Sufficient conditions for non-asymptotic convergence of Riemannian optimization methods

author names withheld

Under Review for OPT 2022

Abstract

Motivated by energy based analyses for descent methods in the Euclidean setting, we investigate a generalisation of such energy based analyses for descent methods over Riemannian manifolds. In doing so, we find that it is possible to derive curvature-free guarantees for such descent methods, improving on work by Zhang and Sra [11]. This analysis allows us to study acceleration of Riemannian gradient descent in the geodesically-convex setting, and improve on an existing result by Alimisis et al. [3]. Finally, extending the analysis by Ahn and Sra [2], we attempt to provide some sufficient conditions for the acceleration of Riemannian descent methods in the strongly geodesically convex setting.

1. Introduction

In this paper, our focus is on minimizing a function f defined over a Riemannian manifold $\mathcal{M}$. Certain functions that are non-convex in the Euclidean sense have shown to be convex in the Riemannian sense over a specific Riemannian manifold – we will refer to this notion as geodesic convexity or g-convexity. This motivates the study of optimization methods over Riemannian manifolds, where considerable progress has been made recently in understanding these methods, and proposing better alternatives. Zhang and Sra [11] gave the first non-asymptotic analysis for Riemannian gradient descent using a modified triangle inequality. This triangle inequality also underscores an accelerated version of Riemannian gradient descent [12], where the authors use the concept of estimate sequences [8] to achieve acceleration within a ball around the minimizer of f for strongly g-convex functions. Alimisis et al. [3] investigate acceleration of first order methods over bounded domains for a broader class of functions which include g-convex functions, and formulate an algorithm which is strictly better than Riemannian gradient descent, and more recently Kim and Yang [5] showed under the same setting that it is possible to match the $\mathcal{O}(1/k^2)$ rate for Euclidean methods, albeit with a curvature factor. Both these prior analyses are energy based, popularised by Wilson et al. [10] in the Euclidean setting. This motivates our work, which is to characterise descent methods and subsequently provide a general energy based analysis for such methods over Riemannian manifolds. Second order methods have also been proposed over Riemannian manifolds, and we refer to Boumal [4, Chapter 6] for a detailed introduction of such methods.

1.1. Background

In this subsection, we introduce key definitions and terminology necessary for this work. A Riemannian manifold $\mathcal{M}$ is a smooth manifold equipped with a Riemannian metric that defines an

© .

inner product between two vectors v, w in the tangent space $\mathcal{T}_x\mathcal{M}$ of x for every $x \in \mathcal{M}$. This induces a norm given by $\|v\|_x = \langle v, v\rangle_x$ for all $v \in \mathcal{T}_x\mathcal{M}$. The gradient of a differentiable function f at a point x is a vector in $\mathcal{T}_x\mathcal{M}$ satisfying $\langle \mathrm{grad} f(x), v\rangle_x = \lim_{t\to 0} \frac{f(c(t))-f(x)}{t}$ for curve $c : [0,1] \to \mathcal{M}$ satisfying $c(0) = x$ and $c'(0) = v$ for every $v \in \mathcal{M}$. A geodesic between two points x and y on the manifold is the length minimizing curve starting at x and ending at y, and the distance $d(x, y)$ between x and y is given by the length of the geodesic. A subset A of $\mathcal{M}$ is a geodesically unique set if for any two points in A, there exists a unique geodesic connecting them. The exponential map at a point x on the manifold maps a tangent vector $v \in \mathcal{T}_x\mathcal{M}$ to a point on the manifold $\mathrm{Exp}_x(v)$ such that $\gamma_{x,v}(1) = \mathrm{Exp}_x(v)$ where $\gamma_{x,v} : [0,1] \to \mathcal{M}$ is a geodesic satisfying $\gamma_{x,v}(0) = x$ and $\gamma'_{x,v}(0) = v$. The inverse of the exponential map if it exists is called the logarithmic map ($\mathrm{Log}_x : y \mapsto v$), which computes the velocity of the geodesic starting from x to reach a point y. The exponential and logarithmic maps can be viewed as manifold analogoues to vector addition and subtraction respectively. A function $f : A \to \mathbb{R}$ is said to be μ-strongly g-convex if for all $x, y \in A$, $f(y) \geq f(x) + \langle \mathrm{grad} f(x), \mathrm{Log}_x(y)\rangle_x + \mu/2\|\mathrm{Log}_y(x)\|_x^2$. When $\mu = 0$, we refer to such a function as being simply g-convex. Similarly, a function $f : A \to \mathbb{R}$ is said to L-g-smooth if for all $x, y \in A$, $f(y) \leq f(x) + \langle \mathrm{grad} f(x), \mathrm{Log}_x(y)\rangle_x + L/2\|\mathrm{Log}_x(y)\|_x^2$. We specify other equivalent definitions of these classes of functions in Appendix A.

2. An energy based analysis of Riemannian descent methods

To study Riemannian descent methods, we first introduce an abstraction that will allow us to study a collection of such algorithms in a unified manner. This abstraction is equivalent to 1-descent algorithms of order p proposed in [9] in the Euclidean case.

Definition 1 (p-descent algorithm) *An iterative algorithm $\mathcal{A}$ is a p-forward descent / p-backward descent algorithm in $\mathcal{M}$ w.r.t. function f if the sequence of iterates $\{x_k\}_{k\geq 1}$ satisfies*

$$f(x_{k+1}) \leq f(x_k) - c\|\mathrm{grad} f(x_{k+1})\|_{x_{k+1}}^{p/p-1}, \qquad k \geq 0 \qquad [p\text{-forward descent}] \tag{1a}$$

$$f(x_{k+1}) \leq f(x_k) - c\|\mathrm{grad} f(x_k)\|_{x_k}^{p/p-1}, \qquad k \geq 0 \qquad [p\text{-backward descent}] \tag{1b}$$

where c is a constant independent of k and $x_0 \in \mathcal{M}$ is the initialisation given to $\mathcal{A}$.

For such descent algorithms, we show that it is possible to obtain rates of convergence to the minimizer $x^\star$ of f where f is g-convex function, analogous to the Euclidean setting. This is possible since the descent property deals with vectors in the tangent space of a point on the manifold, and the tangent space is isomorphic to a Euclidean space. Before stating this theorem, we introduce some assumptions, as made in [12] and [3] that are necessary to prove the theorem.

- ($\mathbf{A}_1$) A is a geodesically unique convex subset of $\mathcal{M}$ with bounded diameter where Exp and Log are well-defined.
- ($\mathbf{A}_2$) $x^\star$ is a minimum of f, which lies inside A.
- ($\mathbf{A}_3$) All the iterates of the algorithm stay within A.

When $\mathcal{M}$ is a Hadamard manifold, an example of A satisfying the assumptions is a sublevel set of f with respect to the initialisation. Hyperbolic spaces and many matrix manifolds are examples of practically relevant Hadamard manifolds. Other examples of A when $\mathcal{M}$ is not a Hadamard manifold include a subset of a sphere of radius R whose diameter is strictly less than πR.

Theorem 2 (Rate for p-descent algorithms over g-convex functions) *Let f be a g-convex function, and let x_0 be the initialisation which belongs in $A \subseteq \mathcal{M}$. If $\{x_k\}$ are the iterates of a p-descent algorithm (forward (Eq. 1a) or backward (Eq. 1b)) then, assuming ($\boldsymbol{A}_1$), ($\boldsymbol{A}_2$) and ($\boldsymbol{A}_3$), they satisfy the following guarantee*

$$f(x_k) - f(x^\star) \leq C_p \frac{\mathrm{diam}(A)^p}{k^{p-1}}, \qquad \forall\, k \geq 0$$

where C_p is a constant dependent on p.

