A Riemannian AdaGrad-Norm Method

Glaydston de C. Bento * Geovani N. Grapiglia † Mauricio S. Louzeiro‡
Daoping Zhang§

September 29, 2025

Abstract

We propose a manifold AdaGrad-Norm method (MAdaGrad), which extends the norm version of AdaGrad (AdaGrad-Norm) to Riemannian optimization. In contrast to line-search schemes, which may require several exponential map computations per iteration, MAdaGrad requires only one. Assuming the objective function f has Lipschitz continuous Riemannian gradient, we show that the method requires at most $\mathcal{O}(\varepsilon^{-2})$ iterations to compute a point x such that $\|\operatorname{grad} f(x)\| \leq \varepsilon$. Under the additional assumptions that f is geodesically convex and the manifold has sectional curvature bounded from below, we show that the method takes at most $\mathcal{O}(\varepsilon^{-1})$ to find x such that $f(x) - f_{low} \leq \epsilon$, where f_{low} is the optimal value. Moreover, if f satisfies the Polyak–Łojasiewicz condition globally on the manifold, we establish a complexity bound of $\mathcal{O}(\log(\varepsilon^{-1}))$, provided that the norm of the initial Riemannian gradient is sufficiently large. For the manifold of symmetric positive definite matrices, we construct a family of nonconvex functions satisfying the PL condition. Numerical experiments illustrate the remarkable performance of MAdaGrad in comparison with Riemannian Steepest Descent equipped with Armijo line-search.

Key words: Riemannian Optimization · Gradient Method · Adaptive Methods · Worst-Case · Complexity Bounds
AMS subject classification: 65K05 · 68Q25 ·90C30 · 49M37

1 Introduction

1.1 Motivation and Contributions

In this work we consider the minimization of a differentiable function $f : M \to \mathbb{R}$, where M is a Riemannian manifold [2, 15, 5]. Problems of this type appear in many important applications such as Low-Rank Matrix Completion [28], Dictionary Learning [7] and Independent Component Analysis [24]. Many optimization algorithms originally developed for the Euclidean setting ($M = \mathbb{R}^n$)

*Universidade Federal de Goiás, IME, Avenida Esperança, s/n, Campus Samambaia, CEP 74690-900, Goiânia, GO, Brazil. e-mail: glaydston@ufg.br.
†Université Catholique de Louvain, ICTEAM/INMA, Avenue Georges Lemaître, 4-6/L4.05.01, B-1348, Louvain-la-Neuve, Belgium. e-mail: geovani.grapiglia@uclouvain.be.
‡Universidade Federal de Goiás, IME, Avenida Esperança, s/n, Campus Samambaia, CEP 74690-900, Goiânia, GO, Brazil. e-mail: mauriciolouzeiro@ufg.br.
§Nankai University, School of Mathematical Sciences and LPMC, Tianjin 300071, China. e-mail: daopingzhang@nankai.edu.cn.

have been extended to Riemannian optimization. Notable examples include variants of the gradient method [25, 8], Newton and conjugate gradient methods [25], quasi-Newton methods [16], trust-region methods [1], and cubic regularization of Newton's method [3], among others. Special attention has been devoted to adaptive methods, which automatically select suitable stepsizes, trust-region radii, or regularization parameters without requiring prior knowledge of the problem-specific constants.

A classical example of an adaptive scheme is the gradient method with Armijo line search, which defines the iterates as

$$x_{k+1} = \exp_{x_k}\left(-\alpha_k \omega^{\ell_k} \operatorname{grad} f(x_k)\right),$$

where ℓ_k is the smallest nonnegative integer ℓ such that

$$f\left(\exp_{x_k}\left(-\alpha_k \omega^{\ell} \operatorname{grad} f(x_k)\right)\right) \leq f(x_k) - \rho\alpha_k\omega^{\ell}\|\operatorname{grad} f(x_k)\|^2, \tag{1}$$

with $\rho, \omega \in (0, 1)$ and $\alpha_0 > 0$ being user-defined parameters. In practice, the values $\ell = 0, 1, 2, \ldots$ are tested sequentially until inequality (1) is satisfied. This backtracking procedure may require multiple evaluations of the exponential map $\exp_{x_k}(\cdot)$, which can make the method computationally expensive. To mitigate this issue, the RWNGrad method was recently proposed in [14] as a Riemannian counterpart of the WNGrad method, originally developed for Euclidean optimization in [29]. Specifically, RWNGrad sets

$$\begin{cases} x_{k+1} &= \exp_{x_k}\left(-\dfrac{1}{\beta_k}\operatorname{grad} f(x_k)\right), \quad \beta_0 > 0 \\ \beta_{k+1} &= \beta_k + \dfrac{\|\operatorname{grad} f(x_k)\|^2}{\beta_k}, \end{cases} \tag{2}$$

thus requiring a single evaluation of the exponential map at each iteration. It was proved in [14] that RWNGrad needs no more than $\mathcal{O}\left(\epsilon^{-2}\right)$ iterations to find x_k such that $\|\operatorname{grad} f(x_k)\| \leq \epsilon$. More importantly, numerical experiments showed that RWNGrad is significantly faster than the Gradient Method with Armijo line search on problems over the manifold $M = \mathbb{P}^n_{++}$ of $(n \times n)$ symmetric and positive definite (SPD) matrices.

Motivated by the encouraging numerical performance of RWNGrad, in this paper we investigate a related yet distinct adaptive strategy, aiming to obtain a Riemannian algorithm with improved numerical performance and stronger theoretical guarantees compared to RWNGrad. Specifically, we propose MAdaGrad (Manifold AdaGrad-Norm), a Riemannian extension of the AdaGrad-Norm method [29]. In contrast to RWNGrad (2), MAdaGrad sets

$$\begin{cases} \beta_{k+1} &= \beta_k + \|\operatorname{grad} f(x_k)\|^2, \quad \beta_0 = 0, \\ x_{k+1} &= \exp_{x_k}\left(-\dfrac{\eta}{\sqrt{\beta_{k+1}}}\operatorname{grad} f(x_k)\right), \end{cases} \tag{3}$$

with $\eta > 0$ being a user-defined parameter. Regarding the theoretical guarantees of MAdaGrad (3), we establish iteration-complexity bounds under various assumptions. When the objective function $f(\cdot)$ is nonconvex, the method achieves a complexity of $\mathcal{O}(\epsilon^{-2})$. This rate improves to $\mathcal{O}(\epsilon^{-1})$ when $f(\cdot)$ is convex and the manifold M has sectional curvature bounded below by a negative constant, or when $f(\cdot)$ is possibly nonconvex but satisfies the Polyak–Łojasiewicz (PL) condition globally. In addition, we provide a family of nonconvex functions over the manifold of symmetric positive definite matrices $\mathbb{P}^n_{++}$ that satisfy the PL condition. Finally, our numerical experiments demonstrate that MAdaGrad can significantly outperform both RWNGrad and the gradient method with Armijo line search.

1.2 Related Literature

In recent years, the challenge of training machine learning models has motivated the development and analysis of numerous adaptive variants of the Stochastic Gradient Descent (SGD) method, including AdaGrad [11], RMSProp [26], Adam [17], and AMSGrad [21]. A key feature of these methods is the use of distinct stepsizes for updating each component of the iterate. For example, the batch version of AdaGrad applied to the minimization of $f : \mathbb{R}^n \to \mathbb{R}$ defines the iterates by

$$\begin{cases} \beta_{k+1} &= \beta_k + \nabla f(x_k) \odot \nabla f(x_k), \quad \beta_k = 0 \in \mathbb{R}^n, \\ [x_{k+1}]_i &= [x_k]_i + \frac{\eta}{\sqrt{[\beta_{k+1}]_i}}[\nabla f(x_k)]_i, \quad i = 1, \ldots, n, \end{cases}$$

where $[\nabla f(x_k) \odot \nabla f(x_k)]_i = [\nabla f(x_k)]_i^2$ for $i = 1, \ldots, n$. The component-wise nature of adaptive methods such as AdaGrad complicates their extension to the Riemannian setting, due to the lack of intrinsic coordinate systems on general manifolds. In [6], this issue was addressed by considering the special case where M is a Cartesian product of Riemannian manifolds. By exploiting this additional structure, the authors proposed Riemannian extensions of AdaGrad, Adam, and AMSGrad. A different approach was considered in [22], where the authors presented a Riemannian adaptive method that encompasses extensions of RMSProp, Adam, and AMSGrad for the case where M is an embedded submanifold of Euclidean space. Therefore, given the component-wise nature of this class of adaptive methods, their generalization to the Riemannian setting typically requires additional assumptions on the manifold M, which restricts their applicability. For this reason, in the present work we focus on generalizing AdaGrad-Norm [29], which, similar to WNGrad [30], does not rely on a coordinate system and can thus be extended to general Riemannian optimization problems.

