM} (V_2 + (s_1 - (s_2/s_1)) V_1) P_{TM} & \text{if } p = 2 \end{cases} \quad (105)$$

as claimed. In either case, evaluating this matrix at $\theta_*^\parallel = y^{1/p} 1_p$ gives the claimed identity

$$\nabla_M^2 \lambda(\theta_*^\parallel) = 4y^{2-4/p} I_{TM}(\theta_*^\parallel). \quad (106)$$

In particular, if $p = 2$, evaluating $\nabla_M^2 \lambda$ on the unit tangent vector field $\hat{\tau} : (\theta_1, \theta_2) \mapsto (-\theta_1, \theta_2)/\sqrt{\theta_1^2 + \theta_2^2}$ yields

$$(\hat{\tau}^T \nabla_M^2 \lambda \hat{\tau})(\theta^\parallel) = \frac{2y^2}{(\theta_1^2 + \theta_2^2)} \left(\theta_1^{-2} + \theta_2^{-2} + \frac{4\theta_1^{-2}\theta_2^{-2}}{\theta_1^{-2} + \theta_2^{-2}} \right) \quad (107)$$

$$= \frac{2y^2}{\theta_1^2 + \theta_2^2} \left(\frac{\theta_1^2 + \theta_2^2}{\theta_1^2 \theta_2^2} + \frac{4}{\theta_1^2 + \theta_2^2} \right) = 2 + \frac{8y^2}{(\theta_1^2 + \theta_2^2)^2} \geq 2 \quad (108)$$

for any $\theta^\parallel = (\theta_1, \theta_2) \in M$, so that λ is 2-geodesically strongly convex over all of M as claimed.

We now assume that $p > 2$ and come to determining a neighbourhood of $\theta_*^\parallel$ on which $\nabla_M^2 \lambda$ is positive definite. It is clear that a sufficient condition for $\nabla_M^2 \lambda$ to be positive-definite is to have all entries of the diagonal matrix $3V_2 - (s_2/s_1)V_1$ be positive; we will show that this sufficient condition can be guaranteed over a geodesic ball in M centred at $\theta_*^\parallel$. Since the geodesic ball of radius r at a point in M is contained in the Euclidean ball of radius r at the same point by the distance-minimising property of geodesics, it suffices to show that this sufficient condition can be guaranteed over a Euclidean ball centred at $\theta_*^\parallel$.

Fix $\delta \in [0, 1]$: for any $\theta = (\theta_1, \dots, \theta_p)$ in the Euclidean ball $B_{\mathbb{R}^p}(\theta_*^\parallel, \delta y^{1/p})$, one has $\theta_i \in [(1 - \delta)y^{1/p}, (1 + \delta)y^{1/p}]$. Consequently, for any $\theta \in B_{\mathbb{R}^p}(\theta_*^\parallel, \delta y^{1/p})$ one has

$$\min_i \theta_i^{-2} \geq (1 + \delta)^{-2} y^{-2/p}, \quad \frac{s_2}{s_1}(\theta) \leq \max_i \theta_i^{-2} \frac{s_1}{s_1}(\theta) \leq (1 - \delta)^{-2} y^{-2/p}. \quad (109)$$

It follows that for all $\theta \in B_{\mathbb{R}^p}(\theta_*^\parallel, \delta y^{1/p})$ one has

$$\lambda_{\min}(3V_2 - (s_2/s_1)V_1)(\theta) \geq \min_i \theta_i^{-2} (3 \min_i \theta_i^{-2} - (s_2/s_1)(\theta)) \quad (110)$$

$$\geq (1 + \delta)^{-2} y^{-2/p} (3(1 + \delta)^{-2} y^{-2/p} - (1 - \delta)^{-2} y^{-2/p}) \quad (111)$$

$$= (1 + \delta)^{-4} (1 - \delta)^{-2} y^{-4/p} (3(1 - \delta)^2 - (1 + \delta)^2), \quad (112)$$

which is strictly greater than zero provided that $3(1 - \delta)^2 - (1 + \delta)^2 > 0$, which holds provided that $0 \leq \delta < 2 - \sqrt{3}$. This proves the result. $\square$

Note that λ is *not* globally convex over M in general. For instance, when $p = 3$, along the family $\theta^\parallel(\epsilon) := (\epsilon, 1, y\epsilon^{-1}) \in M$, elementary calculations show that $\nabla_M^2 \lambda(\theta^\parallel(\epsilon))$ exhibits negative eigenvalues as $\epsilon \rightarrow 0_+$.

Combining Lemmas B.4 and B.5 one obtains explicit bounds under which Assumption 3.4 holds for multilayer scalar factorisation.

Proposition B.6. *If $p = 2$, then λ is 2-geodesically strongly convex over the geodesically convex set M . If $p \geq 3$, then λ is $1.33y^{2-4/p}$ -geodesically strongly convex over the geodesically convex ball $B_M(\theta_*^\parallel, 0.15y^{1/p})$. In all cases, one has*

$$\nabla_M^2 \lambda(\theta_*^\parallel) = 4y^{2-4/p} I_{TM}(\theta_*^\parallel). \quad (113)$$

Proof. Numerical computation reveals that the radius of Lemma B.4 is lower-bounded by $0.15y^{1/p}$. Substituting $\delta = 0.15$ into the strong convexity bound of Lemma B.5 and rounding down to two decimal places yields the claimed result. $\square$

We conclude the section by proving that the additional hypothesis required for Theorem 5.2 is satisfied.

Proposition B.7. *For any $p \geq 2$, with $\nu = 4y^{2-4/p}$ as in Proposition B.6, one has*

$$\frac{\nu}{c(2/\lambda(\theta_*^\parallel))\lambda(\theta_*^\parallel)} \leq \frac{1}{2} < 1, \quad (114)$$

so that the additional hypothesis of Theorem 5.2 holds.

Proof. Since $\lambda(\theta_*^\parallel) = \|Df(\theta_*^\parallel)\|^2 = py^{2-2/p}$, one calculates using Proposition B.3 that

$$c(2/\lambda(\theta_*^\parallel), \theta_*^\parallel)\lambda(\theta_*^\parallel) = y^{2-4/p} \left(\frac{32}{p^2} \binom{p}{2}^2 - \frac{8}{p} \binom{p}{3} \right) \quad (115)$$

$$= y^{2-4/p} 8(p-1)(p-1-(1/6)(p-2)) \quad (116)$$

$$\geq y^{2-4/p} 8(p-1)(p-1-(p-2)) \quad (117)$$

$$\geq 8y^{2-4/p} \quad (118)$$

for any $p \geq 2$. The result follows. $\square$

C The normal form of gradient descent

To prove Theorem 4.1 rigorously requires a number of steps. **First**, one must give leading order formulae for the expression of gradient descent in tubular neighbourhood coordinates. It is essential here that the errors in these leading order formulae are appropriately sharpened from their naive estimates using the structure of the tubular neighbourhood map (Lemma C.2) and the orthogonal stability assumption (Lemma C.4); without this sharpening our convergence theorems are impossible. These considerations culminate in an intermediate coordinate expression for gradient descent in Proposition C.5. **Second**, one must invoke Assumption 3.3 to transform the $\perp$ -dynamics of Proposition C.5 into the easily-analysed normal form quoted in Theorem 4.1. This is formalised in Theorem C.6 and concludes this section.

Throughout this section, given functions $g, h_1, \dots, h_s : \mathbb{R}^k \rightarrow \mathbb{R}^l$ we will use the notation

$$g = O(h_1, \dots, h_s) \quad (119)$$

to mean that there exist constants $C_1, \dots, C_s > 0$ such that

$$\|g(w)\| \leq \sum_{i=1}^s C_i \|h_i(w)\| \quad (120)$$

for all w sufficiently close to zero.

In addition to this asymptotic notation, we will continue the notation of the main body of the paper, denoting by $M := f^{-1}\{y\}$ the submanifold of $\mathbb{R}^p$ obtained as the preimage of a regular value $y \in \mathbb{R}$ of an analytic map $f : \mathbb{R}^p \rightarrow \mathbb{R}$. We will also denote by $n := \nabla f / \|\nabla f\|$ the unit normal along M . Assumptions 3.1, 3.2 and 3.3 are assumed in all of what follows. Assumption 3.4 is *not* needed for this section.

Consider the map $E : M \times \mathbb{R} \rightarrow \mathbb{R}^p$ defined by

$$E(\theta^\parallel, \theta^\perp) := \theta^\parallel + \theta^\perp n(\theta^\parallel). \quad (121)$$

Observe that

$$DE(\theta^\parallel, \theta^\perp) = (I_{TM}(\theta^\parallel) + \theta^\perp Dn(\theta^\parallel), n(\theta^\parallel)), \quad (122)$$

where $I_{TM}(\theta^\parallel) : T_{\theta^\parallel} M \rightarrow T_{\theta^\parallel} M$ is the identity map. The derivative $DE(\theta^\parallel, \theta^\perp)$ acts as the linear map

$$T_{\theta^\parallel} M \oplus T_{\theta^\perp} \mathbb{R} \ni (v^\parallel, v^\perp) \mapsto v^\parallel + \theta^\perp Dn(\theta^\parallel)[v^\parallel] + v^\perp n(\theta^\parallel) \in T_{E(\theta^\parallel, \theta^\perp)} \mathbb{R}^p \quad (123)$$

between tangent spaces, and can equivalently be written in terms of the *shape operator* $S(\theta^\parallel) : T_{\theta^\parallel} M \ni v \mapsto -Dn(\theta^\parallel)[v] \in T_{\theta^\parallel} M$ as

$$DE(\theta^\parallel, \theta^\perp) = (I_{TM}(\theta^\parallel) - \theta^\perp S(\theta^\parallel), n(\theta^\parallel)). \quad (124)$$

Observe that since $n(\theta^\parallel)$ is orthogonal to $T_{\theta^\parallel} M$, $DE(\theta^\parallel, \theta^\perp)$ is invertible whenever $|\theta^\perp| < \|S(\theta^\parallel)\|^{-1}$, where $\|S(\theta^\parallel)\|$ denotes the operator norm of $S(\theta^\parallel) : T_{\theta^\parallel} M \rightarrow T_{\theta^\parallel} M$ with respect to the Riemannian metric. Applying the inverse function theorem proves the following.

Proposition C.1. *The map $E : M \times \mathbb{R} \rightarrow \mathbb{R}^p$ is invertible on an open neighbourhood U of $M \times \{0\}$. The image of U in $\mathbb{R}^p$ is a tubular neighbourhood N of M , and we denote the inverse of E thereon by E^{-1} .*

The coordinates $E : (\theta^\parallel, \theta^\perp) \mapsto E(\theta^\parallel, \theta^\perp)$ are called *tubular neighbourhood coordinates*. The next step is to derive a formula for E^{-1} and thereafter derive a formula for the conjugate of gradient descent in the coordinates $(\theta^\parallel, \theta^\perp)$ induced by E . The following technical lemma will be necessary to get the correct asymptotics for the error terms in our formula for E^{-1} .

Lemma C.2. *Let $U \subset \mathbb{R}^{m_1} \times \mathbb{R}^{m_2}$ be an open set and let $g : U \ni (x, y) \mapsto g(x, y) \in \mathbb{R}^p$ be a C^3 function. Assume that $D_y^k g \equiv 0$ on U for $k = 2, 3$ and that g is invertible on U . Then for any $(x, y) \in U$ and all $v \in \mathbb{R}^p$ sufficiently small, setting $(v_x, v_y) = Dg(x, y)^{-1}[v]$ and $z := g(x, y)$, the point $z + v$ is contained in the domain of g^{-1} and one has*

$$g^{-1}(z + v) = g^{-1}(z) + Dg^{-1}(z)[v] + \frac{1}{2} D^2 g^{-1}(z)[v, v] + O(\|v_x\|^3, \|v_x\|^2 \|v_y\|, \|v_x\| \|v_y\|^2). \quad (125)$$

Proof. Since g is invertible on U and U is open, $g(U) = \text{domain}(g^{-1})$ is also open. Fix $(x, y) \in U$ and set $z := g(x, y)$. Then since $g(U)$ is open, $z + v \in g(U)$ for all sufficiently small v so that $g^{-1}(z + v)$ makes sense. Moreover, for all such v , there exists $u \in \mathbb{R}^{m_1} \times \mathbb{R}^{m_2}$ such that $g((x, y) + u) = z + v$ so that $g^{-1}(z + v) = (x, y) + u = g^{-1}(z) + u$. Hence one has the following formula for the remainder $R(v)$ of the Taylor expansion of $g^{-1}(z + v)$ about z :

$$R(v) := g^{-1}(z + v) - g^{-1}(z) - Dg^{-1}(z)[v] - \frac{1}{2}D^2g^{-1}(z)[v, v] \quad (126)$$

$$= u - Dg^{-1}(z)[v] - \frac{1}{2}D^2g^{-1}(z)[v, v]. \quad (127)$$

To bound $R(v)$, therefore, we seek a formula $u(v)$ for u in terms of v .

From $g((x, y) + u) = z + v$, one obtains

$$v = G(u) := Dg(x, y)[u] + \frac{1}{2}D^2g(x, y)[u, u] + S(u), \quad (128)$$

where the remainder $\tilde{u} \mapsto S(\tilde{u})$ satisfies $S(\tilde{u}) = O(\|\tilde{u}_x\|^3, \|\tilde{u}_x\|^2\|\tilde{u}_y\|)$ since $D_y^k g \equiv 0$ for $k = 2, 3$ by hypothesis. Since $Dg(x, y)$ is invertible by the invertibility of g on U , G is smoothly invertible on a neighbourhood V of 0 by the inverse function theorem. Consider now the second-order approximation

$$u' := Dg(x, y)^{-1} \left(v - \frac{1}{2}D^2g(x, y)[Dg(x, y)^{-1}[v]^{\otimes 2}] \right) \quad (129)$$

to u . Since $u' = O(\|v\|)$, taking v smaller if necessary we are guaranteed that $u' \in V$. Thus there is $C > 0$ such that

$$\|u' - u\| \leq C\|G(u') - G(u)\| = C\|G(u') - v\|. \quad (130)$$

We will show that the right side of (130) admits a sufficiently tight bound $b(v)$ that $u(v) = u'(v) + O(b(v))$ suffices to give the desired decay in $R(v)$.

