A RIEMANNIAN PERSPECTIVE ON GENERALIZED SMOOTHNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

In this work we develop Riemannian optimization formulations for optimization problems posed on Euclidean spaces, and that may not satisfy classical smoothness criteria. We introduce a family of metrics that can confirm global Riemannian smoothness in function classes including certain neural-network training problems, enabling convergence guarantees for Riemannian optimization schemes applied to these objectives. The metrics are conformal scalings and warped products defined using radial scaling functions, and we further derive low-dimensional geodesic equations for these metrics, which can lead to efficient implementations. Finally, we consider connections with (L_0, L_1)-smoothness and ℓ-smoothness criteria.

1 INTRODUCTION

Traditionally, non-asymptotic analysis of first-order gradient methods has relied on smoothness properties, such as requiring that the loss function $f : \mathbb{R}^n \to \mathbb{R}$ has a Lipschitz continuous gradient. While this assumption facilitates analysis, it makes the results challenging to apply to many real world problems where Hessian growth defies the stringent assumptions such as globally Lipschitz gradients that smoothness assumptions make. This has motivated generalizations of smoothness criteria that are more general while still enabling analysis with explicit dependencies on problem data and iteration time. One such notion is (L_0, L_1)-smoothness (Zhang et al., 2020), which states that the Euclidean Hessian is bounded by a constant plus a multiple of the gradient norm. That is, a twice continuously differentiable f is (L_0, L_1)-smooth if there are constants L_0, L_1 such that

$$\| \operatorname{Hess} f(x)\|_2 \leq L_0 + L_1 \| \operatorname{grad} f(x)\|_2 \tag{1}$$

for all x. Note that in the case of $L_1 = 0$ this recovers the usual Euclidean smoothness condition (for functions that are twice continuously differentiable). A generalization of (L_0, L_1)-smoothness was introduced by (Li et al., 2023), who defined ℓ-smoothness in terms of a non-decreasing continuous function $\ell : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ as follows: A function $f : \mathbb{R}^n \to \mathbb{R}$ is ℓ-smooth if for all x,

$$\| \operatorname{Hess} f(x)\| \leq \ell(\| \operatorname{grad} f(x)\|), \tag{2}$$

where (L_0, L_1) smoothness is recovered using $\ell(z) = L_0 + L_1 z$. However, these generalized smoothness conditions are arguably too strong to be applicable to even elementary neural network training problems. This has been noted previously in the context of (L_0, L_1) smoothness (Cooper, 2024), and we provide an example showing that ℓ-smoothness is also too strong for relatively simple MLP training scenarios (Proposition 5.2).

We believe the Riemannian perspective may be useful here: can we define a metric on the parameter space that can reflect the nonlinear, hierarchical structure of a given problem, so that from the Riemannian perspective the function becomes well-behaved, confirming the Riemannian analogue of gradient smoothness? If so, a variety of Riemannian first-order gradient algorithms can then be applied to the problem with corresponding convergence guarantees. In this paper we investigate this question, proposing several Riemannian metrics for function classes that defy Euclidean smoothness. We prove convergence of the schemes, primarily by verifying smoothness, and demonstrate their empirical performance on several statistical and machine learning problems. The Riemannian first-order gradient methods we focus on are SGD and Adagrad-Norm.

Contributions. Our main contributions can be summarized as follows.

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

- We give growth conditions on derivatives that guarantee smoothness in warped and conformal metrics, and show that these conditions are satisfied for a class of neural networks, yielding convergence guarantees for first-order Riemannian stochastic optimization in these scenarios (Theorem 5.3).
- We provide recipes for computing the geodesics of these metrics by solving a low dimensional system of ODEs, independent of model parameter dimension (Proposition C.4).

In addition to these contributions, Theorem 4.5 establishes that (L_0, L_1)-smoothness is equivalent to Riemannian smoothness under a conformal scaling of the Euclidean metric. We show that there are ordinary MLP training problems that are not ℓ smooth for any function ℓ, in Proposition 5.2. Finally, we establish a convergence guarantee for Riemannian Adagrad-Norm in the case of non-convex stochastic optimization on a manifold (Theorem 5.6).

2 Background

We recall notation used for Riemannian manifolds, and refer to the appendix for more detailed definitions. Throughout, M denotes a smooth manifold (Lee, 2018). The tangent space at $x \in M$ is denoted by T_xM. A Riemannian metric on M is a smooth assignment of an inner product $\langle \cdot, \cdot \rangle_x$ to each tangent space T_xM, alternatively denoted $g_x(\cdot, \cdot)$. It induces a norm $\|v\|_x^2 = \langle v, v \rangle_x$ on each tangent space. The Euclidean metric is denoted g^{Euc}. For an interval $J \subseteq \mathbb{R}$, a geodesic is a curve $\gamma : J \to M$ that is locally length minimizing according to the metric. The exponential map Exp maps tangent vectors $v \in T_xM$ to points on M via $\mathrm{Exp}(v) = \gamma(1)$ where γ is a geodesic with $\gamma(0) = x$ and $\gamma'(0) = v$. Given a differentiable function $f : M \to \mathbb{R}$, the Riemannian gradient, denoted $\mathrm{grad}^g f$, is the unique vector field such that $\langle \mathrm{grad}^g f(x), v \rangle_x = df(v)$ for all $(x, v) \in TM$. A retraction is a smooth map $\mathrm{Retr} : TM \to M$ such that $\mathrm{Retr}_x(0) = x$ for all $x \in M$, and $\frac{d}{dt}\big|_{t=0} \mathrm{Retr}_x(tv) = v$ for all $v \in T_xM$. The exponential map is an example of a retraction. A function $f : M \to \mathbb{R}$ is L-smooth with respect to the retraction Retr if for all $(x, v) \in TM$,

$$f(\mathrm{Retr}_x(v)) \leq f(x) + \langle \mathrm{grad}^g f(x), v \rangle_x + \frac{L}{2}\|v\|_x^2 \tag{3}$$

The criterion (3) is standard in Riemannian optimization, and facilitates non-asymptotic analyses similarly to smoothness criteria in the Euclidean setting (Boumal et al., 2018; Hosseini & Sra, 2020a). For a metric g, one may also define the Riemannian Hessian $\mathrm{Hess}^g f$ as the covariant derivative of the gradient: At $x \in M$, $\mathrm{Hess}^g f(x)[u, v] = \langle \nabla_u \mathrm{grad}^g f(x), v \rangle_x$ for $u, v \in T_xM$. A convenient property of the exponential map is that if a function $f : M \to \mathbb{R}$ has an L-Lipschitz gradient, then (3) holds with $\mathrm{Retr} = \mathrm{Exp}$ (Cor. 10.54 in (Boumal, 2023)). Furthermore, if f is twice continuously differentiable, then it has an L-Lipschitz gradient iff $\|\mathrm{Hess}^g f(x)\|_x \leq L$ for all $x \in M$ (Cor. 10.47 in (Boumal, 2023)). For ease of notation, $\mathrm{grad}\, f$ and $\mathrm{Hess}\, f$ without a superscript indicating the metric will always refer to the Euclidean gradient and Hessian respectively, unless otherwise noted.

We focus on two families of Riemannian metrics in particular. The first consists of conformal scalings, whereby an existing metric g is point-wise scaled by a smooth function of the form e^{2h}:

Definition 2.1. Let (M, g) be a Riemannian manifold and let $h : M \to \mathbb{R}$ be smooth. The **conformal scaling of g by h** is the metric $\widetilde{g}$ such that for $p \in M$ and $u, v \in T_pM$, $\widetilde{g}_p(u, v) = e^{2h(p)} g_p(u, v)$.

For example, if $(\mathbb{R}^n, g^{\mathrm{Euc}})$ is Euclidean space with its standard metric, then the metric $g_x(u, v) = (1 + \|x\|^2) u^T v$ is a conformal scaling of g^{Euc} corresponding to $h(x) = \frac{1}{2}\ln(1 + \|x\|^2)$. It follows from the definition of the gradient that the gradient $\mathrm{grad}^{\widetilde{g}}$ is related to g-gradient via (Petersen, 2016),

$$\mathrm{grad}^{\widetilde{g}} f(x) = e^{-2h(x)} \mathrm{grad}^g f(x). \tag{4}$$

The second class of metrics are warped products (O'Neill, 1983). These apply to products $M_1 \times M_2$, and involve a scaling term h applied to the second factor M_2, where the scaling depends only on M_1:

Definition 2.2. Let $(M_1, g^1), (M_2, g^2)$ be manifolds, and let $h : M_1 \to \mathbb{R}_{>0}$ be smooth. The **warped product metric** g on $M_1 \times M_2$ is defined as follows. For $(p, q) \in M_1 \times M_2$, and tangent vectors $u_1, v_1 \in T_pM_1$ and $u_2, v_2 \in T_qM_2$, $g_{(p,q)}((u_1, u_2), (v_1, v_2)) = g_p^1(u_1, v_1) + h(p)^2 g_q^2(u_2, v_2)$.

The notation $M_1 \times_h M_2$ is used to denote the warped product manifold determined by M_1, M_2 and the warping function h. As an example, if $M_1 = M_2 = \mathbb{R}^m$ and $h(p) = \sqrt{1 + \|p\|_2^2}$, then we can

define the warped product metric $g_{(p,q)}((u_1, u_2), (v_1, v_2)) = u_1^T v_1 + (1 + \|p\|_2^2) u_2^T v_2$. It follows from the definition that the warped product gradient is related to the gradients on M_1, M_2 via

$$\operatorname{grad}^g f(p,q) = \Big(\operatorname{grad}^{g_1}_{M_1} f(p,q), \frac{1}{h(p)^2} \operatorname{grad}^{g_2}_{M_2} f(p,q) \Big) \tag{5}$$

Keeping with convention, the definition of the conformal metric typically uses the exponential of the scaling function e^{2h} while the warped product simply uses the square h^2 (Petersen, 2016). Throughout, we assume all manifolds are geodesically complete, meaning their geodesics exist for all time $t \in \mathbb{R}$. Under our definitions, if M_1 and M_2 are geodesically complete, then so is the warped product O'Neill (1983). Furthermore, if (M, g) is geodesically complete, then a conformal scaling $e^{2h} g$ is as well as long as h is bounded from below by a constant $k \in \mathbb{R}$. For all of the scaling functions we construct below, this holds with $k = 0$.

3 Related work

Generalized and relative smoothness Motivated by extending convergence results for L-smooth functions to wider classes of functions, several generalizations of the classic smoothness condition have been proposed. Working in the class of twice continuously differentiable functions for simplicity, several of these generalizations can be framed as criteria on the Hessian that are more general than the classical $\|\operatorname{Hess} f\| \leq L$. One class provides norm-based criteria in terms of upper bounds on $\|\operatorname{Hess} f\|$. An early work in this direction was (L_0, L_1)-smoothness (Zhang et al., 2020). A particularly general condition is given by Li et al. (2023), who defines ℓ-smoothness. See also (Chen et al., 2023) for a related generalization of (L_0, L_1)-smoothness. Applications of generalized smoothness include analysis of adaptive gradient algorithms in the context of ℓ-smooth functions (Hong & Lin, 2024; Wang et al., 2023).

Another set of criteria falls under relative smoothness (Bauschke et al., 2017; Lu et al., 2018), formulated as follows. Let $h : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function. Then $f : \mathbb{R}^n \to \mathbb{R}$ is said to be L-smooth relative to h if $\operatorname{Hess} f(x) \preceq L \operatorname{Hess} h(x)$ for all x. The function h generally is required to be convex but we refer to (Ding et al., 2025)(Lu et al., 2018) for detailed regularity requirements. Investigations into relative smoothness began in convex settings (Bauschke et al., 2017),(Lu et al., 2018) and have been extended to non-convex cases (Bolte et al., 2018; Ding et al., 2025), as well to Riemannian submanifolds of Euclidean space. The possibility of accelerated convergence rates under relative smoothness was explored by Hanzely et al. (2021).

Optimization on manifolds Given a Riemannian manifold M and a continuously differentiable function $f : M \to \mathbb{R}$, consider the optimization problem

$$\min_{x \in M} f(x)$$

A variety of first-order optimization algorithms have been analyzed in the Riemannian setting, including SGD (Bonnabel, 2013; Hosseini & Sra, 2020b), adaptive gradient methods (Becigneul & Ganea, 2019), variance reduction techniques (Zhang et al., 2016), line search methods and others (Hosseini & Sra, 2020a; Boumal, 2023). A natural approach for solving this problem is Riemannian Stochastic Gradient Descent:

$$x_{k+1} = \operatorname{Retr}(-\eta_k v_k), \tag{6}$$

where the $v_k \in T_{x_k} M$ are gradient estimates and η_k is a sequence of step sizes. If f is L-smooth with respect to the retraction (3), and the gradients are unbiased , that is $\mathbb{E}[v_k - \operatorname{grad} f(x_k)] = 0$ with bounded variance $\mathbb{E}[\|v_k - \operatorname{grad} f(x_k)\|_{x_k}^2] \leq \sigma^2$ for all k, then Riemannian SGD with a step size on the order of $\min\{\frac{1}{L}, \frac{1}{\sigma\sqrt{T}}\}$ is known to converge at a rate of $\mathcal{O}(\frac{1}{\sqrt{T}})$ (Hosseini & Sra, 2020a;b).

As in the Euclidean setting, a potential limitation of Riemannian SGD with a fixed step size is that selecting the step size may require knowledge of constant L. Therefore, there has been interest in adaptive gradient methods for manifolds. An early work to explore such methods in the convex setting was (Becigneul & Ganea, 2019), while Kasai et al. (2019) developed a convergence analysis for Adagrad-type algorithms on matrix manifolds. Later, Sakai & Iiduka (2025) describe a general framework for Adaptive optimization algorithms on submanifolds of Euclidean space. Riemannian Adagrad has been analyzed in several settings, including the geodesically convex setting (Becigneul

& Ganea, 2019)(Park et al., 2026). In the non-convex setting, (de C. Bento et al., 2025) analyzed adagrad in the case of full (deterministic) gradients. Convergence analysis of Riemannian Adagrad in the stochastic, non-convex case has been explored in certain cases, including for matrix manifolds (Kasai et al., 2019), and (Sakai & Iiduka, 2025) who consider the setting of Euclidean submanifolds where the Riemannian metric is the restriction of the Euclidean metric. Furthermore, (He et al., 2025) develop Riemannian Bregman gradient methods for relatively smooth optimization problems, with an analysis for the stochastic setting focused on compact submanifolds of Euclidean space. Here, our focus is on constructing metrics for unconstrained problems so that the objective will satisfy Riemannian smoothness and other criteria, such variance conditions on stochastic gradients, needed for convergence analysis.

4 FUNCTION CLASSES

We introduce the function classes (c, h) and (w, h). Each class consists of twice continuously differentiable functions whose Hessian and gradients satisfy growth conditions determined by h and $\|\operatorname{grad} h\|$. As we will show, functions in the first class have bounded Hessian with respect to a conformally scaled Euclidean metric, while the second class is adapted to the product structure of the parameter space, and contains functions with bounded Hessian in a warped product metric.

Definition 4.1. A function $f : \mathbb{R}^m \to \mathbb{R}$ is of class (c, h) if there are constants K_1, K_2 such that for all $p \in \mathbb{R}^m$, $\|\operatorname{Hess} f(p)\| \leq K_2 e^{2h(p)}$ and, whenever $\operatorname{grad} h(p) \neq 0$, it holds that $\|\operatorname{grad} f(p)\| \leq K_1 e^{2h(p)} \|\operatorname{grad} h(p)\|^{-1}$.