1.3 Contents

The remainder of the paper is organized as follows. In Section 2, we introduce the necessary concepts and notations from Riemannian geometry. Section 3 presents the MAdaGrad algorithm and establishes key auxiliary results concerning its iterates and step sizes. In Section 4, we derive iteration-complexity bounds for the nonconvex, convex, and PL cases. We also provide a family of nonconvex functions that satisfy the PL property globally. Finally, in Section 5 we report numerical results.

2 Preliminary

In this section, we recall some concepts, notations, and basic results about Riemannian manifolds. For more details see, for example, [9, 23, 27, 20].

We denote by T_pM the *tangent space* of a Riemannian manifold M at p. The corresponding norm associated to the Riemannian metric $\langle \cdot\ ,\ \cdot \rangle$ is denoted by $\| \ \cdot \ \|$. We use $\ell(\gamma)$ to denote the length of a piecewise smooth curve $\gamma : [a, b] \to M$. The Riemannian distance between p and q in a finite-dimensional Riemannian manifold M is denoted by $d(p, q)$. This distance induces the original topology on M, so that (M, d) becomes a complete metric space. Let $(N, \langle \cdot\ ,\ \cdot \rangle)$ and $(M, \langle\!\langle \cdot\ ,\ \cdot \rangle\!\rangle)$ be Riemannian manifolds and $\Phi : N \to M$ be an isometry, that is, Φ is C^∞, and for all $q \in N$ and $u, v \in T_qN$, we have $\langle u, v\rangle = \langle\!\langle d\Phi_q u, d\Phi_q v\rangle\!\rangle$, where $d\Phi_q : T_qN \to T_{\Phi(q)}M$ is the differential of Φ at $q \in N$. One can verify that Φ preserves geodesics, that is, β is a geodesic in N if and only if $\Phi \circ \beta$ is a geodesic in M. Denote by $\mathcal{X}(M)$, the space of smooth vector fields on M. Let ∇ be the Levi-Civita connection associated to $(M, \langle \cdot\ ,\ \cdot \rangle)$. The Riemannian metric induces a mapping $f \mapsto \operatorname{grad} f$

that associates each differentiable function to its *gradient* via the rule $\langle \operatorname{grad} f, X\rangle = df(X)$, for all $X \in \mathcal{X}(M)$. A vector field V along γ is said to be *parallel* iff $\nabla_{\gamma'} V = 0$. If γ' itself is parallel, we say that γ is a *geodesic*. Given that the geodesic equation $\nabla_{\gamma'}\gamma' = 0$ is a second order nonlinear ordinary differential equation, then the geodesic $\gamma = \gamma_v(\cdot, p)$ is determined by its position p and velocity v at p. It is easy to check that $\|\gamma'\|$ is constant. The restriction of a geodesic to a closed bounded interval is called a *geodesic segment*. A geodesic segment joining p to q in M is said to be *minimal* if its length is equal to $d(p, q)$. For each $t \in [a, b]$, ∇ induces an isometry, relative to $\langle\cdot,\cdot\rangle$, $P_{\gamma,a,t} \colon T_{\gamma(a)}M \to T_{\gamma(t)}M$ defined by $P_{\gamma,a,t}\, v = V(t)$, where V is the unique vector field on γ such that $\nabla_{\gamma'(t)}V(t) = 0$ and $V(a) = v$, the so-called *parallel transport* along the geodesic segment γ joining $\gamma(a)$ to $\gamma(t)$. When there is no confusion, we consider the notation $P_{\gamma,p,q}$ for the parallel transport along the geodesic segment γ joining p to q. A Riemannian manifold is *complete* if the geodesics are defined for any values of $t \in \mathbb{R}$. Hopf-Rinow's theorem asserts that any pair of points in a complete Riemannian manifold M can be joined by a (not necessarily unique) minimal geodesic segment. A set $\Omega \subseteq M$ is said to be *convex* iff any geodesic segment with end points in Ω is contained in Ω. A function $f : M \to \mathbb{R}$ is said to be *convex* on a convex set Ω iff for any geodesic segment $\gamma : [a, b] \to \Omega$, the composition $f \circ \gamma : [a, b] \to \mathbb{R}$ is convex. Owing to the completeness of the Riemannian manifold M, the *exponential map* $\exp_p : T_pM \to M$ can be given by $\exp_p v = \gamma_v(1, p)$, for each $p \in M$. A complete, simply connected Riemannian manifold of non-positive sectional curvature is called a *Hadamard manifold*. For all $p \in M$, the exponential map $\exp_p : T_pM \to M$ is a diffeomorphism and $\exp_p^{-1} : M \to T_pM$ denotes its inverse. In this case, $d(q\,,\,p) = ||\exp_p^{-1} q||$ and the function $d_q^2 : M \to \mathbb{R}$ defined by $d_q^2(p) := d^2(q, p)$ is C^∞ and $\operatorname{grad} d_q^2(p) := -2\exp_p^{-1} q$.

In this paper, all manifolds are assumed to be connected, finite dimensional, and complete.

3 A Riemannian AdaGrad-Norm Method

Consider the problem

$$\min_{x \in M} f(x), \tag{4}$$

where M is a Riemannian manifold and $f \colon M \to \mathbb{R}$ is a differentiable function. As introduced in [8], given $L \geq 0$, the gradient vector fields $\operatorname{grad} f$ is said to be L-Lipschitz continuous if, for any points p and $q \in M$ and γ, a geodesic segment joining p to q, one has $\|P_{\gamma,p,q}\operatorname{grad} f(p) - \operatorname{grad} f(q)\| \leq L d(p, q)$.

Let us assume that:

A1. $\operatorname{grad} f$ is L-Lipschitz continuous;

A2. f has a global minimizer, with optimal value denoted by f^*.

Below we propose a Riemannian generalization of the batch version of method AdaGrad-Norm [29, 31].

Algorithm 1. Riemannian AdaGrad-Norm (MAdaGrad).

Step 0. Given $x_0 \in M$ and $\eta > 0$, set $\beta_0 = 0$ and $k := 0$.

Step 1. If grad $f(x_k) = 0$, then **stop**; otherwise, compute

$$\beta_{k+1} = \beta_k + \|\mathrm{grad} f(x_k)\|^2, \tag{5}$$

$$\alpha_k = \frac{\eta}{\sqrt{\beta_{k+1}}}, \tag{6}$$

$$x_{k+1} = \exp_{x_k}\left(-\alpha_k \mathrm{grad} f(x_k)\right). \tag{7}$$

Step 2. Set $k := k + 1$ and go to Step 1.

The next lemma is a consequence of [8, Lemma 5.1] and has appeared in [4].

Lemma 3.1. *Suppose that A1 holds. Then,*

$$f(\exp_p v) \le f(p) + \langle \mathrm{grad} f(p), v\rangle + \frac{L}{2}\|v\|^2, \qquad p \in M, \quad v \in T_pM. \tag{8}$$

Lemma 3.2. *Suppose that A1 holds and let $\{x_k\}_{k\ge 0}$ be generated by Algorithm 1. Then, the following hold:*

i) $f(x_{k+1}) \le f(x_k) + \dfrac{L\alpha_k^2}{2}(\beta_{k+1} - \beta_k)$ *for all* $k \ge 0$;

ii) if $\alpha_k \le 1/L$ *for some* $k \ge 0$, *then* $f(x_k) - f(x_{k+1}) \ge \dfrac{\alpha_k}{2}\|\mathrm{grad} f(x_k)\|^2$;

iii) if $\beta_{k+1} < \eta^2 L^2$ *for all* $k = 0, \ldots, k_0 - 1$, *where* $k_0 \ge 1$, *then*

$$\alpha_k > \frac{1}{L}, \qquad k = 0, \ldots, k_0 - 1, \tag{9}$$

$$\sum_{k=0}^{k_0-1} \alpha_k \|\mathrm{grad} f(x_k)\|^2 \le \frac{\eta^3 L^2}{\|\mathrm{grad} f(x_0)\|}, \tag{10}$$

$$f(x_{k_0}) \le f(x_0) + \frac{\eta^4 L^3}{2\|\mathrm{grad} f(x_0)\|^2}. \tag{11}$$