For notational convenience, set $w := (1/2)Dg(x, y)^{-1}D^2g(x, y)[Dg(x, y)^{-1}[v]^{\otimes 2}]$. Then $u' = Dg(x, y)^{-1}[v] - w$, and substituting this expression into (128) yields

$$G(u') - v = -D^2g(x, y)[Dg(x, y)^{-1}[v], w] + \frac{1}{2}D^2g(x, y)[w^{\otimes 2}] + S\left(Dg(x, y)^{-1}[v] - \frac{1}{2}w\right). \quad (131)$$

Since $D_y^2g \equiv 0$ by hypothesis, one has $w = O(\|v_x\|^2, \|v_x\|\|v_y\|)$ as well as

$$D^2g(x, y)[Dg(x, y)^{-1}[v], w] = O(\|v_x\|^3, \|v_x\|^2\|v_y\|, \|v_x\|\|v_y\|^2), \quad (132)$$

$$D^2g(x, y)[w^{\otimes 2}] = O(\|w\|^2) = O(\|v_x\|^4, \|v_x\|^3\|v_y\|, \|v_x\|^2\|v_y\|^2) \quad (133)$$

and

$$S\left(Dg(x, y)^{-1}[v] - \frac{1}{2}w\right) = O(\|v_x\|^3, \|v_x\|^2\|v_y\|). \quad (134)$$

Thus, by (130),

$$u = u' + O(\|v_x\|^3, \|v_x\|^2\|v_y\|, \|v_x\|\|v_y\|^2) \quad (135)$$

$$= Df(x, y)^{-1}[v] - \frac{1}{2}D^2f(x, y)^{-1}D^2f(x, y)[Df(x, y)^{-1}[v]^{\otimes 2}] + \quad (136)$$

$$+ O(\|v_x\|^3, \|v_x\|^2\|v_y\|, \|v_x\|\|v_y\|^2). \quad (137)$$

Finally, applying the chain rule to the equation $g^{-1} \circ g = \text{Identity}$ shows that $Dg^{-1}(z) = Dg(x, y)^{-1}$ and $D^2g^{-1}(z) = -Dg(x, y)^{-1}D^2g(x, y)[Dg(x, y)^{-1}, Dg(x, y)^{-1}]$. It thus follows from substituting (137) into (127) that

$$R(v) = u' - Df^{-1}(z)[v] - \frac{1}{2}D^2f^{-1}(z)[v, v] + O(\|v_x\|^3, \|v_x\|^2\|v_y\|, \|v_x\|\|v_y\|^2) \quad (138)$$

$$= O(\|v_x\|^3, \|v_x\|^2\|v_y\|, \|v_x\|\|v_y\|^2) \quad (139)$$

as claimed. $\square$

A formula can now be given for E^{-1} using a Taylor expansion as follows.

Proposition C.3. *For all $\theta^\parallel \in M$ and all $v = v^\parallel + v^\perp n(\theta^\parallel) \in T_{\theta^\parallel} M \oplus \text{span}(n(\theta^\parallel)) = T_{\theta^\parallel} \mathbb{R}^p$ sufficiently small, one has*

$$E^{-1}(\theta^\parallel + v) = \begin{pmatrix} \theta^\parallel + v^\parallel + Dn(\theta^\parallel)[v^\parallel]v^\perp \\ v^\perp \end{pmatrix} + O(\|v^\parallel\|^3, \|v^\parallel\|^2\|v^\perp\|, \|v^\parallel\|\|v^\perp\|^2). \quad (140)$$

Proof. The result follows from a second-order Taylor expansion with remainder:

$$E^{-1}(\theta^\parallel + v) = E^{-1}(\theta^\parallel) + DE^{-1}(\theta^\parallel)[v] + \frac{1}{2}D^2E^{-1}(\theta^\parallel)[v, v] + \text{error}(\theta^\parallel, v). \quad (141)$$

Since the $\theta^\perp$ -derivatives of $E(\theta^\parallel, \theta^\perp) = \theta^\parallel + \theta^\perp n(\theta^\parallel)$ vanish for all orders > 1 , by Lemma C.2 $\text{error}(\theta^\parallel, v)$, satisfies

$$\text{error}(\theta^\parallel, v) = O(\|v^\parallel\|^3, \|v^\parallel\|^2\|v^\perp\|, \|v^\parallel\|\|v^\perp\|^2). \quad (142)$$

It thus remains only to calculate the leading order terms. One easily computes

$$DE(\theta^\parallel, \theta^\perp) \begin{bmatrix} v^\parallel \\ v^\perp \end{bmatrix} = v^\parallel + \theta^\perp Dn(\theta^\parallel)[v^\parallel] + v^\perp n(\theta^\parallel) \quad (143)$$

$$D^2E(\theta^\parallel, \theta^\perp) \begin{bmatrix} v^\parallel \\ v^\perp \end{bmatrix}^{\otimes 2} = D \left(DF(\theta^\parallel, \theta^\perp) \begin{bmatrix} v^\parallel \\ v^\perp \end{bmatrix} \right) \begin{bmatrix} v^\parallel \\ v^\perp \end{bmatrix} \quad (144)$$

$$= D_{\theta^\parallel} (v^\parallel + \theta^\perp Dn(\theta^\parallel)[v^\parallel] + v^\perp n(\theta^\parallel)) [v^\parallel] + \quad (145)$$

$$+ D_{\theta^\perp} (v^\parallel + \theta^\perp Dn(\theta^\parallel)[v^\parallel] + v^\perp n(\theta^\parallel)) [v^\perp] \quad (146)$$

$$= \theta^\perp D^2n(\theta^\parallel)[v^\parallel, v^\parallel] + v^\perp Dn(\theta^\parallel)[v^\parallel] + v^\perp Dn(\theta^\parallel)[v^\parallel] \quad (147)$$

$$= \theta^\perp D^2n(\theta^\parallel)[v^\parallel, v^\parallel] + 2v^\perp Dn(\theta^\parallel)[v^\parallel], \quad (148)$$

for any $(\theta^\parallel, \theta^\perp) \in M \times \mathbb{R}$ and $(v^\parallel, v^\perp) \in T_{\theta^\parallel} M \times \mathbb{R}$. Recalling that E is invertible when restricted to an open neighbourhood U of $M \times \{0\}$ in $M \times \mathbb{R}$, by differentiating the identity $E^{-1} \circ E = \text{Identity}_{U \subset M \times \mathbb{R}}$ one computes

$$DE^{-1}(\theta^\parallel) = DE^{-1} \circ E(\theta^\parallel, 0) = (DE(\theta^\parallel, 0))^{-1} = \begin{pmatrix} P_{TM}(\theta^\parallel) \\ n(\theta^\parallel)^T \end{pmatrix}, \quad (149)$$

where $P_{TM}(\theta^\parallel) : T_{\theta^\parallel} \mathbb{R}^p \rightarrow T_{\theta^\parallel} M$ is the orthogonal projection, so that

$$DE^{-1}(\theta^\parallel)[v] = \begin{pmatrix} v^\parallel \\ v^\perp \end{pmatrix}. \quad (150)$$

Again differentiating the identity $E^{-1} \circ E = \text{Identity}_U$ yields

$$D^2E^{-1}(\theta^\parallel) = D^2E^{-1} \circ E(\theta^\parallel, 0) = -DE(\theta^\parallel, 0)^{-1} D^2E(\theta^\parallel, 0)[DE(\theta^\parallel, 0)^{-1}, DE(\theta^\parallel, 0)^{-1}], \quad (151)$$

so that from (148) one obtains

$$D^2E^{-1}(\theta^\parallel)[v, v] = \begin{pmatrix} P_{TM}(\theta^\parallel) \\ n(\theta^\parallel)^T \end{pmatrix} (2v^\perp Dn(\theta^\parallel)[v^\parallel]) = \begin{pmatrix} 2v^\perp Dn(\theta^\parallel)[v^\parallel] \\ 0 \end{pmatrix}, \quad (152)$$

with the final equality being a consequence of $n^T Dn = -(Dn)^T n = -n^T Dn = 0$ as follows from differentiating the identity $n^T n = 1$. The result follows. $\square$

The next result characterises the scaling of the $\parallel$ -component of $\nabla \ell$ in a neighbourhood of $\theta_\star^\parallel$. It is essential for obtaining the correct asymptotic decay in the error terms.

Lemma C.4. *Assumption 3.2 implies that for all $\theta^\parallel \in M$ sufficiently close to $\theta_\star^\parallel$ and all $\theta^\perp \in \mathbb{R}$ sufficiently small, one has $P_{TM}(\theta^\parallel) \nabla \ell(E(\theta^\parallel, \theta^\perp)) = O(d_M(\theta^\parallel, \theta_\star^\parallel))$, where $P_{TM}(\theta^\parallel) : T_{\theta^\parallel} \mathbb{R}^p \rightarrow T_{\theta^\parallel} M$ is the orthogonal projection onto $T_{\theta^\parallel} M$.*

Proof. By Assumption 3.2, one has

$$P_{TM}(\theta_*^\parallel) \nabla \ell(E(\theta_*^\parallel, \theta^\perp)) = 0 \quad (153)$$

for all $\theta^\perp \in \mathbb{R}$ sufficiently small. Since for all sufficiently small $\theta^\perp$ the map $\theta^\parallel \mapsto P_{TM}(\theta^\parallel) \nabla \ell(E(\theta^\parallel, \theta^\perp))$ is smooth, the result follows. $\square$

We are now able to give the tubular neighbourhood coordinate expression for gradient descent.

Proposition C.5. *In the tubular neighbourhood coordinates $E : (\theta^\parallel, \theta^\perp) \mapsto E(\theta^\parallel, \theta^\perp)$ for a tubular neighbourhood of M , the gradient descent map $\text{GD}_{\mathbb{R}^p}^\ell(\eta, \cdot) : \theta \mapsto \theta - \eta \nabla \ell(\theta)$ is conjugated to the analytic map*

$$\begin{pmatrix} \theta^\parallel \\ \theta^\perp \end{pmatrix} \mapsto \begin{pmatrix} \theta^\parallel - \frac{\eta(\theta^\perp)^2}{2} \nabla_M \lambda(\theta^\parallel) + O((\theta^\perp)^3 \min\{1, d_M(\theta^\parallel, \theta_*^\parallel)\}) \\ (1 - \eta \lambda(\theta^\parallel)) \theta^\perp - \frac{\eta}{2} D^3 \ell[n^{\otimes 3}](\theta^\parallel)(\theta^\perp)^2 - \frac{\eta}{6} D^4 \ell[n^{\otimes 4}](\theta^\parallel)(\theta^\perp)^3 + O((\theta^\perp)^4) \end{pmatrix} \quad (154)$$

Proof. Since f is analytic by Assumption 3.1, so too are $\text{GD}_{\mathbb{R}^p}^\ell(\eta, \cdot)$ and E and hence, since the inverse function theorem preserves analyticity, so too is E^{-1} . Hence the conjugate $E^{-1} \circ \text{GD}_{\mathbb{R}^p}^\ell(\eta, \cdot) \circ E$ is analytic wherever defined. Recalling the open neighbourhood U of $M \times \{0\}$ in $M \times \mathbb{R}$ from Proposition C.1 on which E is a diffeomorphism, given $(\theta^\parallel, \theta^\perp) \in U$

$$\text{GD}_{\mathbb{R}^p}^\ell(\eta, E(\theta^\parallel, \theta^\perp)) = E(\theta^\parallel, \theta^\perp) - \eta \nabla \ell(E(\theta^\parallel, \theta^\perp)) \quad (155)$$

$$= \theta^\parallel + \theta^\perp n(\theta^\parallel) - \eta \nabla \ell(\theta^\parallel + \theta^\perp n(\theta^\parallel)). \quad (156)$$

Applying Proposition C.3 with $v = \theta^\perp n(\theta^\parallel) - \eta \nabla \ell(\theta^\parallel + \theta^\perp n(\theta^\parallel))$, one then obtains

$$E^{-1}(\text{GD}_{\mathbb{R}^p}^\ell(\eta, E(\theta^\parallel, \theta^\perp))) = \begin{pmatrix} \theta^\parallel + v^\parallel + O(\|v^\parallel\|^3, \|v^\parallel\| \|v^\perp\|) \\ v^\perp + O(\|v^\parallel\|^3, \|v^\parallel\|^2 \|v^\perp\|, \|v^\parallel\| \|v^\perp\|^2) \end{pmatrix}. \quad (157)$$

Note the crucial role played by the estimate on the error in Proposition C.3 here: were there to be a pure $O(\|v^\perp\|^k)$ term in the error for some $k \geq 1$, then the same error term would have appeared in the formula (157) and, as we shall now see, would have prevented the $\parallel$ -component of the dynamics from having an $O(d_M(\theta^\parallel, \theta_*^\parallel))$ error term.