As we will see in Theorem 4.3 this definition implies that f has a bounded Riemannian Hessian with respect to the corresponding conformally scaled metric. Likewise, (w, h) is the class of functions $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ whose partial derivatives satisfy the following growth conditions in relation to h:

Definition 4.2. A function $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ is of class (w, h) if there are constants $K_p, K_q, K_{pp}, K_{pq}, K_{qq}$ such that for all $(p, q) \in \mathbb{R}^m \times \mathbb{R}^n$, $\|\frac{\partial^2 f}{\partial p^2}(p, q)\| \leq K_{pp}$, $\|\frac{\partial^2 f}{\partial q^2}(p, q)\| \leq K_{qq} h(p)^2$, $\|\frac{\partial^2 f}{\partial p \partial q}(p, q)\| \leq K_{pq} h(p)$, and, whenever $\operatorname{grad} h(p) \neq 0$, it holds that $\|\frac{\partial f}{\partial p}(p, q)\| \leq K_p h(p) \|\operatorname{grad} h(p)\|^{-1}$, and $\|\frac{\partial f}{\partial q}(p, q)\| \leq K_q h(p)^2 \|\operatorname{grad} h(p)\|^{-1}$.

In the following theorem we establish that the functions of class (w, h) are smooth with respect to a warped product metric, while those in (c, h) are smooth in a conformally scaled Euclidean metric.

Theorem 4.3. *Let $h : \mathbb{R}^m \to \mathbb{R}$ be a smooth function. Then a twice continuously differentiable function $f : \mathbb{R}^m \to \mathbb{R}$ of class (c, h) has a bounded Hessian with respect to the metric*

$$g^{\mathrm{c}}_p(u, v) = e^{2h(p)} u^T v. \tag{7}$$

Specifically, $\|\operatorname{Hess}^{\mathrm{c}} f\|_{\mathrm{c}} \leq K_2 + 3K_1$. Likewise, for a smooth function $h : \mathbb{R}^m \to \mathbb{R}_{>0}$, a twice continuously differentiable function $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ of class (w, h) has a bounded Hessian with respect to the metric

$$g^{\mathrm{w}}_{(p,q)}((u_1, u_2), (v_1, v_2)) = u_1^T v_1 + h(p)^2 u_2^T v_2. \tag{8}$$

Namely, $\|\operatorname{Hess}^{\mathrm{w}} f\|_{\mathrm{w}} \leq \max\{K_{pp}, K_{qq} + K_p\} + K_{pq} + K_q$.

We note that twice continuously differentiable functions $f : \mathbb{R}^n \to \mathbb{R}$ that have a Hessian norm which grows at most polynomially in their parameter are in the class (c, h) for a suitable choice of h:

Proposition 4.4. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function, and let there be a $C \geq 0$ and an even integer $k \geq 2$ such that $\|\operatorname{Hess} f(x)\| \leq C(1 + \|x\|^k)$. Then f is of class (c, h) for $h(x) = \frac{1}{2} \ln(1 + \|x\|^k)$.*

Note that the requirement that k be even is a technicality to ensure that h is smooth over $\mathbb{R}^n$.

Relation with (L_0, L_1)-smoothness. Recall that a function $f : \mathbb{R}^n \to \mathbb{R}$ is (L_0, L_1)-smooth if (1) holds for all $x \in \mathbb{R}^n$ (Zhang et al., 2020). The condition of being (L_0, L_1)-smooth is equivalent to a function having a bounded Riemannian Hessian in a suitable metric:

Theorem 4.5. *A C^∞ function $f : \mathbb{R}^n \to \mathbb{R}$ is (L_0, L_1) smooth iff it has a bounded Riemannian Hessian in the conformally scaled Euclidean metric $g_x(u, v) = (1 + \|\operatorname{grad} f(x)\|_2^2)^{1/2} u^T v$.*

Note that the assumption of f being C^∞ is used to establish smoothness of the scaling function, in order that the resulting conformal metric is smooth. We briefly remark on how Theorem 4.5 is proved. Intuitively, the value of Riemannian hessian $\mathrm{Hess}\, f[X, X]$ at the tangent vector $X \in T_pM$ is $(f \circ \gamma)''(0)$ where γ is a geodesic through p whose velocity vector at $t = 0$ is X. In the Euclidean setting, this is simply $\mathrm{Hess}^{\mathrm{Euc}}\, f(p) = \frac{\partial^2 f}{\partial x^2}(p)[X, X]$. In the setting of a conformally scaled Euclidean metric, since the curve γ is no longer a (Euclidean) straight line, the Hessian is expressed in terms of the underlying Euclidean one $\mathrm{Hess}^{\mathrm{Euc}}\, f$ plus additional terms involving $\frac{\partial f}{\partial x}$ and $\frac{\partial h}{\partial x}$. Specifically, in a conformal scaling of the Euclidean metric,

$$\mathrm{Hess}^{\mathrm{c}}\, f[X, X] = \mathrm{Hess}\, f[X, X] - 2\langle \mathrm{grad}\, h, X\rangle\langle \mathrm{grad}\, f, X\rangle + \langle \mathrm{grad}\, f, \mathrm{grad}\, h\rangle \|X\|_2^2 \tag{9}$$

The proof proceeds by specializing to $h(p) = \frac{1}{2}\ln\sqrt{1 + \|\mathrm{grad}\, f(p)\|^2}$ and showing that (L_0, L_1) smoothness implies that the right hand side is bounded, and vice-versa.

5 Model smoothness and optimization analysis

We illustrate these classes in the context of a two layer network for classification. The parameters are weight matrices w_1, w_2 and bias vectors b_1, b_2. Inputs are vectors z and targets are encoded in one-hot vectors v. An activation function σ is used at the hidden layer. The parameters are grouped together as $p = (w_2, b_2), \quad q = (w_1, b_1)$. We consider classification problems with C classes using the cross entropy loss. The optimization problem is the empirical risk:

$$\begin{aligned} &\min_{(p,q)\in\mathbb{R}^{m+n}} \quad f(p, q), \\ \text{where} \quad &f(p, q) = \frac{1}{N}\sum\nolimits_{i=1}^{N} f_i(p, q), \\ &f_i(p, q) = -v_i^T x_2(z_i) + \log \sum\nolimits_{k=1}^{C} \exp(x_{2,k}(z_i)), \\ &x_2(z) = w_2 x_1(z) + b_2, \\ &x_1(z) = \sigma(w_1 z + b_1). \end{aligned} \tag{10}$$

We employ the following assumption on σ:

Assumption 5.1. *The activation σ is twice continuously differentiable, and $|\sigma|, |\sigma'|$ and $|\sigma''|$ are globally bounded.*

It has been noted previously that (L_0, L_1)-smoothness criteria may be too strong to apply to neural networks (Cooper, 2024). We verify this for the function f above, by showing that the generalized notion of ℓ-smoothness (2) is too strong. Recall that $f : \mathbb{R}^n \to \mathbb{R}$ is ℓ-smooth if $\ell : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ is a non-decreasing continuous function such that (2) holds for all $x \in \mathbb{R}^n$ (Li et al., 2023). We show in Proposition 5.2 that there can be no such ℓ for certain instances of the problem defined in (10).

Proposition 5.2. *Consider problem 10 in the context of binary classification where the dataset is $D = \{(z, v)\}$ with $z = 1$ and $v = (0, 1)$, and the* tanh *activation function is used. Then there is no function ℓ such that* (2) *holds for all x.*

This contrasts with the following result, which states that under Assumption 5.1, the function f of in (10) has a bounded Riemannian Hessian and gradient under both conformal scaling and warped product metrics.

Theorem 5.3. *Let f be the loss function in* (10) *and let Assumption 5.1 hold. Let $h_1(p, q) = \frac{1}{2}\ln(1 + \|p\|^2)$ and let $h_2(p) = \sqrt{1 + \|p\|^2}$. Then f is of class (c, h_1) and of class (w, h_2). Additionally, $\mathrm{grad}^g\, f$ is globally bounded in each metric. Consequently, there are L_g, K_g such that for all (p, q), $\|\mathrm{Hess}^g\, f(p, q)\|_g \leq L_g$ and $\|\mathrm{grad}^g\, f(p, q)\|_g^2 \leq K_g$ for $g \in \{g^{\mathrm{c}}, g^{\mathrm{w}}\}$.*

Theorem 5.3 verifies that problem (10) meets the convergence criteria for Riemannian SGD. The corresponding convergence result for Adagrad-Norm, is stated below in Corollary 5.7. The idea behind the proof of Theorem 5.3 is as follows. For the conformal scaling (7), we start from the formula (9), which, for $X \in \mathbb{R}^n$ provides an expression for the Hessian value $\mathrm{Hess}^{\mathrm{c}}\, f(X, X)$ in terms of the Euclidean Hessian of f and gradient of h. We then specialize to a unit vector for the conformally scaled metric, meaning $\|X\|_g^2 = (1 + \|p\|^2)\|X\|_2^2 = 1$, and use the assumptions on f and standard properties of norms to conclude the bound in the Theorem. Similar logic holds for the warped product.

5.1 Optimization analysis

While the metrics we have introduced enable convergence analysis for standard algorithms like Riemannian SGD, an adaptive algorithm like Adagrad is of interest as it does not require knowledge of the smoothness constant for achieving convergence step size, while still incorporating a tunable parameter. Therefore, We now proceed to a proof of convergence for Riemannian Adagrad-Norm applied to nonconvex functions. Let $\mathcal{F}_n = \sigma(x_1, v_1, v_2, \ldots, v_{n-1})$ denote the history of the algorithm up to time n. We introduce the following assumptions on the stochastic gradients:

Assumption 5.4. *The v_k are conditionally unbiased gradient estimates: $\mathbb{E}[v_k \,|\, \mathcal{F}_k] = \operatorname{grad}^g f(x_k)$ for all $k \geq 1$.*

We also assume that the stochastic gradients are bounded with respect to the Riemannian metric:

Assumption 5.5. *There exists a constant $C \geq 0$ such that $\|v_k\|_{x_k}^2 \leq C$ for all $k \geq 1$.*

In the Euclidean setting the assumption that gradient estimates are uniformly bounded is common in studies of non-convex Adaptive gradient methods (Liang et al., 2025; Défossez et al., 2022; Chen et al., 2019) (along with relaxations including (Hong & Lin, 2024; Wang et al., 2023; Ward et al., 2019)). Note that Assumption 5.5 can be significantly weaker as the bound is in terms of the Riemannian metric. In particular, Problem 10 under Assumption 5.1 satisfies this criteria.

The variant of Adagrad we consider uses delayed normalization, where c_k^2 is the sum of squared gradient norms up to but not including v_k. This is a similar setting explored in other variants of Adagrad (Li & Orabona, 2018) including for Riemannian manifolds (Park et al., 2026). Furthermore, we consider the Norm variant of Adagrad, where adaptation happens via a single parameter c_k, as opposed to per-parameter rates. Starting from $x_1 \in M$, a learning rate $\eta > 0$, it proceeds as follows:

$$\begin{aligned} x_{k+1} &= \operatorname{Retr}_{x_k}\big(-\frac{\eta}{c_k} v_k\big) \\ c_{k+1}^2 &= c_k^2 + \|v_k\|_{x_k}^2. \end{aligned} \tag{11}$$

where c_1 is initialized at a constant value. Our proof uses $c_1 = 1$ for simplicity. The following convergence result considers the expected squared (Riemannian) gradient norm when the algorithm is run for T steps and a random iterate x_R is chosen with $R \sim \operatorname{Unif}\{1, 2, \ldots, T\}$. This is a standard modality for convergence analysis in non-convex settings (Ghadimi & Lan, 2013).

Theorem 5.6. *Suppose that f is L-smooth with respect to* Retr (3)*, and that f is bounded from below by $f^* \in \mathbb{R}$. Further suppose that Assumptions 5.4 and 5.5 hold. Let $\Delta = f(x_1) - f^*$, and for $T > 0$ let $R \sim \operatorname{Unif}\{1, 2, \ldots, T\}$. Then for $c_1 = 1$ and a step size $\eta > 0$,*

$$\mathbb{E}[\|\operatorname{grad}^g f(x_R)\|_{x_R}^2] \leq \frac{\Delta\sqrt{1 + (T-1)C}}{\eta T} + \frac{\eta L}{2}\sqrt{1 + (T-1)C}\frac{C + \log(1 + CT)}{T}.$$

This implies a convergence rate of $\mathcal{O}\big(\frac{\log T}{\sqrt{T}}\big)$.

The result above establishes approximate stationarity in the given metric, and this cannot be translated into approximate stationarity for the Euclidean gradient without further assumptions, such as bounded iterates or uniform equivalence between the Riemannian and Euclidean metrics, the latter of which does not hold for the particular scaling and warping functions we apply.

Consider the model training problem introduced above in Problem 10, where gradient estimates are provided by minibatch gradients with membership selected uniformly at random at each step:

$$v_t = \frac{1}{B}\sum_{i \in b_t} \operatorname{grad}^g f_i(x_t) \tag{12}$$

Corollary 5.7. *Consider applying Adagrad-Norm* (11) *to Problem 10 under assumption 5.1, using the exponential retraction* $\operatorname{Retr} = \operatorname{Exp}$ *corresponding to $g \in \{g^{\mathrm{c}}, g^{\mathrm{w}}\}$, and using minibatch gradients* (12)*. For each of the two metrics, Assumption 5.5 and L-smoothness are satisfied, implying a convergence rate of $\mathcal{O}(\frac{\log T}{\sqrt{T}})$.*

The metrics that we presented enable us to establish smoothness as well as bounds on the stochastic gradients, without assuming global bounds on the Euclidean variance. This is in contrast to analyses that assume Euclidean variance is bounded uniformly (Ding et al., 2025).

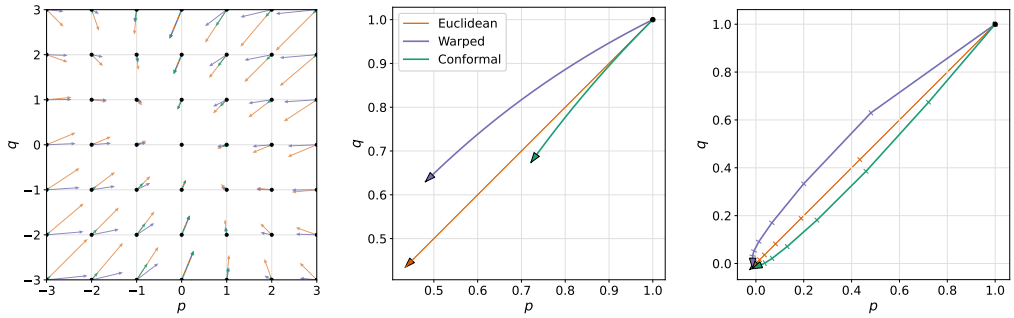

Figure 1: Comparison of warped, conformal, and Euclidean optimization for the function $f(p, q) = 2p^2 + 2q^2 + 4(\sqrt{2} - 1)pq + 1$. Left: The (negative) gradient fields for each geometry. Middle: The geodesic curves starting from $(p, q) = (1, 1)$, with initial velocity $-\operatorname{grad}^g f(p, q)$, from $t = 0$ to 1. Right: Riemannian gradient descent trajectories starting from $(1, 1)$, with a step size of 0.1.

6 GEODESICS

In general, computing the exponential map in an n-dimensional manifold involves solving a set of n nonlinear differential equations. However, the metrics we are concerned with have structure that can be exploited. Specifically, the scaling and warping functions in (7), (8) are *radial* functions that only depend on the norm $\|p\|$ of their argument. As a result, for a geodesic $\gamma(t) = (p(t), q(t))$, we find that, firstly, $p(t)$ lives in a 2D subspace of $\mathbb{R}^n$, while secondly, the curve $q(t)$ is a pregeodesic (a composition of geodesic and a re-scaling of time). In warped products, this pregeodesic structure is guaranteed (O'Neill, 1983); for the conformal scaling it follows from the scaling function only depending on p. Here we present the equations for the conformal setting and we refer to the appendix for detailed derivations including the treatment for the warped product case, which is similar.