Proof. By using (8) with $p = x_k$ and $v = -\alpha_k \,\mathrm{grad}\, f(x_k)$, and (7), we obtain

$$\begin{aligned} f(x_{k+1}) &\le f(x_k) - \alpha_k\|\mathrm{grad} f(x_k)\|^2 + \frac{L\alpha_k^2}{2}\|\mathrm{grad} f(x_k)\|^2 \\ &\le f(x_k) + \frac{L\alpha_k^2}{2}\|\mathrm{grad} f(x_k)\|^2 \end{aligned} \tag{12}$$

for all $k \ge 0$. The proof of item *i*) follows by combining the last inequality in (12) with (5), and the proof of item *ii*) follows by using the first inequality in (12) along with the fact that $-L\alpha_k \ge -1$

for some $k \geq 0$. Now, let us assume that $\beta_{k+1} < \eta^2 L^2$ for $k = 0, \ldots, k_0 - 1$, for some $k_0 \geq 1$. Consequently, we have $\sqrt{\beta_{k+1}} < \eta L$ for $k = 0, \ldots, k_0 - 1$, and the proof of (9) is an immediate consequence of (6). On the other hand, by using (6) again, we obtain

$$\begin{aligned}\sum_{k=0}^{k_0-1} \alpha_k \|\mathrm{grad} f(x_k)\|^2 &= \sum_{k=0}^{k_0-1} \frac{\eta}{\sqrt{\beta_{k+1}}} \|\mathrm{grad} f(x_k)\|^2 \\ &\leq \frac{\eta}{\|\mathrm{grad} f(x_0)\|} \sum_{k=0}^{k_0-1} \|\mathrm{grad} f(x_k)\|^2, \end{aligned} \tag{13}$$

where the last inequality follows from the fact that

$$\beta_{k+1} \geq \beta_1 = \beta_0 + \|\mathrm{grad} f(x_0)\|^2 = \|\mathrm{grad} f(x_0)\|^2,$$

which is an immediate consequence of (5) and $\beta_0 = 0$. Combining the inequality in (13) with (5), we get

$$\sum_{k=0}^{k_0-1} \alpha_k \|\mathrm{grad} f(x_k)\|^2 \leq \frac{\eta}{\|\mathrm{grad} f(x_0)\|} \sum_{k=0}^{k_0-1} (\beta_{k+1} - \beta_k) = \frac{\eta}{\|\mathrm{grad} f(x_0)\|} \beta_{k_0}. \tag{14}$$

Hence, (10) is obtained directly by combining (14) with the inequality $\beta_{k_0} < \eta^2 L^2$, which follows by taking $k = k_0 - 1$ in $\sqrt{\beta_{k+1}} < \eta L$. To conclude the proof of item *iii*), note that

$$f(x_{k_0}) - f(x_0) = \sum_{k=0}^{k_0-1} (f(x_{k+1}) - f(x_k)) \leq \sum_{k=0}^{k_0-1} \frac{L\alpha_k^2}{2} (\beta_{k+1} - \beta_k),$$

where the last inequality is obtained from item *i*). On the other hand, from (6), we have $\alpha_k^2 = \eta^2/\beta_{k+1}$, which, combined with the last inequality and using again that $\beta_{k+1} \geq \|\mathrm{grad} f(x_0)\|^2$ for all $k \in \mathbb{N}$, yields

$$\begin{aligned} f(x_{k_0}) - f(x_0) &\leq \frac{L\eta^2}{2} \sum_{k=0}^{k_0-1} \frac{1}{\beta_{k+1}} (\beta_{k+1} - \beta_k) \\ &\leq \frac{L\eta^2}{2\|\mathrm{grad} f(x_0)\|^2} \sum_{k=0}^{k_0-1} (\beta_{k+1} - \beta_k) \\ &= \frac{L\eta^2}{2\|\mathrm{grad} f(x_0)\|^2} \beta_{k_0}. \end{aligned}$$

Therefore, (11) follows by using the fact that $\beta_{k_0} < \eta^2 L^2$, concluding the proof. □

Lemma 3.3. *Suppose that A1 and A2 hold and let $\{x_k\}_{k\geq 0}$ be generated by Algorithm 1. If*

$$k_0 = \inf \left\{k \in \mathbb{N} \colon \beta_{k+1} \geq \eta^2 L^2\right\} < +\infty, \tag{15}$$

then

$$\sum_{k=k_0}^{T-1} \alpha_k \|\,\mathrm{grad}\, f(x_k)\|^2 \leq 2 \left(f(x_0) - f^* + \frac{\eta^4 L^3}{2\|\,\mathrm{grad}\, f(x_0)\|^2} \right), \qquad \forall\, T > k_0, \tag{16}$$

and

$$\alpha_k \geq \left(L + \frac{2(f(x_0) - f^*)}{\eta^2} + \frac{\eta^2 L^3}{\|\,\mathrm{grad}\, f(x_0)\|^2} \right)^{-1}, \qquad \forall\, k \geq k_0. \tag{17}$$

Proof. In view of (6), (5), and (15), we have

$$\alpha_k = \frac{\eta}{\sqrt{\beta_{k+1}}} \le \frac{\eta}{\sqrt{\beta_{k_0+1}}} \le \frac{1}{L}, \qquad k \ge k_0.$$

Thus, by Lemma 3.2 (ii), it follows that

$$f(x_k) - f(x_{k+1}) \ge \frac{\alpha_k}{2}\|\operatorname{grad} f(x_k)\|^2, \qquad k \ge k_0.$$

Summing up these inequalities for $k = k_0, \ldots, T-1$, and using A2, we get

$$\sum_{k=k_0}^{T-1} \frac{\alpha_k}{2}\|\operatorname{grad} f(x_k)\|^2 \le f(x_{k_0}) - f^*. \tag{18}$$

If $k_0 = 0$, then it follows from (18) that (16) is true. If $k_0 \ge 1$, then $\beta_{k+1} < \eta^2 L^2$ for $k = 0, \ldots, k_0 - 1$. Thus, by inequality (11) in Lemma 3.2 we have

$$f(x_{k_0}) - f^* \le f(x_0) - f^* + \frac{\eta^4 L^3}{2\|\operatorname{grad} f(x_0)\|^2}. \tag{19}$$

Therefore, combining (18) and (19) we see that (16) is also true when $k_0 \ge 1$. On the other hand, notice that

$$\begin{aligned}\frac{\alpha_k}{2}\|\operatorname{grad} f(x_k)\|^2 &= \frac{\eta}{2}\frac{\beta_{k+1} - \beta_k}{\sqrt{\beta_{k+1}}} \\ &= \frac{\eta}{2}\frac{(\sqrt{\beta_{k+1}} - \sqrt{\beta_k})(\sqrt{\beta_{k+1}} + \sqrt{\beta_k})}{\sqrt{\beta_{k+1}}} \ge \frac{\eta}{2}(\sqrt{\beta_{k+1}} - \sqrt{\beta_k}).\end{aligned} \tag{20}$$

Now, combining (16) and (20), it follows that

$$\begin{aligned}\frac{\eta}{2}(\sqrt{\beta_T} - \sqrt{\beta_{k_0}}) &= \sum_{k=k_0}^{T-1} \frac{\eta}{2}(\sqrt{\beta_{k+1}} - \sqrt{\beta_k}) \le \sum_{k=k_0}^{T-1} \frac{\alpha_k}{2}\|\operatorname{grad} f(x_k)\|^2 \\ &\le f(x_0) - f^* + \frac{\eta^4 L^3}{2\|\operatorname{grad} f(x_0)\|^2},\end{aligned}$$

which implies

$$\begin{aligned}\sqrt{\beta_T} &\le \sqrt{\beta_{k_0}} + \frac{2(f(x_0) - f^*)}{\eta} + \frac{\eta^3 L^3}{\|\operatorname{grad} f(x_0)\|^2} \\ &\le \eta L + \frac{2(f(x_0) - f^*)}{\eta} + \frac{\eta^3 L^3}{\|\operatorname{grad} f(x_0)\|^2}.\end{aligned}$$

Since T is an arbitrary integer bigger than k_0, we have

$$\sqrt{\beta_{k+1}} \le \eta L + \frac{2(f(x_0) - f^*)}{\eta} + \frac{\eta^3 L^3}{\|\operatorname{grad} f(x_0)\|^2}, \qquad k \ge k_0.$$

Consequently,

$$\alpha_k = \frac{\eta}{\sqrt{\beta_{k+1}}} \ge \left(L + \frac{2(f(x_0) - f^*)}{\eta^2} + \frac{\eta^2 L^3}{\|\operatorname{grad} f(x_0)\|^2}\right)^{-1}, \qquad k \ge k_0,$$

which implies that (17) is true. $\square$

Lemma 3.4. *Suppose that A1 and A2 hold and let $\{x_k\}_{k\geq 0}$ be generated by Algorithm 1. Then,*

$$\alpha_k \geq \left(L + \frac{2(f(x_0)-f^*)}{\eta^2} + \frac{\eta^2 L^3}{\|\operatorname{grad} f(x_0)\|^2}\right)^{-1} \equiv \alpha_{\min}, \qquad k \geq 0, \tag{21}$$

and

$$\sum_{k=0}^{T-1} \alpha_k \|\operatorname{grad} f(x_k)\|^2 \leq \frac{\eta^3 L^2}{\|\operatorname{grad} f(x_0)\|} + 2(f(x_0)-f^*) + \frac{\eta^4 L^3}{\|\operatorname{grad} f(x_0)\|^2}, \qquad T \geq 1. \tag{22}$$

Proof. Let us divide the proof in two cases.