While not as involved, it is possible to obtain rates when f is non-convex or when it is gradient dominated. We include these results in the Appendix B, and the proof of Theorem 2 and the proofs for rates in the non-convex and gradient dominated settings in Appendix C.

Examples of Riemannian p-descent algorithms and their rates.

Theorem 2 allows us to immediately infer rates of convergence for popular Riemannian methods when used to optimize g-convex functions. We give instances of such methods below.

1. The Riemannian gradient descent algorithm which generates a sequence of iterates according to the recursion $x_{k+1} = \mathrm{Exp}_x(-\eta \mathrm{grad} f(x))$ for $k \geq 0$, is a 2-backward descent method when f is L-g-smooth and $0 < \eta < 2/L$.

2. The Riemannian proximal descent algorithm which generates a sequence of iterates according to the recursion $x_{k+1} = \mathrm{argmin}_{y \in \mathcal{M}} f(y) + \frac{1}{2\eta} d^2(y, x_k)$ for $k \geq 0$, is a 2-forward descent method for any $\eta > 0$.

3. The Riemannian cubic-regularized Newton algorithm which generates a sequence of iterates according to the recursion $x_{k+1} = \mathrm{Exp}_{x_k}(s_k)$ where s_k satisfies $m_k(s_k) \leq m_k(0)$, $\|\nabla m_k(s_k)\|_{x_k} \leq \theta \|s_k\|_{x_k}^2$ with $m_k(s) := f(x_k) + \langle s, \mathrm{grad} f(x_k)\rangle_{x_k} + \frac{1}{2}\langle s, \mathrm{Hess} f(x_k)[s]\rangle_{x_k} + \frac{M}{3}\|s\|_{x_k}^3$ is 3-forward descent algorithm when f has ρ-Lipschitz continuous Hessians and $M > \rho/2$.

Notably, the rates of convergence that we attain for these methods are *curvature-independent* and specifically improves the popular result for Riemannian gradient descent by Zhang and Sra [11]. This theorem allows us to give the first known rates for a cubic-regularized Newton type algorithm over Riemannian manifolds; previously known guarantees were established in the in the non-convex setting by Agarwal et al. [1] for a slightly complicated version of the algorithm. We include complete proofs of these propositions in Appendix C.1.

3. Accelerating descent methods for g-convex and strongly g-convex functions

The energy based analysis provides an effective way to analyse accelerated versions of descent methods, as studied by Wilson et al. [10] in the Euclidean case. In this section, we derive guarantees for an accelerated version of a 2-backward descent algorithm. The algorithm is composed of three updates defined below.

$$x_{k+1} = \mathrm{Exp}_{y_k}(\tau_{k+1} \mathrm{Log}_{y_k}(z_k)) \tag{2a}$$

$$y_{k+1} = G_c(x_{k+1}) \tag{2b}$$

$$z_{k+1} = \mathrm{Exp}_{x_{k+1}}\left((\alpha_{k+1} + \beta_{k+1})^{-1}\left\{\beta_{k+1}\mathrm{Log}_{x_{k+1}}(z_k) - \mathrm{grad} f(x_{k+1})\right\}\right) \tag{2c}$$

with $y_0 = z_0 \in A \subseteq \mathcal{M}$. $\tau_k \in (0,1)$, $\alpha_k, \beta_k > 0$ for all $k \geq 0$. G_c is a mapping which ensures that for all x, $f(G_c(x)) - f(x) \leq -c\|\mathrm{grad} f(x)\|_x^2$. To proceed with the energy based analysis, we first define the energy function. We use a combination of the function optimality gap and the variant of the distance to the optimum, formally defined below.

$$E_k = A_k \cdot (f(y_k) - f(x^\star)) + B_k \cdot \left[\|\mathrm{Log}_{x_k}(z_k) - \mathrm{Log}_{x_k}(x^\star)\|_{x_k}^2\right]. \tag{3}$$

This choice of the energy function was previously used by Ahn and Sra [2] to study acceleration of Riemannian gradient descent for strongly g-convex functions. Since the distance term of the energy is explicitly dependent on x_k, the analysis is not straightforward, as we cannot directly compare $\|\mathrm{Log}_{x_k}(y)\|_{x_k}$ and $\|\mathrm{Log}_{x_{k+1}}(y)\|_{x_{k+1}}$ for y in general. To aid us in proceeding with the analysis, we use the notion of a *valid distortion rate* which was originally proposed by Ahn and Sra [2].

Definition 3 (Valid Distortion Rate Ahn and Sra [2, Definition 3.2]) *δ_k is a* valid distortion rate *at iteration k if* $\|\mathrm{Log}_{x_k}(z_{k-1}) - \mathrm{Log}_{x_k}(x^\star)\|_{x_k}^2 \leq \delta_k \|\mathrm{Log}_{x_{k-1}}(z_{k-1}) - \mathrm{Log}_{x_{k-1}}(x^\star)\|_{x_{k-1}}^2$.

Ahn and Sra [2] provide a computable forms of δ_k based on the iterates x_k and z_k at each iteration k depending on whether $\mathcal{M}$ is a Hadamard manifold. Assuming the existence of such valid distortion rates, we can use Equation 3 with specific setting of parameters A_k, B_k, and show that the accelerated method in Equation 2 with algorithmic parameters $\tau_k, \alpha_k, \beta_k$ is strictly better than Riemannian gradient descent.

Theorem 4 (Guarantees for g-convex functions) *Let $\{y_k\}$ be the sequence of iterates generated by the algorithm described in Equation 2 with parameters*

$$\begin{cases} \tau_{k+1} = \frac{2\overline{A}_k B_k}{A_{k+1}\delta_{k+1}B_{k+1} + 2B_k\overline{A}_k}; \;\; \alpha_{k+1} = \frac{B_{k+1} - B_k/\delta_k}{\overline{A}_k}; \;\; \beta_{k+1} = \frac{B_k/\delta_{k+1}}{\overline{A}_{k+1}} \\ \overline{A}_{k+1} = A_{k+1} - A_k; \;\; A_{k+1} = \frac{(k+1)(k+2)}{2}; \;\; B_{k+1} = \frac{4}{c} \end{cases}$$

when given a g-convex and initialisation $y_0 = z_0 \in A \subseteq \mathcal{M}$. Under ($\boldsymbol{A}_1$), ($\boldsymbol{A}_2$) and ($\boldsymbol{A}_3$) and assuming the existence of a valid distortion rate at every iteration $k \geq 0$, this sequence satisfies for every $k \geq 0$

$$f(y_k) - f(x^\star) \leq \frac{E_0}{k^2} + \frac{{}^4\!/_c \cdot \mathrm{diam}(A)^2 \cdot (1 - {}^1\!/_{\delta_{\max}})}{k}, \quad \delta_{\max} := \max_{t \leq k} \delta_t.$$

Remark Recently, Alimisis et al. [3] proposed a slightly different version of algorithm 2 which involves a geodesic search that consequently incurs an error. In contrast, the algorithm in Eq. 2 does not involve such a search step, and consequently, the rate guarantee derived in Thm. 4 is free of a search error. Next, from the above theorem, when $\delta_{\max} = 1$ (for instance: when $\mathcal{M}$ is Euclidean), we recover the $O({}^1\!/_{k^2})$ rate shown by Nesterov [7]. When $\delta_{\max} \to \infty$, then we achieve the same rate as Riemannian gradient descent / proximal descent. Our guarantee states that we achieve acceleration initially (which is when ${}^1\!/_{k^2}$ rate term dominates), and hence we are strictly better than Riemannian gradient descent.