Proposition 6.1. *Let N be a manifold with exponential map Exp^N and let $h : \mathbb{R}^m \to \mathbb{R}$ and $c : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ be smooth functions such that $h(p) = \frac{1}{2} \ln c(\|p\|^2)$. Let g be the product metric on $M = \mathbb{R}^m \times N$ and let $\widetilde{g}$ be the conformal scaling $\widetilde{g} = e^{2h} g$. Then the exponential map on $(M, \widetilde{g})$ is $\operatorname{Exp}_{(p(0),q(0))}(p'(0), q'(0)) = (p(1), q(1))$ where $p(t), q(t)$ are defined as follows. For $m > 1$, set u_1, u_2 to orthonormal vectors such that $\operatorname{span}\{p(0), p'(0)\} \subseteq \operatorname{span}\{u_1, u_2\}$, while if $m = 1$, set $u_1 = 1$ and $u_2 = 0$. Let $C = c(\|p(0)\|^2) \left(\|p'(0)\|_2^2 + \|q'(0)\|_{q(0)}^2 \right)$, and let $\eta : [0, 1] \to \mathbb{R}$ and $y_i : [0, 1] \to \mathbb{R}$ for $i = 1, 2$ be solutions to*

$$
\begin{aligned}
\eta'(t) &= \frac{c(s(0))}{c(s(t))}, \\
y_i''(t) &= C \frac{c'(s(t))}{c(s(t))^2} y_i(t) - 2 \frac{c'(s(t))}{c(s(t))} r(t) y_i'(t), \quad i = 1, 2, \\
s(t) &= y_1^2(t) + y_2^2(t), \\
r(t) &= y_1(t) y_1'(t) + y_2(t) y_2'(t),
\end{aligned}
$$

with initial conditions $\eta(0) = 0$, and $y_i(0) = p(0)^T u_i, y_i'(0) = p'(0)^T u_i$ for $i = 1, 2$. Then $p(1) = y_1(1) u_1 + y_2(1) u_2$ and $q(1) = \operatorname{Exp}_{q(0)}^N(\eta(1) q'(0))$.

It is worth noting the following properties of the geodesics, which hold for both warped and conformal metrics. Firstly, in case c is a constant function e.g. $c(p) = 1$, then $y'' = 0$ and $p(1) = p(0) + p'(0)$. Secondly, for any choice of manifold for the second component q, the p component is always a linear combination of $p(0)$ and $p'(0)$, with coefficients determined by the scalar functions y_1, y_2. Finally, in the major cases of interest in this paper, the same is true for q, since for $N = \mathbb{R}^n$, the exponential map is addition of vectors $\operatorname{Exp}_q(q') = q + q'$.

For a sense of how optimization steps differ between warped and conformal metrics, we refer to Figure 1. Here, we illustrate the differing gradient fields (left), and the differing retractions (middle) and optimization trajectories (right) starting from a point $(1, 1)$. In our implementation, the equations

are integrated using a 4th order Runge-Kutta method. We evaluate the overhead of performing integration at each step in Appendix D.

7 Experimental results

We evaluated warped and conformal metrics in three optimization settings. These included a quartic regression task, robust regression with noise scale estimation, and image classification using MLP and CNN models. We defer details on data generation and further experimental results to Appendix E.

Least fourth regression. Here we evaluate the conformal geometry for solving a quartic regression. Given data $\{(x_i, y_i)\}_{i=1}^N$ where each $(x_i, y_i) \in \mathbb{R}^n \times \mathbb{R}$, we define $f : \mathbb{R}^n \to \mathbb{R}$ as

$$f(p) = \frac{1}{4N} \sum_{i=1}^{N} (x_i^T p - y_i)^4. \tag{13}$$

We include this convex problem to demonstrate differing optimization dynamics between the conformal and Euclidean schemes. The Euclidean hessian of f is not globally bounded, yet f has a bounded Hessian for the conformally scaled Euclidean metric with $h(p) = \frac{1}{2}\log(1 + \|p\|^2)$. This is a sufficient condition for convergence of Riemannian gradient descent at small enough step sizes. We set the number of features to $n = 32$ and the number of examples to be $N = 2048$. The problem data (x_i, y_i), coefficients p^* and initial p_0 are generated as detailed in the Appendix. We applied Riemannian GD to (13) and find that conformal optimization is generally more robust under varying initial scaling, as shown in the left side of Figure 2. In this plot, empty regions of space correspond to divergence during optimization. We see that for the Euclidean case, the stability of optimization is heavily dependent on the radius. The conformal method remains stable across a broader set of initialization scales.

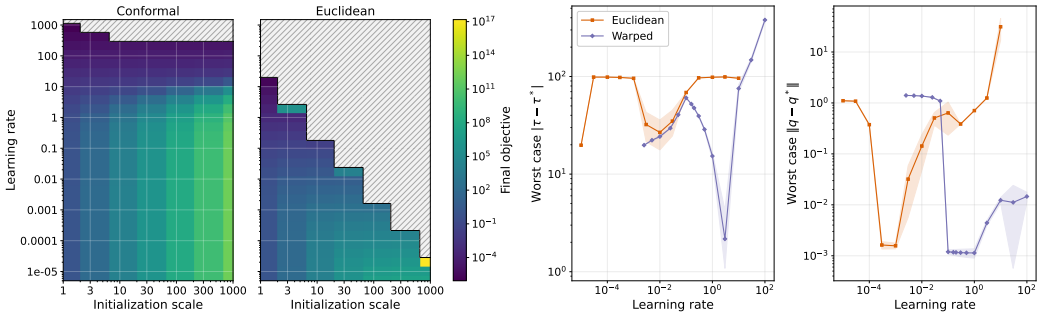

Figure 2: Results for the quartic and robust regression tasks. Left (Quartic): Final loss across step sizes and initialization scales, for conformal (far left) and Euclidean (near left). The shaded area indicates divergence. Right (Robust): For each learning rate, we show the value of the final gap between τ and τ^* (near right) as well as q and q^* (far right), at the initialization scale with largest final NLL.

Robust regression with scale estimation. We demonstrate the warped geometry on a statistical robust regression problem. Given $N = 1024$ observations $\{(x_i, y_i)\}_{i=1}^N$, we seek coefficients $q \in \mathbb{R}^n$ for $n = 8$ as well as a precision parameter $\tau > 0$ assuming the linear regression model $y_i = x_i^T q + \frac{1}{\tau}\xi_i$ where the ξ_i are i.i.d. with mean zero. Assuming a Student's t-distribution on the noise ξ and taking a maximum likelihood approach results in the following objective function:

$$f(p, q) = -\log\sqrt{1+p^2} + \frac{1}{N}\sum_{i=1}^{N} \rho_\nu\left(\sqrt{1+p^2}(x_i^T q - y_i)\right) \tag{14}$$

where ρ_ν is the loss function $\rho_\nu(z) = \frac{\nu+1}{2}\log(1 + \frac{z^2}{\nu})$ and ν is the degrees of freedom, which we fix at $\nu = 3$. Further details on problem setup, including a derivation of Riemannian smoothness for

(14), may be found in Appendix E. We find that across a learning rate sweep, while both methods are able to recover the regression coefficients q with similar accuracy, the warped geometry enables substantially smaller precision recovery error,

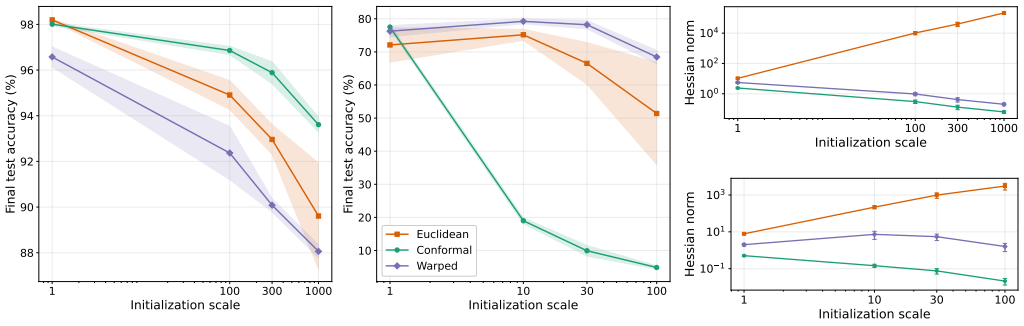

Figure 3: Left and Middle: Final test accuracy across initialization scales for Euclidean, Conformal, and Warped optimizers, on MNIST (Left) and GTSRB (Middle) tasks. Right: the approximate Hessian norm of the objective at initialization, across scales for each geometry, for MNIST (upper right) and GTSRB (bottom right).

Image classification. In this section we explore neural network models for image classification, focusing on the MNIST and GTSRB datasets Stallkamp et al. (2011). For MNIST, we used an MLP model with the $\tanh$ activation function, and 1 hidden layer. For GTSRB, we employed a model with a convolutional input layer, followed by average pooling and down-sampling. We use the erf activation function and a linear output layer. Note that this small CNN fits within our smoothness framework, as it is a case of the generic feature extractor pattern of Proposition B.6. We studied the sensitivity of the models to the geometry and initialization scale. After carrying out sweeps across learning rates and scales using a validation set, for each method we identified the learning rate with the best AUC (Area Under the Curve) in a plot of accuracy versus scale, and use this learning rate for evaluations across scales. The results are shown in Figure 3. We see that for MNIST, choice of metric has a large impact on the sensitivity to the initialization scales. On the other hand, for GTSRB, while the conformal Hessian estimate was relatively small, this did not translate into favorable empirical performance. In each case, we also plot the norm of the Hessian at model initialization. This is estimated following a power iteration approach as described in the appendix. We see that, while the Euclidean Hessian grows with the initialization scale, the Riemannian Hessians remain bounded, in line with our theoretical results.

8 Conclusion

This paper introduced Riemannian formulations for optimization problems based on conformal scalings and warped products. With these geometries, several classes of functions that defy classical Euclidean smoothness become smooth with respect to the Riemannian metric. This enables the application of Riemannian optimization algorithms utilizing the exponential mapping to obtain convergence. While computing the exponential map is non-trivial and requires solving a system of differential equations, the low dimensional nature of the dynamics enables an efficient integration scheme. The results show how metric design can be a useful tool for obtaining smoothness and first-order convergence guarantees.

References

Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

Gary Becigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=r1eiqi09K7.

Jérôme Bolte, Shoham Sabach, Marc Teboulle, and Yakov Vaisbourd. First Order Methods Beyond Convexity and Lipschitz Gradient Continuity with Applications to Quadratic Inverse Problems. *SIAM J. Optim.*, July 2018. URL https://epubs.siam.org/doi/10.1137/17M1138558.

Silvère Bonnabel. Stochastic Gradient Descent on Riemannian Manifolds. *IEEE Trans. Autom. Control*, 58(9):2217–2229, March 2013. doi: 10.1109/TAC.2013.2254619.

N. Boumal, P.-A. Absil, and C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, To appear, 2018.

Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, Cambridge, England, UK, March 2023. ISBN 978-1-00916616-4. doi: 10.1017/9781009166164.

Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of adam-type algorithms for non-convex optimization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1x-x309tm.

Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Y Cooper. A theoretical study of the (l_0, l_1)-smoothness condition in deep learning. In *OPT 2024: Optimization for Machine Learning*, 2024.

Glaydston de C. Bento, Geovani N. Grapiglia, Mauricio S. Louzeiro, and Daoping Zhang. A Riemannian AdaGrad-Norm Method, September 2025. URL https://arxiv.org/html/2509.19682v1. [Online; accessed 10. Nov. 2025].

Alexandre Défossez, Leon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=ZPQhzTSWA7.

Kuangyu Ding, Jingyang Li, and Kim-Chuan Toh. Nonconvex stochastic bregman proximal gradient method with application to deep learning. *Journal of Machine Learning Research*, 26(39):1–44, 2025.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Filip Hanzely, Peter Richtárik, and Lin Xiao. Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. *Comput. Optim. Appl.*, 79(2):405–440, June 2021. ISSN 1573-2894. doi: 10.1007/s10589-021-00273-8.

Chang He, Jiaxiang Li, Bo Jiang, Shiqian Ma, and Shuzhong Zhang. On relatively smooth optimization over riemannian manifolds. *arXiv preprint arXiv:2508.03048*, 2025.

Yusu Hong and Junhong Lin. On Convergence of Adam for Stochastic Optimization under Relaxed Assumptions. *Advances in Neural Information Processing Systems*, 37:10827–10877, December 2024. doi: 10.52202/079017-0346.

Reshad Hosseini and Suvrit Sra. Recent Advances in Stochastic Riemannian Optimization. In *Handbook of Variational Methods for Nonlinear Geometric Data*, pp. 527–554. Springer, Cham, Switzerland, April 2020a. ISBN 978-3-030-31351-7. doi: 10.1007/978-3-030-31351-7_19.

Reshad Hosseini and Suvrit Sra. An alternative to EM for Gaussian mixture models: batch and stochastic Riemannian optimization. *Math. Program.*, 181(1):187–223, May 2020b. ISSN 1436-4646. doi: 10.1007/s10107-019-01381-4.

Hiroyuki Kasai, Pratik Jawanpuria, and Bamdev Mishra. Riemannian adaptive stochastic gradient algorithms on matrix manifolds. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3262–3271. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kasai19a.html.

Ali Kavis, Kfir Y Levy, and Volkan Cevher. High probability bounds for a class of non-convex algorithms with adagrad stepsize. In *International Conference on Learning Representations*, 2022.

John M Lee. *Introduction to Riemannian manifolds*, volume 2. Springer, 2018.

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=8aunGrXdkl.

Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. *arXiv preprint arXiv:1805.08114*, 2018.

Yuqing Liang, Meixuan He, Jinlan Liu, and Dongpo Xu. Convergence of Adam for non-convex objectives: relaxed hyperparameters and non-ergodic case. *Mach. Learn.*, 114(3):1–42, March 2025. ISSN 1573-0565. doi: 10.1007/s10994-025-06737-w.

H. Lu, R. Freund, and Y. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

Barrett O'Neill. *Semi-Riemannian Geometry With Applications to Relativity, 103, Volume 103 (Pure and Applied Mathematics)*. Academic Press, 1983. ISBN 0125267401.

Jiyoung Park, Jaewook J. Suh, Bofan Wang, Anirban Bhattacharya, and Shiqian Ma. Adaptive gradient descent on riemannian manifolds and its applications to gaussian variational inference. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=2TTQjRkgFn.

Barak A. Pearlmutter. Fast exact multiplication by the hessian. *Neural Computation*, 6(1):147–160, 01 1994. ISSN 0899-7667. doi: 10.1162/neco.1994.6.1.147. URL https://doi.org/10.1162/neco.1994.6.1.147.

Peter Petersen. *Riemannian Geometry*. Springer International Publishing, Cham, Switzerland, 2016. ISBN 978-3-319-26654-1. URL https://link.springer.com/book/10.1007/978-3-319-26654-1.

Hiroyuki Sakai and Hideaki Iiduka. A general framework of riemannian adaptive optimization methods with a convergence analysis. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=knv4lQFVoE.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: A multi-class classification competition. In *The 2011 International Joint Conference on Neural Networks*, pp. 1453–1460, 2011. doi: 10.1109/IJCNN.2011.6033395.

Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of AdaGrad for Non-convex Objectives: Simple Proofs and Relaxed Assumptions. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 161–190. PMLR, July 2023. URL https://proceedings.mlr.press/v195/wang23a.html.

Rachel Ward, Xiaoxia Wu, and Leon Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6677–6686, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W. Mahoney. Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE International Conference on Big Data (Big Data)*, pp. 581–590, 2020. doi: 10.1109/BigData50022.2020.9378171.

Hongyi Zhang, Sashank J. Reddi, and Suvrit Sra. Riemannian svrg: Fast stochastic optimization on riemannian manifolds. In *Advances in Neural Information Processing Systems 29*, pp. 4592–4600. 2016.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020.