Case 1: $k_0 = \inf\left\{k \in \mathbb{N} \colon \beta_{k+1} \geq \eta^2 L^2\right\} = +\infty$.

In this case, we have $\beta_{k+1} < \eta^2 L^2$ for all $k \geq 0$. Thus, it follows from Lemma 3.2 (iii) that

$$\alpha_k > \frac{1}{L}, \quad \forall k \geq 0 \quad \text{and} \quad \sum_{k=0}^{T-1} \alpha_k \|\operatorname{grad} f(x_k)\|^2 \leq \frac{\eta^3 L^2}{\|\operatorname{grad} f(x_0)\|}, \qquad T \geq 1.$$

Therefore, (21) and (22) hold.

Case 2: $k_0 = \inf\left\{k \in \mathbb{N} \colon \beta_{k+1} \geq \eta^2 L^2\right\} < +\infty$.

In this case, if $k_0 = 0$, then (21) and (22) follow directly from Lemma 3.3. If $k_0 \geq 1$, it follows from Lemmas 3.2 (iii) and 3.3 that $\alpha_k > 1/L$ for $k = 0, \ldots, k_0 - 1$, and $\alpha_k \geq \alpha_{\min}$ for $k \geq k_0$. Therefore, (21) is true for all $k \geq 0$. Moreover, given $T \geq 1$, we have two possibilities.

Subcase 2.1: $T \leq k_0$.

In this subcase, it follows from Lemma 3.2 that

$$\sum_{k=0}^{T-1} \alpha_k \|\operatorname{grad} f(x_k)\|^2 \leq \sum_{k=0}^{k_0-1} \alpha_k \|\operatorname{grad} f(x_k)\|^2 \leq \frac{\eta^3 L^2}{\|\operatorname{grad} f(x_0)\|},$$

and so (22) is true.

Subcase 2.2: $T > k_0$.

In this subcase, by Lemma 3.2 (iii) and Lemma 3 3.3 we have

$$\begin{aligned}\sum_{k=0}^{T-1} \alpha_k \|\operatorname{grad} f(x_k)\|^2 &= \sum_{k=0}^{k_0-1} \alpha_k \|\operatorname{grad} f(x_k)\|^2 + \sum_{k=k_0}^{T-1} \alpha_k \|\operatorname{grad} f(x_k)\|^2 \\ &\leq \frac{\eta^3 L^2}{\|\operatorname{grad} f(x_0)\|} + 2\left(f(x_0) - f^*\right) + \frac{\eta^4 L^3}{\|\operatorname{grad} f(x_0)\|^2},\end{aligned}$$

that is, (22) is true. □

4 Worst-Case Complexity Bounds

In this section, we establish iteration-complexity bounds for MAdaGrad (3). We show that the method achieves a complexity of $\mathcal{O}(\epsilon^{-2})$, which improves to $\mathcal{O}(\epsilon^{-1})$ both when the objective function is convex and when it globally satisfies the Polyak–Łojasiewicz (PL) condition. It is worth mentioning that, in the convex case, we assume that the manifold M has sectional curvature bounded below by a negative constant.

4.1 General Case

Theorem 4.1. *Suppose that A1-A3 hold and let* $\{x_k\}_{k\geq 0}$ *be generated by Algorithm 1. Given* $\epsilon > 0$*, let*

$$T_g(\epsilon) = \inf\{k \in \mathbb{N} : \|\operatorname{grad} f(x_k)\| \leq \epsilon\}.$$

Then

$$T_g(\epsilon) \leq \left[\frac{\eta^3 L^2}{\alpha_{\min}\|\operatorname{grad} f(x_0)\|} + \frac{2(f(x_0) - f^*)}{\alpha_{\min}} + \frac{\eta^4 L^3}{\alpha_{\min}\|\operatorname{grad} f(x_0)\|^2}\right]\epsilon^{-2}, \tag{23}$$

where $\alpha_{\min}$ *is defined in (21).*

Proof. If $T_g(\epsilon) = 0$, then (23) is true. Thus, let us assume that $T_g(\epsilon) \geq 1$. By Lemma 3.4, we have

$$\begin{aligned}
&\frac{\eta^3 L^2}{\|\operatorname{grad} f(x_0)\|} + 2(f(x_0) - f^*) + \frac{\eta^4 L^3}{\|\operatorname{grad} f(x_0)\|^2} \\
&\geq \sum_{k=0}^{T_g(\epsilon)-1} \alpha_k \|\operatorname{grad} f(x_k)\|^2 \\
&\geq \alpha_{\min} \sum_{k=0}^{T_g(\epsilon)-1} \|\operatorname{grad} f(x_k)\|^2 \\
&\geq \alpha_{\min} T_g(\epsilon)\epsilon^2.
\end{aligned}$$

Then, isolating $T_g(\epsilon)$, we conclude that (23) also holds in this case. □

4.2 Convex Case

Lemma 4.2. *Suppose that A1 and A2 hold, and let* $\{x_k\}_{k\geq 0}$ *be generated by Algorithm 1. Then,*

$$\sum_{k=0}^{\infty} \alpha_k^2 \|\operatorname{grad} f(x_k)\|^2 \leq \rho \equiv \frac{\eta}{\|\operatorname{grad} f(x_0)\|}\left[\frac{\eta^3 L^2}{\|\operatorname{grad} f(x_0)\|} + 2(f(x_0) - f^*) + \frac{\eta^4 L^3}{\|\operatorname{grad} f(x_0)\|^2}\right]. \tag{24}$$

Proof. From (5) and (6), we have $\alpha_k \leq \eta/\|\operatorname{grad} f(x_0)\|$ for all $k \geq 0$, which implies

$$\sum_{k=0}^{\infty} \alpha_k^2 \|\operatorname{grad} f(x_k)\|^2 \leq \frac{\eta}{\|\operatorname{grad} f(x_0)\|} \sum_{k=0}^{\infty} \alpha_k \|\operatorname{grad} f(x_k)\|^2.$$

Thus, the proof of (24) follows from Lemma 3.4. □

Now, let us consider the following assumptions:

A3. M has sectional curvature bounded below by a negative constant, i.e., $K \geq \kappa$ with $\kappa < 0$;

A4. $f : M \to \mathbb{R}$ is convex on M and admits a minimizer q with $f^* = f(q)$.

Taking into account Lemma 4.2, the next lemma follows from [13, Lemma 3.6].

Lemma 4.3. *Suppose that A3 and A4 hold, and let $\{x_k\}_{k\geq 0}$ be the sequence generated by Algorithm 1. Then, for each $k \geq 0$, the following inequality holds:*

$$d^2(x_{k+1}, q) \leq d^2(x_k, q) + \mathcal{K}^q_{\rho,\kappa}\alpha_k^2 \|\mathrm{grad}\, f(x_k)\|^2 + 2\alpha_k[f^* - f(x_k)],$$

where

$$\mathcal{K}^q_{\rho,\kappa} := \frac{\sinh\left(\hat{\kappa}\sqrt{\rho}\right)}{\hat{\kappa}\sqrt{\rho}} \frac{\mathcal{C}^q_{\rho,\kappa}}{\tanh \mathcal{C}^q_{\rho,\kappa}} \qquad \mathcal{C}^q_{\rho,\kappa} := \cosh^{-1}\left(\cosh(\hat{\kappa} d(x_0, q)) e^{\frac{1}{2}\left(\hat{\kappa}\sqrt{\rho}\right)\sinh\left(\hat{\kappa}\sqrt{\rho}\right)}\right), \tag{25}$$

with ρ is defined in (24) *and $\hat{\kappa} \equiv \sqrt{|\kappa|}$.*

Theorem 4.4. *Suppose that A1–A4 hold, and let $\{x_k\}_{k\geq 0}$ be the sequence generated by Algorithm 1. Given $\epsilon > 0$, let*

$$T_f(\epsilon) = \inf\left\{k \in \mathbb{N} \,:\, f(x_k) - f^* \leq \epsilon\right\}. \tag{26}$$