We now compute formulae for $v^\parallel$ and $v^\perp$. Using the fact that $\nabla \ell(\theta^\parallel) = 0$, one has the Taylor expansion

$$\nabla \ell(E(\theta^\parallel, \theta^\perp)) = \theta^\perp \nabla^2 \ell[n](\theta^\parallel) + \frac{(\theta^\perp)^2}{2} \nabla^3 \ell[n^{\otimes 2}](\theta^\parallel) + \frac{(\theta^\perp)^3}{6} \nabla^4 \ell[n^{\otimes 3}](\theta^\parallel) + O((\theta^\perp)^4) \quad (158)$$

$$= \theta^\perp \lambda n(\theta^\parallel) + \frac{(\theta^\perp)^2}{2} \nabla^3 \ell[n^{\otimes 2}](\theta^\parallel) + \frac{(\theta^\perp)^3}{6} \nabla^4 \ell[n^{\otimes 3}](\theta^\parallel) + O((\theta^\perp)^4), \quad (159)$$

where the second line follows from the fact that $n(\theta^\parallel)$ is an eigenvector of $\nabla^2 \ell(\theta^\parallel)$ with eigenvalue $\lambda(\theta^\parallel)$. Taking the inner-product of (159) with $n(\theta^\parallel)$ then reveals that

$$v^\perp = \langle n(\theta^\parallel), v \rangle \quad (160)$$

$$= (1 - \eta \lambda(\theta^\parallel)) \theta^\perp - \frac{\eta}{2} D^3 \ell[n^{\otimes 3}](\theta^\parallel)(\theta^\perp)^2 - \frac{\eta}{6} D^4 \ell[n^{\otimes 4}](\theta^\parallel)(\theta^\perp)^3 + O((\theta^\perp)^4). \quad (161)$$

One obtains $v^\parallel$ by projecting (159) onto $T_{\theta^\parallel} M$:

$$v^\parallel = P_{TM}(\theta^\parallel) v = -\eta P_{TM}(\theta^\parallel) \nabla \ell(E(\theta^\parallel, \theta^\perp)) \quad (162)$$

$$= -\frac{\eta(\theta^\perp)^2}{2} P_{TM} \nabla^3 \ell[n^{\otimes 2}](\theta^\parallel) + O((\theta^\perp)^3 \min\{1, d_M(\theta^\parallel, \theta_*^\parallel)\}) \quad (163)$$

$$= -\frac{\eta(\theta^\perp)^2}{2} \nabla_M (\nabla^2 \ell[n^{\otimes 2}])(\theta^\parallel) + O((\theta^\perp)^3 \min\{1, d_M(\theta^\parallel, \theta_*^\parallel)\}) \quad (164)$$

$$= -\frac{\eta(\theta^\perp)^2}{2} \nabla_M \lambda(\theta^\parallel) + O((\theta^\perp)^3 \min\{1, d_M(\theta^\parallel, \theta_*^\parallel)\}), \quad (165)$$

where the second line follows from Lemma C.4, the third line follows from the definition $\nabla_M = P_{TM} \nabla$ of the Levi-Civita connection on $M \subset \mathbb{R}^p$, and the final line follows from the identity $\nabla^2 \ell[n, n] \equiv \lambda$ that holds along M . Substituting these expressions into (157) yields the result. $\square$

By finally invoking Assumption 3.3, we may now transform the tubular neighbourhood coordinate expression of gradient descent from Proposition C.5 into the normal form expression of Theorem 4.1.

Theorem C.6. *Recall the function $c : \mathbb{R} \times M \rightarrow \mathbb{R}_{>0}$*

$$c(\eta, \theta^\parallel) := \left(\frac{\eta}{2} D^3 \ell[n^{\otimes 3}](\theta^\parallel) \right)^2 - \frac{\eta}{6} D^4 \ell[n^{\otimes 4}](\theta^\parallel) > 0 \quad (166)$$

from Assumption 3.3. There is an open neighbourhood W of $\{(2/\lambda(\theta^\parallel), \theta^\parallel, 0) : \theta^\parallel \in M\}$ in $\mathbb{R} \times M \times \mathbb{R}$ on which $\varphi : W \rightarrow \mathbb{R} \times M \times \mathbb{R}$ defined by

$$\varphi(\eta, \theta^\parallel, \theta^\perp) = (\varphi^\eta(\eta, \theta^\parallel, \theta^\perp), \varphi^\parallel(\eta, \theta^\parallel, \theta^\perp), \varphi^\perp(\eta, \theta^\parallel, \theta^\perp)) \quad (167)$$

$$:= \left(\eta, \theta^\parallel, \frac{\theta^\perp}{c(\eta, \theta^\parallel)^{1/2}} - \frac{\frac{\eta}{2} D^3 \ell[n^{\otimes 3}](\theta^\parallel)}{(1 - \eta \lambda(\theta^\parallel))^2 - (1 - \eta \lambda(\theta^\parallel))} \frac{(\theta^\perp)^2}{c(\eta, \theta^\parallel)} \right) \quad (168)$$

is an analytic diffeomorphism onto its image. Moreover, the composite $(I_{\mathbb{R}} \times E) \circ \varphi : W \rightarrow \mathbb{R}^p \times \mathbb{R}$, where $I_{\mathbb{R}}$ is the identity map, conjugates $(I_{\mathbb{R}}, \text{GD}_{\mathbb{R}^p}^\ell) : \mathbb{R} \times \mathbb{R}^p \rightarrow \mathbb{R} \times \mathbb{R}^p$ to the map

$$\begin{pmatrix} \eta \\ \theta^\parallel \\ \theta^\perp \end{pmatrix} \mapsto \begin{pmatrix} \eta \\ \text{GD}_M^\lambda \left(\frac{\eta(\theta^\perp)^2}{2c(\eta, \theta^\parallel)}, \theta^\parallel \right) + O((\theta^\perp)^3 \min\{1, d_M(\theta^\parallel, \theta_*^\parallel)\}) \\ (1 - \eta \lambda(\theta^\parallel)) \theta^\perp + (\theta^\perp)^3 + O((\theta^\perp)^4) \end{pmatrix}. \quad (169)$$

Proof. That an open neighbourhood W of $\{(2/\lambda(\theta^\parallel), \theta^\parallel, 0) : \theta^\parallel \in M\}$ exists on which φ is a diffeomorphism follows from the fact that $c(\eta, \theta^\parallel)^{-1/2} \neq 0$ for all $\theta^\parallel \in M$ and η in a neighbourhood of $2/\lambda(\theta^\parallel)$ and the inverse function theorem. That φ is analytic is a consequence of the analyticity of f and the functions used in defining φ in terms of $\ell = (1/2)(f - y)^2$.

The formula (169) pertains to the conjugation

$$\varphi^{-1} \circ (I_{\mathbb{R}} \times E^{-1}) \circ \text{GD}_{\mathbb{R}^p}^\ell \circ (I_{\mathbb{R}} \times E) \circ \varphi \quad (170)$$

of $\text{GD}_{\mathbb{R}^p}^\ell$. Proposition C.5 already gives us a formula for $(I_{\mathbb{R}} \times E^{-1}) \circ \text{GD}_{\mathbb{R}^p}^\ell \circ (I_{\mathbb{R}} \times E)$, so it suffices merely to conjugate this formula by φ .

Since $\varphi^\parallel(\eta, \theta^\parallel, \theta^\perp) = \theta^\parallel$, the $\parallel$ -component of (170) is given by substituting $\varphi^\perp(\eta, \theta^\parallel, \theta^\perp)$ in place of $\theta^\perp$ in the $\parallel$ -component of (154), which yields the map

$$\begin{pmatrix} \eta \\ \theta^\parallel \\ \theta^\perp \end{pmatrix} \mapsto \theta^\parallel - \frac{\eta(\theta^\perp)^2}{2c(\eta, \theta^\parallel)} \nabla_M \lambda(\theta^\parallel) + O((\theta^\perp)^3 \min\{1, d_M(\theta^\parallel, \theta_*^\parallel)\}). \quad (171)$$

As demonstrated in [33], however, one has

$$\exp_{\theta^\parallel} \left(- \frac{\eta(\theta^\perp)^2}{2c(\eta, \theta^\parallel)} \nabla_M \lambda(\theta^\parallel) \right) = \theta^\parallel - \frac{\eta(\theta^\perp)^2}{2c(\eta, \theta^\parallel)} \nabla_M \lambda(\theta^\parallel) + O((\theta^\perp)^4 \|\nabla_M \lambda(\theta^\parallel)\|^2) \quad (172)$$

$$= \theta^\parallel - \frac{\eta(\theta^\perp)^2}{2c(\eta, \theta^\parallel)} \nabla_M \lambda(\theta^\parallel) + O((\theta^\perp)^4 \min\{1, d_M(\theta^\parallel, \theta_*^\parallel)^2\}) \quad (173)$$

since $\nabla_M \lambda(\theta_*^\parallel) = 0$. Thus the $\parallel$ -component of (170) may be written as

$$\begin{pmatrix} \eta \\ \theta^\parallel \\ \theta^\perp \end{pmatrix} \mapsto \text{GD}_M^\lambda \left(\frac{\eta(\theta^\perp)^2}{2c(\eta, \theta^\parallel)}, \theta^\parallel \right) + O((\theta^\perp)^3 \min\{1, d_M(\theta^\parallel, \theta_*^\parallel)\}) \quad (174)$$

as claimed.

Finally, the $\perp$ -component of (169) follows from the same argument used in [27, Theorem 4.3]. $\square$

D Convergence theorems

The purpose of this section is to provide proofs of the main convergence theorems presented in the paper. In addition to assuming Assumptions 3.1, 3.2 and 3.3 as in the previous section, in this section we will also assume Assumption 3.4, which states that there is a geodesically convex ball $B_M(\theta_*^\parallel, r)$ centred on $\theta_*^\parallel$ over which λ is geodesically smooth, geodesically strongly convex, and has unique global minimum at $\theta_*^\parallel$ with Riemannian Hessian at $\theta_*^\parallel$ being a multiple of the identity.

D.1 Subcritical regime

Our convergence theorem in the subcritical regime has two components. First, it is proved that gradient descent implicitly descends $\lambda(\theta_t^\parallel)$ from being initially $> 2/\eta$ to eventually being $< 2/\eta$. For this, it is necessary to have a descent lemma for the perturbed Riemannian gradient descent component of Theorem 4.1 (Lemma D.1) as well as upper and lower bounds on $|\theta_t^\perp|$ during this phase (Lemma D.3). Following this transient phase, convergence is easily proved using the fact that $|\theta_t^\perp|$ contracts exponentially to zero.

Lemma D.1 (Descent lemma for perturbed Riemannian gradient descent). *For the μ -geodesically strongly convex, L -geodesically smooth function λ on the geodesically convex ball $B_M(\theta_*^\parallel, r)$ centred on $\theta_*^\parallel$ in M , consider the map*

$$\text{GD}^\parallel(\eta, \theta^\parallel, \theta^\perp) = \exp_{\theta^\parallel} \left(\frac{-\eta(\theta^\perp)^2}{2c(\eta, \theta^\parallel)} \nabla_M \lambda(\theta^\parallel) \right) + O((\theta^\perp)^3 \min\{1, d_M(\theta^\parallel, \theta_*^\parallel)\}) \quad (175)$$

of Theorem C.6. Then for any $\eta > 0$ and all $\theta^\perp$ sufficiently close to zero, $\text{GD}^\parallel(\eta, \cdot, \theta^\perp)$ maps $B_M(\theta_*^\parallel, r)$ into $B_M(\theta_*^\parallel, r)$. Moreover, for any $\eta > 0$, any $\theta^\parallel \in B_M(\theta_*^\parallel, r)$ and all $\theta^\perp$ sufficiently close to zero one has

$$\lambda(\text{GD}^\parallel(\eta, \theta^\parallel, \theta^\perp)) - \lambda(\theta_*^\parallel) \leq \left(1 - \frac{\mu\eta(\theta^\perp)^2}{4c(\eta, \theta^\parallel)} \right) (\lambda(\theta^\parallel) - \lambda(\theta_*^\parallel)). \quad (176)$$

Proof. For $\theta^\perp$ sufficiently small, there is unique $\delta(\theta^\parallel, \theta^\perp) = O((\theta^\perp)^3 \min\{1, d_M(\theta^\parallel, \theta_*^\parallel)\})$ such that

$$\text{GD}^\parallel(\eta, \theta^\parallel, \theta^\perp) = \exp_{\theta^\parallel} \left(-\frac{\eta(\theta^\perp)^2}{2c(\eta, \theta^\parallel)} \nabla_M \lambda(\theta^\parallel) + \delta(\theta^\parallel, \theta^\perp) \right). \quad (177)$$

To see that $\text{GD}^\parallel(\eta, \cdot, \theta^\perp)$ preserves the ball $B_M(\theta_*^\parallel, r)$, consider the function $\rho(\theta^\parallel) := (1/2)d_M(\theta^\parallel, \theta_*^\parallel)^2$. By the Gauss lemma [29, Corollary 6.10], $\nabla_M \rho(\theta^\parallel) = -\log_{\theta^\parallel}(\theta_*^\parallel)$. Fixing $\theta^\parallel \in B_M(\theta_*^\parallel, r)$, for any vector $w \in T_{\theta^\parallel} M$ one therefore has

$$\left. \frac{d}{dt} \right|_{t=0} \rho \left(\exp_{\theta^\parallel} \left(-t \nabla_M \lambda(\theta^\parallel) + t^{3/2} w \right), \theta_*^\parallel \right) = -\langle \nabla_M \rho(\theta^\parallel), \nabla_M \lambda(\theta^\parallel) \rangle \quad (178)$$

$$= \langle \log_{\theta^\parallel}(\theta_*^\parallel), \nabla_M \lambda(\theta^\parallel) \rangle \quad (179)$$

$$\leq -2\mu\rho(\theta^\parallel) < 0, \quad (180)$$

where the third line follows from Lemma E.2. Thus, for all $\theta^\perp$ sufficiently small, we are assured that $\rho(\text{GD}^\parallel(\eta, \theta^\parallel, \theta^\perp)) < \rho(\theta^\parallel)$, implying that $\text{GD}^\parallel(\eta, \theta^\parallel, \theta^\perp) \in B_M(\theta_*^\parallel, r)$.