We can also show that for μ-strongly g-convex functions, 2-backward descent methods can be accelerated in a similar manner. This is direct consequence of Ahn and Sra [2, Theorem 3.1].

Proposition 5 (Guarantees for μ-strongly g-convex functions) *Consider the algorithm described in Equation 2 where c in G_c satisfies $2\mu c < 1$ for a μ-strongly g-convex function. Let $\{y_k\}$ be the sequence of iterates generated by such an algorithm with parameters*

$$\begin{cases} \tau_{k+1} = \frac{\xi_{k+1}-2\mu c}{1-2\mu c}; \;\; \alpha_{k+1} = \mu; \;\; \beta_{k+1} = \frac{\xi_{k+1}-2\mu c}{2c} \\ A_{k+1} = \frac{A_k}{1-\xi_{k+1}}; \;\; B_{k+1} = \frac{\xi_{k+1}^2}{1-\xi_{k+1}} \cdot \frac{A_k}{4c}. \end{cases},$$

where ξ_{k+1} is the solution to the equation $\frac{\xi_{k+1}(\xi_{k+1}-2\mu c)}{1-\xi_{k+1}} = \frac{\xi_k^2}{\delta_{k+1}}$ in $[2\mu c, 1)$ with $A_0, B_0 > 0$. Under ($\mathbf{A}_1$), ($\mathbf{A}_2$) and ($\mathbf{A}_3$) and assuming the existence of a valid distortion rate at each iteration $k \geq 0$, this sequence satisfies for every $k \geq 0$

$$f(y_k) - f(x^\star) \leq \prod_{j=1}^{k}(1-\xi_j)\left[f(y_0) - f(x^\star) + \frac{\xi_0^2}{4c}\|\mathrm{Log}_{x_0}(z_0) - \mathrm{Log}_{x_0}(x^\star)\|_{x_0}^2\right].$$

Remark Note that the rate of convergence directly depends on values taken by ξ_j. Let $\delta_{\max} = \max_{t\leq k}\delta_k$. When $\delta_{\max} = 1$ (for instance: when $\mathcal{M}$ is Euclidean), Ahn and Sra [2] show in their Lemma 2.1 that the sequence $\{\xi_k\}_{k\geq 0}$ converges to $\sqrt{2\mu c}$; specifically, when $\xi_0 \geq \sqrt{2\mu c}$, $\{\xi_k\}_{k\geq 0}$ converges to $\sqrt{2\mu c}$, giving us the rate $\mathcal{O}(\exp(-\sqrt{2\mu c}\cdot k))$ compared to $\mathcal{O}(\exp(-2\mu c \cdot k))$ for standard gradient descent. When $\delta_{\max} \to \infty$, then $\xi_k = 2\mu c$, and we achieve the same rate as Riemannian gradient descent. This implies that the algorithm is strictly better than the non-accelerated version; this was also noted by Ahn and Sra [2] but for a gradient descent step.

3.1. Some sufficient conditions for eventual full acceleration of Riemannian descent methods over strongly g-convex functions

As noted earlier, there exists a computable sequence of valid distortion rates $\{\delta_{k+1}\}$ dependent on the iterates $\{(x_k, z_k)\}$ generated by the algorithm in Equation 2. More precisely, for Hadamard manifolds with sectional curvature lower bounded by $-\kappa < 0$, the valid distortion rate at the kth iteration is given by $\delta_{k+1} = T_\kappa(d(x_k, z_k))$ where $T_\kappa(0) = 1$. Therefore, it would be instructive to analyse the rate at which the sequence $\{d(x_k, z_k)\}$ converges to 0, and translate that analysis to a rate at which the sequence $\{\xi_k\}$ converges to $\sqrt{2\mu c}$. This is the technique adopted in [2] for their analysis. In this subsection, we attempt to provide an extension of their analysis for gradient descent that would be applicable more generally when G_c is an arbitrary 2-backward descent method. We begin by giving the following lemma, which is a generalisation of Lemma 4.2 in Ahn and Sra [2].

Lemma 6 *Define $D_0 = f(y_0) - f(x^\star) + \frac{\xi_0^2}{4c}\|\mathrm{Log}_{z_0}(x^\star)\|_{z_0}^2$ and $\mu > 0$. Let $\{(x_k, y_k, z_k)\}$ be the sequence of iterates obtained from Algorithm 2, where the descent constant c of G_c satisfies $c < \min\{1/6L, 1/2\mu\}$ with parameter settings stated in Prop. 5. If $\mathcal{M}$ is a Hadamard manifold, $\xi_0 \in (2\mu c, \sqrt{2\mu c}]$ and the iterates satisfy $(d(x_{k+1}, y_{k+1}) \leq \mathcal{C}'_{L,\mu,c}\sqrt{\prod_{j=1}^{k}(1-\xi_j)\cdot D_0}$ for all $k \geq 0$, then $d(x_{k+1}, z_{k+1}) \leq \mathcal{C}_{L,\mu,c}\sqrt{\prod_{j=1}^{k}(1-\xi_j)\cdot D_0}$ for all $k \geq 0$ as well, where $\mathcal{C}_{L,\mu,c}$ and $\mathcal{C}'_{L,\mu,c}$ are constants only depending of L, μ, c.*

Remark The above lemma states that if any descent method that descends sufficiently and causes the sequence of distances $\{d(x_k, y_k)\}$ to decrease at a geometric rate, then the sequence of distances $\{d(x_k, z_k)\}$ decreases at the same rate. The original analysis by Ahn and Sra [2] provides

such a result when G_c is a gradient update, along with an interesting requirement that the step size be strictly greater than $1/L$. Recall that for a gradient update, $c = c(\gamma) := \gamma(1 - L\gamma/2)$ and $\operatorname{argmax}_\gamma c(\gamma) = 1/L$. Our lemma states that a small enough descent is sufficient for a similar geometric convergence property.

With the above lemma, we can provide a general convergence result due to a careful analysis of the evolution of the sequence $\{\xi_k\}$ by Ahn and Sra [2].

Proposition 7 (Eventual acceleration of the Algorithm in Eq. 2) *Consider the algorithm in Eq. 2 with where c satisfies $c < \min\{1/6L, 1/2\mu\}$ with parameter settings stated in Prop. 5 and $\xi_0 \in (2\mu c, \sqrt{2\mu c}]$. Then, when $\mathcal{M}$ is a Hadamard manifold with sectional curvature lower bounded by $-\kappa < 0$, the sequence of iterates $\{y_k\}$ generated by this algorithm satisfies the following guarantee*

$$f(y_k) - f(x^\star) \leq \prod_{j=1}^{k}(1 - \xi_j) \cdot D_0$$

where D_0 is a constant depending on the initial parameters and initial values of y_0 and z_0. Moreover, if $d(x_{k+1}, y_{k+1}) \leq \mathcal{C}'_{L,\mu,c}\sqrt{\prod_{j=1}^{k}(1 - \xi_j) \cdot D_0}$ for all $k \geq 0$, the sequence $\{\xi_k\}$ satisfies $|\xi_k - \sqrt{2\mu c}| \leq \epsilon$ when $k \geq \mathcal{C}_{\kappa,L,\mu,c}\log(1/\epsilon)$ where $\mathcal{C}_{\kappa,L,\mu,c}$ is a constant depending on κ, L, μ, c.