Then

$$T_f(\epsilon) \leq \left(\frac{d^2(x_0, q) + \rho\mathcal{K}^q_{\rho,\kappa}}{2\alpha_{\min}}\right)\epsilon^{-1}, \tag{27}$$

where $\alpha_{\min}$, ρ and $\mathcal{K}^q_{\rho,\kappa}$ are defined in (21), (24) *and* (25)*, respectively.*

Proof. If $T_f(\epsilon) = 0$, then (27) is true. Thus, let us assume that $T_f(\epsilon) \geq 1$. By combining Lemma 4.3 with (21), we obtain

$$f(x_k) - f^* \leq \frac{d^2(x_k, q) - d^2(x_{k+1}, q) + \mathcal{K}^q_{\rho,\kappa}\alpha_k^2 \|\,\mathrm{grad}\, f(x_k)\|^2}{2\alpha_{\min}},$$

for all $k \geq 0$. Summing this inequality over $k = 0, \ldots, T_f(\epsilon) - 1$ and applying Lemma 4.2, we obtain

$$\begin{aligned}
\epsilon &< \min\left\{f(x_k) - f^* \colon k = 0, \ldots, T_f(\epsilon) - 1\right\} \leq \frac{1}{T_f(\epsilon)} \sum_{k=0}^{T_f(\epsilon)-1} \left(f(x_k) - f^*\right) \\
&\leq \frac{1}{T_f(\epsilon)} \sum_{k=0}^{T_f(\epsilon)-1} \frac{d^2(x_k, q) - d^2(x_{k+1}, q) + \mathcal{K}^q_{\rho,\kappa}\alpha_k^2 \|\,\mathrm{grad}\, f(x_k)\|^2}{2\alpha_{\min}} \\
&\leq \frac{1}{T_f(\epsilon)} \left(\frac{d^2(x_0, q) + \rho\mathcal{K}^q_{\rho,\kappa}}{2\alpha_{\min}}\right).
\end{aligned}$$

Therefore, isolating $T_f(\epsilon)$ we conclude that (27) is true. $\square$

4.3 μ-Polyak-Lojasiewicz Case

Throughout this section, the results are established taking into account the following assumption:

> **A5.** $f : M \to \mathbb{R}$ has a minimizer $q \in M$, with $f^* = f(q)$, and there exists $\mu > 0$ such that
>
> $$f(x) - f^* \leq \frac{1}{\mu}\|\,\mathrm{grad}\, f(x)\|^2, \qquad x \in M.$$

To the best of our knowledge, the inequality in A5 was first introduced by Polyak in [19] within the context of linear optimization. In this seminal work, the inequality played an important role in the asymptotic convergence analysis of the classical gradient method.

Lemma 4.5. *Suppose that A1 and A5 hold and let $\{x_k\}_{k\geq 0}$ be generated by Algorithm 1. If*

$$\beta_{k+1} < \eta^2 L^2, \quad \text{for } k = 0, \dots, T-1 \tag{28}$$

and

$$T \geq 1 + \left[\frac{\eta^2 L^2}{\mu}\epsilon^{-1} + 1\right] \log\left(\frac{\eta^2 L^2}{\|\operatorname{grad} f(x_0)\|^2}\right) \tag{29}$$

for some $\epsilon > 0$, then $\min\{f(x_k) - f^* \colon k = 0, \dots, T-1\} \leq \epsilon$.

Proof. By (5), we have

$$\begin{aligned}
&\beta_1 = \|\operatorname{grad} f(x_0)\|^2 \\
&\beta_2 = \beta_1\left(1 + \frac{\|\operatorname{grad} f(x_1)\|^2}{\beta_1}\right) = \|\operatorname{grad} f(x_0)\|^2\left(1 + \frac{\|\operatorname{grad} f(x_1)\|^2}{\beta_1}\right) \\
&\beta_3 = \beta_2\left(1 + \frac{\|\operatorname{grad} f(x_1)\|^2}{\beta_2}\right) = \|\operatorname{grad} f(x_0)\|^2\left(1 + \frac{\|\operatorname{grad} f(x_1)\|^2}{\beta_1}\right)\left(1 + \frac{\|\operatorname{grad} f(x_2)\|^2}{\beta_2}\right) \\
&\quad\vdots \\
&\beta_T = \|\operatorname{grad} f(x_0)\|^2 \prod_{k=1}^{T-1}\left(1 + \frac{\|\operatorname{grad} f(x_k)\|^2}{\beta_k}\right).
\end{aligned} \tag{30}$$

Using (28), (30) and A5, it follows that

$$\begin{aligned}
\eta^2 L^2 > \beta_T &= \|\operatorname{grad} f(x_0)\|^2 \prod_{k=1}^{T-1}\left[1 + \frac{\|\operatorname{grad} f(x_k)\|^2}{\beta_k}\right] \\
&\geq \|\operatorname{grad} f(x_0)\|^2 \prod_{k=1}^{T-1}\left[1 + \frac{\mu(f(x_k) - f^*)}{\eta^2 L^2}\right] \\
&\geq \|\operatorname{grad} f(x_0)\|^2 \prod_{k=1}^{T-1}\left[1 + \frac{\mu}{\eta^2 L^2}\min\{f(x_k) - f^* \colon k = 1, \dots, T-1\}\right] \\
&= \|\operatorname{grad} f(x_0)\|^2 \left[1 + \frac{\mu}{\eta^2 L^2}\min\{f(x_k) - f^* \colon k = 1, \dots, T-1\}\right]^{T-1}.
\end{aligned}$$

Now, suppose by contradiction that $\min\{f(x_k) - f^* \colon k = 0, \dots, T-1\} > \epsilon$. Then, applying the previous inequality and using the fact that the logarithm function is increasing, we obtain

$$\begin{aligned}
\log\left(\frac{\eta^2 L^2}{\|\operatorname{grad} f(x_0)\|^2}\right) &> (T-1)\log\left(1 + \frac{\mu\epsilon}{\eta^2 L^2}\right) \\
&\geq (T-1)\frac{\frac{\mu\epsilon}{\eta^2 L^2}}{1 + \frac{\mu\epsilon}{\eta^2 L^2}} = (T-1)\left[\frac{\eta^2 L^2}{\mu}\epsilon^{-1} + 1\right]^{-1},
\end{aligned}$$

which contradicts (29). This completes the proof. $\square$

Lemma 4.6. *Suppose that A1 and A5 hold and let* $\{x_k\}_{k\geq 0}$ *be generated by Algorithm 1. If*

$$k_0 = \min\left\{k \in \mathbb{N}\colon \beta_{k+1} \geq \eta^2 L^2\right\} < +\infty \tag{31}$$

and

$$T_0 \geq \frac{\left|\log\left(\left[f(x_0) - f^* + \frac{\eta^4 L^3}{2\|\operatorname{grad} f(x_0)\|^2}\right]\epsilon^{-1}\right)\right|}{\left|\log\left(1 - \frac{\mu\alpha_{\min}}{2}\right)\right|} \tag{32}$$

for some $\epsilon > 0$ *and for* $\alpha_{\min}$ *defined in (21), then*

$$f(x_{k_0+T_0}) - f^* \leq \epsilon. \tag{33}$$

Proof. By (31), (6), and (5), we have $\alpha_k = \eta/\sqrt{\beta_{k+1}} \leq 1/L$ for all $k \geq k_0$. Thus, by Lemma 3.2 (ii), A5, and (21), we have

$$f(x_k) - f(x_{k+1}) \;\geq\; \frac{\alpha_k}{2}\|\operatorname{grad} f(x_k)\|^2 \;\geq\; \frac{\mu\alpha_{\min}}{2}\left(f(x_k) - f^*\right), \qquad \forall k \geq k_0. \tag{34}$$

From this, it follows that

$$1 - \frac{\mu\alpha_{\min}}{2} \in (0,1). \tag{35}$$

Furthermore, (34) implies that

$$f(x_{k+1}) - f^* \leq \left(1 - \frac{\mu\alpha_{\min}}{2}\right)\left(f(x_k) - f^*\right), \qquad \forall k \geq k_0.$$

Hence,

$$f(x_{k+1}) - f^* \leq \left(1 - \frac{\mu\alpha_{\min}}{2}\right)^{k-k_0+1}\left(f(x_{k_0}) - f^*\right), \qquad \forall k \geq k_0. \tag{36}$$