We now come to the descent lemma. Since $B_M(\theta_*, r)$ is geodesically convex, we may perform a covariant Taylor expansion and invoke the geodesic smoothness of λ to obtain

$$\lambda(\text{GD}^\parallel(\eta, \theta^\parallel, \theta^\perp)) \leq \lambda(\theta^\parallel) + \left\langle \nabla_M \lambda(\theta^\parallel), -\frac{\eta(\theta^\perp)^2}{2c(\eta, \theta^\parallel)} \nabla_M \lambda(\theta^\parallel) + \delta(\theta^\parallel, \theta^\perp) \right\rangle + \quad (181)$$

$$+ \frac{L}{2} \left\| \frac{\eta(\theta^\perp)^2}{2c(\eta, \theta^\parallel)} \nabla_M \lambda(\theta^\parallel) - \delta(\theta^\parallel, \theta^\perp) \right\|^2 \quad (182)$$

$$\leq \lambda(\theta^\parallel) - \frac{\eta(\theta^\perp)^2}{2c(\eta, \theta^\parallel)} \left(1 - \frac{\eta(\theta^\perp)^2 L}{4c(\eta, \theta^\parallel)} \right) \|\nabla_M \lambda(\theta^\parallel)\|^2 + \quad (183)$$

$$+ \left(1 - \frac{\eta(\theta^\perp)^2 L}{2c(\eta, \theta^\parallel)} \right) \langle \nabla_M \lambda(\theta^\parallel), \delta(\theta^\parallel, \theta^\perp) \rangle + \frac{L}{2} \|\delta(\theta^\parallel, \theta^\perp)\|^2. \quad (184)$$

From Cauchy-Schwarz and the estimate $0 \leq (\epsilon^{1/2}a - \epsilon^{-1/2}b)^2$ (implying that $ab \leq (1/2)(\epsilon a^2 + \epsilon^{-1}b^2)$) with $\epsilon = \eta(\theta^\perp)^2/(4c(\eta, \theta^\parallel))$, one obtains

$$\langle \nabla_M \lambda(\theta^\parallel), \delta(\theta^\parallel, \theta^\perp) \rangle \leq \|\nabla_M \lambda(\theta^\parallel)\| \|\delta(\theta^\parallel, \theta^\perp)\| \quad (185)$$

$$\leq \frac{\eta(\theta^\perp)^2}{8c(\eta, \theta^\parallel)} \|\nabla_M \lambda(\theta^\parallel)\|^2 + \frac{2c(\eta, \theta^\parallel)}{\eta(\theta^\perp)^2} \|\delta(\theta^\parallel, \theta^\perp)\|^2. \quad (186)$$

Taking $|\theta^\perp|$ sufficiently small that $\eta(\theta^\perp)^2 L/(4c(\eta, \theta^\parallel)) \leq 1/4$, therefore, one obtains

$$\lambda(\text{GD}^\parallel(\eta, \theta^\parallel, \theta^\perp)) \leq \lambda(\theta^\parallel) - \frac{\eta(\theta^\perp)^2}{4c(\eta, \theta^\parallel)} \|\nabla_M \lambda(\theta^\parallel)\|^2 + \left(\frac{2c(\eta, \theta^\parallel)}{\eta(\theta^\perp)^2} + \frac{L}{2} \right) \|\delta(\theta^\parallel, \theta^\perp)\|^2 \quad (187)$$

$$= \lambda(\theta^\parallel) - \frac{\eta(\theta^\perp)^2}{4c(\eta, \theta^\parallel)} \|\nabla_M \lambda(\theta^\parallel)\|^2 + O((\theta^\perp)^4 \min\{1, d_M(\theta^\parallel, \theta_*^\parallel)\}^2). \quad (188)$$

Subtracting $\lambda(\theta_*^\parallel)$ from both sides and applying geodesic strong convexity ($\|\nabla_M \lambda(\theta^\parallel)\|^2 \geq 2\mu(\lambda(\theta^\parallel) - \lambda(\theta_*^\parallel))$) by Lemma E.3) yields

$$\lambda(\text{GD}^\parallel(\eta, \theta^\parallel, \theta^\perp)) - \lambda(\theta_*^\parallel) \leq \left(1 - \frac{\mu\eta(\theta^\perp)^2}{2c(\eta, \theta^\parallel)} \right) (\lambda(\theta^\parallel) - \lambda(\theta_*^\parallel)) + O((\theta^\perp)^4 \min\{1, d_M(\theta^\parallel, \theta_*^\parallel)\}^2). \quad (189)$$

Finally, invoking geodesic strong convexity once more to estimate $d_M(\theta^\parallel, \theta_*^\parallel)^2 \leq (2/\mu)(\lambda(\theta^\parallel) - \lambda(\theta_*^\parallel))$ as in Lemma E.1 and taking $\theta^\perp$ sufficiently small allows one to absorb the remainder term on the left side into the $\lambda(\theta^\parallel) - \lambda(\theta_*^\parallel)$ factor and yields the result.. $\square$

The next result will be used in giving uniform bounds on the magnitudes of the $\theta^\perp$ iterates.

Lemma D.2. For $\alpha \in \mathbb{R}$, define $f_\alpha : \mathbb{R} \rightarrow \mathbb{R}$ by

$$f_\alpha(x) := -(1 + \alpha)x + x^3 + O(x^4). \quad (190)$$

For $\alpha_0, \alpha_1 \in \mathbb{R}$, consider the composite $f_{\alpha_1\alpha_0} := f_{\alpha_1} \circ f_{\alpha_0}$. Then for all $\gamma > 0$ sufficiently small and all $\alpha_0, \alpha_1 \in [0, \gamma]$:

1. $f_{\alpha_1\alpha_0}$ is monotonically increasing on $[-2\sqrt{\gamma}, 2\sqrt{\gamma}]$.
2. $f_{\alpha_1\alpha_0}$ admits the sole fixed points 0,

$$\xi_- = -\sqrt{\frac{\alpha_0 + \alpha_1}{2}} + O(\alpha_0 + \alpha_1), \quad \xi_+ = \sqrt{\frac{\alpha_0 + \alpha_1}{2}} + O(\alpha_0 + \alpha_1) \quad (191)$$

in the interval $[-2\sqrt{\gamma}, 2\sqrt{\gamma}]$.

Proof. For $\gamma > 0$ small and all $\alpha_0, \alpha_1 \in [0, \gamma]$, one computes

$$f_{\alpha_1\alpha_0}(x) = (1 + \alpha_1 + \alpha_0)x - 2x^3 + O(\alpha_0\alpha_1x, \alpha_1x^3, \alpha_0x^3, x^4), \quad (192)$$

$$Df_{\alpha_1\alpha_0}(x) = 1 + \alpha_1 + \alpha_0 - 6x^2 + O(\alpha_0\alpha_1, \alpha_1x^2, \alpha_0x^2, x^3). \quad (193)$$

In particular, for all $|x| \leq 2\sqrt{\gamma}$, one has

$$Df_{\alpha_1\alpha_0}(x) \geq 1 + \alpha_1 + \alpha_0 - 24\gamma + O(\gamma^{3/2}) > 0 \quad (194)$$

for all γ sufficiently small, so that $f_{\alpha_1\alpha_0}|_{[-2\sqrt{\gamma}, 2\sqrt{\gamma}]}$ is monotonically increasing. Set $\beta := (\alpha_0 + \alpha_1)/2$ for notational convenience, and denote

$$\Phi(x) := f_{\alpha_1\alpha_0}(x) - x = 2\beta x - 2x^3 + O(\beta^2x, \beta x^3, x^4). \quad (195)$$

Fixed points of $f_{\alpha_1\alpha_0}$ are precisely roots of Φ , one of which is clearly $x = 0$. Moreover, if $\beta = 0$, then $\Phi(x) = -2x^3 + O(x^4)$ so that in this case $x = 0$ is the only root of Φ on $[-2\sqrt{\gamma}, 2\sqrt{\gamma}]$ for all γ sufficiently small. Suppose now that $\beta > 0$. We will demonstrate the existence of a unique root ξ_+ of Φ in $(0, 2\sqrt{\gamma}]$ with the claimed formula. Since $2\beta x - 2x^3 > 0$ for all $0 < x \leq 2^{-1/2}\sqrt{\beta}$, we see that for all γ sufficiently small, $\Phi(x)$ is strictly positive on $(0, 2^{-1/2}\sqrt{\beta}]$ and so cannot admit any roots therein. However, observe that for all γ sufficiently small one has

$$D\Phi(x) = 2\beta - 6x^2 + O(\beta^2, \beta x^2, x^3) < 0 \quad (196)$$

for all $2^{-1/2}\sqrt{\beta} \leq x \leq 2\sqrt{\gamma}$, implying that Φ is monotonically decreasing on $[2^{-1/2}\sqrt{\beta}, 2\sqrt{\gamma}]$, and thus admits *at most* one root therein. Since moreover $\Phi(2^{-1/2}\sqrt{\beta}) > 0$ and $\Phi(2\sqrt{\gamma}) = -12\beta^{3/2} + O(\beta^2) < 0$, by the intermediate value theorem Φ admits *precisely* one root ξ_+ in $[2^{-1/2}\sqrt{\beta}, 2\sqrt{\gamma}]$. By the mean value theorem, there is z between $\sqrt{\beta}$ and $\xi_+ \leq 2\sqrt{\beta}$ such that

$$0 = \Phi(\xi_+) = \Phi(\sqrt{\beta}) + D\Phi(z)(\xi_+ - \sqrt{\beta}), \quad (197)$$

so that, since $|D\Phi(z)| \geq 4\beta + O(\beta^{3/2}) = \Omega(\beta)$ and $\Phi(\sqrt{\beta}) = O(\beta^2)$,

$$\xi_+ = \sqrt{\beta} - \frac{\Phi(\sqrt{\beta})}{D\Phi(z)} = \sqrt{\beta} + O(\beta) \quad (198)$$

as claimed. A similar argument can be used to show there exists a unique root $\xi_- = -\sqrt{\beta} + O(\beta)$ of Φ in $[-2\sqrt{\gamma}, 0)$, thus proving the lemma. $\square$

Lemma D.3 (Upper and lower bounds for $T^\perp$ iterates). *Let $\{\lambda_t\}_{t \in \mathbb{N}}$ be a monotonically decreasing sequence of numbers. Then for all $\eta > 0$ such that $\eta\lambda_0 > 2$ is sufficiently small, there is a constant $C > 0$ such that for any $\theta_0^\perp$ sufficiently small, the iterates defined by*

$$\theta_{t+1}^\perp = (1 - \eta\lambda_t)\theta_t^\perp + (\theta_t^\perp)^3 + O((\theta_t^\perp)^4) \quad (199)$$

satisfy the bounds

$$\frac{|\theta_0^\perp|}{\sqrt{1 + 3(\theta_0^\perp)^2 t}} \leq |\theta_t^\perp| \leq 2\sqrt{\eta\lambda(\theta_0^\perp) - 2} \quad (200)$$

for all $t \in \mathbb{N}$ such that $\eta\lambda_t \geq 2$.

Proof. We prove the upper bound first. For each $s \in \mathbb{N}$ such that $\eta\lambda_s > 2$, denote $\alpha_s := \eta\lambda_s - 2$ and denote by f_s the map

$$f_s : \theta^\perp \mapsto (1 - \eta\lambda_s)\theta^\perp + (\theta^\perp)^3 + O((\theta^\perp)^4) \quad (201)$$

$$= -(1 + \alpha_s)\theta^\perp + (\theta^\perp)^3 + O((\theta^\perp)^4). \quad (202)$$

Supposing that $s \in \mathbb{N}$ is such that both $\eta\lambda_s, \eta\lambda_{s+1} \geq 2$, since $\{\lambda_t\}_{t \in \mathbb{N}}$ is monotonically decreasing Lemma D.2 guarantees that $f_{s+1,s} := f_{s+1} \circ f_s$ admits attractive fixed points $\xi_{s+1,s,\pm} = \pm\sqrt{(\alpha_{s+1} + \alpha_s)/2} + O(\alpha_{s+1} + \alpha_s)$ and is monotonically increasing on $[-2\sqrt{\alpha_0}, 2\sqrt{\alpha_0}]$ provided $\alpha_0 = \eta\lambda_0 - 2$ is sufficiently small. Suppose now that $\theta_0^\perp \in [-2\sqrt{\alpha_0}, 2\sqrt{\alpha_0}] \setminus \{0\}$. The iterates $\theta_t^\perp$ as defined in (199) satisfy

$$\theta_{2t+2}^\perp = f_{2t+1,2t}(\theta_{2t}^\perp), \quad \theta_{2t+1}^\perp = f_{2t,2t-1}(\theta_{2t-1}^\perp), \quad t \in \mathbb{N}. \quad (203)$$

Assuming without loss of generality that $\theta_0^\perp > 0$, we will prove by induction that $\theta_{2t}^\perp \leq 2\sqrt{\alpha_0}$ for all $t \in \mathbb{N}$ such that $\alpha_s \geq 0$ for all $s \leq 2t - 1$. Clearly $\theta_0^\perp \leq 2\sqrt{\alpha_0}$ by assumption. Suppose as an inductive hypothesis that $0 < \theta_{2t}^\perp \leq 2\sqrt{\alpha_0}$ and that $\alpha_{2t+1}, \alpha_{2t} \geq 0$. Using the monotonicity of $f_{2t+1,2t}$ from Lemma D.2 and the fact that $\xi_{2t+1,2t,+}$ is its sole fixed point in $(0, 2\sqrt{\alpha_0}]$, either

$$\theta_{2t}^\perp \leq \xi_{2t+1,2t,+} \Rightarrow \theta_{2t}^\perp \leq f_{2t+1,2t}(\theta_{2t}^\perp) = \theta_{2t+2}^\perp \leq \xi_{2t+1,2t,+} \leq 2\sqrt{\alpha_0}, \quad (204)$$

or

$$\theta_{2t}^\perp \geq \xi_{2t+1,2t,+} \Rightarrow \xi_{2t+1,2t,+} \leq f_{2t+1,2t}(\theta_{2t}^\perp) = \theta_{2t+2}^\perp \leq \theta_{2t}^\perp \leq 2\sqrt{\alpha_0} \quad (205)$$

by the inductive hypothesis; in either case $\theta_{2t+2}^\perp \leq 2\sqrt{\alpha_0}$, thus proving the claim. A symmetric argument shows that the odd iterates satisfy $|\theta_{2t+1}^\perp| \leq 2\sqrt{\alpha_0}$ for all $t \in \mathbb{N}$ such that $\alpha_s \geq 0$ for all $s \leq 2t$, thus proving the upper-bound in (200).