Remark To achieve full acceleration, we would require $\xi_k = \sqrt{2\mu c}$ for all $k \geq 0$. This theorem states that while we might not be able to have $\xi_k = \sqrt{2\mu c}$ for all $k \geq 0$, we can still get arbitrarily close as the algorithm proceeds, and consequently achieve acceleration. We conjecture that this analysis will also extend to non-Hadamard manifolds under suitable assumptions ($(\mathbf{A}_1)$, $(\mathbf{A}_2)$, $(\mathbf{A}_3)$) as discussed in [2, Section D].

4. Conclusion

In this work, we presented a general analysis of Riemannian optimization methods using energy-based analysis framework that has gained popularity in the Euclidean setting and more recently in the Riemannian setting. Such an analysis is also conducive to a study of accelerated first order Riemannian descent methods. To this end, we showed that we can obtain a accelerated algorithms for first order descent methods in a straightforward manner in the g-convex and strongly g-convex setting, and present an analysis for the latter case which extends an existing analysis. Some open questions remain: can we achieve (eventual) acceleration for a fully proximal point method, or other higher order methods such as cubic-regularized Newton even on bounded domains?

References

[1] Naman Agarwal, Nicolas Boumal, Brian Bullins, and Coralia Cartis. Adaptive regularization with cubics on manifolds. *Mathematical Programming*, 188(1):85–134, 2021.

[2] Kwangjun Ahn and Suvrit Sra. From nesterov's estimate sequence to riemannian acceleration. In *Conference on Learning Theory*, pages 84–118. PMLR, 2020.

[3] Foivos Alimisis, Antonio Orvieto, Gary Becigneul, and Aurelien Lucchi. Momentum improves optimization on riemannian manifolds. In *International Conference on Artificial Intelligence and Statistics*, pages 1351–1359. PMLR, 2021.

[4] Nicolas Boumal. An introduction to optimization on smooth manifolds. *Available online, May*, 3, 2020.

[5] Jungbin Kim and Insoon Yang. Nesterov acceleration for riemannian optimization. *arXiv preprint arXiv:2202.02036*, 2022.

[6] Yu Nesterov. Accelerating the cubic regularization of newton's method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008.

[7] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence o (1/kˆ 2). In *Doklady AN USSR*, volume 269, pages 543–547, 1983.

[8] Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer.

[9] Ashia C Wilson, Lester Mackey, and Andre Wibisono. Accelerating rescaled gradient descent: Fast optimization of smooth functions. *Advances in Neural Information Processing Systems*, 32, 2019.

[10] Ashia C Wilson, Ben Recht, and Michael I Jordan. A lyapunov analysis of accelerated methods in optimization. *J. Mach. Learn. Res.*, 22:113–1, 2021.

[11] Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *Conference on Learning Theory*, pages 1617–1638. PMLR, 2016.

[12] Hongyi Zhang and Suvrit Sra. Towards riemannian accelerated gradient methods. *arXiv preprint arXiv:1806.02812*, 2018.

Appendix A. More definitions

A function $f : A \to \mathbb{R}$ is g-convex if for any two points x, y in A,

$$f(\gamma_{x,y}(t)) \leq tf(y) + (1-t)f(y)$$

for all $t \in [0,1]$ where $\gamma_{x,y} : [0,1] \to A$ is the geodesic with end points given by x and y. This is the zeroth order definition of g-convexity. When f is differentiable, then we obtain the canonical definition of g-convexity which states that for any x, y in A

$$f(y) \geq f(x) + \langle \mathrm{grad} f(x), \mathrm{Log}_x(y) \rangle_x.$$

With the notion of curves and geodesics, one can transport vectors in a tangent space at one point to the tangent space at another point. This is made possible via the concept of parallel transports. The parallel transport between $\mathcal{T}_x\mathcal{M}$ and $\mathcal{T}_y\mathcal{M}$ for $x, y \in \mathcal{M}$ along curve c is denoted by $\Gamma(c)_x^y : \mathcal{T}_x\mathcal{M} \to \mathcal{T}_y\mathcal{M}$. When c is a geodesic between x and y we omit the c in the notation and use Γ_x^y simply. A key property of parallel transports is that it is norm-preserving: for any $v \in \mathcal{T}_x\mathcal{M}$, $\|v\|_x = \|\Gamma_x^y v\|_y$ where $\Gamma_x^y v \in \mathcal{T}_y\mathcal{M}$ per the definition of Γ_x^y. A function $f : A \to \mathbb{R}$ is said to have L-Lipschitz gradients if for any $x, y \in A$,

$$\|\mathrm{grad} f(x) - \Gamma_y^x \mathrm{grad} f(y)\|_x \leq L \cdot d(x, y).$$

Similar to the Euclidean case, $f : A \to \mathbb{R}$ is L-g-smooth iff f has L-Lipschitz gradients.

Appendix B. Rates for p-descent algorithms for non-convex functions and gradient-dominated functions

Theorem 8 (Rate for p-descent over non-convex functions) *Let f a non-convex function, and let $x_0 \in \mathcal{M}$ be the initialisation. Then, a p-descent algorithm (forward (Eq. 1a) or backward (Eq. 1b)) satisfies the following guarantee*

$$\min_{t \leq k} \|\mathrm{grad} f(x_t)\|_{x_t} \leq \left(\frac{f(x_0) - f(x^\star)}{ck} \right)^{(p-1)/p}$$

Definition 9 ((τ, p)-gradient dominated functions) *A differentiable function $f : \mathcal{M} \to \mathbb{R}$ is said to be (τ, p)-gradient dominated if $x^\star$ is the global minimizer of f and for all x*

$$f(x) - f(x^\star) \leq \tau \|\mathrm{grad} f(x)\|_x^{p/p-1}.$$

Theorem 10 (Rate for p-descent algorithms over (τ, p)-gradient dominated functions) *Let f be a (τ, p)-gradient dominated function, and let $x_0 \in \mathcal{M}$ be the initialisation. Then, a p-descent algorithm satisfies the following guarantees*

$$f(x_k) - f(x^\star) \leq \left(1 + \frac{c}{\tau}\right)^{-k} (f(x_0) - f(x^\star)), \qquad \forall\, k \geq 0 \qquad [p\text{-forward descent}]$$

$$f(x_k) - f(x^\star) \leq \left(1 - \frac{c}{\tau}\right)^{k} (f(x_0) - f(x^\star)), \qquad \forall\, k \geq 0 \qquad [p\text{-backward descent}]$$

Appendix C. Proofs for the rate theorems in Section 2 and B

Proof [Proof of Theorem 2] We begin by noting that under the assumptions, the exponential map and its inverse exists at every $v \in \mathcal{T}_x\mathcal{M}$ for every $x \in A$. Consider an energy function

$$E_k = A_k(f(x_k) - f(x^\star)).$$

Here, $\{A_k\}_{k\geq 1}$ is a sequence satisfying $A_{k+1} = A_k + a_k$ and $x^\star$ is the minimizer of f. The difference between E_{k+1} and E_k is

$$\begin{aligned}
E_{k+1} - E_k &= (A_k + a_k)(f(x_{k+1}) - f(x^\star)) - A_k(f(x_k) - f(x^\star)) \\
&= A_k(f(x_{k+1}) - f(x_k)) + a_k(f(x_{k+1}) - f(x^\star)) \qquad \text{(4a)} \\
&= (A_k + a_k)(f(x_{k+1}) - f(x^\star)) - (A_k + a_k)(f(x_k) - f(x^\star)) + a_k(f(x_k) - f(x^\star)) \\
&= (A_k + a_k)(f(x_{k+1}) - f(x_k)) + a_k(f(x_k) - f(x^\star)). \qquad \text{(4b)}
\end{aligned}$$