If $k_0 \geq 1$, it follows from inequality (11) in Lemma 3.2 and from (31) that

$$f(x_{k_0}) - f^* \leq f(x_0) - f^* + \frac{\eta^4 L^3}{2\|\operatorname{grad} f(x_0)\|^2}. \tag{37}$$

Clearly (37) is also true when $k_0 = 0$. If $T_0 = 0$, then it follows from (32) and (37) that (33) is true. Now consider the case $T_0 \geq 1$. By combining (36) with $k = k_0 + T_0 - 1$ and (37), we obtain

$$f(x_{k_0+T_0}) - f^* \leq \left(1 - \frac{\mu\alpha_{\min}}{2}\right)^{T_0}\left(f(x_0) - f^* + \frac{\eta^4 L^3}{2\|\operatorname{grad} f(x_0)\|^2}\right). \tag{38}$$

Thus, it follows from (32), (35) and (38) that (33) holds. Indeed, otherwise, we would have

$$\epsilon < \left(1 - \frac{\mu\alpha_{\min}}{2}\right)^{T_0}\left(f(x_0) - f^* + \frac{\eta^4 L^3}{2\|\operatorname{grad} f(x_0)\|^2}\right),$$

which, by (35) and the properties of the logarithm, leads to

$$T_0\left|\log\left(1 - \frac{\mu\alpha_{\min}}{2}\right)\right| < \log\left(\left(f(x_0) - f^* + \frac{\eta^4 L^3}{2\|\operatorname{grad} f(x_0)\|^2}\right)\epsilon^{-1}\right),$$

contradicting (32). □

Theorem 4.7. *Suppose that A1 and A5 hold and let $\{x_k\}_{k\geq 0}$ be generated by Algorithm 1. For each $\epsilon > 0$, define $T_f(\epsilon) = \inf\{k \in \mathbb{N} : f(x_k) - f^* \leq \epsilon\}$. If $\|\operatorname{grad} f(x_0)\| \geq \eta L$, then*

$$T_f(\epsilon) < 1 + \frac{\left|\log\left(\left[f(x_0) - f^* + \frac{\eta^4 L^3}{2\|\operatorname{grad} f(x_0)\|^2}\right]\epsilon^{-1}\right)\right|}{\left|\log\left(1 - \frac{\mu\alpha_{\min}}{2}\right)\right|}, \tag{39}$$

where $\alpha_{\min}$ is defined in (21). Otherwise,

$$\begin{aligned} T_f(\epsilon) \quad &< \quad 1 + \left[\frac{\eta^2 L^2}{\mu}\epsilon^{-1} + 1\right]\log\left(\frac{\eta^2 L^2}{\|\operatorname{grad} f(x_0)\|^2}\right) \\ &\quad + \frac{\left|\log\left(\left[f(x_0) - f^* + \frac{\eta^4 L^3}{2\|\operatorname{grad} f(x_0)\|^2}\right]\epsilon^{-1}\right)\right|}{\left|\log\left(1 - \frac{\mu\alpha_{\min}}{2}\right)\right|}. \end{aligned} \tag{40}$$

Proof. First, let us consider the case $\|\operatorname{grad} f(x_0)\| \geq \eta L$. Then, it follows from (5) with $k = 0$ that $\beta_1 = \|\operatorname{grad} f(x_0)\|^2 \ \geq \ \eta^2 L^2$, which guarantees the equality

$$k_0 \ := \ \inf\left\{k \in \mathbb{N} \ : \ \beta_{k+1} \geq \eta^2 L^2\right\} = 0. \tag{41}$$

If $T_f(\epsilon) = 0$, then (39) holds. Therefore, suppose that $T_f(\epsilon) \geq 1$. In this case, by the definition of $T_f(\epsilon)$, we have

$$f(x_{T_f(\epsilon)-1}) - f^* > \epsilon. \tag{42}$$

Thus, in view of (41) and (42), it follows from the contrapositive of Lemma 4.6 that we must have

$$T_f(\epsilon) - 1 < \frac{\left|\log\left(\left[f(x_0) - f^* + \frac{\eta^4 L^3}{2\|\operatorname{grad} f(x_0)\|^2}\right]\epsilon^{-1}\right)\right|}{\left|\log\left(1 - \frac{\mu\alpha_{\min}}{2}\right)\right|},$$

which establishes (39).

Now, suppose that $\|\operatorname{grad} f(x_0)\| < \eta L$. If $T_f(\epsilon) = 0$, then (40) holds. So, assume that $T_f(\epsilon) \geq 1$. Let us divide the rest of the analysis in two cases.

Case 1: $k_0 := \inf\left\{k \in \mathbb{N} \ : \ \beta_{k+1} \geq \eta^2 L^2\right\} = +\infty$.

In this case, in particular, we have

$$\beta_{k+1} < \eta^2 L^2, \quad \text{for } k = 0, \ldots, T_f(\epsilon) - 1. \tag{43}$$

Moreover, by the definition of $T_f(\epsilon)$ we also have

$$\min\{f(x_k) - f^* \colon k = 0, 1, \ldots, T_f(\epsilon) - 1\} > \epsilon. \tag{44}$$

Thus, in view of (43) and (44), it follows from that contrapositive of Lemma 4.5 that we must have

$$T_f(\epsilon) < 1 + \left[\frac{\eta^2 L^2}{\mu}\epsilon^{-1} + 1\right]\log\left(\frac{\eta^2 L^2}{\|\operatorname{grad} f(x_0)\|^2}\right).$$

Therefore, (40) is true in this case.

Case 2: $k_0 := \inf\left\{k \in \mathbb{N} \ : \ \beta_{k+1} \geq \eta^2 L^2\right\} < +\infty$.

From the definition of k_0 we have

$$\beta_{k+1} < \eta^2 L^2, \quad \text{for } k = 0, \ldots, k_0 - 1. \tag{45}$$

Regarding the relation between $T_f(\epsilon)$ and k_0, there are only two possibilities.

Subcase 2.1: $T_f(\epsilon) \le k_0$.

In this subcase, by (45) we have

$$\beta_{k+1} < \eta^2 L^2, \quad \text{for } k = 0, \ldots, T_f(\epsilon) - 1.$$

Therefore, as in Case 1, we conclude that (40) is true.

Subcase 2.2: $T_f(\epsilon) = k_0 + T_0$ for some $T_0 \ge 1$.

In this case, in addition to (45), we also have

$$\min\{f(x_k) - f^* \colon k = 0, 1, \ldots, k_0\} \ge \min\{f(x_k) - f^* \colon k = 0, 1, \ldots, T_f(\epsilon) - 1\} > \epsilon.$$

Thus, by the contrapositive of Lemma 4.5 com $T = k_0 + 1$, it follows that

$$k_0 < \left[\frac{\eta^2 L^2}{\mu}\epsilon^{-1} + 1\right] \log\left(\frac{\eta^2 L^2}{\| \operatorname{grad} f(x_0)\|^2}\right). \tag{46}$$

If $T_0 = 1$, then it follows from (46) that

$$T_f(\epsilon) = k_0 + T_0 < 1 + \left[\frac{\eta^2 L^2}{\mu}\epsilon^{-1} + 1\right] \log\left(\frac{\eta^2 L^2}{\| \operatorname{grad} f(x_0)\|^2}\right),$$

and so (40) is true. On the other hand, if $T_0 \ge 2$, then

$$f(x_{k_0+T_0-1}) - f^* = f(x_{T_f(\epsilon)-1}) - f^* > \epsilon.$$

Thus, by the contrapositive of Lemma 4.6 we must have

$$T_0 - 1 < \frac{\left|\log\left(\left[f(x_0) - f^* + \frac{\eta^4 L^3}{2\| \operatorname{grad} f(x_0)\|^2}\right]\epsilon^{-1}\right)\right|}{\left|\log\left(1 - \frac{\mu\alpha_{\min}}{2}\right)\right|}. \tag{47}$$

By combining (46) and (47) with the fact that $T_f(\epsilon) = k_0 + T_0$, we conclude that (40) also holds in this subcase. □

4.3.1 A Class of Nonconvex PL functions on SPD matrices

In this section, we provide a class of nonconvex functions that satisfy Assumption A5 on a particular Hadamard manifold. Let $\mathbb{R}^{n\times n}$ be the set of real matrices of order $n \times n$, $\mathbb{P}^n \subset \mathbb{R}^{n\times n}$ the set of symmetric matrices, and $\mathbb{P}^n_{++} \subset \mathbb{R}^{n\times n}$ the cone of symmetric positive definite matrices. Define

$$\langle U, V\rangle_X := \operatorname{tr}(VX^{-1}UX^{-1}), \quad X \in \mathbb{P}^n_{++},\ U, V \in \mathbb{P}^n, \tag{48}$$

where $\operatorname{tr}(\cdot)$ denotes the trace operator. It is well known that $M = (\mathbb{P}^n_{++}, \langle\cdot,\cdot\rangle)$ is a Hadamard manifold (see, for example, [18, Theorem 1.2, Page 325]), and that $T_X M$ can be identified with $\mathbb{P}^n$

for every $X \in M$. The Riemannian gradient and Riemannian Hessian of $f : \mathbb{P}^n_{++} \to \mathbb{R}$ are given, respectively, by

$$\mathrm{grad} f(X) = X f'(X) X, \tag{49}$$

$$\mathrm{hess}\, f(X)V = X f''(X) V X + \frac{1}{2}\left[V f'(X) X + X f'(X) V\right], \tag{50}$$

where $V \in T_X M$, and $f'(X)$ and $f''(X)$ denote the Euclidean gradient and Hessian of f at X, respectively, with respect to the Frobenius metric.