We now prove the lower bound. Invoking the upper bound, α_0 can be taken sufficiently small such that the recursive estimate

$$|\theta_{t+1}^\perp| = |\theta_t^\perp|(1 + \alpha_t - (\theta_t^\perp)^2 + O((\theta_t^\perp)^3)) \quad (206)$$

$$\geq |\theta_t^\perp|(1 - 2(\theta_t^\perp)^2) \quad (207)$$

holds as long as $\alpha_t \geq 0$. By taking α_0 even smaller if necessary, squaring, taking the reciprocal of both sides and applying $1/(1-x)^2 \leq 1+3x$ for x sufficiently small one deduces that

$$\frac{1}{(\theta_{t+1}^\perp)^2} \leq \frac{1}{(\theta_t^\perp)^2}(1 + 6(\theta_t^\perp)^2) \Rightarrow \frac{1}{(\theta_{t+1}^\perp)^2} - \frac{1}{(\theta_t^\perp)^2} \leq 6, \quad (208)$$

from which it follows by a telescoping sum that

$$|\theta_t^\perp| \geq \frac{|\theta_0^\perp|}{\sqrt{1 + 6(\theta_0^\perp)^2 t}} \quad (209)$$

for all t such that $\alpha_t \geq 0$. This proves the claim. $\square$

Theorem D.4. Assume that $\eta < 2/\lambda(\theta_*^\parallel)$, and let $0 < c, C < \infty$ satisfy $c \leq c(\eta, \theta^\parallel) \leq C$ for all $\theta^\parallel \in B_M(\theta_*^\parallel, r)$. Then for all $\theta_0^\perp$ sufficiently close to zero and all $\theta_0^\parallel \neq \theta_*^\parallel \in B_M(\theta_*^\parallel, r)$, if $\eta > 2/\lambda(\theta_0^\parallel)$ is sufficiently small, then there is $\tau \in \mathbb{N}$, with

$$\tau \leq \tau(\theta_0^\parallel, \theta_0^\perp) := \left\lceil \frac{1}{6(\theta_0^\perp)^2} \left(\frac{2/\eta - \lambda(\theta_*^\parallel)}{\lambda(\theta_0^\parallel) - \lambda(\theta_*^\parallel)} \right)^{-\frac{24C}{\mu\eta}} \right\rceil, \quad (210)$$

such that $\eta < 2/\lambda(\theta_t^\parallel)$ for all $t \geq \tau$ and $\eta \geq 2/\lambda(\theta_t^\parallel)$ for all $t < \tau$. Consequently, setting $\beta := 1 - (2 - \eta\lambda(\theta_\tau^\parallel)) < 1$, the iterates $(\theta_t^\parallel, \theta_t^\perp)$ converge to a global minimum $(\theta_\infty^\parallel, 0)$ of ℓ in M with suboptimal sharpness

$$\lambda(\theta_\infty^\parallel) - \lambda(\theta_*^\parallel) \geq \exp\left(-\frac{4\eta L^2}{c\mu} \frac{(\theta_\tau^\perp)^2}{1 - \beta^2}\right) (\lambda(\theta_\tau^\parallel) - \lambda_*). \quad (211)$$

at a rate $O(\beta^t)$.

Proof. For any $\theta_0^\parallel \neq \theta_*^\parallel$, take $\theta_0^\perp \neq 0$ sufficiently close to zero that the hypotheses of Lemma D.1 and Lemma D.3 are satisfied, with λ_0 in Lemma D.3 taken to be equal to $\lambda(\theta_0^\parallel)$, and assume that $\eta\lambda_0 > 0$ is sufficiently small as in Lemma D.3. Then we may assume that $\lambda_t := \lambda(\theta_t^\parallel)$ is monotonically decreasing and the bounds of Lemma D.3 apply to $\theta_t^\perp$. Since $\{\lambda_t\}_{t \in \mathbb{N}}$ is monotonically decreasing, any τ satisfying $\eta < 2/\lambda(\theta_t^\parallel)$ for all $t \geq \tau$ and $\eta \geq 2/\lambda(\theta_t^\parallel)$ for all $t < \tau$ is necessarily unique. We claim that this τ satisfies the upper bound $\tau \leq \tau(\theta_0^\parallel, \theta_0^\perp)$. Indeed, suppose for a contradiction that $\tau > \tau(\theta_0^\parallel, \theta_0^\perp)$. Then for all $t < \tau$, denoting $\lambda_* := \lambda(\theta_*^\parallel)$, one would have

$$\lambda_{t+1} - \lambda_* \leq \left(1 - \frac{\mu\eta}{4c(\eta, \theta_t^\parallel)} \frac{(\theta_0^\perp)^2}{(1 + 6(\theta_0^\perp)^2 t)}\right) (\lambda_t - \lambda_*) \quad (212)$$

by Lemmas D.1 and D.3. Invoking $c(\eta, \theta_t^\parallel) \leq C$, taking logarithms, applying the upper-bound $\ln(1-x) \leq -x$ and applying a telescoping sum then yield

$$\ln(\lambda_t) - \ln(\lambda_0) \leq -\frac{\mu\eta}{24C} \sum_{s=0}^{t-1} \frac{1}{\frac{1}{6(\theta_0^\perp)^2} + s} \quad (213)$$

$$\leq -\frac{\mu\eta}{24C} \ln(1 + 6(\theta_0^\perp)^2 t) \quad (214)$$

by Lemma E.4, so that

$$\lambda_t \leq (\lambda_0 - \lambda_*)(1 + 6(\theta_0^\perp)^2 t)^{-\frac{\mu\eta}{24C}} + \lambda_*. \quad (215)$$

However, plugging in $t = \tau(\theta_0^\perp, \theta_0^\parallel)$ then reveals that $\lambda_t < 2/\eta$, thus contradicting the assumption that $\tau > \tau(\theta_0^\perp, \theta_0^\parallel)$ and implying that $\tau \leq \tau(\theta_0^\perp, \theta_0^\parallel)$ as claimed.

We now come to the final convergence and sharpness suboptimality claim. For all $t \geq \tau$, Lemma D.1 continues to imply that $\lambda_t < 2/\eta$ decreases monotonically, while $\theta_t^\perp$ contracts exponentially according to the estimate

$$|\theta_{t+1}^\perp| = |\theta_t^\perp| |1 + (\eta\lambda_t - 2) - (\theta_t^\perp)^2 + O(\theta_t^\perp)^3| \leq |\theta_t^\perp| |1 - (2 - \eta\lambda_t)| \leq |\theta_t^\perp| |1 - (2 - \eta\lambda_\tau)|. \quad (216)$$

This proves the claim that the iterates $(\theta_t^\parallel, \theta_t^\perp)$ converge to a global minimum $\theta_\infty^\parallel$ of ℓ at rate $O((1 - (2 - \eta\lambda_\tau))^t)$. We now turn to proving the suboptimality of the sharpness at $\theta_\infty^\parallel$. The geodesic strong convexity of λ in the form of Lemma E.2 implies that

$$\mu d_M(\theta^\parallel, \theta_*^\parallel)^2 \leq \langle \Pi_{\theta^\parallel \rightarrow \theta_*^\parallel} \nabla_M \lambda(\theta^\parallel), \log_{\theta_*^\parallel}(\theta^\parallel) \rangle \leq \|\nabla_M \lambda(\theta^\parallel)\| d_M(\theta^\parallel, \theta_*^\parallel) \quad (217)$$

for any $\theta^\parallel \in B_M(\theta_*^\parallel, r)$, so that the $\parallel$ -update can be written

$$\text{GD}^\parallel(\eta, \theta^\parallel, \theta^\perp) = \exp_{\theta^\parallel} \left(-\frac{\eta(\theta^\perp)^2}{2c(\eta, \theta^\parallel)} \nabla_M \lambda(\theta^\parallel) + O((\theta^\perp)^3 \min\{1, d_M(\theta^\parallel, \theta_*^\parallel)\}) \right) \quad (218)$$

$$= \exp_{\theta^\parallel} \left(-\frac{\eta(\theta^\perp)^2}{2c(\eta, \theta^\parallel)} \nabla_M \lambda(\theta^\parallel) + O((\theta^\perp)^3 \min\{1, \|\nabla_M \lambda(\theta^\parallel)\|\}) \right). \quad (219)$$

Consequently, invoking the geodesic strong convexity of λ once again in the form of Lemma E.1,

$$\lambda_{t+1} \geq \lambda_t + \langle \nabla_M \lambda(\theta_t^\parallel), \log_{\theta_t^\parallel}(\theta_{t+1}^\parallel) \rangle + \frac{\mu}{2} \|\log_{\theta_t^\parallel}(\theta_{t+1}^\parallel)\|^2 \quad (220)$$

$$= \lambda_t - \left\langle \nabla_M \lambda(\theta_t^\parallel), \frac{\eta(\theta_t^\perp)^2}{2c(\eta, \theta_t^\parallel)} \nabla_M \lambda(\theta_t^\parallel) + O((\theta_t^\perp)^3 \min\{1, \|\nabla_M \lambda(\theta_t^\parallel)\|\}) \right\rangle \quad (221)$$

$$+ \frac{\mu}{2} \left\| \frac{\eta(\theta_t^\perp)^2}{2c(\eta, \theta_t^\parallel)} \nabla_M \lambda(\theta_t^\parallel) + O((\theta_t^\perp)^3 \min\{1, \|\nabla_M \lambda(\theta_t^\parallel)\|\}) \right\|^2 \quad (222)$$

$$= \lambda_t - \frac{\eta(\theta_t^\perp)^2}{2c(\eta, \theta_t^\parallel)} \|\nabla_M \lambda(\theta_t^\parallel)\|^2 (1 + O(\theta_t^\perp)) \quad (223)$$

$$\geq \lambda_t - \frac{\eta(\theta_t^\perp)^2}{c(\eta, \theta_t^\parallel)} \|\nabla_M \lambda(\theta_t^\parallel)\|^2 \quad (224)$$

for sufficiently small $2\sqrt{\eta\lambda_0 - 2} > |\theta_t^\perp|$. Subtracting λ_* from both sides and invoking the Lipschitz gradients ($\|\nabla_M \lambda(\theta^\parallel)\| \leq L d_M(\theta^\parallel, \theta_*^\parallel)$) and strong convexity ($(\mu/2) d_M(\theta^\parallel, \theta_*^\parallel)^2 \leq \lambda(\theta^\parallel) - \lambda_*$ from Lemma E.1) properties of λ then yield

$$\lambda_{t+1} - \lambda_* \geq \lambda_t - \lambda_* - \frac{\eta(\theta_t^\perp)^2}{c(\eta, \theta_t^\parallel)} L^2 d_M(\theta_t^\parallel, \theta_*^\parallel)^2 \quad (225)$$

$$\geq (\lambda_t - \lambda_*) \left(1 - \frac{2\eta(\theta_t^\perp)^2 L^2}{c(\eta, \theta_t^\parallel) \mu} \right) \quad (226)$$

$$\geq (\lambda_t - \lambda_*) \left(1 - \frac{2\eta L^2 (\theta_\tau^\perp)^2}{c\mu} \beta^{2t} \right), \quad (227)$$

where in the final line we have invoked $c(\eta, \theta_t^\parallel) \geq c$. It follows that

$$\lambda_t - \lambda_* \geq \prod_{s=\tau}^{t-1} \left(1 - \frac{2\eta L^2 (\theta_\tau^\perp)^2}{c\mu} \beta^{2s} \right) (\lambda_\tau - \lambda_*). \quad (228)$$

Taking $2\sqrt{\eta\lambda_0 - 2} > |\theta_\tau^\perp|$ smaller if necessary and using the bound $\ln(1-x) \geq -2x$ valid for small x one has

$$\ln \left(\prod_{s=0}^{t-1} \left(1 - \frac{2\eta L^2 (\theta_\tau^\perp)^2}{c\mu} \beta^{2s} \right) \right) \geq -\frac{4\eta L^2 (\theta_\tau^\perp)^2}{c\mu} \sum_{s=0}^{t-1} \beta^{2s} \quad (229)$$

$$\geq -\frac{4\eta L^2 (\theta_\tau^\perp)^2}{c\mu} \frac{1}{1 - \beta^2} \quad (230)$$

from which the result follows. $\square$

D.2 Critical regime

The convergence theorem in the critical regime is the most difficult to prove. This is because in this regime, the fixed point to which the iterates converge is merely *parabolic* rather than *hyperbolic* as is (at least asymptotically) true in the subcritical and supercritical regimes. The convergence theorem in the critical regime will follow from a careful analysis of the following parabolic system.

Fix $0 < a < c$ and $b > 0$. Let $A(x, y) = O(x, y)$, $B(x) = O(x)$ and $C(y) = O(y)$ be analytic functions, and consider the analytic functions

$$T_x(x, y) = (1 - (a + A(x, y))y^2)x, \quad T_y(x, y) = (1 + (b + B(x))x^2 - (c + C(y))y^2)y. \quad (231)$$

Denote by $T = (T_x, T_y) : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ the corresponding analytic map. Clearly 0 is a fixed point of T ; we will be concerned with proving convergence of the iterates of T to this fixed point. The following result guarantees that T admits an invariant manifold approaching the origin. This result is essential, since our final convergence proof is a two-stage proof; first, convergence to a neighbourhood of the invariant manifold is proved, following which, second, convergence to the origin is easy.