Since f is g-convex,

$$\begin{aligned}
f(x_k) - f(x^\star) &\leq \langle \mathrm{grad} f(x_k), -\mathrm{Log}_{x_k}(x^\star)\rangle_{x_k}, \text{ and} \\
f(x_{k+1}) - f(x^\star) &\leq \langle \mathrm{grad} f(x_{k+1}), -\mathrm{Log}_{x_{k+1}}(x^\star)\rangle_{x_{k+1}}.
\end{aligned}$$

If $\mathcal{A}$ is a p-forward-descent algorithm w.r.t. f, we can use Equation 1a and bound the difference in energies

$$\begin{aligned}
E_{k+1} - E_k &\leq -cA_k \|\mathrm{grad} f(x_{k+1})\|_{x_{k+1}}^{p/p-1} + a_k \langle \mathrm{grad} f(x_{k+1}), -\mathrm{Log}_{x_{k+1}}(x^\star)\rangle_{x_{k+1}} \\
&= \frac{cA_k p}{p-1}\left(\left\langle \mathrm{grad} f(x_{k+1}), -\frac{a_k}{cA_k}\frac{p-1}{p}\mathrm{Log}_{x_{k+1}}(x^\star)\right\rangle_{x_{k+1}} - \frac{\|\mathrm{grad} f(x_{k+1})\|_{x_{k+1}}^{p/p-1}}{p/p-1}\right)
\end{aligned} \tag{5}$$

If $\mathcal{A}$ is a p-backward-descent algorithm w.r.t. f, we can use Equation 1b and bound the difference in energies

$$\begin{aligned}
E_{k+1} - E_k &\leq -c(A_k + a_k)\|\mathrm{grad} f(x_k)\|_{x_k}^{p/p-1} + a_k \langle \mathrm{grad} f(x_k), -\mathrm{Log}_{x_k}(x^\star)\rangle_{x_k} \\
&= \frac{c(A_k + a_k)p}{p-1}\left(\left\langle \mathrm{grad} f(x_k), -\frac{a_k}{c(A_k + a_k)}\frac{p-1}{p}\mathrm{Log}_{x_k}(x^\star)\right\rangle_{x_k} - \frac{\|\mathrm{grad} f(x_k)\|_{x_k}^{p/p-1}}{p/p-1}\right).
\end{aligned} \tag{6}$$

To bound the quantity inside the brackets in Equations 5 and 6, we use Lemma 11. Specifically, we invoke the lemma with $q = p/p-1$ and

- $\alpha = -\frac{a_k}{cA_k}\frac{p-1}{p}$ for Equation 5,
- $\alpha = -\frac{a_k}{c(A_k+a_k)}\frac{p-1}{p}$ for Equation 6

to get

$$
\begin{aligned}
(5) \Rightarrow E_{k+1} - E_k &\leq \frac{cA_k p}{p(p-1)} \cdot \left(\frac{a_k}{A_k}\right)^p \cdot \left(\frac{p-1}{p}\right)^p \|\mathrm{Log}_{x_{k+1}}(x^\star)\|^p_{x_{k+1}} \\
&= c'_p \frac{a_k^p}{A_k^{p-1}} \|\mathrm{Log}_{x_{k+1}}(x^\star)\|^p_{x_{k+1}} \qquad (7) \\
(6) \Rightarrow E_{k+1} - E_k &\leq \frac{c(A_k + a_k)p}{p(p-1)} \cdot \left(\frac{a_k}{A_k + a_k}\right)^p \cdot \left(\frac{p-1}{p}\right)^p \|\mathrm{Log}_{x_k}(x^\star)\|^p_{x_k} \\
&= c'_p \frac{a_k^p}{(A_k + a_k)^{p-1}} \|\mathrm{Log}_{x_k}(x^\star)\|^p_{x_k}. \qquad (8)
\end{aligned}
$$

where $c'_p = \frac{c^{1-p}}{p}\left(\frac{p-1}{p}\right)^{p-1}$. By definition of the exponential map, $d(x,z) = \|\mathrm{Exp}_x^{-1}(z)\|_x$ for all $x, z \in A$. Also, by ($\mathbf{A}_2$), $x^\star \in A$. Therefore, $d(z, x^\star) = \|\mathrm{Exp}_z^{-1}(x^\star)\|_z \leq \mathrm{diam}(A)$ for any $z \in A$. This further bounds of the difference in energy as

$$
(7) \Rightarrow E_{k+1} - E_k \leq c'_p \frac{a_k^p}{A_k^{p-1}} \mathrm{diam}(A)^p \qquad (9)
$$

$$
(8) \Rightarrow E_{k+1} - E_k \leq c'_p \frac{a_k^p}{(A_k + a_k)^{p-1}} \mathrm{diam}(A)^p \qquad (10)
$$

Choose $A_k = \frac{k(k+1)\ldots(k+p-1)}{p!}$. This gives $a_k = A_{k+1} - A_k = \frac{(k+1)\ldots(k+p-1)}{(p-1)!}$. Furthermore,

$$
\begin{aligned}
\frac{a_k^p}{A_k^{p-1}} &= \underbrace{\frac{(k+1)\ldots(k+p-1)}{k^{p-1}}}_{\leq p^{p-1}} \frac{(p-1)!^{p-1}(p^{p-1})}{(p-1)!^{p-1}(p-1)!} \leq \frac{p^{2(p-1)}}{(p-1)!} \\
\frac{a_k^p}{(A_k + a_k)^{p-1}} &= \underbrace{\frac{(k+1)\ldots(k+p-1)}{(k+p)^{p-1}}}_{\leq 1} \frac{(p-1)!^{p-1}(p^{p-1})}{(p-1)!^{p-1}(p-1)!} \leq \frac{p^{p-1}}{(p-1)!}
\end{aligned}
$$

We finally have

$$
\begin{aligned}
(9) \Rightarrow E_{k+1} - E_k &\leq \underbrace{\frac{c^{1-p}}{p}\left(\frac{p-1}{p}\right)^{p-1} \cdot \frac{p^{2(p-1)}}{(p-1)!}}_{c''_{p,\mathrm{fwd}}} \cdot \mathrm{diam}(A)^p \\
(10) \Rightarrow E_{k+1} - E_k &\leq \underbrace{\frac{c^{1-p}}{p}\left(\frac{p-1}{p}\right)^{p-1} \cdot \frac{p^{p-1}}{(p-1)!}}_{c''_{p,\mathrm{bwd}}} \cdot \mathrm{diam}(A)^p.
\end{aligned}
$$

Summing both sides from $k = 0$ to $k = T - 1$, we get

$$\begin{aligned}
&E_T - E_0 \leq c''_{p,\text{fwd}} \cdot \operatorname{diam}(A)^p \cdot T \\
&\quad \Rightarrow E_T \leq c''_{p,\text{fwd}} \cdot \operatorname{diam}(A)^p \cdot T + E_0 \Rightarrow f(x_T) - f(x^\star) \leq c''_{p,\text{fwd}} \cdot \frac{T}{A_T} \cdot \operatorname{diam}(A)^p, \text{ and} \\
&E_T - E_0 \leq c''_{p,\text{bwd}} \cdot \operatorname{diam}(A)^p \cdot T \\
&\quad \Rightarrow E_T \leq c''_{p,\text{bwd}} \cdot \operatorname{diam}(A)^p \cdot T + E_0 \Rightarrow f(x_T) - f(x^\star) \leq c''_{p,\text{bwd}} \cdot \frac{T}{A_T} \cdot \operatorname{diam}(A)^p.
\end{aligned}$$