Consider the class of functions $f : \mathbb{P}^n_{++} \to \mathbb{R}$ defined by

$$f(X) = a \ln^4(\det(X)) - b \ln^3(\det(X)) - \frac{b^3}{a^2} \ln(\det(X)), \tag{51}$$

where $a, b > 0$. Since

$$f'(X) = \left[4a \ln^3(\det(X)) - 3b \ln^2(\det(X)) - \frac{b^3}{a^2}\right] X^{-1}, \tag{52}$$

it follows from (49) that

$$\mathrm{grad}\, f(X) = \left[4a \ln^3(\det(X)) - 3b \ln^2(\det(X)) - \frac{b^3}{a^2}\right] X. \tag{53}$$

Therefore, the set of critical points of f is $\Omega \equiv \left\{X \in \mathbb{P}^n_{++} :\ \det(X) = e^{b/a}\right\}$. Moreover, (51) implies that $f(X) = -b^4/a^3$ for all $X \in \Omega$. Given this and the coercivity of f, we conclude that $f^* = -b^4/a^3$. Consequently, using (48), (51), and (53), along with appropriate algebraic manipulations, we obtain

$$\begin{aligned}
\frac{\|\,\mathrm{grad}\, f(X)\|^2}{f(X) - f^*} \quad &= \quad \frac{\left[4a \ln^3(\det(X)) - 3b \ln^2(\det(X)) - b^3/a^2\right]^2 n}{a \ln^4(\det(X)) - b \ln^3(\det(X)) - (b^3/a^2) \ln(\det(X)) + (b^4/a^3)} \\
&= \quad \frac{\left[[\ln(\det(X)) - b/a]\left[4a \ln^2(\det(X)) + b \ln(\det(X)) + b^2/a\right]\right]^2 n}{[\ln(\det(X)) - b/a]^2 \left[a \ln^2(\det(X)) + b \ln(\det(X)) + b^2/a\right]} \\
&= \quad \frac{9na^2 \ln^4(\det(X))}{a \ln^2(\det(X)) + b \ln(\det(X)) + b^2/a} \\
&\quad\quad + n\left(7a \ln^2(\det(X)) + b \ln(\det(X)) + b^2/a\right) \\
&\geq \quad n\left(7a \ln^2(\det(X)) + b \ln(\det(X)) + b^2/a\right) \geq (27nb^2)/(28a)
\end{aligned}$$

for all $X \in \mathbb{P}^n_{++} \setminus \Omega$, which shows that the function f defined in (51) satisfies A5 for every $0 < \mu \leq (27nb^2)/(28a)$. On the other hand, denoting the Euclidean norm by $\|\cdot\|_e$, it follows from (52) that

$$\begin{aligned}
&\frac{\|f'(X)\|_e^2}{f(X) - f^*} \\
&\quad = \left[\frac{9a^2 \ln^4(\det(X))}{a \ln^2(\det(X)) + b \ln(\det(X)) + b^2/a} + 7a \ln^2(\det(X)) + b \ln(\det(X)) + b^2/a\right] \mathrm{tr}\left(X^{-2}\right),
\end{aligned}$$

and, denoting by I_n the $n \times n$ identity matrix, we obtain

$$\begin{aligned}
\inf_{X \in \mathbb{P}^n_{++}} \frac{\|f'(X)\|_e^2}{f(X) - f^*} &= \lim_{t \to +\infty} \frac{\|f'(tI_n)\|_e^2}{f(tI_n) - f^*} \\
&= \lim_{t \to +\infty} \left[\frac{9a^2 \ln^4(t^n)}{a \ln^2(t^n) + b \ln(t^n) + b^2/a} + 7a \ln^2(t^n) + b \ln(t^n) + b^2/a\right] \frac{n}{t^2} \\
&= 0,
\end{aligned}$$

which implies that there exists no $\mu > 0$ such that the function f defined in (51) satisfies A5 in the Euclidean setting.

Now, observe that

$$f''(X)V = \left[12a\ln^2(\det(X)) - 6b\ln(\det(X))\right]\operatorname{tr}(X^{-1}V)X^{-1} \\ - \left[4a\ln^3(\det(X)) - 3b\ln^2(\det(X)) - \frac{b^3}{a^2}\right]X^{-1}VX^{-1},$$

for all $X \in \mathbb{P}^n_{++}$ and $V \in \mathbb{P}^n$. By combining this equality with (52) and (50), we obtain

$$\operatorname{hess} f(X)V = \left[12a\ln^2(\det(X)) - 6b\ln(\det(X))\right]\operatorname{tr}(X^{-1}V)X,$$

for all $X \in \mathbb{P}^n_{++}$ and $V \in \mathbb{P}^n$, which implies that

$$\langle \operatorname{hess} f(X)V, V\rangle = -3b^2/(4a)\|V\|^2 < 0$$

for all $X \in \left\{X \in \mathbb{P}^n_{++} \colon \det(X) = e^{b/(4a)}\right\}$ and $V \in \mathbb{P}^n$. Therefore, f is not convex.

5 Numerical Results

We evaluated the relative performance of Algorithm 1 by testing its Matlab implementation with $\eta = 10$ (MAdaGrad) against the Riemannian Gradient Method with Armijo line search [12] and the RWNGrad [14]. The experiments were conducted on two classes of test problems.

All implementations were performed in Matlab R2022a on a MacBook Pro equipped with an Apple M1 Pro processor and 16 GB of RAM. To ensure reproducibility, we fixed the randomness using Matlab's built-in function `rng(2025)`.

5.1 Problem Class 1

We considered class of problems of the form

$$\min_{X \in \mathbb{P}^n_{++}} f(X) \equiv \ln(\det(X))^2 - \ln(\det(X)). \tag{54}$$

For $n = 10$, the codes were run from 100 starting points, randomly generated with eigenvalues in the interval $(0, 20)$. Following [12, 14], each starting point X_0 was constructed as $X_0 = Q^T\Gamma Q$, where Γ is a diagonal matrix whose entries are independent random variables uniformly distributed in $(0, 20)$, and Q is obtained from the QR decomposition of a matrix with entries uniformly generated in $(0, 1)$. Figure 1 shows the performance profiles [10] with respect to CPU time for finding X such that $\|\operatorname{grad} f(X)\| \leq 10^{-4}$, with each code allowed a maximum of $1,000$ iterations. As can be seen, MAdaGrad is significantly faster than the other two methods.

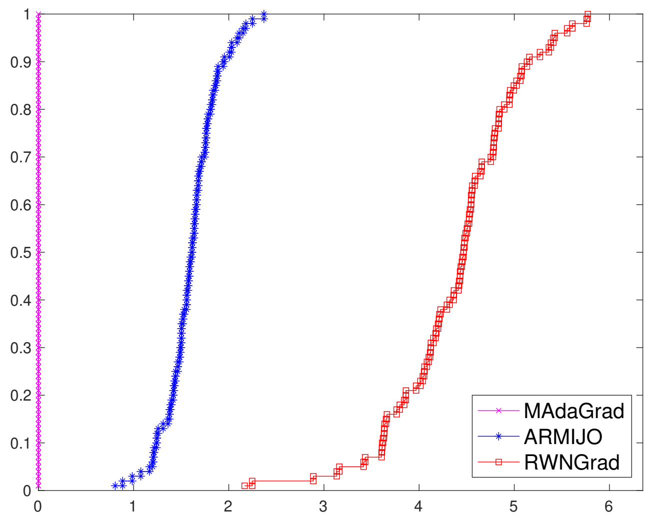

Figure 1: The Performance profiles (in $\log_2$ scale) with respect to CPU time for Problem 1. The magenta line corresponds to MAdaGrad, the blue line to ARMIJO, and the red line to RWNGrad.