9 AI USE STATEMENT

In this work, we used generative AI tools for assisting with providing ingredients for proving mathematical claims. In particular, AI was used for simplifying the geodesic equations, and for streamlining the construction used in Proposition 5.2. We have not used generative AI tools for developing theoretical models or conceptual frameworks, or to formulate the mathematical claims, or to draft any part of this manuscript, and the rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. Code generated with LLM assistance was rigorously validated by careful inspection and manual unit testing by the authors. Likewise, any suggested ingredients for mathematical claims were validated by manual verification and cross referenced with existing literature. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Appendix

A Background on Riemannian Manifolds

An n-dimensional smooth manifold M is a topological space that is locally homeomorphic to an open subset of $\mathbb{R}^n$ (Lee, 2018). At each $x \in M$, the tangent space T_xM can be thought of as equivalence classes of smooth curves on M through x, and these intuitively represent the directions in which one can move from x while remaining on the manifold. A tangent vector X at p acts on functions defined in a neighborhood of p as a directional derivative: $X(f) = df(X)$. A vector field on a Riemannian Manifold is a smooth assignment of an element of T_xM for each x in M. A Riemannian metric on M is an assignment of an inner product $\langle \cdot, \cdot \rangle_x$, alternatively denoted by $g_x(\cdot, \cdot)$, to each tangent space T_xM that varies smoothly with x. It induces a norm $\|\xi\|_x^2 = \langle \xi, \xi \rangle_x$ on each tangent space. A Riemannian manifold is a pair of a manifold M and metric g, denoted (M, g). Note that by definition of the gradient, if $X \in T_pM$ and f is a smooth function, then $df(X) = \langle \operatorname{grad} f(p), X \rangle_p$. Consistent with the notation introduced for tangent vectors previously, this could be expressed equivalently as $X(f) = \langle \operatorname{grad} f(p), X \rangle_p$. A Riemannian metric also determines a method for differentiating vector fields, or covariant derivative, known as the Levi-Civita Connection. The derivative of Y along X is $\nabla_X Y$. Important properties of the covariant derivative include that it is bilinear in the fields X, Y. Furthermore, if $f : \mathbb{R}^m \to \mathbb{R}$ is a smooth function then $\nabla_{fX} Y = f \nabla_X Y$ and

$$\nabla_X (fY) = X(f)Y + f \nabla_X Y \tag{15}$$

The Riemannian metric enables calculating the lengths of curves $\gamma : J \to M$ where $J \subseteq \mathbb{R}$ is an interval via $L(\gamma) = \int_J \|\gamma'(t)\|_{\gamma(t)} dt$ and this leads to a metric (distance function) on M, denoted $d(x, y)$, defined as the infimum of the lengths of all smooth curves between x and y. A geodesic is a curve $\gamma : J \to M$ on M that is locally length minimizing, or equivalently where $D_t \gamma' = 0$ where D_t represents the covariant derivative along a curve. The exponential map $\operatorname{Exp} : TM \to M$ maps tangent vectors $\xi \in T_xM$ to points on M via $\operatorname{Exp}(\xi) = \gamma(1)$ where γ is a geodesic with $\gamma(0) = x$ and $\gamma'(0) = \xi$. Intuitively, Exp generalizes the notion of a straight line in a given direction emanating from a point. We employ the Einstein summation convention, where a term with a repeating upper and lower index indicates summation over that index: $a^i b_i = \sum_i a^i b_i$. For further background see (Lee, 2018), (Boumal, 2023).

Properties of connections for conformal scalings and warped products. Evaluating the Hessian and determining geodesics in the various manifolds involves using the Riemannian connection in those spaces. Therefore we recall the necessary identities for conformal and warped connections. Beginning with conformal scalings, let (M, g) be a Riemannian manifold, $h : M \to \mathbb{R}$ be smooth, and let $\widetilde{g} = e^{2h} g$ be the conformal scaling of g by h. The connection $\nabla^{\widetilde{g}}$ is related to ∇^g via (Lee, 2018)

$$\nabla^{\widetilde{g}}_X Y = \nabla^g_X Y + dh(X)Y + dh(Y)X - \langle X, Y \rangle_g \operatorname{grad}^g h. \tag{16}$$

Next, we turn to warped products, Let $(M_1, g^1), (M_2, g^2)$ be manifolds. If $p \in M^1$ and $v \in T_pM^1$, then for any $q \in M_2$, the tangent vector v can be lifted to the tangent vector $\widehat{v} = (v, 0) \in T_{(p,q)}(M_1 \times M_2)$. Similarly, a vector field X on M^1 can be lifted to the vector field $\widehat{X}_{(p,q)} = \widehat{X_p}$ on $M_1 \times M_2$. The following technical result from (O'Neill, 1983) summarizes the formulas relevant for warped products:

Proposition A.1 (Ch. 7, Prop. 35 in (O'Neill, 1983))**.** *Consider a warped product $M \times_h N$ and let X, Y be vector fields on M and let V, W be vector fields on N. Let $\widehat{X}, \widehat{Y}, \widehat{V}, \widehat{W}$ denote their lifts onto $M \times N$ and let ∇^M, ∇^N denote the connections on M and N respectively. Then the following equations characterize the connection on the warped product:*

$$\nabla_{\widehat{X}} \widehat{Y} = \widehat{\nabla^M_X Y}, \tag{17}$$

$$\nabla_{\widehat{X}} \widehat{V} = \nabla_{\widehat{V}} \widehat{X} = \frac{X(h)}{h} \widehat{V}, \tag{18}$$

$$\nabla_{\widehat{V}} \widehat{W} = -h \langle V, W \rangle_N (\operatorname{grad}^M h) + \widehat{\nabla^N_V W}.$$

B PROOFS FOR SECTION 4

B.1 PROOF OF THEOREM 4.3

First, we derive generic sufficient conditions on a function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ to have a bounded Hessian when a warped product or conformal metric is placed on its input space. These conditions are growth requirements on the (Euclidean) gradient and Hessian of f. This is done for the conformal scaling scenario in Lemma B.1. For warped products, we derive a formula for the Hessian in Prop B.2 and translate this into criteria for boundedness of the Hessian in Corollary B.3. We recall some elementary bounds on the gradient and Hessian of the models we are interested in Proposition B.6. Finally, we confirm that these bounds match the growth rate criteria we derived for each metric.

Lemma B.1. *Let (M, g) be a manifold, and let $f : M \to \mathbb{R}$ and $h : M \to \mathbb{R}$ be smooth functions. Let $\widetilde{g} = e^{2h} g$ be a conformal scaling of g. Let there be constants K_1, K_2 such that, $\forall p \in M$, $\| \operatorname{Hess}^g f(p)\|_g \leq K_2 e^{2h(p)}$ and, whenever $\operatorname{grad}^g h(p) \neq 0$, it holds that $\| \operatorname{grad}^g f(p)\|_g \leq K_1 e^{2h(p)} \| \operatorname{grad} h(p)\|_g^{-1}$. Then $\| \operatorname{Hess}^{\widetilde{g}} f\|_{\widetilde{g}} \leq 3K_1 + K_2$.*

Proof. Let $p \in M$ and let $X \in T_pM$. Using the definition of the Riemannian Hessian, and the formula for the covariant derivative for a conformally scaled metric(16), it follows that the $\operatorname{Hess}^{\widetilde{g}}$ is related to $\operatorname{Hess} g$ and the derivatives of the scaling function h via ,

$$\operatorname{Hess}^{\widetilde{g}} f[X, X] = \operatorname{Hess}^g f[X, X] - 2g(\operatorname{grad}^g f, X) g(\operatorname{grad}^g h, X) + g(\operatorname{grad}^g f, \operatorname{grad}^g h)\|X\|_g^2 \quad (19)$$

Further assume that $\|X\|_{\widetilde{g}} = 1$. It suffices to show $|\operatorname{Hess}^{\widetilde{g}} f(p)[X, X]| \leq 3K_1 + K_2$. Taking absolute values in (19), followed by standard norm inequalities,

$$|\operatorname{Hess}^{\widetilde{g}} f[X, X]| \leq \| \operatorname{Hess}^g f\|_g \|X\|_g^2 + 3\| \operatorname{grad}^g f\|_g \| \operatorname{grad}^g h\|_g \|X\|_g^2$$

Note that $\|X\|_{\widetilde{g}}^2 = e^{2h(p)} \|X\|_g^2$, so if X is a $\widetilde{g}$-unit vector, then $\|X\|_g^2 = e^{-2h(p)}$ and

$$|\operatorname{Hess}^{\widetilde{g}} f[X, X]| \leq K_2 + 3\| \operatorname{grad}^g f\|_g \| \operatorname{grad}^g h\|_g e^{-2h(p)}$$

If $\operatorname{grad}^g h = 0$ at p, then we are done. Else, apply the assumption on $\| \operatorname{grad}^g f(p)\|_g$. □

For the case of the warped product, we start with a formula for the Hessian. We provide the derivation of it for completeness.

Proposition B.2. *Let $M \times_h N$ be a warped product manifold. For a function $u : M \times N \to \mathbb{R}$, the Hessian $\operatorname{Hess}^{\mathrm{w}}$ is related to the ordinary product Hessian $\operatorname{Hess}^{\mathrm{p}} u$, as follows. Let $Y = (Y_M, Y_N)$ and $X = (X_M, X_N)$ be tangent vectors at a point $(p, q) \in M \times N$. Then*

$$\operatorname{Hess}^{\mathrm{w}} u[Y, X] = \operatorname{Hess}^{\mathrm{p}} u[Y, X] - \frac{1}{h} Y_M(h) \langle \operatorname{grad}^N u, X_N \rangle - \frac{1}{h} X_M(h) \langle \operatorname{grad}^N u, Y_N \rangle + h \langle \operatorname{grad}^M u, \operatorname{grad} h \rangle \langle Y_N, X_N \rangle.$$

In the case $X = Y$, this simplifies to

$$\operatorname{Hess}^{\mathrm{w}} u[X, X] = \operatorname{Hess}^{\mathrm{p}} u[X, X] - \frac{2}{h} X_M(h) \langle \operatorname{grad}^N u, X_N \rangle + h \langle \operatorname{grad}^M u, \operatorname{grad} h \rangle \|X_N\|^2. \quad (20)$$

Proof. At $(p, q) \in M \times N$, the gradient $\operatorname{grad}^M u(p, q)$ is the tangent vector $(\operatorname{grad}^M u(\cdot, q))(p) \in T_pM$ and likewise $\operatorname{grad}^N u(p, q) = (\operatorname{grad}^N u(p, \cdot))(q) \in T_qN$. Using the natural identification of $T_{(p,q)} M \times N$ with the direct sum $T_pM \oplus T_qN$, we treat $\operatorname{grad}^M u$ and $\operatorname{grad}^N u$ as tangent vectors on $M \times N$ with $\operatorname{grad}^M u$ tangent to M and $\operatorname{grad}^N u$ tangent to N. Therefore we express the gradient in the warped product as $\operatorname{grad}^{\mathrm{w}} u = \operatorname{grad}^M u + \frac{1}{h^2} \operatorname{grad}^N u$. Recall the formula for the Riemannian Hessian:

$$\operatorname{Hess} u[Y, X] = \langle \nabla_Y \operatorname{grad} u, X \rangle$$

The ordinary product hessian is

$$\operatorname{Hess}^{\mathrm{p}} u[Y, X] =$$
$$\langle \nabla^{\mathrm{p}}_{Y_M} \operatorname{grad}^M u, X_M \rangle_M + \langle \nabla^{\mathrm{p}}_{Y_M} \operatorname{grad}^N u, X_N \rangle_N + \langle \nabla^{\mathrm{p}}_{Y_N} \operatorname{grad}^M u, X_M \rangle_M + \langle \nabla^{\mathrm{p}}_{Y_N} \operatorname{grad}^N u, X_N \rangle_N$$

where the superscript p denotes the standard product connection. For the warped product, we will see that the Hessian evaluates to the sum of $\mathrm{Hess}^p\, u$ and additional terms coming from the warped structure. We start from the formula

$$\mathrm{Hess}^{\mathrm{w}}\, u[Y,X] = \langle \nabla^{\mathrm{w}}_{Y_M+Y_N}(\mathrm{grad}^M\, u + \frac{1}{h^2}\,\mathrm{grad}^N\, u), X_M + X_N\rangle_{\mathrm{w}} \tag{21}$$

From bilinearity of the covariant derivative and inner product, the right side of (21) breaks out into eight separate terms, with several redundancies due to symmetry of the Hessian. In each case, we rely on the standard properties of the covariant derivative (including (15)) and the identities for the connection on the warped product in Proposition A.1. We present the detailed derivations for one of the cases, and the remaining cases follow similarly. Note that $\mathrm{grad}\, u$ can be written locally as

$$\mathrm{grad}\, u = \phi^i U_i, + \psi^j V_j. \tag{22}$$

where the ϕ^i are real-valued functions of (p,q) and U_i is the lift onto $M\times N$ of a vector field on M, and likewise for real-valued ψ^j and vector fields V_j tangent to N. The term $\langle \nabla_{Y_M}(h^{-2}\,\mathrm{grad}^N\, u), X_N\rangle_{\mathrm{w}}$ evaluates to a standard Hessian factor and a warped correction:

$$\begin{aligned}
\langle \nabla_{Y_M}(h^{-2}\,\mathrm{grad}^N\, u), X_N\rangle_{\mathrm{w}} &= \langle \nabla_{Y_M}(h^{-2}\psi^j V_j), X_N\rangle_{\mathrm{w}}\\
&= \langle Y_M(h^{-2}\psi^j)V_j + h^{-2}\psi^j \nabla_{Y_M}(V_j), X_N\rangle_{\mathrm{w}}\\
&= \langle Y_M(h^{-2}\psi^j)V_j + h^{-2}\psi^j \frac{1}{h}Y_M(h)V_j, X_N\rangle_{\mathrm{w}}\\
&= \langle (h^{-2}Y_M(\psi^j) + \psi^j Y_M(h^{-2}))V_j + h^{-2}\psi^j\frac{1}{h}Y_M(h)V_j, X_N\rangle_{\mathrm{w}}\\
&= h^{-2}h^2\langle \nabla^{\mathrm{p}}_{Y_M}\,\mathrm{grad}^N, X_N\rangle_N + h^2\langle \psi^j Y_M(h^{-2})V_j, X_N\rangle_N + h^{-2}h^2\langle \psi^j\frac{1}{h}Y_M(h)V_j, X_N\rangle_N\\
&= \langle \nabla^{\mathrm{p}}_{Y_M}\,\mathrm{grad}^N, X_N\rangle_N + h^2\langle \psi^j Y_M(h^{-2})V_j, X_N\rangle_N + \langle \psi^j\frac{1}{h}Y_M(h)V_j, X_N\rangle_N\\
&= \langle \nabla^{\mathrm{p}}_{Y_M}\,\mathrm{grad}^N, X_N\rangle_N + h^2\langle \psi^j\frac{-2}{h^3}Y_M(h)V_j, X^N\rangle_N + \frac{1}{h}Y_M(h)\langle \mathrm{grad}^N\, u, X_N\rangle_N\\
&= \langle \nabla^{\mathrm{p}}_{Y_M}\,\mathrm{grad}^N, X_N\rangle_N - \frac{1}{h}Y_M(h)\langle \mathrm{grad}^N\, u, X_N\rangle_N.
\end{aligned}$$

Next, $\langle \nabla_{Y_M}\,\mathrm{grad}^M\, u, X_M\rangle$ evaluates to a standard Hessian factor:

$$\langle \nabla_{Y_M}\,\mathrm{grad}^M\, u, X_M\rangle_{\mathrm{w}} = \langle \nabla^{\mathrm{p}}_{Y_M}\,\mathrm{grad}^M\, u, X_M\rangle_{\mathrm{M}}.$$

The term $\langle \nabla_{Y_M}(h^{-2}\,\mathrm{grad}^N\, u), X_M\rangle$ vanishes as in the standard Hessian case:

$$\langle \nabla_{Y_M}(h^{-2}\,\mathrm{grad}^N\, u), X_M\rangle_{\mathrm{w}} = 0,$$

and the term $\langle \nabla_{Y_M}\,\mathrm{grad}^M\, u, X_N\rangle$ also vanishes:

$$\langle \nabla_{Y_M}\,\mathrm{grad}^M\, u, X_N\rangle_{\mathrm{w}} = 0.$$

The term $\langle \nabla_{Y_N}(h^{-2}\,\mathrm{grad}^N\, u), X_N\rangle$ evaluates to a standard Hessian factor:

$$\langle \nabla_{Y_N}(h^{-2}\,\mathrm{grad}^N\, u), X_N\rangle_{\mathrm{w}} = \langle \nabla^{\mathrm{p}}_{Y_N}\,\mathrm{grad}^N\, u, X_N\rangle_N,$$

while the term $\langle \nabla_{Y^N}\,\mathrm{grad}^M\, u, X_N\rangle_{\mathrm{w}}$ evaluates to a pure warped factor:

$$\langle \nabla_{Y_N}\,\mathrm{grad}^M\, u, X_N\rangle_{\mathrm{w}} = h\langle \mathrm{grad}^M\, u, \mathrm{grad}\, h\rangle_M\langle Y_N, X_N\rangle_N.$$

By symmetry of the Hessian, the sum of the final two terms can be computed from existing derivations:

$$\langle \nabla_{Y_N}\,\mathrm{grad}^M\, u, X_M\rangle_{\mathrm{w}} + \langle \nabla_{Y_N}(h^{-2}\,\mathrm{grad}^N\, u), X_M\rangle_{\mathrm{w}} = \langle \nabla^{\mathrm{p}}_{X_M}\,\mathrm{grad}^N, Y_N\rangle_{\mathrm{N}} - \frac{1}{h}X_M(h)\langle \mathrm{grad}^N\, u, Y_N\rangle_{\mathrm{N}}$$

Adding the terms together, we obtain the claimed equation:

$$\mathrm{Hess}^{\mathrm{w}}\, u[Y,X] = \mathrm{Hess}^{\mathrm{p}}\, u[Y,X] - \frac{1}{h}Y^M(h)\langle \mathrm{grad}^N\, u, X^N\rangle - \frac{1}{h}X^M(h)\langle \mathrm{grad}^N\, u, Y^N\rangle + h\langle \mathrm{grad}^M\, u, \mathrm{grad}\, h\rangle\langle Y_N, X^N\rangle.$$

□

Using the formula derived in the PropositionB.2, we can show that the functions in (w, h) have a bounded Hessian in a warped product metric.