Lemma D.5. *There exists $r > 0$ and a function $u : \{x > 0 : |x^2 - r| < r\} \rightarrow \mathbb{R}$ whose graph $\Gamma_u := \{(x, u(x)) : x \in \text{dom}(u)\}$ is invariant under T , and for which $u(x) = \sqrt{\frac{b}{c-a}}x + O(x^2)$.*

Proof. The analytic map $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ extends uniquely to a holomorphic map $\tilde{T}$ defined on a neighbourhood of 0 in $\mathbb{C}^2$. Like T , at the origin the map $\tilde{T}$ is the identity to first order, vanishes at order two, and has third order Taylor coefficients given by

$$\partial_{xyy}^3 T_x(0, 0) = -2a, \quad \partial_{yxx}^3 T_y(0, 0) = 2b, \quad \partial_{yyy}^3 T_y(0, 0) = -6c, \quad (232)$$

with all other third order coefficients being zero. Observe that on the vector v given by

$$v = \left(1, \sqrt{\frac{b}{c-a}} \right) \quad (233)$$

one has

$$\partial^3 T(0, 0)[v, v, v] = \left(-6a \frac{b}{c-a}, 6b \sqrt{\frac{b}{c-a}} - 6c \sqrt{\frac{b}{c-a}}^3 \right) \quad (234)$$

$$= -6a \sqrt{\frac{b}{c-a}}^3 v, \quad (235)$$

so that v is a non-degenerate characteristic direction of $\tilde{T}$ [6, Definition 4.1], which is moreover attracting since $-6a\sqrt{b/(c-a)} < 0$. By [6, Section 6], therefore, there exists $r > 0$ and a holomorphic function $u : \{x \in \mathbb{C} : |x^2 - r| < r\} \rightarrow \mathbb{C}$ such that $u(x) = \sqrt{\frac{b}{c-a}}x + O(x^2)$ whose graph is invariant under T . Restricting this u to the intersection of its domain with the positive real axis yields the result. $\square$

The next lemma will be key in establishing convergence to the invariant manifold.

Lemma D.6. *Let $u(x) = \sqrt{\frac{b}{c-a}}x + O(x^2)$ be the function from Lemma D.5, and define $\Delta(x, y) := y - u(x)$. Then $\tilde{\Delta}(x, y) := \Delta(x, y)/x$ satisfies*

$$|\tilde{\Delta} \circ T(x, y)| \leq |\tilde{\Delta}(x, y)| \left(1 - \frac{c-a}{2}y^2\right) \quad (236)$$

for all sufficiently small $x > 0$, $y \geq 0$.

Proof. We estimate $\frac{\tilde{\Delta} \circ T}{\tilde{\Delta}}$. Define

$$D(x, y) := \frac{T_x(x, y)}{x}, \quad N(x, y) := \frac{\Delta \circ T(x, y)}{\Delta(x, y)}, \quad (x, y) \in \text{dom}(u) \times \mathbb{R}_{\geq 0}. \quad (237)$$

Observe that both D and N are analytic on the domain of u . Observe moreover that

$$\frac{\tilde{\Delta} \circ T(x, y)}{\tilde{\Delta}(x, y)} = \frac{\Delta \circ T(x, y)}{T_x(x, y)} \cdot \frac{x}{\Delta(x, y)} = \frac{N(x, y)}{D(x, y)} = 1 - \frac{D(x, y) - N(x, y)}{D(x, y)}, \quad (238)$$

so that estimation of $\frac{\tilde{\Delta} \circ T}{\tilde{\Delta}}$ is reduced to estimating $\frac{D-N}{D}$. We first estimate $D - N$. For this, observe that since $T(x, 0) = (x, 0)$, one has $D(x, 0) - N(x, 0) = 0$ for all $x \in \text{dom}(u)$. Thus, by analyticity of D and N , there is a unique analytic function H on $\text{dom}(u) \times \mathbb{R}$ such that

$$D(x, y) - N(x, y) = yH(x, y). \quad (239)$$

We compute $H(x, y)$ to second order. For notational convenience, set

$$\kappa := \sqrt{\frac{b}{c-a}} \quad (240)$$

so that $u(x) = \kappa x + O(x^2)$. Differentiate (239) with respect to y and evaluate at $y = 0$ to obtain

$$H(x, 0) = \partial_y D(x, 0) - \partial_y N(x, 0), \quad (241)$$

Observe that

$$\partial_y D(x, 0) = \partial_y (1 - ay^2 + A(x, y)y^2)|_{y=0} = 0 \quad (242)$$

while, from $N = (\Delta \circ T)/\Delta$, one has

$$\partial_y N(x, 0) = \frac{\Delta \partial_y (\Delta \circ T) - (\Delta \circ T) \partial_y \Delta}{\Delta^2}(x, 0). \quad (243)$$

From $\Delta(x, y) = y - u(x)$, is easily seen that

$$\Delta(x, 0) = -u(x), \quad \partial_y \Delta(x, 0) = 1, \quad \Delta \circ T(x, 0) = -u(T_x(x, 0)) = -u(x) \quad (244)$$

while

$$\partial_y (\Delta \circ T)(x, 0) = \partial_y T_y(x, 0) - u'(x) \partial_y T_x(x, 0) = 1 + bx^2 + O(x^3) \quad (245)$$

so that (243) yields

$$H(x, 0) = -\partial_y N(x, 0) = \frac{bx^2 + O(x^3)}{u(x)} = (c-a)\kappa x + O(x^2). \quad (246)$$

To compute the second order component of H , differentiate (239) twice with respect to y and evaluate at $y = 0$ to obtain

$$\partial_y H(x, 0) = \frac{1}{2}(\partial_y^2 D(x, 0) - \partial_y^2 N(x, 0)). \quad (247)$$

One sees that

$$\partial_y^2 D(x, 0) = \partial_y^2 (1 - ay^2 + A(x, y)y^2)|_{y=0} = -2a + O(x), \quad (248)$$

while

$$\partial_y^2 N(x, 0) = \left(\frac{\partial_y^2(\Delta \circ T)}{\Delta} - \frac{(\Delta \circ T) \partial_y^2 \Delta}{\Delta^2} - 2 \frac{\partial_y(\Delta \circ T) \partial_y \Delta}{\Delta^2} + 2 \frac{(\Delta \circ T) (\partial_y \Delta)^2}{\Delta^3} \right)(x, 0). \quad (249)$$

Since

$$\partial_y^2(\Delta \circ T)(x, 0) = \partial_y^2 T_y(x, 0) - u''(x)(\partial_y T_x(x, 0))^2 - u'(x) \partial_y^2 T_x(x, 0) = 2a\kappa x + O(x^2) \quad (250)$$

one obtains

$$\partial_y^2 N(x, 0) = -\frac{2a\kappa x + O(x^2)}{\kappa x + O(x^2)} - 2 \frac{1 + bx^2 + O(x^3)}{\kappa^2 x^2 + O(x^3)} + 2 \frac{1}{\kappa^2 x^2 + O(x^3)} \quad (251)$$

$$= -2a - 2b/\kappa^2 + O(x) = -2c + O(x). \quad (252)$$

Thus

$$\partial_y H(x, 0) = \frac{1}{2}(-2a + 2c + O(x)) = (c - a) + O(x). \quad (253)$$

We deduce from (246), (253) and (239) that

$$D(x, y) - N(x, y) = (c - a)(y^2 + \kappa yx) + O(yx^2, y^2x, y^3). \quad (254)$$

It follows that for all sufficiently small $x > 0, y \geq 0$, one has $D(x, y) - N(x, y) \geq \frac{(c-a)}{2}y^2$ and $1 \geq D(x, y) > 0$ so that

$$\frac{\tilde{\Delta} \circ T(x, y)}{\tilde{\Delta}(x, y)} = 1 - \frac{D(x, y) - N(x, y)}{D(x, y)} \leq 1 - \frac{c-a}{2}y^2 \quad (255)$$

as claimed. $\square$

Lemma D.7 below gives uniform upper-bounds on the magnitudes of iterates, which are used in enabling certain estimates in the convergence theorem.

Lemma D.7. Define $W(x, y) := bx^2 + \frac{a}{2}y^2$. Then for all (x_0, y_0) in a sufficiently small neighbourhood of 0, with $y_0 > 0$, the sequence x_t is monotonically non-increasing and the sequence y_t satisfies the uniform bound $y_t^2 \leq 2W(x_0, y_0)/a$ for all $t \in \mathbb{N}$.

Proof. In any sufficiently small neighbourhood U of $0 \in \mathbb{R}_{\geq 0}^2$, one has

$$W(T(x, y)) = bx^2(1 - ay^2 + O(xy^2, y^3))^2 + \frac{1}{2}ay^2(1 + bx^2 - cy^2 + O(x^3, y^3))^2 \quad (256)$$

$$= bx^2(1 - 2ay^2) + O(x^3y^2, x^2y^3) + \frac{1}{2}ay^2(1 + 2bx^2 - 2cy^2) + O(x^3y^2, y^5) \quad (257)$$

$$= W(x, y) - abx^2y^2 - acy^4 + O(x^3y^2, x^2y^3, y^5) \quad (258)$$

$$\leq W(x, y) - \frac{1}{2}y^2(abx^2 + acy^2) \quad (259)$$

for all $(x, y) \in U$, so that W is non-increasing as a function of the iterates of the map close to zero.

Now fix (x_0, y_0) sufficiently small that, with $W_0 := W(x_0, y_0)$ one has $(x_0, \sqrt{2W_0/a})$ contained in U . Then in particular (x_0, y_0) is contained in U since $y_0 \leq \sqrt{2W_0/a}$. Suppose as an inductive hypothesis that $y_t \leq \sqrt{2W_0/a}$, $x_t \leq x_0$ and $W_t \leq W_0$. Shrinking U yet further if necessary, one sees that

$$1 > 1 - y_t^2 + O(x_t y_t^2, y_t^3) > 0 \quad (260)$$

thus yielding $x_{t+1} < x_t \leq x_0$. Moreover, the inductive hypothesis implies that $(x_t, y_t) \in U$; thus, denoting $W_t := W(x_t, y_t)$ and $W_{t+1} := W(x_{t+1}, y_{t+1})$, one sees that

$$W_{t+1} \leq W_t - \frac{1}{2}ay_t^2(bx_t^2 + cy_t^2) \leq W_0, \quad (261)$$

so that $y_{t+1} \leq \sqrt{2W_{t+1}/a} \leq \sqrt{2W_0/a}$. This completes the proof. $\square$

Finally we can prove convergence of the iterates of T toward zero.

Theorem D.8. *Set $\kappa := \sqrt{b/(c-a)}$. For all sufficiently small $x_0 \geq 0$ and $y_0 > 0$ and all $0 < \epsilon < \kappa$, there is*

$$\tau \equiv \tau(y_0, \epsilon) = O(y_0^{-2} \epsilon^{-3c/(c-a)}) \quad (262)$$

so that for all $t \geq \tau$, one has

$$\frac{x_\tau}{\sqrt{1 + 3a(\kappa + \epsilon)^2 x_\tau^2 (t - \tau)}} \leq x_t \leq \frac{x_\tau}{\sqrt{1 + 2a(\kappa - \epsilon)^2 x_\tau^2 (t - \tau)}} \quad (263)$$

and

$$(\kappa - \epsilon/2)x_t \leq y_t \leq (\kappa + \epsilon/2)x_t. \quad (264)$$

In particular, $x_t, y_t = \Theta(t^{-1/2})$.

Proof. Using Lemma D.7, we may take (x_0, y_0) sufficiently small that the quantity $\tilde{\Delta}_t := \tilde{\Delta}(x_t, y_t)$ of Lemma D.6 satisfies the recursion

$$|\tilde{\Delta}_{t+1}| \leq |\tilde{\Delta}_t| \left(1 - \frac{c-a}{2} y_t^2 \right) \quad (265)$$

for all $t \in \mathbb{N}$. By taking (x_0, y_0) yet smaller if necessary we may also assume that y_t satisfies the lower-bound

$$y_{t+1} \geq y_t(1 - 2cy_t^2). \quad (266)$$

Squaring and taking the reciprocal yields

$$\frac{1}{y_{t+1}^2} \leq \frac{1}{y_t^2} \frac{1}{1 - 2cy_t^2}, \quad (267)$$

and, taking (x_0, y_0) yet smaller if necessary so that $1/(1 - 2cy_t^2) \leq 1 + 6y_t^2$ for all t , one obtains the recursive estimate

$$\frac{1}{y_{t+1}^2} - \frac{1}{y_t^2} \leq 6c, \quad (268)$$

from which one obtains the bound

$$y_t^2 \geq \frac{y_0^2}{1 + 6cy_0^2 t} \quad (269)$$

for all $t \in \mathbb{N}$ by telescoping. Plugging this into the recursion for $\tilde{\Delta}$ and taking logarithms, one has

$$\ln |\tilde{\Delta}_{t+1}| - \ln |\tilde{\Delta}_t| \leq -\frac{c-a}{2} \frac{y_0^2}{1 + 6cy_0^2 t}. \quad (270)$$

Taking a telescoping sum yields

$$\ln |\tilde{\Delta}_t| - \ln |\tilde{\Delta}_0| \leq -\frac{c-a}{12c} \sum_{s=0}^{t-1} \frac{1}{\frac{1}{6cy_0^2} + s} \leq -\frac{c-a}{12c} \ln(1 + 6cy_0^2 t) \quad (271)$$

by Lemma E.4. Thus

$$|\tilde{\Delta}_t| \leq |\tilde{\Delta}_0| (1 + 6cy_0^2 t)^{-(c-a)/(12c)}, \quad (272)$$

and for any $0 < \epsilon < \sqrt{b/(c-a)}$ one sees that for any t greater than or equal to

$$\tau(y_0, \epsilon) := \left\lceil \frac{1}{6cy_0^2} \left(\left(\frac{\epsilon}{4|\tilde{\Delta}_0|} \right)^{-12c/(c-a)} - 1 \right) \right\rceil \quad (273)$$

one has $|\tilde{\Delta}_t| < \epsilon/4$. It follows from the definition of $\tilde{\Delta}(x, y) := (y - \kappa x + O(x^2))/x$ that $t \geq \tau(y_0, \epsilon)$ implies

$$y_t \in [\kappa x - (\epsilon/4)x + O(x^2), \kappa x + (\epsilon/4)x + O(x^2)] \subset [(\kappa - \epsilon/2)x, (\kappa + \epsilon/2)x], \quad (274)$$

where, shrinking x_0, y_0 if necessary, the final containment is obtained using Lemma D.7. With $t \geq \tau(y_0, \epsilon)$ and the bounds $(\kappa + \epsilon/2)x_t \geq y_t \geq (\kappa - \epsilon/2)x_t$ in hand, one obtains

$$(1 - a(\kappa + \epsilon)^2 x_t^2)x_t \leq x_{t+1} \leq (1 - a(\kappa - \epsilon)^2 x_t^2)x_t, \quad (275)$$

again shrinking x_0, y_0 and invoking Lemma D.7 if necessary to deal with the higher-order terms. Inverting and squaring yield the bounds

$$\frac{1}{x_{t+1}^2} \leq \frac{1}{x_t^2} \frac{1}{(1 - a(\kappa + \epsilon)^2 x_t^2)^2} \leq \frac{1}{x_t^2} (1 + 3a(\kappa + \epsilon)^2 x_t^2) \quad (276)$$

and

$$\frac{1}{x_{t+1}^2} \geq \frac{1}{x_t^2} \frac{1}{(1 - a(\kappa - \epsilon)^2 x_t^2)^2} \geq \frac{1}{x_t^2} (1 + 2a(\kappa - \epsilon)^2 x_t^2), \quad (277)$$

again shrinking x_0, y_0 and invoking Lemma D.7 if necessary to obtain the final upper bound in (276), so that

$$3a(\kappa + \epsilon)^2 \geq \frac{1}{x_{t+1}^2} - \frac{1}{x_t^2} \geq 2a(\kappa - \epsilon)^2, \quad (278)$$

from which telescoping summation yields the bounds (263). The corresponding bounds for y_t now follow from the limits of (274). This completes the proof. $\square$

Finally, we may prove our convergence theorem in the critical regime using Theorem D.8.