5.2 Problem Class 2

We also considered the class of problems of the form

$$\min_{X\in\mathbb{P}^n_{++}} f(X) \equiv \frac{1}{2}\sum_{j=1}^{m}\left\|\ln\left(X^{-1/2}A_jX^{-1/2}\right)\right\|_F^2, \tag{55}$$

for fixed $A_1,\ldots,A_m \in \mathbb{P}^n_{++}$. For $n = 20$ and $m = 5$, we ran 100 test problems generated by randomly constructing 100 sets of matrices $A_1,\ldots,A_m \in \mathbb{P}^n_{++}$. As in [12, 14], each matrix A_j was constructed as $A_j = Q_j^T\Lambda_jQ_j$, where Λ_j is a diagonal matrix whose entries are independent random variables uniformly distributed in $(0, 20)$, and Q_j is obtained from the QR decomposition of a matrix with entries uniformly generated in $(0, 1)$. For each problem, the initial point X_0 was chosen as

$$X_0 = \exp\left(\frac{1}{m}\sum_{j=1}^{m}\ln(A_j)\right).$$

Figure 2 shows the performance profiles [10] with respect to CPU time for finding X such that $\|\operatorname{grad} f(X)\| \le 10^{-4}$, with each code allowed a maximum of 1,000 iterations. Once again, our method MAdaGrad is considerably faster than the other two methods.

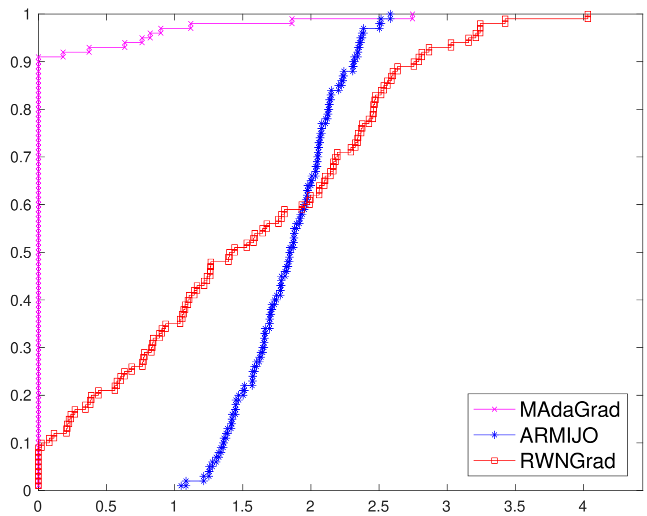

Figure 2: The Performance profiles (in $\log_2$ scale) with respect to CPU time for Problem 2. The magenta line corresponds to MAdaGrad, the blue line to ARMIJO, and the red line to RWNGrad.

6 Conclusion

In this paper, we have introduced MAdaGrad, a novel generalization of AdaGrad-Norm to Riemannian optimization. We established iteration complexity guarantees in several regimes: $\mathcal{O}(\varepsilon^{-2})$ for finding ε-stationary points under Lipschitz continuous Riemannian gradients; $\mathcal{O}(\varepsilon^{-1})$ for geodesically convex objectives on manifolds with sectional curvature bounded from below; and $\mathcal{O}(\log(\varepsilon^{-1}))$ under a global Polyak–Łojasiewicz condition. Furthermore, we constructed nonconvex functions on the manifold $\mathbb{P}^n_{++}$ of symmetric positive definite matrices that satisfy the PL condition. Numerical experiments confirmed the efficiency of MAdaGrad, showing consistent improvements over Riemannian Steepest Descent with Armijo line-search [12] and the RWNGrad method [14] on optimization problems over $\mathbb{P}^n_{++}$.

References

[1] P.-A. Absil, C. Baker, and K. Gallivan. Trust-region methods on riemannian manifolds. *Foundations of Computational Mathematics*, 7:303–330, 2007.

[2] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, Princeton, NJ, 2008. With a foreword by Paul Van Dooren.

[3] N. Agarwal, N. Boumal, B. Bullins, and C. Cartis. Adaptive regularization with cubics on manifolds. *Math. Program.*, 188(1):85–134, 2021.

[4] G. C. Bento, O. P. Ferreira, and J. G. Melo. Iteration-complexity of gradient, subgradient and proximal point methods on riemannian manifolds. *Journal of Optimization Theory and Applications*, 173(2):548–562, 2017.

[5] N. Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023.

[6] G. Bécigneul and O.-E. Ganea. Riemannian adaptive optimization methods. In *7th International Conference on Learning Representations (ICLR)*, 2019.

[7] A. Cherian and S. Sra. Dictionary learning and sparse coding for positive definite matrices. *IEEE Transactions on Neural Networks and Learning Systems*, 28, 2017.

[8] J. Cruz Neto, L. Lima, and P. Oliveira. Geodesic algorithms in riemannian geometry. *Balkan J. Geom. Appl*, 3(2):89–100, 1998.

[9] M. P. do Carmo. *Riemannian geometry*. Mathematics: Theory & Applications. Birkhäuser Boston, Inc., Boston, MA, 1992. Translated from the second Portuguese edition by Francis Flaherty.

[10] E. Dolan and J. Moré. Benchmarking optimization software with performance profiles. *Mathematical Programming*, 91:201–2013, 2002.

[11] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, pages 257–269. JMLR Workshop and Conference Proceedings, 2011.

[12] O. P. Ferreira, M. S. Louzeiro, and L. F. Prudente. Gradient method for optimization on Riemannian manifolds with lower bounded curvature. *SIAM J. Optim.*, 29(4):2517–2541, 2019.

[13] O. P. Ferreira, M. S. Louzeiro, and L. F. Prudente. Iteration-complexity and asymptotic analysis of steepest descent method for multiobjective optimization on riemannian manifolds. *Journal of Optimization Theory and Applications*, 184:507–533, 2020.

[14] G. N. Grapiglia and G. F. Stella. An adaptive riemannian gradient method without function evaluations. *Journal of Optimization Theory and Applications*, 197(3):1140–1160, 2023.

[15] J. Hu, X. Liu, and Y.-x. Yuan. A brief introduction to manifold optimization. *Journal of the Operations Research Society of China*, 8:199–248, 2020.

[16] W. Huang, K. A. Gallivan, and P.-A. Absil. A broyden class of quasi-newton methods for riemannian optimization. *SIAM Journal on Optimization*, 25(3):1660–1685, 2015.

[17] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015.

[18] S. Lang. *Fundamentals of differential geometry*, volume 191 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1999.

[19] B. T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

[20] T. Rapcsák. *Smooth nonlinear optimization in* $\mathbf{R}^n$, volume 19 of *Nonconvex Optimization and its Applications*. Kluwer Academic Publishers, Dordrecht, 1997.

[21] S. J. Reddi, S. Kale, and S. Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations (ICLR)*, 2018.

[22] H. Sakai and H. Iiduka. A general framework of riemannian adaptive optimization methods with a convergence analysis. *Transactions on Machine Learning Research (TMLR)*, 2025. Accepted for publication.

[23] T. Sakai. *Riemannian geometry*, volume 149 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 1996. Translated from the 1992 Japanese original by the author.

[24] S. E. Selvan, U. Amato, K. A. Gallivan, C. Qi, M. F. Carfora, M. Larobina, and B. Alfano. Descent algorithms on oblique manifold for source-adaptive ica contrast. *IEEE Transactions on Neural Networks and Learning Systems*, 23(12):1930–1947, 2012.

[25] S. T. Smith. Optimization techniques on Riemannian manifolds. In *Hamiltonian and gradient flows, algorithms and control*, volume 3 of *Fields Inst. Commun.*, pages 113–136. Amer. Math. Soc., Providence, RI, 1994.

[26] T. Tieleman and G. Hinton. Lecture 6.5—rmsprop: Divide the gradient by a running average of its recent magnitude, 2012. COURSERA: Neural Networks for Machine Learning.

[27] C. Udrişte. *Convex functions and optimization methods on Riemannian manifolds*, volume 297 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1994.

[28] B. Vandereycken. Low-rank matrix completion by riemannian optimization. *SIAM Journal on Optimization*, 23:1214–1236, 2013.

[29] R. Ward, X. Wu, and L. Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research*, 21(219):1–30, 2020.

[30] L. Wu, R. Ward, and L. Bottou. Wngrad: Learn the learning rate in gradient descent. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 5290–5298. PMLR, 2018.

[31] Y. Xie, X. Wu, and R. Ward. Linear convergence of adaptive stochastic gradient descent. In *International conference on artificial intelligence and statistics*, pages 1475–1485. PMLR, 2020.