Corollary B.3. *Let $u : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function of class (w, h) 4.2, and let $\mathbb{R}^m \times \mathbb{R}^n$ have the warped product metric* (8)*. Then $\|\operatorname{Hess}^{\mathrm{w}} u\|_{\mathrm{w}} \leq \max\{K_{pp}, K_{qq} + K_p\} + K_{pq} + K_q$. Furthermore, if there are constants L_p, L_q such that $\|\frac{\partial u}{\partial p}(p,q)\| \leq L_p$ and $\|\frac{\partial u}{\partial q}(p,q)\| \leq L_q h(p)$ then $\|\operatorname{grad}^{\mathrm{w}} u\|_{\mathrm{w}} \leq \sqrt{L_p^2 + L_q^2}$.*

Proof. For the gradient bound, observe that $\operatorname{grad}^{\mathrm{w}}$ takes the form

$$\operatorname{grad}^{\mathrm{w}} u = \Big(\frac{\partial u}{\partial p}, \frac{1}{h^2}\frac{\partial u}{\partial q}\Big)$$

Taking the (squared) norm and applying the assumptions on $\|\frac{\partial u}{\partial p}\|, \|\frac{\partial u}{\partial q}\|$ yields the result. For the Hessian bound, we start by specializing B.2 to the case of $M = \mathbb{R}^m$ and $N = \mathbb{R}^N$. In this case the $\operatorname{Hess}^{\mathrm{p}}, \operatorname{grad}^N, \operatorname{grad}^M$ are the ordinary Euclidean objects. Let $X = (X_M, X_N) \in \mathbb{R}^m \times \mathbb{R}^n$. Then we may express $\operatorname{Hess}^{\mathrm{w}} u[X, X]$ in terms of the Euclidean partial derivatives of u as follows

$$\operatorname{Hess}^{\mathrm{w}} u[X, X] = A(X_M, X_M) + 2B(X_M, X_N) + C(X_N, X_N)$$

where A, B, C are the bilinear maps

$$\begin{aligned}
A[X_M, X_M] &= \frac{\partial^2 u}{\partial p^2}(p,q)[X_M, X_M] \\
B[X_M, X_N] &= \frac{\partial^2 u}{\partial p \partial q}(p,q)[X_M, X_N] - \frac{1}{h(p)}\langle \nabla h(p), X_M\rangle \left\langle \frac{\partial u}{\partial q}(p,q), X_N \right\rangle \\
C[X_N, X_N] &= \frac{\partial^2 u}{\partial q^2}(p,q)[X_N, X_N] + h(p)\left\langle \frac{\partial u}{\partial p}(p,q), \nabla h(p)\right\rangle \|X_N\|^2
\end{aligned}$$

Suppose that $\|X\|_{\mathrm{w}} = 1$. Then $\|X_M\|^2 + h(p)^2\|X_N\|^2 = 1$, and

$$\begin{aligned}
\|A[X_M, X_M]\| &\leq \left\|\frac{\partial^2 u}{\partial p^2}(p,q)\right\|_2 \|X_M\|^2 \\
\|B[X_M, X_N]\| &\leq \left(\left\|\frac{\partial^2 u}{\partial p \partial q}(p,q)\right\|_2 + \frac{1}{h(p)}\left\|\frac{\partial h}{\partial p}(p)\right\|\left\|\frac{\partial u}{\partial q}(p,q)\right\|\right)\|X_M\|\|X_N\| \\
\|C[X_N, X_N]\| &\leq \left(\left\|\frac{\partial^2 u}{\partial q^2}(p,q)\right\| + h(p)\left\|\frac{\partial u}{\partial p}(p,q)\right\|\left\|\frac{\partial h}{\partial p}(p)\right\|\right)\|X_N\|^2
\end{aligned}$$

And it follows that

$$\begin{aligned}
\|\operatorname{Hess}^{\mathrm{w}} u\| \leq \max\left\{\left\|\frac{\partial^2 u}{\partial p^2}(p,q)\right\|, \frac{1}{h(p)^2}\left\|\frac{\partial^2 u}{\partial q^2}(p,q)\right\| + \frac{1}{h(p)}\left\|\frac{\partial u}{\partial p}(p,q)\right\|\left\|\frac{\partial h}{\partial p}(p,q)\right\|\right\} \\
+ \frac{1}{h(p)}\left\|\frac{\partial^2 u}{\partial p \partial q}(p,q)\right\|_2 + \frac{1}{h(p)^2}\left\|\frac{\partial h}{\partial p}(p)\right\|\left\|\frac{\partial u}{\partial q}(p,q)\right\|.
\end{aligned} \tag{23}$$

□

Proof of Proposition 4.4. We apply Lemma B.1. Noting that $\|\operatorname{grad} h(x)\| = \frac{k}{2}\frac{\|x\|^{k-1}}{1+\|x\|^k}$, It suffices to exhibit constants L_1, L_2 such that

$$\|\operatorname{Hess} f(x)\| \leq L_1(1 + \|x\|^k), \tag{24a}$$

$$\|\operatorname{grad} f(x)\| \leq L_2\frac{(1+\|x\|^k)^2}{\|x\|^{k-1}}. \tag{24b}$$

Clearly we can take $L_1 = C$, as condition (24a) is exactly our assumption. For (24b),

$$\begin{aligned}
\|\operatorname{grad} f(x)\| &\leq \|\operatorname{grad} f(0)\| + \int_0^1 \|\operatorname{Hess} f(tx)\|\|x\|dt \\
&\leq \|\operatorname{grad} f(0)\| + \int_0^1 C(1 + t^k\|x\|^k)\|x\|dt \\
&= \|\operatorname{grad} f(0)\| + C\|x\| + C\|x\|^{k+1}\frac{1}{k+1} \leq D(1 + \|x\|^{k+1})
\end{aligned}$$

for some D (for instance, $D = \|\frac{\partial f}{\partial x}(0)\| + C + 1$). Letting $r = \|x\|$, it suffices to show that $1 + r^{k+1} \leq \frac{(1+r^k)^2}{r^{k-1}}$. The term on the RHS is $\frac{(1+r^k)^2}{r^{k-1}} = (r^{1-k} + 2r + r^{k+1})$ Since $1 + r^{k+1} \leq r^{1-k} + 2r + r^{k+1}$, we are done. □

B.2 Proof of Theorem 4.5

The following auxiliary Lemma will be needed in the proof of Theorem 4.5.

Lemma B.4. *Let $y \in \mathbb{R}^n$ and let $A \in \mathbb{R}^{n\times n}$ be symmetric. Then for any unit vector x, $-\frac{\|y\|\|Ay\|}{2} \leq (y^T x)y^T Ax - \frac{y^T Ay}{2} \leq \frac{\|y\|\|Ay\|}{2}$*

Proof. For any vectors z, y, it can be verified that the matrix $B = \frac{1}{2}(yz^T + zy^T)$ has eigenvalues $\lambda_1 = \frac{1}{2}(z^T y - \|z\|\|y\|)$ and $\lambda_2 = \frac{1}{2}(z^T y + \|z\|\|y\|)$, corresponding to Eigenvectors $v_1 = \|z\|y - \|y\|z$ and $v_2 = \|z\|y + \|y\|z$, respectively. Specializing this to $z = Ay$, for any unit vector x we obtain that $\frac{1}{2}(y^T Ay - \|Ay\|\|y\|) \leq x^T y(Ay)^T x \leq \frac{1}{2}(y^T Ay + \|Ay\|\|y\|)$. Subtracting $\frac{y^T Ay}{2}$ from each side of this equation yields the result. □

Theorem B.5 (Theorem 4.5)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is (L_0, L_1) smooth as in Definition 1 iff it has a bounded Hessian in the conformally scaled Euclidean metric $g_x(u, v) = (1 + \|\operatorname{grad} f(x)\|_2^2)^{1/2} u^T v$.*

Proof. The definition of g stated in the Theorem corresponds to a conformal scaling $\widetilde{g} = e^{2h} g^{\mathrm{Euc}}$ where h is

$$h(p) = \frac{1}{2}\log((1 + \|\operatorname{grad} f(p)\|^2)^{1/2}). \tag{25}$$

Firstly, suppose that f is (L_0, L_1) smooth. We show that the conditions of Lemma B.1 are satisfied for $\widetilde{g}$. We have

$$\begin{aligned}
\|\operatorname{Hess} f(p)\| &\leq (L_0 + L_1\|\operatorname{grad} f(p)\|) \\
&\leq \max\{L_0, L_1\}(1 + \|\operatorname{grad} f(p)\|) \\
&\leq \max\{L_0, L_1\}\sqrt{2}(1 + \|\operatorname{grad} f(p)\|^2)^{1/2} \\
&= K_2 e^{2h(p)}
\end{aligned}$$

where $K_2 = \sqrt{2}\max\{L_0, L_1\}$. Next, suppose that $\operatorname{grad} h(p) \neq 0$. We have

$$\begin{aligned}
\|\operatorname{grad} h(p)\| &= \frac{1}{2}\frac{1}{1 + \|\operatorname{grad} f(p)\|^2}\|\operatorname{Hess} f(p)[\operatorname{grad} f(p)]\| \\
&\leq \frac{1}{2}\frac{1}{1 + \|\operatorname{grad} f(p)\|^2}(L_0 + L_1\|\operatorname{grad} f(p)\|)\|\operatorname{grad} f(p)\| \\
&\leq \frac{\max\{L_0, L_1\}}{2}\frac{1 + \|\operatorname{grad} f(p)\|}{1 + \|\operatorname{grad} f(p)\|^2}\|\operatorname{grad} f(p)\|
\end{aligned} \tag{26}$$

For ease of notation write $u = \|\operatorname{grad} f(p)\|$. Then (26) implies

$$\begin{aligned}
e^{2h(p)}\|\operatorname{grad} h(p)\|^{-1} &\geq \frac{2}{\max\{L_0, L_1\}}\frac{(1 + u^2)^{3/2}}{(1 + u)u} \\
&\geq \frac{2}{\max\{L_0, L_1\}}\frac{1}{2}u
\end{aligned}$$

Therefore we may take $K_1 = \max\{L_0, L_1\}$.

Next, suppose that for some K the function f has $\|\operatorname{Hess}^{\widetilde{g}} f(p)\|_{\widetilde{g}} \leq K$ for all p, where $\widetilde{g} = e^{2h} g^{\mathrm{Euc}}$, for h the function (25). Fix a $p \in M$ and let $X \in T_p M$ be a Euclidean unit vector $\|X\|_2 = 1$. As p is fixed in the remainder, for ease of notation, we omit indication of it below, noting here that all terms $\operatorname{Hess} f$ and $\operatorname{grad} f$ are evaluated at p. Specializing (19) to the case that g is the Euclidean metric, we obtain,

$$-K\|X\|_{\widetilde{g}}^2 \leq \operatorname{Hess} f[X, X] - \frac{\langle \operatorname{grad} f, X\rangle\langle \operatorname{Hess} f \operatorname{grad} f, X\rangle}{(1 + \|\operatorname{grad} f\|^2)} + \frac{\operatorname{Hess} f[\operatorname{grad} f, \operatorname{grad} f]}{2(1 + \|\operatorname{grad} f\|^2)} \leq K\|X\|_{\widetilde{g}}^2 \tag{27}$$

Applying Lemma B.4 with $y = \operatorname{grad} f(p)$ and $A = \operatorname{Hess} f(p)$, we see that

$$\begin{aligned} -\frac{\|\operatorname{grad} f\|^2 \|\operatorname{Hess} f(p)\|}{2(1+\|\operatorname{grad} f\|^2)} &\leq \frac{\langle \operatorname{grad} f, X\rangle \langle \operatorname{Hess} f \operatorname{grad} f, X\rangle}{(1+\|\operatorname{grad} f\|^2)} - \frac{\operatorname{Hess} f[\operatorname{grad} f, \operatorname{grad} f]}{2(1+\|\operatorname{grad} f\|^2)} \\ &\leq \frac{\|\operatorname{grad} f\|^2 \|\operatorname{Hess} f\|}{2(1+\|\operatorname{grad} f\|^2)} \end{aligned} \tag{28}$$

Adding (28) to (27), we obtain

$$|\operatorname{Hess} f[X,X]| \leq K\|X\|_g^2 + \frac{\|\operatorname{grad} f\|^2 \|\operatorname{Hess} f\|}{2(1+\|\operatorname{grad} f\|^2)}$$

Noting that $\|\operatorname{grad} f\|^2/(1+\|\operatorname{grad} f\|^2) \leq 1$, and that by our definition of h, for a g-unit vector X we have $\|X\|_g^2 = (1+\|\operatorname{grad} f\|^2)^{1/2} \leq (1+\|\operatorname{grad} f\|)$ We obtain

$$|\operatorname{Hess} f[X,X]| \leq K(1+\|\operatorname{grad} f\|) + \frac{\|\operatorname{Hess} f\|}{2}$$

Since X was an arbitrary unit vector we obtain that $\|\operatorname{Hess} f\| \leq 2K(1+\|\operatorname{grad} f\|)$. That is, f is (L_0, L_1) smooth with $L_0 = L_1 = 2K$. □

B.3 Proof of Proposition 5.2

Proof. Without loss of generality, we consider the setting of binary classification with classes $\{0,1\}$ and the binary cross entropy loss. Consider mapping an input $z = 1$ to the target label $y = 1$. Then, the output of the network is

$$x = p\tanh(q), \tag{29}$$

and the loss function is

$$f = \log(1+e^{-2x}). \tag{30}$$

For any value $c > 0$, we show there is a sequence of model parameters $\{(p_n, q_n)\}_{n\geq 1}$ such that $\|\operatorname{grad} f(p_n,q_n)\| = c$ while $\|\operatorname{Hess} f(p_n,q_n)\| \to \infty$. To start, the first derivatives of f are

$$\begin{aligned} \frac{\partial f}{\partial p}(p,q) &= -\frac{1}{1+e^{2x}} 2\tanh(q), \\ \frac{\partial f}{\partial q}(p,q) &= -\frac{1}{1+e^{2x}} 2p\tanh'(q). \end{aligned}$$