Since $A_T \geq {}^{T^p}/_{p!}$, ${}^{T}/_{A_T} \leq {}^{p!}/_{T^{p-1}}$. Consequently,

$$\begin{aligned}
p\text{-fwd-descent} \Rightarrow f(x_T) - f(x^\star) &\leq \frac{c^{1-p}}{p}\left(\frac{p-1}{p}\right)^{p-1} \cdot \frac{p^{2(p-1)}}{(p-1)!}\frac{p!}{T^{p-1}} \cdot \operatorname{diam}(A)^p \\
&= \frac{c^{1-p} \cdot (p^2 - p)^{p-1} \cdot \operatorname{diam}(A)^p}{T^{p-1}}, \\
p\text{-bwd-descent} \Rightarrow f(x_T) - f(x^\star) &\leq \frac{c^{1-p}}{p}\left(\frac{p-1}{p}\right)^{p-1} \cdot \frac{p^{p-1}}{(p-1)!}\frac{p!}{T^{p-1}} \cdot \operatorname{diam}(A)^p \\
&= \frac{c^{1-p} \cdot (p-1)^{p-1} \cdot \operatorname{diam}(A)^p}{T^{p-1}}.
\end{aligned}$$

■

Proof [Proof of Theorem 8] If $\mathcal{A}$ is a p-forward descent algorithm w.r.t. f, we can use Equation 1a to get

$$\begin{aligned}
c\|\operatorname{grad} f(x_{k+1})\|_{x_{k+1}}^{p/p-1} &\leq f(x_k) - f(x_{k+1}) \\
\sum_{k=0}^{T-1} c\|\operatorname{grad} f(x_{k+1})\|_{x_{k+1}}^{p/p-1} &\leq f(x_0) - f(x_T) \\
&\leq f(x_0) - f(x^\star) \\
\Rightarrow \min_{k \leq T} \|\operatorname{grad} f(x_k)\|_{x_k}^{p/p-1} &\leq \frac{f(x_0) - f(x^\star)}{cT} \\
\Rightarrow \min_{k \leq T} \|\operatorname{grad} f(x_k)\|_{x_k} &\leq \left(\frac{f(x_0) - f(x^\star)}{cT}\right)^{(p-1)/p}.
\end{aligned}$$

If $\mathcal{A}$ is a p-backward descent algorithm w.r.t. f, we can use Equation 1b to get

$$\begin{aligned}
c\|\operatorname{grad} f(x_k)\|_{x_k}^{p/p-1} &\leq f(x_k) - f(x_k) \\
\sum_{k=0}^{T-1} c\|\operatorname{grad} f(x_k)\|_{x_k}^{p/p-1} &\leq f(x_0) - f(x_T) \\
&\leq f(x_0) - f(x^\star) \\
\Rightarrow \min_{k \leq T} \|\operatorname{grad} f(x_k)\|_{x_k}^{p/p-1} &\leq \frac{f(x_0) - f(x^\star)}{cT} \\
\Rightarrow \min_{k \leq T} \|\operatorname{grad} f(x_k)\|_{x_k} &\leq \left(\frac{f(x_0) - f(x^\star)}{cT}\right)^{(p-1)/p}.
\end{aligned}$$

■

Proof [Proof of Theorem 10] Consider the energy function

$$E_k = f(x_k) - f(x^\star).$$

Then, we obtain

$$E_{k+1} - E_k = f(x_{k+1}) - f(x_k).$$

If $\mathcal{A}$ is a p-forward descent algorithm w.r.t. f, then using Eq. 1a

$$E_{k+1} - E_k = f(x_{k+1}) - f(x_k) \leq -c\|\mathrm{grad} f(x_{k+1})\|_{x_{k+1}}^{p/p-1} \leq -\frac{c}{\tau}(f(x_{k+1}) - f(x^\star)) = -\frac{c}{\tau}E_{k+1}.$$

As a result,

$$E_{k+1} \leq \left(1 + \frac{c}{\tau}\right)^{-1} E_k \Rightarrow E_T \leq \left(1 + \frac{c}{\tau}\right)^{-T} E_0.$$

If $\mathcal{A}$ is a p-backward descent algorithm w.r.t. f, then using Eq. 1b

$$E_{k+1} - E_k = f(x_{k+1}) - f(x_k) \leq -c\|\mathrm{grad} f(x_k)\|_{x_k}^{p/p-1} \leq -\frac{c}{\tau}(f(x_k) - f(x^\star)) = -\frac{c}{\tau}E_k.$$

As a result,

$$E_{k+1} \leq \left(1 - \frac{c}{\tau}\right) E_k \Rightarrow E_T \leq \left(1 - \frac{c}{\tau}\right)^{T} E_0.$$

■

C.1. Proofs for the examples of descent methods

Proof [Rate for Riemannian gradient descent] Using the assumptions, we have for $k \geq 0$ that

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \mathrm{grad} f(x_k), \mathrm{Log}_{x_k}(x_{k+1})\rangle_{x_k} + \frac{L}{2}\|\mathrm{Log}_{x_k}(x_{k+1})\|_{x_k}^2 \\ &= f(x_k) - \eta\|\mathrm{grad} f(x_k)\|_{x_k}^2 + \frac{\eta^2 L}{2}\|\mathrm{grad} f(x_k)\|_{x_k}^2. \end{aligned}$$

When $\eta = 1/L$, we get a simplified bound as

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\mathrm{grad} f(x_k)\|_{x_k}^2.$$

Using Theorem 2 in the backward case with $c = 1/2L, p = 2$, we directly obtain

$$f(x_k) - f(x^\star) \leq \frac{2L \cdot \mathrm{diam}(A)^2}{k}.$$

■

Proof [Rate for Riemannian proximal descent] Using the assumptions, we have for $k \geq 0$ that

$$\begin{aligned} f(x_{k+1}) + \frac{1}{2\eta}\|\mathrm{Log}_{x_{k+1}}(x_k)\|_{x_{k+1}}^2 &\leq f(x_k) \\ \Rightarrow f(x_{k+1}) &\leq f(x_k) - \frac{1}{2\eta}\|\mathrm{Log}_{x_{k+1}}(x_k)\|_{x_{k+1}}^2. \end{aligned}$$

The proximal update also satisfies $\mathrm{Log}_{x_{k+1}}(x_k) = \eta \mathrm{grad} f(x_{k+1})$ leading to

$$f(x_{k+1}) \leq f(x_k) - \frac{\eta}{2}\|\mathrm{grad} f(x_{k+1})\|^2_{x_{k+1}}.$$

Using Theorem 2 in the forward case with $c = \frac{\eta}{2}$, $p = 2$, we directly obtain

$$f(x_k) - f(x^\star) \leq \frac{4\eta^{-1} \cdot \mathrm{diam}(A)^2}{k}.$$

■

Proof [Rate for Cubic-regularized Newton] Before we state the proof, we introduce the notion of Hessian Lipschitz functions.

A twice differentiable function $f : \mathcal{M} \to \mathbb{R}$ is said to have ρ-Lipschitz continuous Riemannian Hessians, when for all x, s in the domain of the exponential map,

$$\left| f(\mathrm{Exp}_x(s)) - f(x) - \langle s, \mathrm{grad} f(x)\rangle_x - \frac{1}{2}\langle s, \mathrm{Hess} f(x)[s]\rangle_x \right| \leq \frac{\rho}{6}\|s\|^3_x.$$

Equivalently,

$$\left\| \left\{\Gamma_x^{\mathrm{Exp}_x(s)}\right\}^{-1} \mathrm{grad} f(\mathrm{Exp}_x(s)) - \mathrm{grad} f(x) - \mathrm{Hess} f(x)[s] \right\|_x \leq \frac{\rho}{2}\|s\|^2_x,$$

For convenience, we will denote $\Gamma_x^{\mathrm{Exp}_x(s)}$ by P_s when the choice of x is obvious.