Theorem D.9. *Assume that $\eta = 2/\lambda(\theta_*^\parallel)$. Assume in addition that $\nu/(c(2/\lambda(\theta_*^\parallel), \theta_*^\parallel)\lambda(\theta_*^\parallel)) < 1$, where ν is defined in Assumption 3.4 and c is defined in Assumption 3.3. Then for all $\theta_0^\parallel$ sufficiently close to $\theta_*^\parallel$ and all $\theta^\perp \neq 0$ sufficiently small, one has $d_M(\theta_t^\parallel, \theta_*^\parallel) = \Theta(t^{-1/2})$ and $|\theta_t^\perp| = \Theta(t^{-1/2})$.*

Proof. Denote $\lambda_* := \lambda(\theta_*^\parallel)$ and $c_* := c(2/\lambda_*, \theta_*^\parallel)$ for notational convenience. By Theorem 4.1, for all $(\theta^\parallel, \theta^\perp)$ sufficiently close to $(\theta_*^\parallel, 0)$, GD has the coordinate expression

$$\begin{pmatrix} \text{GD}^\parallel(\eta, \theta^\parallel, \theta^\perp) \\ \text{GD}^\perp(\eta, \theta^\parallel, \theta^\perp) \end{pmatrix} = \begin{pmatrix} \exp_{\theta^\parallel} \left(-\frac{(\theta^\perp)^2}{c(2/\lambda(\theta^\parallel), \theta_*^\parallel)\lambda_*} \nabla_M \lambda(\theta^\parallel) \right) + O((\theta^\perp)^3 d_M(\theta^\parallel, \theta_*^\parallel)) \\ \left(1 - \frac{2\lambda(\theta^\parallel)}{\lambda_*}\right) \theta^\perp + (\theta^\perp)^3 + O((\theta^\perp)^4) \end{pmatrix}. \quad (279)$$

We first derive a formula for $d_M(\text{GD}^\parallel(\eta, \theta^\parallel, \theta^\perp), \theta_*^\parallel) = \|\log_{\theta_*^\parallel}(\text{GD}^\parallel(\eta, \theta^\parallel, \theta^\perp))\|$. Using a covariant Taylor expansion (29), observe that:

$$\nabla_M \lambda(\theta^\parallel) = \Pi_{\theta_*^\parallel \rightarrow \theta^\parallel} (\nabla_M \lambda(\theta_*^\parallel) + \nabla_M^2 \lambda(\theta_*^\parallel) [\log_{\theta_*^\parallel}(\theta^\parallel)] + O(d_M(\theta^\parallel, \theta_*^\parallel)^2)) \quad (280)$$

$$= \Pi_{\theta_*^\parallel \rightarrow \theta^\parallel} (\nu \log_{\theta_*^\parallel}(\theta^\parallel) + O(d_M(\theta^\parallel, \theta_*^\parallel)^2)). \quad (281)$$

by Assumption 3.4, where $\Pi_{\theta_*^\parallel \rightarrow \theta^\parallel}$ is parallel transport from $\theta_*^\parallel$ to $\theta^\parallel$. Similarly, observe that

$$\frac{(\theta^\perp)^2}{c(2/\lambda(\theta^\parallel), \lambda_*)} = \frac{(\theta^\perp)^2}{c_* \lambda_*} + O((\theta^\perp)^2 d_M(\theta^\parallel, \theta_*^\parallel)). \quad (282)$$

We may thus write

$$\log_{\theta_*^\parallel}(\text{GD}^\parallel(\eta, \theta^\parallel, \theta^\perp)) = \log_{\theta_*^\parallel} \left(\exp_{\theta^\parallel} \left(-\frac{\nu (\theta^\perp)^2}{c_* \lambda_*} \Pi_{\theta_*^\parallel \rightarrow \theta^\parallel} (\log_{\theta_*^\parallel}(\theta^\parallel)) \right) \right) \quad (283)$$

$$+ O((\theta^\perp)^3 d_M(\theta^\parallel, \theta_*^\parallel), (\theta^\perp)^2 d_M(\theta^\parallel, \theta_*^\parallel)^2) \quad (284)$$

$$= \left(1 - \frac{\nu}{c_* \lambda_*} (\theta^\perp)^2\right) \log_{\theta_*^\parallel}(\theta^\parallel) \quad (285)$$

$$+ O((\theta^\perp)^3 d_M(\theta^\parallel, \theta_*^\parallel), (\theta^\perp)^2 d_M(\theta^\parallel, \theta_*^\parallel)^2). \quad (286)$$

Hence:

$$d_M(\text{GD}^\perp(\eta, \theta^\parallel, \theta^\perp), \theta_*^\parallel) = \|\log_{\theta_*^\parallel}(\text{GD}^\perp(\eta, \theta^\parallel, \theta^\perp))\| \quad (287)$$

$$= \left(1 - \frac{\nu}{c_* \lambda_*} (\theta^\perp)^2 + O((\theta^\perp)^3 + (\theta^\perp)^2 d_M(\theta^\parallel, \theta_*^\parallel))\right) d_M(\theta^\parallel, \theta_*^\parallel). \quad (288)$$

We now derive a formula for $|\text{GD}^\perp(\eta, \theta^\parallel, \theta^\perp)|$. Perform a covariant Taylor expansion (29) on $\lambda(\theta^\parallel)$ to obtain

$$\lambda(\theta^\parallel) = \lambda_* + \langle \nabla_M \lambda(\theta_*^\parallel), \log_{\theta_*^\parallel}(\theta^\parallel) \rangle + \frac{1}{2} \nabla_M^2 \lambda(\theta_*^\parallel) [\log_{\theta_*^\parallel}(\theta^\parallel)^{\otimes 2}] + O(d_M(\theta^\parallel, \theta_*^\parallel)^3) \quad (289)$$

$$= \lambda_* + \frac{1}{2} \nu d_M(\theta^\parallel, \theta_*^\parallel)^2 + O(d_M(\theta^\parallel, \theta_*^\parallel)^3) \quad (290)$$

by Assumption 3.4. Then one sees that

$$\text{GD}^\perp(\eta, \theta^\parallel, \theta^\perp) = -\left(1 + \frac{\nu}{\lambda_*} d_M(\theta^\parallel, \theta_*^\parallel)^2 + O(d_M(\theta^\parallel, \theta_*^\parallel)^3)\right) \theta^\perp + (\theta^\perp)^3 + O((\theta^\perp)^4) \quad (291)$$

$$= -\left(1 + \frac{\nu}{\lambda_*} d_M(\theta^\parallel, \theta_*^\parallel)^2 - (\theta^\perp)^2 + O(d_M(\theta^\parallel, \theta_*^\parallel)^3, (\theta^\perp)^3)\right) \theta^\perp. \quad (292)$$

so that

$$|\text{GD}^\perp(\eta, \theta^\parallel, \theta^\perp)| = \left(1 + \frac{\nu}{\lambda_*} d_M(\theta^\parallel, \theta_*^\parallel)^2 - (\theta^\perp)^2 + O(d_M(\theta^\parallel, \theta_*^\parallel)^3, (\theta^\perp)^3)\right) |\theta^\perp| \quad (293)$$

provided $d_M(\theta^\parallel, \theta_*^\parallel)$ and $|\theta^\perp|$ are sufficiently small. The result now follows by invoking Theorem D.8. $\square$

D.3 Supercritical regime

Our convergence theorem in the supercritical regime is the easiest of the three. The orbit to which the iterates are attracted is hyperbolic unlike in the critical case, and the transient phenomena are relatively simple compared with those of the subcritical case.

Theorem D.10. *Assume that $\eta > 2/\lambda(\theta_*^\parallel)$ is sufficiently small. Then there are positive constants C_1, C_2, C_3 and a stable orbit of period two of the form $(\theta_*^\parallel, \pm(\eta\lambda(\theta_*^\parallel) - 2)^{1/2} + O(\eta\lambda(\theta_*^\parallel) - 2))$ to which the iterates $(\theta_t^\parallel, \theta_t^\perp)$ starting from any $(\theta_0^\parallel, \theta_0^\perp)$ satisfying $0 < |\theta_0^\perp| \leq C_1(\eta\lambda(\theta_*^\parallel) - 2)^{1/2}$ and $d_M(\theta_0^\parallel, \theta_*^\parallel) \leq C_2(\eta\lambda(\theta_*^\parallel) - 2)^{1/2}$ converge with rate $O((1 - C_3(\eta\lambda(\theta_*^\parallel) - 2))^t)$.*

Proof. Arguing using covariant Taylor expansions as in the proof of Theorem D.9, one sees that

$$d_M(\text{GD}^\perp(\eta, \theta^\parallel, \theta^\perp), \theta_*^\parallel) = \left(1 - \frac{\eta\nu}{2c(\eta, \theta_*^\parallel)} (\theta^\perp)^2 + O((\theta^\perp)^3, (\theta^\perp)^2 d_M(\theta^\parallel, \theta_*^\parallel))\right) d_M(\theta^\parallel, \theta_*^\parallel) \quad (294)$$

and

$$\text{GD}^\perp(\eta, \theta^\parallel, \theta^\perp) = -\left(1 + (\eta\lambda(\theta_*^\parallel) - 2) + \frac{\eta\nu}{2} d_M(\theta^\parallel, \theta_*^\parallel)^2 - (\theta^\perp)^2 + O(d_M(\theta^\parallel, \theta_*^\parallel)^3, (\theta^\perp)^3)\right) \theta^\perp. \quad (295)$$

Consequently, the result follows from Theorem D.11 below. $\square$

Theorem D.11. *Given $\alpha > 0$ and constants $a > 0$ and $b > 0$, consider the functions*

$$F_x(x, y) := (1 - ay^2 + O(y^3, xy^2))x, \quad F_y(x, y) := -(1 + \alpha + bx^2 - y^2 + O(x^3, y^3))y. \quad (296)$$

Then, there are constants $C_1, C_2, C_3 > 0$ such that for all α sufficiently small, the map $F = (F_x, F_y)$ admits a stable, period-2 orbit $\{(0, \xi_+), (0, \xi_-)\}$ with $\xi_\pm = \pm\sqrt{\alpha} + O(\alpha)$, and the iterates $(x_{t+1}, y_{t+1}) := F(x_t, y_t)$ starting from any (x_0, y_0) with $0 < |y_0| \leq C_1\sqrt{\alpha}$ and $|x_0| \leq C_2\sqrt{\alpha}$ converge to this orbit at a rate of $O((1 - C_3\alpha)^t)$.

Proof. The line $x = 0$ is preserved by F , and thereon one sees that F_y takes the form

$$F_y(0, \cdot) : y \mapsto -(1 + \alpha)y + y^3 + O(y^4) \quad (297)$$

which, by Lemma D.2, admits a period-two orbit $\{\xi_+, \xi_-\}$ of the form $\xi_\pm = \pm\sqrt{\alpha} + O(\alpha)$. We will show that the iterates of the square F^2 of F converge to $(0, \xi_\pm)$ for all initial points (x_0, y_0) sufficiently small, $y_0 \neq 0$. This will be achieved by showing that $\|DF^2(x, y)\| < 1$ uniformly over a neighbourhood of $(0, \xi_\pm)$, followed by proving a guarantee of convergence to this neighbourhood in finite time.

Via a routine calculation one sees that

$$F^2(x, y) = \begin{pmatrix} x(1 - a(1 + (1 + \alpha)^2)y^2 + O(xy^2, y^3)) \\ y((1 + \alpha)^2 + 2(1 + \alpha)bx^2 - (1 + \alpha)(1 + (1 + \alpha)^2)y^2 + O(x^3, x^2y, xy^2, y^3)) \end{pmatrix}. \quad (298)$$

The derivative of F^2 is then given by

$$DF^2(x, y) = \begin{pmatrix} 1 - a(1 + (1 + \alpha)^2)y^2 & -2a(1 + (1 + \alpha)^2)xy \\ 4(1 + \alpha)bx^2 & (1 + \alpha)^2 + 2(1 + \alpha)bx^2 - 3(1 + \alpha)(1 + (1 + \alpha)^2)y^2 \end{pmatrix} + O(x^3, x^2y, xy^2, y^3). \quad (299)$$

$$(300)$$

In particular, for all y satisfying

$$\frac{\alpha}{(1 + \alpha)(1 + (1 + \alpha)^2)} \leq y^2 \leq 2\alpha, \quad (301)$$

the matrix $DF^2(x, y)$ has entries with magnitudes upper-bounded by

$$\begin{pmatrix} 1 - \frac{a\alpha}{1 + \alpha} & O(x\sqrt{\alpha}) \\ O(x\sqrt{\alpha}) & 1 - (1/2)\alpha + O(x^2) \end{pmatrix} + O(x^3, \sqrt{\alpha}x^2, \alpha x, \alpha^{3/2}) \quad (302)$$

for all α sufficiently small. Using the bound $\|A\|_2 \leq \sqrt{\|A\|_1 \|A\|_\infty}$ for a matrix A , where $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are the maximum column 1-norm and maximum row 1-norm respectively, one sees then that there are $C_2, C_3 > 0$ such that for all α sufficiently small, all x satisfying

$$|x| \leq C_2\sqrt{\alpha}. \quad (303)$$

and all y satisfying (301), one has $\|DF^2(x, y)\|_2 \leq 1 - C_3\alpha$.