We let G denote $\|\operatorname{grad} f(p,q)\|^2$:

$$G(p,q) = \left(\frac{\partial f}{\partial p}(p,q)^2 + \frac{\partial f}{\partial q}(p,q)^2\right)$$

For each $z > 0$ consider values of (p,q) such that $p > 0$ and $q = \operatorname{atanh}(\frac{z}{p})$. Introducing the shorthand $t = \tanh(q)$, on this curve we have $e^{2pt} = e^{2z}$ and $t = \frac{z}{p}$ and

$$G\left(p, \operatorname{atanh}\left(\frac{z}{p}\right)\right) = \frac{4}{(1+e^{2z})^2}\left(\frac{z^2}{p^2} + p^2\left(1-\frac{z^2}{p^2}\right)^2\right)$$

where we have also used $\tanh'(u) = 1-\tanh(u)^2$. Define $H_z(p)$ for $p > z > 0$ as $H_z(p) = G(p, \operatorname{atanh}(\frac{z}{p}))$. Observe that

$$\lim_{p\to z} H_z(p) = \frac{4}{(1+e^{2z})^2},$$

while

$$\lim_{p\to\infty} H_z(p) = \infty.$$

For any $c > 0$ let z be large enough so that $\frac{4}{(1+e^{2z})^2} < c^2$. By continuity there must then be a $p_z > z$ such that $H_z(p_z) = c^2$. That is, for any $c > 0$ and all z sufficiently large there is a $p_z > z$ such that

$$G\left(p_z, \operatorname{atanh}\left(\frac{z}{p_z}\right)\right) = c^2 \tag{31}$$

The second derivative $\frac{\partial^2 f}{\partial q^2}$ is

$$\frac{\partial^2 f}{\partial q^2}(p,q) = -2t\frac{\partial f}{\partial q}(p,q) + e^{2pt}\left(\frac{\partial f}{\partial q}(p,q)\right)^2 \tag{32}$$

Let $q_z = \operatorname{atanh}(\frac{z}{p_z})$. As $p_z > 0$ and $q_z > 0$, the first term on the right side of (32) is positive at (p_z, q_z), so

$$\frac{\partial^2 f}{\partial q^2}(p_z,q_z) \geq e^{2pt}\left(\frac{\partial f}{\partial q}(p_z,q_z)\right)^2.$$

By (31),

$$\left(\frac{\partial f}{\partial q}(p_z,q_z)\right)^2 = c^2 - \left(\frac{\partial f}{\partial p}(p_z,q_z)\right)^2 = c^2 - \frac{4\tanh(q_z)^2}{(1+e^{2z})^2} \geq c^2 - \frac{4}{(1+e^{2z})^2}$$

Therefore,

$$\lim_{z\to\infty}\frac{\partial^2 f}{\partial q^2}(p_z,q_z) \geq \lim_{z\to\infty} e^{2z}\left(c^2 - \frac{4}{(1+e^{2z})^2}\right) = \infty.$$

□

Proposition B.6. *Let $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ be defined as follows. For parameters $p = (w_2, b_2) \in \mathbb{R}^{C\times m_2} \times \mathbb{R}^C$ and $q \in \mathbb{R}^n$, and a twice continuously differentiable function $y : \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}^{m_2}$ and input $z \in \mathbb{R}^r$, and a one-hot vector $v \in \mathbb{R}^C$, $f(p,q)$ is*

$$\begin{aligned} f(p,q) &= -v^T x_2 + \log\textstyle\sum_{k=1}^C \exp(x_{2,k}), \\ \textit{where}\quad x_2 &= w_2 x_1 + b_2, \\ x_1 &= y(q,z). \end{aligned} \tag{33}$$

Then the norms of the first and second derivatives of f may be bounded as follows:

$$\begin{aligned} &\left\|\frac{\partial f}{\partial p}(p,q)\right\| \leq \sqrt{2(\|y(q)\|^2+1)}, \qquad \left\|\frac{\partial f}{\partial q}(p,q)\right\| \leq \sqrt{2}\left\|\frac{\partial y}{\partial q}(p,q)\right\|\|w_2\|, \\ &\left\|\frac{\partial^2 f}{\partial p^2}(p,q)\right\| \leq 1 + \|y(q)\|^2 \\ &\left\|\frac{\partial^2 f}{\partial q^2}(p,q)\right\| \leq \left(\sqrt{2}\left\|\frac{\partial^2 y}{\partial q^2}(p,q)\right\| + \frac{1}{2}\left\|\frac{\partial y}{\partial q}(p,q)\right\|^2\right)(\|w_2\|^2+1) \\ &\left\|\frac{\partial^2 f}{\partial p \partial q}(p,q)\right\| \leq \sqrt{2}\left\|\frac{\partial y}{\partial q}(p,q)\right\| + \left(\frac{1}{2}\left\|\frac{\partial y}{\partial q}(p,q)\right\| + \frac{1}{2}\|y(q)\|\left\|\frac{\partial y}{\partial q}(p,q)\right\|\right)\|w_2\|. \end{aligned} \tag{34}$$

Proof. Define $r : \mathbb{R}^C \to \mathbb{R}^C$ via $r(x)_i = \frac{\exp(x_i)}{\sum_k \exp(x_k)}$. The derivatives of f may be calculated as:

$$\begin{aligned} \frac{\partial f}{\partial w_2}(p,q) &= (-v + r(x_2))y(q)^T, \\ \frac{\partial f}{\partial b_2}(p,q) &= -v + r(x_2), \\ \frac{\partial f}{\partial q}(p,q) &= (-v + r(x_2))^T w_2 \frac{\partial y}{\partial q}(q). \end{aligned} \tag{35}$$

For the second derivatives, noting that $\frac{\partial r}{\partial x}(x) = \operatorname{diag}(r(x)) - r(x)r(x)^T$, we have

$$\begin{aligned}
\frac{\partial^2 f}{\partial w_2^2}(p,q)[\Delta_1,\Delta_2] &= (\Delta_2 y(q))^T \left(\operatorname{diag}(r(x)) - r(x)r(x)^T\right)\Delta_1 y(q),\\
\frac{\partial^2 f}{\partial w_2 \partial b_2}(p,q)[\Delta_b,\Delta_w] &= \Delta_b^T \left(\operatorname{diag}(r(x)) - r(x)r(x)^T\right)\Delta_w y(q)\\
\frac{\partial^2 f}{\partial b_2^2}(p,q)[\Delta_1,\Delta_2] &= \Delta_1^T(\operatorname{diag} r(x) - r(x)r(x)^T)\Delta_2\\
\frac{\partial^2 f}{\partial q^2}(p,q)[\Delta_1,\Delta_2] &= (-v + r(x))w_2\frac{\partial^2 y}{\partial q^2}(q)[\Delta_1,\Delta_2]\\
&\quad + \left((\operatorname{diag} r(x) - r(x)r(x)^T\right) w_2\frac{\partial y}{\partial q}(q)\Delta_2)^T w_2\frac{\partial y}{\partial q}(q)\Delta_1\\
\frac{\partial^2 f}{\partial q \partial b_2}(p,q)[\Delta_b,\Delta_q] &= \Delta_b^T(\operatorname{diag} r(x) - r(x)r(x)^T)w_2\frac{\partial y}{\partial q}(q)\Delta_q\\
\frac{\partial^2 f}{\partial w_2 \partial q}(p,q)[\Delta_q,\Delta_w] &= \left((\operatorname{diag} r(x) - r(x)r(x)^T)\Delta w y(q)\right)^T w_2\frac{\partial y}{\partial q}(q)\Delta_q\\
&\quad + (r(x) - v)^T(\Delta w)\frac{\partial y}{\partial q}(q)[\Delta_q]
\end{aligned}$$

For bounding norms in the above, note the following:

$$\begin{aligned}
\|\operatorname{diag} r(x) - r(x)r(x)^T\|_2 &\le \frac{1}{2}\\
\|r(x) - v\|_2 &\le \sqrt{2}
\end{aligned} \tag{36}$$

The above leads to

$$\begin{aligned}
\left\|\frac{\partial f}{\partial w_2}(p,q)\right\| &\le \sqrt{2}\|y(q)\|\\
\left\|\frac{\partial f}{\partial b_2}(p,q)\right\| &\le \sqrt{2}
\end{aligned}$$

Grouping together w_2 and b_2 as $p = (w_2, b_2)$, this implies

$$\left\|\frac{\partial f}{\partial p}(p,q)\right\| \le \sqrt{2\|y(q)\|^2 + 2} \tag{37}$$

Furthermore,

$$\left\|\frac{\partial f}{\partial q}(p,q)\right\| \le \sqrt{2}\|w_2\|\left\|\frac{\partial y}{\partial q}(q)\right\| \quad \text{by } 35,\ 36 \tag{38}$$

For second derivatives:

$$\begin{aligned}
\left\|\frac{\partial^2 f}{\partial w_2^2}(p,q)\right\| &\le \frac{1}{2}\|y(q)\|^2\\
\left\|\frac{\partial^2 f}{\partial w_2 \partial b_2}(p,q)\right\| &\le \frac{1}{2}\|y(q)\|\\
\left\|\frac{\partial^2 f}{\partial b_2^2}(p,q)\right\| &\le \frac{1}{2}
\end{aligned}$$

Again, grouping (w_2, b_2) together into a parameter $p = (w_2, b_2)$, we have

$$\begin{aligned}
\left\|\frac{\partial^2 f}{\partial p^2}\right\| &\le \max\left\{\left\|\frac{\partial^2 f}{\partial w_2^2}(p,q)\right\|, \left\|\frac{\partial^2 f}{\partial b_2^2}(p,q)\right\|\right\} + \left\|\frac{\partial^2 f}{\partial w_2 \partial b_2}(p,q)\right\|\\
&\le \frac{1}{2}\max\{\|y(q)\|^2, 1\} + \frac{1}{2}\|y(q)\|\\
&\le 1 + \|y(q)\|^2
\end{aligned} \tag{39}$$

For the term $\frac{\partial^2 f}{\partial q^2}$ we have

$$\begin{aligned}\left\|\frac{\partial^2 f}{\partial q^2}(p,q)\right\| &\le \sqrt{2}\|w_2\|\left\|\frac{\partial^2 y}{\partial q^2}(q)\right\| + \frac{1}{2}\|w_2\|^2\left\|\frac{\partial y}{\partial q}(q)\right\|^2 \\ &= \sqrt{2}\left\|\frac{\partial^2 y}{\partial q^2}(q)\right\|\|w_2\| + \frac{1}{2}\left\|\frac{\partial y}{\partial q}(q)\right\|^2\|w_2\|^2 \\ &\le \left(\sqrt{2}\left\|\frac{\partial^2 y}{\partial q^2}(q)\right\| + \frac{1}{2}\left\|\frac{\partial y}{\partial q}(q)\right\|^2\right)(\|w_2\|^2+1)\end{aligned} \tag{40}$$

Finally

$$\begin{aligned}\left\|\frac{\partial^2 f}{\partial q \partial b_2}(p,q)\right\| &\le \frac{1}{2}\|w_2\|\left\|\frac{\partial y}{\partial q}(q)\right\| \\ \left\|\frac{\partial^2 f}{\partial w_2 \partial q}(p,q)\right\| &\le \frac{1}{2}\|y(q)\|\|w_2\|\left\|\frac{\partial y}{\partial q}(q)\right\| + \sqrt{2}\left\|\frac{\partial y}{\partial q}(q)\right\|\end{aligned}$$

Therefore

$$\begin{aligned}\left\|\frac{\partial^2 f}{\partial p \partial q}(p,q)\right\| &\le \left\|\frac{\partial^2 f}{\partial q \partial b_2}(p,q)\right\| + \left\|\frac{\partial^2 f}{\partial w_2 \partial q}(p,q)\right\| \\ &\le \sqrt{2}\left\|\frac{\partial y}{\partial q}(q)\right\| + \left(\frac{1}{2}\left\|\frac{\partial y}{\partial q}(q)\right\| + \frac{1}{2}\|y(q)\|\left\|\frac{\partial y}{\partial q}(q)\right\|\right)\|w_2\|\end{aligned} \tag{41}$$

□

Proof of Theorem 5.3. Note that Proposition B.6 can be applied to Problem 10 by setting $y(q,z) = y((w_1,b_1),z) = \sigma(w_1 z + b_1)$. Under this definition of y, and using Assumption 5.1 on the activation, it is evident that there is a constant K such that, for all p, q,

$$\max\left\{\|y(q)\|, \left\|\frac{\partial y}{\partial q}(q)\right\|, \left\|\frac{\partial^2 y}{\partial q^2}(q)\right\|\right\} \le K. \tag{42}$$

Combining this with (34) we see that

$$\begin{aligned}&\left\|\frac{\partial f}{\partial p}(p,q)\right\| \le \sqrt{2(K^2+1)}, \qquad \left\|\frac{\partial f}{\partial q}(p,q)\right\| \le \sqrt{2}K\|w_2\|, \\ &\left\|\frac{\partial^2 f}{\partial p^2}(p,q)\right\| \le 1 + K^2 \\ &\left\|\frac{\partial^2 f}{\partial q^2}(p,q)\right\| \le \left(\sqrt{2}K + \frac{1}{2}K^2\right)(\|w_2\|^2+1) \\ &\left\|\frac{\partial^2 f}{\partial p \partial q}(p,q)\right\| \le \frac{K}{\sqrt{2}}\max\{4, 1+K\}\sqrt{1+\|w_2\|^2}\end{aligned} \tag{43}$$

First consider the conformal scaling case $g = \mathrm{g}^c$. The gradient norm in g^c will be

$$\|\operatorname{grad}^{\mathrm{c}} f(p,q)\|_{\mathrm{c}}^2 = \frac{1}{1+\|p\|^2}\left(\left\|\frac{\partial f}{\partial p}(p,q)\right\|^2 + \left\|\frac{\partial f}{\partial q}(p,q)\right\|^2\right)$$

From (43) we have

$$\sqrt{\left\|\frac{\partial f}{\partial p}(p,q)\right\|^2 + \left\|\frac{\partial f}{\partial q}(p,q)\right\|^2} \le \sqrt{2(K^2+1+K^2\|w_2\|^2)} \le \sqrt{2(K^2+1)}\sqrt{1+\|w_2\|^2}$$

Therefore

$$\|\operatorname{grad} f(p,q)\|_{\mathrm{c}} \le \sqrt{2(K^2+1)}$$

To show that the Hessian is bounded in the conformally scaled metric, we apply Lemma B.1 with $M = \mathbb{R}^m \times \mathbb{R}^n$ and the scaling function $h(p,q) = \frac{1}{2}\ln(1+\|p\|^2)$. Noting that $\|\operatorname{grad} h(p)\|^{-1} = \frac{1+\|p\|^2}{\|p\|}$, the conditions of the lemma are

$$\|\operatorname{Hess} f(p,q)\| \leq K_2(1+\|p\|^2), \tag{44a}$$

$$\|\operatorname{grad} f(p,q)\| \leq K_1 \frac{(1+\|p\|^2)^2}{\|p\|}. \tag{44b}$$

For the condition on the Hessian (44a), we have

$$\|\operatorname{Hess} f\| \leq \max\left\{1+K^2, (\sqrt{2}K + \frac{1}{2}K^2)(\|w_2\|^2+1)\right\} + \frac{K}{\sqrt{2}}\max\{4, 1+K\}\sqrt{1+\|w_2\|^2}$$

therefore we can take $K_2 = \max\{1+K^2, \sqrt{2}K + \frac{1}{2}K^2\} + \frac{K}{\sqrt{2}}\max\{4,1+K\}$. For condition (44b), note that $\sqrt{1+\|p\|^2} \leq (1+\|p\|^2)^2/\|p\|$, so we can simply take $K_1 = \sqrt{2(K^2+1)}$. This shows that f has a bounded Hessian for the conformally scaled Euclidean metric.