Under our assumptions, the domain of the exponential map when restricted to $x \in A$ is the tangent space at every point. At iteration k, the update velocity satisfies

$$f(x_k) + \langle s_k, \mathrm{grad} f(x_k)\rangle_{x_k} + \frac{1}{2}\langle s_k, \mathrm{Hess} f(x_k)[s_k]\rangle_{x_k} + \frac{M}{3}\|s_k\|^3_{x_k} \leq f(x_k).$$

Using the fact that f has ρ-Lipschitz continuous Riemannian Hessians, we get

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle s_k, \mathrm{grad} f(x_k)\rangle_{x_k} + \frac{1}{2}\langle s_k, \mathrm{Hess} f(x_k)[s_k]\rangle_{x_k} + \frac{\rho}{6}\|s_k\|^3_{x_k} \\ &\leq f(x_k) - \left(\frac{M}{3} - \frac{\rho}{6}\right)\|s_k\|^3_{x_k}. \end{aligned}$$

From Agarwal et al. [1, Theorem 3], the gradient of m_k at s_k can be computed as

$$\begin{aligned} \nabla m_k(s_k) &= \mathrm{grad} f(x_k) + \mathrm{Hess} f(x_k)[s_k] + M\|s_k\|_{x_k} s_k \\ &= P_{s_k}^{-1}\mathrm{grad} f(x_{k+1}) + \mathrm{grad} f(x_k) + \mathrm{Hess} f(x_k)[s_k] - P_{s_k}^{-1}\mathrm{grad} f(x_{k+1}) \\ &\quad + M\|s_k\|_{x_k} s_k \\ \Rightarrow \|\nabla m_k(s_k)\|_{x_k} &= \|P_{s_k}^{-1}\mathrm{grad} f(x_{k+1}) + \mathrm{grad} f(x_k) + \mathrm{Hess} f(x_k)[s_k] - P_{s_k}^{-1}\mathrm{grad} f(x_{k+1}) \\ &\quad + M\|s_k\|_{x_k} s_k\|_{x_k} \\ &\geq \|P_{s_k}^{-1}\mathrm{grad} f(x_{k+1})\|_{x_k} - \|P_{s_k}^{-1}\mathrm{grad} f(x_{k+1}) - \mathrm{grad} f(x_k) - \mathrm{Hess} f(x_k)[s_k]\|_{x_k} \\ &\quad - M\|s_k\|^2_{x_k} \\ &\geq \|\mathrm{grad} f(x_{k+1})\|_{x_{k+1}} - \frac{\rho}{2}\|s_k\|^2_{x_k} - M\|s_k\|^2_{x_k} \\ \Rightarrow \theta\|s_k\|^2_{x_k} &\geq \|\mathrm{grad} f(x_{k+1})\|_{x_{k+1}} - \frac{\rho}{2}\|s_k\|^2_{x_k} - M\|s_k\|^2_{x_k} \\ \Rightarrow \|\mathrm{grad} f(x_{k+1})\|_{x_{k+1}} &\leq \left(\theta + \frac{\rho}{2} + M\right)\|s_k\|^2_{x_k}. \end{aligned}$$

Combining this with the descent statement previously, we get

$$f(x_{k+1}) \leq f(x_k) - \left(\frac{M}{3} - \frac{\rho}{6}\right)\left(\theta + \frac{\rho}{2} + M\right)^{-3/2} \|\mathrm{grad} f(x_{k+1})\|_{x_{k+1}}^{3/2}.$$

When $\theta = \rho/2$, $M = \rho$, we get a concise inequality

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{12\sqrt{2}\sqrt{\rho}} \|\mathrm{grad} f(x_{k+1})\|_{x_{k+1}}^{3/2}.$$

Using Theorem 2 in the forward case with $c = \frac{1}{12\sqrt{2}\sqrt{\rho}}, p = 3$, we directly obtain

$$f(x_k) - f(x^\star) \leq \frac{10368\rho \cdot \mathrm{diam}(A)^3}{k^2}.$$

■

Appendix D. Proof for results in Section 3

In the proof that follow, we denote $\|\mathrm{Log}_x(w) - \mathrm{Log}_x(v)\|_x$ by $d_x(w, v)$ for convenience. With this notation, the δ_k is a valid distortion rate at iteration k if

$$d_{x_k}(z_{k-1}, x^\star)^2 \leq \delta_k d_{x_{k-1}}(z_{k-1}, x^\star)^2.$$

D.1. Proof for convergence guarantees of Algorithm 2

Proof [Proof of Theorem 4] We analyse the difference in energy functions at iterations k and $k+1$. We assume that f is μ-strongly g-convex, and at the end make the substitution $\mu = 0$.

$$\begin{aligned}
&E_{k+1} - E_k \\
&= \underbrace{A_{k+1} \cdot (f(y_{k+1}) - f(x^\star)) - A_k \cdot (f(y_k) - f(x^\star))}_{\Delta E_k^F} + \underbrace{B_{k+1} \cdot d_{x_{k+1}}(z_{k+1}, x^\star)^2 - B_k \cdot d_{x_k}(z_k, x^\star)^2}_{\Delta E_k^D}
\end{aligned}$$

We begin by simplifying ΔE_k^D.

$$\begin{aligned}
\Delta E_k^D &\overset{(i)}{\leq} B_{k+1} d_{x_{k+1}}(z_{k+1}, x^\star)^2 - \frac{B_k}{\delta_{k+1}} d_{x_{k+1}}(z_k, x^\star)^2 \\
&= \underbrace{\left(B_{k+1} - \frac{B_k}{\delta_{k+1}}\right)}_{\overline{B}_k} d_{x_{k+1}}(z_{k+1}, x^\star)^2 + \frac{B_k}{\delta_{k+1}}(d_{x_{k+1}}(z_{k+1}, x^\star)^2 - d_{x_{k+1}}(z_k, x^\star)^2) \\
&\overset{(ii)}{=} \overline{B}_k d_{x_{k+1}}(z_{k+1}, x^\star)^2 - \frac{B_k}{\delta_{k+1}} d_{x_{k+1}}(z_k, z_{k+1})^2 \\
&\quad + \frac{2B_k}{\delta_{k+1}} \left(\langle \mathrm{Log}_{x_{k+1}}(z_{k+1}) - \mathrm{Log}_{x_{k+1}}(x^\star), \mathrm{Log}_{x_{k+1}}(z_{k+1}) - \mathrm{Log}_{x_{k+1}}(z_k)\rangle_{x_{k+1}}\right)
\end{aligned}$$