Letting V be the neighbourhood of the two-point set $\{(0, \xi_\pm)\}$ enclosed by the bounds (301) and (303), one sees that the iterates (x_{2t}, y_{2t}) of F^2 starting at any point $(x_0, y_0) \in V$ converge toward $(0, \xi_{\text{sign}(y_0)})$ with rate $\|(x_{2t}, y_{2t} - \xi_\pm)\| \leq (1 - C_3\alpha)^t \|(x_0, y_0 - \xi_{\text{sign}(y_0)})\|$.

We complete the proof by showing that the iterates starting from any x_0 satisfying (303) and any y_0 satisfying

$$0 < y_0^2 \leq \frac{\alpha}{(1 + \alpha)(1 + (1 + \alpha)^2)} \quad (304)$$

are eventually drawn into V . This, however, is easy: for any x satisfying (303) and any y satisfying $y^2 \leq \alpha/((1 + \alpha)(1 + (1 + \alpha)^2))$, one has

$$|(F^2)_y(x, y)| \geq |y|((1 + \alpha)^2 - (1 + \alpha)(1 + (1 + \alpha)^2)y^2 + O(\alpha^{3/2})) \quad (305)$$

$$\geq |y|(1 + (1/2)\alpha) \quad (306)$$

for all α sufficiently small. It follows that the iterates $y_{2t} = (F^{2t})_y(x_0, y_0)$ satisfy the recursion $|y_{2(t+1)}| \geq (1 + (1/2)\alpha)|y_{2t}|$ as long as $y_{2t}^2 \leq \alpha/((1 + \alpha)(1 + (1 + \alpha)^2))$. We claim now that there exists $\tau \in \mathbb{N}$ satisfying

$$\tau \leq \frac{1}{2} \left\lceil \ln \left(\frac{\alpha}{y_0^2(1 + \alpha)(1 + (1 + \alpha)^2)} \right) \ln \left(1 + \frac{1}{2}\alpha \right)^{-1} \right\rceil \quad (307)$$

such that $y_{2\tau} \geq \alpha/((1 + \alpha)(1 + (1 + \alpha)^2))$. Suppose for a contradiction that this were not the case; then

$$\frac{\alpha}{(1 + \alpha)(1 + (1 + \alpha)^2)} > y_0^2(1 + (1/2)\alpha)^{2\tau} \geq \frac{\alpha}{(1 + \alpha)(1 + (1 + \alpha)^2)}, \quad (308)$$

which is a contradiction. This proves the result. $\square$

E Elementary lemmas

In this section, for ease of reference, we collect several well-known facts and their elementary proofs. Recall that a subset M' of a Riemannian manifold M is *geodesically convex* if any two points of M' can be connected by a unique geodesic contained entirely in M' .

Lemma E.1. *Let $\lambda : M' \rightarrow \mathbb{R}$ be a C^2 function on a geodesically convex subset M' of a Riemannian manifold M . If λ is μ -geodesically strongly convex in the sense that $\nabla_M^2 \lambda \succeq \mu I_{TM}$ at all points on M' , then*

$$\lambda(y) \geq \lambda(x) + \langle \nabla_M \lambda(x), \log_x(y) \rangle + \frac{\mu}{2} d_M(x, y)^2 \quad (309)$$

for all $x, y \in M'$.

Proof. Fix $x, y \in M'$. Since M' is geodesically convex, there exists a unique geodesic $\gamma : [0, 1] \rightarrow M'$ such that $\gamma(0) = x$ and $\gamma(1) = y$. This γ satisfies

$$\dot{\gamma}(t) = \Pi(\gamma)_t \dot{\gamma}(0) = \Pi(\gamma)_t \log_x(y), \quad \|\dot{\gamma}(t)\| = d_M(x, y) \quad (310)$$

for all $t \in [0, 1]$, where $\Pi(\gamma)_t : T_{\gamma(0)}M \rightarrow T_{\gamma(t)}M$ is parallel transport. Define $\varphi(t) := \lambda(\gamma(t))$. Then using (23) and the chain rule,

$$\dot{\varphi}(t) = \langle \nabla_M \lambda(\gamma(t)), \dot{\gamma}(t) \rangle, \quad \ddot{\varphi}(t) = \langle \dot{\gamma}(t), \nabla_M^2 \lambda(\gamma(t)) \dot{\gamma}(t) \rangle \quad (311)$$

for all $t \in [0, 1]$. Thus, by Taylor's theorem, there is $s \in [0, 1]$ such that

$$\lambda(y) = \varphi(1) = \varphi(0) + \dot{\varphi}(0) + \frac{1}{2} \ddot{\varphi}(s) \quad (312)$$

$$= \lambda(x) + \langle \nabla_M \lambda(x), \log_x(y) \rangle + \frac{1}{2} \langle \dot{\gamma}(s), \nabla_M^2 \lambda(\gamma(s)) \dot{\gamma}(s) \rangle \quad (313)$$

$$\geq \lambda(x) + \langle \nabla_M \lambda(x), \log_x(y) \rangle + \frac{\mu}{2} \|\dot{\gamma}(s)\|^2 \quad (314)$$

$$= \lambda(x) + \langle \nabla_M \lambda(x), \log_x(y) \rangle + \frac{\mu}{2} d_M(x, y)^2 \quad (315)$$

as claimed. $\square$

Lemma E.2. *Let $\lambda : M' \rightarrow \mathbb{R}$ be a C^2 function on a geodesically convex subset M' of a Riemannian manifold M . If λ is μ -geodesically strongly convex in the sense that $\nabla_M^2 \lambda \succeq \mu I_{TM}$ at all points on M' , then*

$$\langle \Pi_{x \rightarrow y} \nabla_M \lambda(x) - \nabla_M \lambda(y), \log_y(x) \rangle \geq \mu d_M(x, y)^2 \quad (316)$$

for all $x, y \in M'$, where $\Pi_{x \rightarrow y} : T_x M \rightarrow T_y M$ is parallel transport.

Proof. Fix $x, y \in M'$. From Lemma E.1, one has the estimates

$$\lambda(y) \geq \lambda(x) + \langle \nabla_M \lambda(x), \log_x(y) \rangle + \frac{\mu}{2} d_M(x, y)^2, \quad (317)$$

$$\lambda(x) \geq \lambda(y) + \langle \nabla_M \lambda(y), \log_y(x) \rangle + \frac{\mu}{2} d_M(x, y)^2. \quad (318)$$

Since parallel transport is an orthogonal transformation, one has

$$\langle \nabla_M \lambda(x), \log_x(y) \rangle = -\langle \nabla_M \lambda(x), \Pi_{y \rightarrow x} \log_y(x) \rangle = -\langle \Pi_{x \rightarrow y} \nabla_M \lambda(x), \log_y(x) \rangle. \quad (319)$$

Substituting this into (317) and rearranging yields

$$\langle \Pi_{x \rightarrow y} \nabla_M \lambda(x), \log_y(x) \rangle \geq \frac{\mu}{2} d_M(x, y)^2 + \lambda(x) - \lambda(y) \quad (320)$$

while (318) may be rearranged to give

$$-\langle \nabla_M \lambda(y), \log_y(x) \rangle \geq \frac{\mu}{2} d_M(x, y)^2 + \lambda(y) - \lambda(x). \quad (321)$$

Adding these estimates then yields the result. $\square$

Lemma E.3. Let $\lambda : M' \rightarrow \mathbb{R}$ be a C^2 function on a geodesically convex subset M' of a Riemannian manifold M . If λ is μ -geodesically strongly convex in the sense that $\nabla_M^2 \lambda \succeq \mu I_{TM}$ at all points on M' and admits a critical point x_* in M' , then x_* is the unique global minimum of λ in M' and

$$\|\nabla_M \lambda(x)\|^2 \geq 2\mu(\lambda(x) - \lambda(x_*)) \quad (322)$$

for all $x \in M'$.

Proof. That the global minimum x_* is unique follows from Lemma E.1: for any $x \neq x_*$ in M , one has

$$\lambda(x) \geq \lambda(x_*) + \frac{\mu}{2} d_M(x, x_*)^2 > \lambda(x_*), \quad (323)$$

thus proving the first claim. To prove the second claim, fix $x \in M'$. Lemma E.1 then applies to yield

$$\lambda(x) - \lambda(x_*) \leq -\langle \nabla_M \lambda(x), \log_x(x_*) \rangle - \frac{\mu}{2} \|\log_x(x_*)\|^2 \quad (324)$$

$$= -\frac{\mu}{2} \left\| \frac{1}{\mu} \nabla_M \lambda(x) + \log_x(x_*) \right\|^2 + \frac{1}{2\mu} \|\nabla_M \lambda(x)\|^2 \quad (325)$$

$$\leq \frac{1}{2\mu} \|\nabla_M \lambda(x)\|^2, \quad (326)$$

from which the result follows. $\square$

Lemma E.4. For any $c > 0$, and any integers $0 \leq s < t$, one has

$$\ln \left(\frac{c+t+1}{c+s} \right) \leq \sum_{k=s}^t \frac{1}{c+k} \leq \frac{1}{c+s} + \ln \left(\frac{c+t}{c+s} \right). \quad (327)$$

Proof. Since $x \mapsto \frac{1}{(c+x)}$ is monotonically decreasing along the positive reals, one has

$$\int_n^{n+1} f(x) dx \leq \int_n^{n+1} f(n) dx = f(n) = \int_{n-1}^n f(n) dx \leq \int_{n-1}^n f(x) dx \quad (328)$$

for any natural number n , from which it follows that

$$\int_s^{t+1} \frac{1}{(c+x)} dx \leq \sum_{k=s}^t \frac{1}{(c+k)} = \frac{1}{(c+s)} + \sum_{k=s+1}^t \frac{1}{(c+k)} \quad (329)$$

$$\leq \frac{1}{(c+s)} + \int_s^t \frac{1}{(c+x)} dx. \quad (330)$$

Integrating both sides now yields the result. $\square$

F Experimental supplement

Our experiments were conducted for a depth 5, width 1 scalar factorisation problem, corresponding to the function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ defined by

$$f(\theta_1, \dots, \theta_p) := \theta_p \cdots \theta_1, \quad (\theta_1, \dots, \theta_p) \in \mathbb{R}^p, \quad (331)$$

with target value $y = 1$. Our code³ runs gradient descent on this problem, and can be executed in seconds on a single cpu. At each iterate, the projection onto the solution manifold $M := f^{-1}\{y\}$ is computed as follows.

The KKT conditions for the constrained optimisation problem

$$\min_{\theta \geq 0} \frac{1}{2} \|\theta\| - \theta\|^2 \quad \text{such that} \quad \prod_{i=1}^p \theta_i = y \quad (332)$$

³Available at <https://github.com/lemacdonald/eos-convergence-rates-codimension-1>

are given by

$$\theta_i^{\parallel} - \theta_i^{\parallel} \theta_i + \alpha \frac{y}{\theta_i^{\parallel}} = 0, \quad i = 1, \dots, p, \quad (333)$$

where α is the constraint parameter. These equations admit the quadratic solutions

$$\theta_i^{\parallel}(\alpha) = \frac{\theta_i \pm \sqrt{\theta_i^2 - 4\alpha y}}{2}, \quad i = 1, \dots, p. \quad (334)$$

It is easily verified $\alpha \mapsto \theta_i^{\parallel}(\alpha)$, is monotonically decreasing (when (334) is taken with the $+$ sign) or monotonically increasing (when (334) is taken with the $-$ sign); consequently, the map $\phi : \alpha \mapsto \prod_{i=1}^p \theta_i^{\parallel}(\alpha)$ is also monotonic provided the signs are consistent among the $\theta_i^{\parallel}(\alpha)$.

The sign conventions we use are as follows. If $\prod_{i=1}^p \theta_i < y$, then we must take the $+$ sign for all $\theta_i^{\parallel}(\alpha)$. If $\prod_{i=1}^p \theta_i \geq y$, then, noting that $2^{-p} \prod_{i=1}^p \theta_i$ is the minimum value possible for $\prod_{i=1}^p \theta_i^{\parallel}(\alpha)$ when all $\theta_i^{\parallel}(\alpha)$ are taken with the positive sign in (334), either:

1. $y \geq 2^{-p} \prod_{i=1}^p \theta_i$, in which case we take the positive sign in (334) for all i .
2. $y < 2^{-p} \prod_{i=1}^p \theta_i$, in which case we take the negative sign in (334) for all i .

Having determined the sign to use uniformly over all $i = 1, \dots, p$ in (334), we are assured that $\alpha \mapsto \phi(\alpha) = \prod_{i=1}^p \theta_i^{\parallel}(\alpha)$ is monotonic, hence that $\phi(\alpha) = y$ admits a unique solution which we obtain using the bisection method.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: See the "Paper contributions" section in the Introduction, which outlines our contributions in the context of codimension 1 least squares problems.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: The limitations are discussed in detail in the Discussion/Conclusion section under several bold headings.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: Complete proofs of all theoretical results are provided in several appendices with appropriate referencing.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We included the code to produce our plots with the submission.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide open access to our code; see experimental supplement.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All relevant hyperparameters etc are contained in the code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Our experiments are deterministic.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: See the “Experimental supplement” appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn’t make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: We have reviewed and conformed with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: Our work is primarily theoretical and concerned with understanding in simple settings what practitioners have already been doing at scale for over 10 years. No foreseeable societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We considered only toy models in our experiments.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: No existing assets used.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: No new assets introduced.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No crowdsourcing or research with human subjects.

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs were not used outside of creation of code for toy experiments and minimal assistance with proving theorems. All material was thoroughly checked and revised if necessary.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.