For the case of the warped product $g = g^{\mathrm{w}}$, we apply Corollary B.3, using the warping function $h(p) = \sqrt{1+\|p\|^2}$. In this case, $\|\frac{\partial h}{\partial p}(p)\| = \frac{\|p\|}{\sqrt{1+\|p\|^2}}$ and the conditions for being of class (w, h) are as follows 4.2: There are constants $K_{pp}, K_{qq}, K_{pq}, K_p$ and K_q such that

$$\begin{gathered}
\left\|\frac{\partial^2 f}{\partial p^2}(p,q)\right\| \leq K_{pp}, \qquad \left\|\frac{\partial^2 f}{\partial q^2}(p,q)\right\| \leq K_{qq}(1+\|p\|^2), \\
\left\|\frac{\partial^2 f}{\partial p \partial q}(p,q)\right\| \leq K_{pq}\sqrt{1+\|p\|^2}, \\
\left\|\frac{\partial f}{\partial p}(p,q)\right\| \leq K_p \frac{1+\|p\|^2}{\|p\|}, \qquad \left\|\frac{\partial f}{\partial q}(p,q)\right\| \leq K_q \frac{(1+\|p\|^2)^{3/2}}{\|p\|}.
\end{gathered} \tag{45}$$

From (43) we may take $K_p = \sqrt{2(K^2+1)}$, $K_{pp} = 1+K^2$, $K_q = \sqrt{2}K$, $K_{qq} = \sqrt{2}K + \frac{1}{2}K^2$, and $K_{pq} = \frac{K}{\sqrt{2}}\max\{4, 1+K\}$. To show that the gradient is bounded in the warped geometry, we require there are L_p, L_q such that $\|\frac{\partial f}{\partial p}(p,q)\| \leq L_p$ and $\|\frac{\partial f}{\partial q}(p,q)\| \leq L_q\sqrt{1+\|p\|^2}$. Clearly, we can take $L_p = K_p$ and $L_q = K_q$. Therefore f has bounded Hessian and gradient with respect to the Riemannian metric g^{w}. □

B.4 Optimization analysis

The proof of the Riemannian Adagrad-Norm uses the following Lemma, which is similar to other results used in online learning and the analysis of Adagrad; see e.g. Lemma 4.2 in (Kavis et al., 2022).

Lemma B.7. *Let $a_1, a_2, \ldots$ be a sequence of non-negative numbers. Let $c_1 > 0$ be given and for $n > 1$ define $c_n = c_{n-1} + a_{n-1}$. Then*

$$\sum_{n=1}^{T} \frac{a_n}{c_n} \leq \log\frac{c_{T+1}}{c_1} + \left(\frac{1}{c_1} - \frac{1}{c_{T+1}}\right)\max_{1\leq i\leq T} a_i \tag{46}$$

Proof of Proposition 5.6. For the Adagrad updates (11), the smoothness condition (3) implies

$$f(x_{k+1}) \leq f(x_k) - \frac{\eta}{c_k}\langle \operatorname{grad} f(x_k), v_k\rangle_{x_k} + \frac{L}{2}\frac{\eta^2}{c_k^2}\|v_k\|_{x_k}^2$$

Taking expectations,

$$\mathbb{E}[f(x_{k+1})] \leq \mathbb{E}[f(x_k)] - \mathbb{E}\left[\frac{\eta}{c_k}\|\operatorname{grad}^g f(x_k)\|_{x_k}^2\right] + \frac{L}{2}\mathbb{E}\left[\frac{\eta^2}{c_k^2}\|v_k\|_{x_k}^2\right]$$

Rearranging and summing,

$$\sum_{k=1}^{T}\mathbb{E}\left[\frac{\eta}{c_k}\|\operatorname{grad}^g f(x_k)\|_{x_k}^2\right] \leq f(x_1) - \mathbb{E}[f(x_{T+1})] + \frac{\eta^2 L}{2}\mathbb{E}\left[\sum_{k=1}^{T}\frac{\|v_k\|_{x_k}^2}{c_k^2}\right] \tag{47}$$

Using Lemma B.7, the sum on the right of (46) is bounded by

$$\sum_{k=1}^{T} \frac{\|v_k\|_{x_k}^2}{c_k^2} \leq \log \frac{c_{T+1}^2}{c_1^2} + \left(\frac{1}{c_1^2} - \frac{1}{c_{T+1}^2}\right) \max_{1\leq i\leq T} \|v_i\|_{x_i}^2 \tag{48}$$

Using that $\|v_k\|^2 \leq C$ and $c_1 = 1$, the RHS may be bounded by

$$\sum_{k=1}^{T} \frac{\|v_k\|_{x_k}^2}{c_k^2} \leq \log \frac{c_1^2 + CT}{c_1^2} + \frac{C}{c_1^2} = \log(1 + CT) + C$$

For the sum on the left of (47), note that $c_k^2 \leq 1 + (k-1)C \leq 1 + (T-1)C$ so $\frac{1}{c_k} \geq \frac{1}{\sqrt{1+(T-1)C}}$ and therefore

$$\sum_{k=1}^{T} \mathbb{E}\left[\frac{1}{c_k} \|\operatorname{grad}^g f(x_k)\|_{x_k}^2\right] \geq \frac{1}{\sqrt{1+(T-1)C}} \sum_{k=1}^{T} \mathbb{E}[\|\operatorname{grad}^g f(x_k)\|_{x_k}^2] \tag{49}$$

Combining the above (47)(48),(49), we get

$$\frac{\eta}{\sqrt{1+(T-1)C}} \sum_{k=1}^{T} \mathbb{E}[\|\operatorname{grad}^g f(x_k)\|_{x_k}^2] \leq f(x_1) - f^* + \frac{\eta^2 L}{2}(C + \log(1+CT))$$

Dividing each side of this equation by ηT and using the definitions of Δ and R,

$$\frac{1}{\sqrt{1+(T-1)C}} \mathbb{E}[\|\operatorname{grad}^g f(x_R)\|_{x_R}^2] \leq \frac{\Delta}{\eta T} + \frac{\eta L}{2} \frac{C + \log(1+CT)}{T}$$

Multiplying each side by $\sqrt{1+(T-1)C}$,

$$\mathbb{E}[\|\operatorname{grad}^g f(x_R)\|_{x_R}^2] \leq \frac{\Delta\sqrt{1+(T-1)C}}{\eta T} + \frac{\eta L}{2}\sqrt{1+(T-1)C}\frac{C + \log(1+CT)}{T}.$$

□

C Proofs for Section 6

We derive geodesics for the warped and conformal metrics. These are based on specializing existing equations for geodesics to the particular manifold and warping / scaling function we are using. We work in the general setting of a product $\mathbb{R}^n \times M_2$, where M_2 is an arbitrary Riemannian manifold. We assume M_2 is complete to avoid any issues with geodesics being defined, and we assume that the scaling / warping function is a radial function of its first component.

The idea in both cases is as follows. In general the geodesic is a curve $\gamma(t) = (p(t), q(t))$ and in both cases $p(t)$ satisfies an equation of the form

$$p''(t) = \xi_1(t)p(t) + \xi_2(t)p'(t) \tag{50}$$

for some real-valued $\xi_i(t)$. In general, equations of the form (50) have solutions that stay in the 2D subspace $\operatorname{span}\{p(0), p'(0)\}$. This is captured in the following:

Lemma C.1. *Consider a second order ODE on $\mathbb{R}^n$ for $n \geq 1$ of the form $p''(t) = \xi_1(t)p(t) + \xi_2(t)p'(t)$ where $p(0), p'(0)$ are given, and ξ_1, ξ_2 are real-valued functions. Let $V = \operatorname{span}\{p(0), p'(0)\}$ and define u_1, u_2 as follows. If $n = 1$, then $u_1 = 1$ and $u_2 = 0$. If $n \geq 2$, let u_1, u_2 be orthonormal vectors such that $V \subseteq \operatorname{span}\{u_1, u_2\}$. Then a solution $p(t)$ is given by $p(t) = y_1(t)u_1 + y_2(t)u_2$ where each y_i is the solution to*

$$y_i''(t) = \xi_1(t)y_i(t) + \xi_2(t)y_i'(t)$$

with initial conditions

$$y_i(0) = p(0)^T u_i$$
$$y_i'(0) = p'(0)^T u_i$$

Furthermore,

$$\|p(t)\|^2 = y_1(t)^2 + y_2(t)^2$$
$$\|p'(t)\|^2 = y_1'(t)^2 + y_2'(t)^2$$
$$p(t)^T p'(t) = y_1(t)y_1'(t) + y_2(t)y_2'(t)$$

For warped products, geodesics satisfy the following:

Proposition C.2 (Ch. 7, Prop 38 in (O'Neill, 1983))**.** *Let (M_1, g^1) and (M_2, g^2) be Riemannian manifolds and let $h : M_1 \to \mathbb{R}_{>0}$ be a smooth function. A curve $\gamma(t) = (p(t), q(t))$ on $M_1 \times M_2$ is a geodesic on the warped product manifold $M_1 \times_h M_2$ iff*

$$D_t p'(t) = \|q'(t)\|^2_{q(t)} (h \circ p)(t) \operatorname{grad} h(p(t)), \tag{51a}$$

$$D_t q'(t) = -2 \frac{(h \circ p)'(t)}{(h \circ p)(t)} q'(t). \tag{51b}$$

The first equation is a relation between vector fields along the curve p, which is in M, while the second involves vector fields along the curve q in N

The following further properties for warped product geodesics will also be used.

Proposition C.3 (Ch. 7, (O'Neill, 1983))**.** *Let $M \times_h N$ be a warped product manifold and let $\gamma(t) = (p(t), q(t))$ be a geodesic. Define $K = \|q'(0)\|^2 (h \circ p)(0)^4$. Then for all t,*

$$\|q'(t)\|^2_{q(t)} = \frac{K}{(h \circ p)(t)^4}. \tag{52}$$

Furthermore, $q(t)$ is a pregeodesic; that is, a reparameterization of a geodesic on M_2: $q(t) = (\tilde{q} \circ \eta)(t)$ where $\tilde{q}$ is a geodesic starting at $q(0)$ with velocity $q'(0)$, and $\eta(t)$ is the solution to the following ODE:

$$\eta'(t) = \frac{(h \circ p)(0)^2}{(h \circ p)(t)^2},$$

with initial conditions $\eta(0) = 0$.

The next proposition gives the equations for the exponential map on a warped product of the form $\mathbb{R}^n \times N$.

Proposition C.4. *Let N be a Riemannian manifold with exponential map Exp and let $h : \mathbb{R}^n \to \mathbb{R}_{>0}$ and $c : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ be smooth functions such that $h(p) = c(\|p\|_2^2)$. Let $M = \mathbb{R}^n \times N$ be the product manifold, and let $(p(0), q(0)) \in M$ and $(p'(0), q'(0)) \in T_{(p(0),q(0))} M$. If $n = 1$, set $u_1 = 1$ and $u_2 = 0$. For $n \geq 2$, set u_1, u_2 to orthonormal vectors such that $\operatorname{span}\{p(0), p'(0)\} \subseteq \operatorname{span}\{u_1, u_2\}$. Let $K = \|q'(0)\|^2_{q(0)} c(\|p(0)\|_2^2)^4$. Let $\eta : [0, 1] \to \mathbb{R}$ and $y_i : [0, 1] \to \mathbb{R}$ for $i = 1, 2$ be solutions to*

$$\begin{aligned}
\eta'(t) &= \frac{c(s(0))^2}{c(s(t))^2}, \\
y_i''(t) &= 2K \frac{c'(s(t))}{c(s(t))^3} y_i(t), \quad i = 1, 2, \\
s(t) &= y_1(t)^2 + y_2(t)^2,
\end{aligned}$$

with initial conditions $\eta(0) = 0$ and $y_i(0) = p(0)^T u_i, y_i'(0) = p'(0)^T u_i$ for $i = 1, 2$. Then the exponential map on the warped product $\mathbb{R}^n \times_h N$, is $\operatorname{Exp}_{(p(0),q(0))}(p'(0), q'(0)) = (p(1), q(1))$ where $p(1) = y_1(1)u_1 + y_2(1)u_2$ and $q(1) = \operatorname{Exp}^N_{q(0)}(\eta(1) q'(0))$.

Proof. Specializing (51a) to the setting of $\mathbb{R}^n \times N$ we obtain

$$p''(t) = \|q'(t)\|^2_{q(t)} (h \circ p)(t) \operatorname{grad} h(p(t)) \tag{53}$$

Applying (52),

$$p''(t) = \frac{K}{(h \circ p)(t)^3} \operatorname{grad} h(p(t)) \tag{54}$$

Using $h(p) = c(\|p\|_2^2)$ we have $\operatorname{grad} h(p(t)) = 2c'(\|p(t)\|^2) p(t)$ and therefore

$$p''(t) = 2K \frac{c'(\|p(t)\|^2)}{c(\|p(t)\|^2)^3} p(t) \tag{55}$$

Define $s(t) = \|p(t)\|^2$. The equation for p follows by applying Lemma C.1, with $\xi_1(t) = 2K\frac{c'(s(t))}{c(s(t))^3}$ Note that according to the Lemma $s(t) = y_1(t)^2 + y_2(t)^2$. For q we appeal to Proposition C.3, which tells us that $q = \widetilde{q} \circ \eta$, where $\widetilde{q}$ is an M_2 geodesic starting from $q(0)$ and initial velocity $q'(0)$ and η satisfies the ODE

$$\eta'(t) = \frac{c(s(0))^2}{c(s(t))^2}$$

with initial condition $\eta(0) = 0$.

□

The following proposition states the equations for the exponential on the conformally scaled products that we use.

Proposition C.5. *Let N be a manifold with exponential map Exp^N and let $h : \mathbb{R}^m \to \mathbb{R}$ and $c : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ be smooth functions such that $h(p) = \frac{1}{2}\ln c(\|p\|^2)$. Let g be the product metric on $M = \mathbb{R}^m \times N$ and let $\widetilde{g}$ be the conformal scaling $\widetilde{g} = e^{2h}g$. Then the exponential map on $(M, \widetilde{g})$ is $\operatorname{Exp}_{(p(0),q(0))}(p'(0), q'(0)) = (p(1), q(1))$ where $p(t), q(t)$ are defined as follows. For $m > 1$, set u_1, u_2 to orthonormal vectors such that $\operatorname{span}\{p(0), p'(0)\} \subseteq \operatorname{span}\{u_1, u_2\}$, while if $m = 1$, set $u_1 = 1$ and $u_2 = 0$. Let $C = c(\|p(0)\|^2)\left(\|p'(0)\|_2^2 + \|q'(0)\|_{q(0)}^2\right)$, and let $\eta : [0,1] \to \mathbb{R}$ and $y_i : [0,1] \to \mathbb{R}$ for $i = 1, 2$ be solutions to*

$$\begin{aligned}
\eta'(t) &= \frac{c(s(0))}{c(s(t))}, \\
y_i''(t) &= C\frac{c'(s(t))}{c(s(t))^2}y_i(t) - 2\frac{c'(s(t))}{c(s(t))}r(t)y_i'(t), \quad i = 1, 2, \\
s(t) &= y_1^2(t) + y_2^2(t), \\
r(t) &= y_1(t)y_1'(t) + y_2(t)y_2'(t),
\end{aligned}$$

with initial conditions $\eta(0) = 0$, and $y_i(0) = p(0)^T u_i, y_i'(0) = p'(0)^T u_i$ for $i = 1, 2$. Then $p(1) = y_1(1)u_1 + y_2(1)u_2$ and $q(1) = \operatorname{Exp}^N_{q(0)}(\eta(1)q'(0))$.

Proof. In general, for a conformal scaling $\widetilde{g} = e^{2h}g$ of a metric g, the geodesic equation is

$$D_t\gamma' = \|\gamma'\|_g^2 \operatorname{grad}^g h - 2dh(\gamma')\gamma'.$$

This follows by combining (19) with the equation $D_t\gamma' = 0$ (Lee, 2018). Let M be a product manifold $M = M_1 \times M_2$. Writing $(p, q) \in M_1 \times M_2$ for a general point of M, suppose that h depends only on p. The equation for the geodesic $\gamma(t) = (p(t), q(t))$ becomes

$$\begin{aligned}
D_t p' &= \|\gamma'\|_g^2 \operatorname{grad}^g h - 2dh(\gamma')p' \\
D_t q' &= -2dh(\gamma')q'
\end{aligned}$$

Suppose $M_1 = \mathbb{R}^m$. $M_2 = M$ and $h = \frac{1}{2}\ln c(\|p\|^2)$. Then $\operatorname{grad}^g h(p) = \frac{c'(\|p\|^2)}{c(\|p\|^2)}p$ and the equations become

$$\begin{aligned}
p''(t) &= \|\gamma'(t)\|_g^2 \frac{c'(\|p(t)\|^2)}{c(\|p(t)\|^2)}p(t) - 2\langle \operatorname{grad} h, \gamma'\rangle_g p'(t) \\
D_t q'(t) &= -2\langle \operatorname{grad} h, \gamma'\rangle_{\gamma(t)} q'(t)
\end{aligned}$$

Using again that h depends only on p,

$$\begin{aligned}
p''(t) &= \|\gamma'(t)\|_g^2 \frac{c'(\|p(t)\|^2)}{c(\|p(t)\|^2)}p(t) - 2\frac{c'(\|p(t)\|^2)}{c(\|p(t)\|^2)}\langle p(t), p'(t)\rangle_2 p'(t) \\
D_t q'(t) &= -2\frac{c'(\|p(t)\|^2)}{c(\|p(t)\|^2)}p'(t)^T p(t) q'(t)
\end{aligned}$$

Note that since $\|\cdot\|_{\widetilde{g}}^2 = e^{2h}\|\cdot\|_g^2$ and geodesics have constant velocity $\|\gamma'\|_{\widetilde{g}}^2 = C$, we have $\|\gamma'\|_g^2 = Ce^{-2h} = \frac{C}{c(\|p(t)\|^2)}$. This implies

$$p''(t) = C\frac{c'(\|p(t)\|^2)}{c(\|p(t)\|^2)^2}p(t) - 2\frac{c'(\|p\|^2)}{c(\|p\|^2)}\langle p(t), p'(t)\rangle_2 p'(t)$$

Furthermore, for q,

$$D_t q'(t) = -2\frac{c'(\|p(t)\|^2)}{c(\|p(t)\|^2)}p(t)^T p'(t)q'(t) = -\frac{(c\circ p)'(t)}{(c\circ p)(t)}q'(t)$$

In the equation for p we recognize the structure of (50) and we can apply Lemma C.1. For q, we recognize a pregeodesic structure similar to the case of the warped product (51b). Hence q takes the form $q = \widetilde{q}\circ\eta$ where $\eta'(t) = \frac{(c\circ p)(0)}{(c\circ p)(t)}$ and $\eta(0) = 0$. □

D Overhead for retraction computation

Riemannian optimization results in potentially higher time-per-iteration compared to Euclidean optimization due to the need for calculating a retraction. There are three stages of the update. Firstly, we need to compute the initial conditions and coefficient data of the ODEs. Second, the system of ODEs must be integrated. Thirdly, the next optimization state must be read out from the ODE solution. For the ODE computation, both conformal and warped cases reduce to two second order equations and one first order equation, which is expressed as five first order equations for our implementation. We solve the equations using a fourth order Runge-Kutta integrator over 32 steps. Note that since the ODE itself has a dimensionality independent of the model size, it can be expected to take up a diminishing portion of training time as model size increases.

As can be gleaned from the statements of Proposition C.4 and C.5, computing the initial conditions and performing the read out can be done with a small number of inner products and vector operations. Specifically, in terms of AXPBY operations (vector computations of the form $\alpha X + \beta Y$ for scalars α, β and vectors X, Y) our implementations of both conformal and warped geodesics require eight inner products and three AXPBY steps.

	Iteration time (ms)		
Model size	Euclidean	Warped	Conformal
12.7K - 814K	2.23 ± 0.23	$3.43 \pm 0.31\,(+54\%)$	$3.23 \pm 0.31\,(+44\%)$
6.51M	8.32 ± 0.06	$9.72 \pm 0.14\,(+17\%)$	$9.73 \pm 0.15\,(+17\%)$
52.1M	64.25 ± 0.08	$70.51 \pm 0.14\,(+10\%)$	$70.58 \pm 0.12\,(+10\%)$
208M	347.60 ± 0.25	$369.38 \pm 0.23\,(+6\%)$	$369.50 \pm 0.37\,(+6\%)$

Table 1: Time-per-iteration under Euclidean, Warped, and Conformal geometries. In parentheses is the slow down compared to the Euclidean algorithm of the same model size class. The first row is an aggregate of model sizes 12.7K, 102K and 814K.

We evaluated the overhead for the exponential computation as follows. We compute the time needed for models to train under Euclidean, warped, and conformal algorithms respectively, at increasing hidden sizes, and report the time per iteration (measured in ms) in Table 1. Time is measured for 200 minibatches, after a 100 minibatch warm up period. We find that the constant low-dimension leads to a reduction in overhead as model size increases. While for small models this is significant, at larger models this reduces to a single digit percentage overhead.

E Experiment details

All of our experiment configurations involve setting a learning rate and an scaling value for initialization of the optimization variable, where the exact method of scaling differs between experiment. Across the three tasks, all experiments at a given configuration are repeated across 5 random seeds.

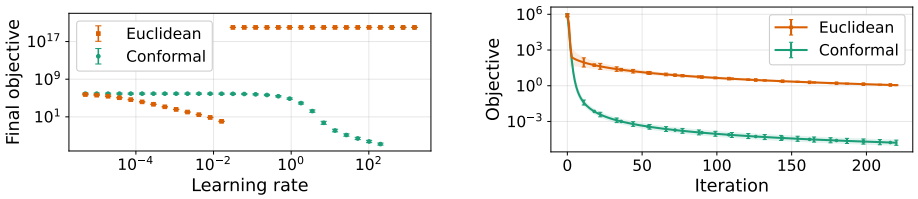

Figure 4: Diagnostic curves from the least-fourth regression experiment. Focusing on the initialization scale $r = 30$, the left plot shows the final objective values across learning rates. The plot on the right shows the trajectory of the objective at $r = 30$ at the best performing learning rates for each method, which were $\eta \approx 0.01$ for Euclidean and $\eta \approx 200$ for Conformal.

Least fourth regression. The data (x_i, y_i) were generated as follows. For each seed, the $N = 2048$ input vectors $x_i \in \mathbb{R}^n$ with $n = 32$ are sampled from the normal distribution with coordinate-wise variance $1/32$. We set p^* to be an n dimensional random vector, whence $y_i = x_i^T p^*$.

We initialize $p_0 = p^* + r\xi$ where ξ is a normal random vector and r is a tunable scaling parameter that varies over $r \in \{1, 3, 10, 30, 100, 300, 1000\}$.

At each scaling value, we sweep across 30 logarithmically spaced learning rates $\eta \in [5 \times 10^{-6}, 1.5 \times 10^3]$. Both Euclidean and Conformal use the same learning rate set. After the initial data, radius, and learning rate are determined, full batch gradient descent proceeds for 220 steps. If an objective value above 10^{20} is achieved, that seed is discarded. If more than half of seeds cross this boundary, the configuration is considered divergent. We plot the final loss values across learning rates for radius $r = 30$, as well as optimization trajectories at the best performing learning rates at that radius in Figure 4.

Robust regression with scale estimation. We assume a linear regression model $y_i = x_i^T q + \frac{1}{\tau}\xi_i$, where the ξ_i are distributed according to Student's t-distribution with ν degrees of freedom, where we fix $\nu = 3$. This results in the conditional density of y given x, q and τ as

$$g(y|x, q, \tau) = \tau \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}} \left(1 + \frac{\tau^2 (y - x^T q)^2}{\nu}\right)^{-(\nu+1)/2}$$

where the Γ function provides normalization constants. We seek to recover τ and q by minimizing the negative log likelihood (NLL) under the above distributional assumption. Parameterizing the precision as $\tau = \sqrt{1 + p^2}$, and dropping terms that are independent of q or τ results in the objective function (14).

The robust regression objective 14 is in class (w, h) for $h(p) = \sqrt{1 + p^2}$, hence it is smooth for the warped product metric. To see this, note that with the definition $\rho_3(z) = 2\log(1 + \frac{z^2}{3})$, the objective f can be written as

$$f(p, q) = -\log h(p) + \frac{1}{N} \sum_{i=1}^{N} \rho_3 \left(h(p)(x_i^T q - y_i)\right)$$

Next, note that $|\rho_3'(z)| \leq \frac{2}{\sqrt{3}}$ and $|\rho_3''(z)| \leq \frac{4}{3}$. Furthermore, relating to the redescending property of the loss ρ, we have $|z\rho_3'(z)| \leq 4$, $|z\rho_3''(z)| \leq \frac{2}{\sqrt{3}}$ and $|z^2\rho_3''(z)| \leq 4$. Using these bounds leads to the following: There are constants K_p, K_q, K_{pp}, K_{pq} and K_{qq} such that

$$\|\tfrac{\partial f}{\partial p}(p, q)\| \leq K_p, \quad \|\tfrac{\partial f}{\partial q}(p, q)\| \leq h(p) K_q$$

$$\|\tfrac{\partial^2 f}{\partial p^2}(p, q)\| \leq K_{pp}, \quad \|\tfrac{\partial^2 f}{\partial q^2}(p, q)\| \leq h(p)^2 K_{qq}$$

$$\|\tfrac{\partial^2 f}{\partial p \partial q}(p, q)\| \leq K_{p,q}.$$

Since $0 < |h'(p)| < 1$ when $p \neq 0$ and $h(p) > 1$ it is evident that $\|\frac{\partial f}{\partial p}(p, q)\| \leq K_p h(p) |\frac{\partial h}{\partial p}(p)|^{-1}$. The condition on $\|\frac{\partial f}{\partial q}(p, q)\|$ required by Definition 4.2 holds similarly.

Throughout all experiments, the target scale is $\tau^* = 100$, and q^* is fixed $q^* = (-0.8, -0.6, \ldots, 0.4, 0.6) \in \mathbb{R}^8$. At each seed, the data x_i are sampled from a standard normal and q_0 is initialized as $q_0 = q^* + u$ where u is a standard normal random vector. In order to evaluate the sensitivity of the geometries to their starting point, p_0 is initialized at values $p_0 \in \{1, 20, 40, 60, 80\}$, which we term the "initial scale".

In our experiments, we sweep across learning rates η. In the Euclidean case, η ranges across 13 logarithmically spaced values in $[10^{-5}, 10]$, while in the warped case, η ranges over 15 logarithmically spaced values in $[2.5 \times 10^{-3}, 10^2]$. At each learning rate and seed value, we perform batch gradient descent for 10^4 steps starting from each possible initial scale, and record the final NLL as well as error metrics $|\tau - \tau^*|$ and $\|q - q^*\|$. At each learning rate and scale, we average the final NLL across seeds, and then, for each learning rate, identify the scale resulting in the worst NLL. From this worst performing scale, we plot the error metrics for τ and q in the right plot of Figure 2.

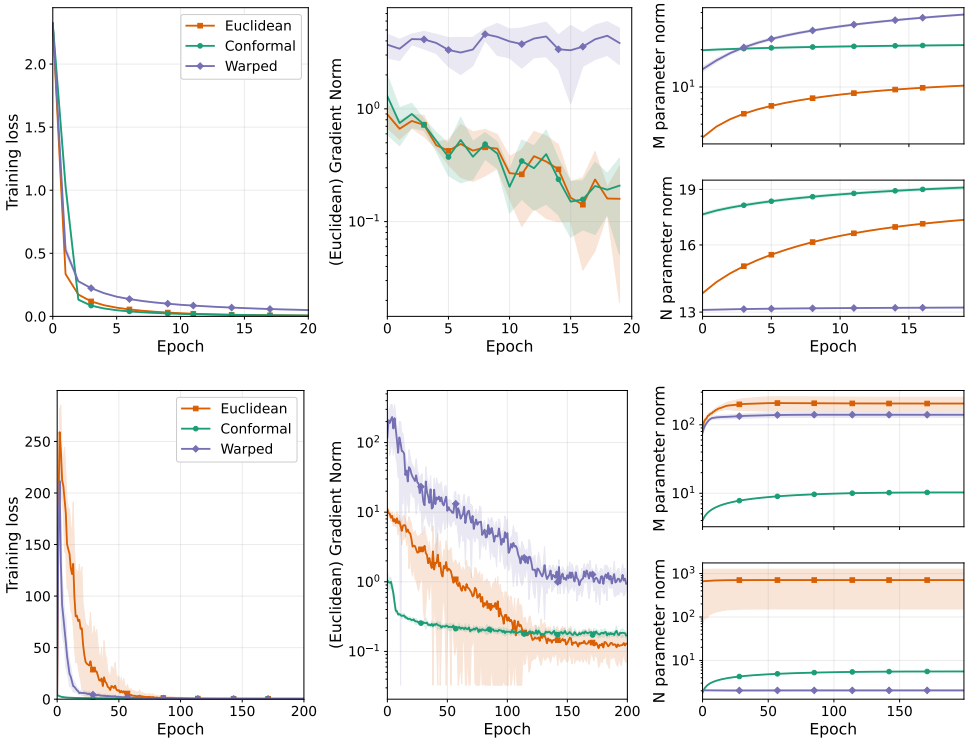

Figure 5: Training loss, gradient norms, and parameter norms for MNIST (top) and GTSRB (bottom),

Image classification. For MNIST, our model is an MLP with one hidden layer of size 512 and we use the $\tanh$ activation, and we perform standard preprocessing for that dataset. For the 43 class GTSRB dataset, we first resize the variable-sized images to 32×32 and each channel is standardized. The model for GTSRB is a single hidden layer network, where the first layer is a convolutional layer with a 5×5 kernel and 8 output channels, and we employ a 4×4 average pooling after the nonlinear activation. The activation function use is the error function erf. In both cases, the standard cross entropy loss is used.

To evaluate optimization behavior across scales, we scale up the output weights and biases of each model after it is randomly initialized. For MNIST, our scales range over $\{1, 100, 300, 1000\}$ while for GTSRB they range over $\{1, 10, 30, 100\}$.

For MNIST, we use a batch size of 256 and sweep over 10 approximately logarithmically spaced learning rates in $[10^{-2}, 3 \times 10^2]$. At each step size and initialization scale we run optimization for 20. For GTSRB, we employ a batch size of 4096 and sweep over 10 rates distributed in $[10^{-1}, 6]$. For each step size and scale, we train for 200 epochs. In the right plots of Figure 5, the impact of the

metric can be seen in the trajectory of parameter norms. In both datasets, the warped metric has the effect of suppressing the input layer parameters (component N).

To estimate the Hessian norm, we use a power-iteration based approach. After initialization, a fixed minibatch of size 512 is chosen, and within the power iteration loop, two applications of automatic differentiation are used to evaluate the Hessian-vector products needed Pearlmutter (1994); Yao et al. (2020). We note that while in the Euclidean case, the bilinear map that needs to be evaluated is $v^T Av$ where $A = \mathrm{Hess}^{\mathrm{Euc}}$ is the Euclidean Hessian, in the warped and conformal cases, A is a combination of the Euclidean Hessian and other terms; specifically, in the conformal case it is given by (19) while the warped case follows from Prop B.2. A total of 30 power iteration steps are used in each case.