$$
\begin{aligned}
&\overset{(iii)}{=} \overline{B}_k d_{x_{k+1}}(z_{k+1}, x^\star)^2 - \frac{B_k}{\delta_{k+1}} d_{x_{k+1}}(z_k, z_{k+1})^2 \\
&\quad + \frac{2B_k}{\beta_{k+1}\delta_{k+1}} \left(\left\langle \mathrm{Log}_{x_{k+1}}(z_{k+1}) - \mathrm{Log}_{x_{k+1}}(x^\star), -\alpha_{k+1}\mathrm{Log}_{x_{k+1}}(z_{k+1}) - \mathrm{grad} f(x_{k+1}) \right\rangle_{x_{k+1}} \right) \\
&\overset{(iv)}{=} \overline{B}_k d_{x_{k+1}}(z_{k+1}, x^\star)^2 - \frac{B_k}{\delta_{k+1}} d_{x_{k+1}}(z_k, z_{k+1})^2 \\
&\quad + \frac{2B_k}{\beta_{k+1}\delta_{k+1}} \left(\left\langle \mathrm{Log}_{x_{k+1}}(z_{k+1}) - \mathrm{Log}_{x_{k+1}}(x^\star), -\alpha_{k+1}\mathrm{Log}_{x_{k+1}}(z_{k+1}) + \alpha_{k+1}\mathrm{Log}_{x_{k+1}}(x_{k+1}) \right\rangle_{x_{k+1}} \right. \\
&\qquad \left. - \left\langle \mathrm{Log}_{x_{k+1}}(z_{k+1}) - \mathrm{Log}_{x_{k+1}}(x^\star), \mathrm{grad} f(x_{k+1}) \right\rangle_{x_{k+1}} \right) \\
&\overset{(iv)}{=} \overline{B}_k d_{x_{k+1}}(z_{k+1}, x^\star)^2 \\
&\quad + \frac{2B_k\alpha_{k+1}}{\delta_{k+1}\beta_{k+1}} \left\langle \mathrm{Log}_{x_{k+1}}(x^\star) - \mathrm{Log}_{x_{k+1}}(z_{k+1}), \mathrm{Log}_{x_{k+1}}(z_{k+1}) - \mathrm{Log}_{x_{k+1}}(x_{k+1}) \right\rangle_{x_{k+1}} \\
&\qquad + \frac{2B_k}{\delta_{k+1}\beta_{k+1}} \left\langle \mathrm{Log}_{x_{k+1}}(x^\star) - \mathrm{Log}_{x_{k+1}}(z_{k+1}), \mathrm{grad} f(x_{k+1}) \right\rangle_{x_{k+1}} - \frac{B_k}{\delta_{k+1}} d_{x_{k+1}}(z_k, z_{k+1})^2 \\
&\overset{(v)}{=} \overline{B}_k d_{x_{k+1}}(z_{k+1}, x^\star)^2 \\
&\quad + \frac{B_k\alpha_{k+1}}{\delta_{k+1}\beta_{k+1}} \left(d_{x_{k+1}}(x_{k+1}, x^\star)^2 - d_{x_{k+1}}(z_{k+1}, x_{k+1})^2 - d_{x_{k+1}}(z_{k+1}, x^\star)^2 \right) \\
&\qquad + \frac{2B_k}{\delta_{k+1}\beta_{k+1}} \left\langle \mathrm{Log}_{x_{k+1}}(x^\star) - \mathrm{Log}_{x_{k+1}}(z_{k+1}), \mathrm{grad} f(x_{k+1}) \right\rangle_{x_{k+1}} - \frac{B_k}{\delta_{k+1}} d_{x_{k+1}}(z_k, z_{k+1})^2 \\
&= \underbrace{\left(\overline{B}_k - \frac{B_k\alpha_{k+1}}{\delta_{k+1}\beta_{k+1}} \right) d_{x_{k+1}}(z_{k+1}, x^\star)^2}_{T_1} - \frac{B_k}{\delta_{k+1}} \left(\frac{\alpha_{k+1}}{\beta_{k+1}} d_{x_{k+1}}(z_{k+1}, x_{k+1})^2 + d_{x_{k+1}}(z_{k+1}, z_k)^2 \right) \\
&\quad + \frac{2B_k}{\delta_{k+1}\beta_{k+1}} \left\langle \mathrm{Log}_{x_{k+1}}(x^\star) - \mathrm{Log}_{x_{k+1}}(z_{k+1}), \mathrm{grad} f(x_{k+1}) \right\rangle_{x_{k+1}} + \frac{B_k\alpha_{k+1}}{\delta_{k+1}\beta_{k+1}} d_{x_{k+1}}(x_{k+1}, x^\star)^2 \\
&= T_1 - \frac{B_k(\alpha_{k+1} + \beta_{k+1})}{\delta_{k+1}\beta_{k+1}} \left(\frac{\alpha_{k+1}}{\alpha_{k+1} + \beta_{k+1}} d_{x_{k+1}}(z_{k+1}, x_{k+1})^2 + \frac{\beta_{k+1}}{\alpha_{k+1} + \beta_{k+1}} d_{x_{k+1}}(z_k, z_{k+1})^2 \right) \\
&\quad + \frac{2B_k}{\delta_{k+1}\beta_{k+1}} \left\langle \mathrm{Log}_{x_{k+1}}(x^\star) - \mathrm{Log}_{x_{k+1}}(z_{k+1}), \mathrm{grad} f(x_{k+1}) \right\rangle_{x_{k+1}} + \frac{B_k\alpha_{k+1}}{\delta_{k+1}\beta_{k+1}} d_{x_{k+1}}(x_{k+1}, x^\star)^2 \\
&\overset{(vi)}{=} T_1 - \frac{B_k(\alpha_{k+1} + \beta_{k+1})}{\delta_{k+1}\beta_{k+1}} \|\mathrm{Log}_{x_{k+1}}(z_{k+1}) - w_{k+1}\|_{x_{k+1}}^2 \\
&\quad + \frac{2B_k}{\delta_{k+1}\beta_{k+1}} \left\langle \mathrm{Log}_{x_{k+1}}(x^\star) - \mathrm{Log}_{x_{k+1}}(z_{k+1}), \mathrm{grad} f(x_{k+1}) \right\rangle_{x_{k+1}} + \frac{B_k\alpha_{k+1}}{\delta_{k+1}\beta_{k+1}} d_{x_{k+1}}(x_{k+1}, x^\star)^2 \\
&\overset{(vii)}{=} T_1 - \frac{B_k(\alpha_{k+1} + \beta_{k+1})}{\delta_{k+1}\beta_{k+1}} \|\mathrm{Log}_{x_{k+1}}(z_{k+1}) - w_{k+1}\|_{x_{k+1}}^2 + \frac{B_k\alpha_{k+1}}{\delta_{k+1}\beta_{k+1}} d_{x_{k+1}}(x_{k+1}, x^\star)^2 \\
&\quad + \frac{2B_k}{\delta_{k+1}\beta_{k+1}} \langle \mathrm{Log}_{x_{k+1}}(x^\star), \mathrm{grad} f(x_{k+1}) \rangle_{x_{k+1}} - \frac{2B_k}{\delta_{k+1}\beta_{k+1}} \langle w_{k+1}, \mathrm{grad} f(x_{k+1}) \rangle_{x_{k+1}} \\
&\qquad + \frac{2B_k}{\delta_{k+1}\beta_{k+1}} \langle w_{k+1} - \mathrm{Log}_{x_{k+1}}(z_{k+1}), \mathrm{grad} f(x_{k+1}) \rangle_{x_{k+1}} \\
&\overset{(viii)}{\leq} T_1 - \frac{B_k(\alpha_{k+1} + \beta_{k+1})}{\delta_{k+1}\beta_{k+1}} \|\mathrm{Log}_{x_{k+1}}(z_{k+1}) - w_{k+1}\|_{x_{k+1}}^2 + \frac{B_k\alpha_{k+1}}{\delta_{k+1}\beta_{k+1}} d_{x_{k+1}}(x_{k+1}, x^\star)^2
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{2B_k}{\delta_{k+1}\beta_{k+1}}(f(x^\star) - f(x_{k+1}) - \mu/2 \cdot d(x_{k+1}, x^\star)) + \frac{2B_k}{\delta_{k+1}\beta_{k+1}(\alpha_{k+1}+\beta_{k+1})}\|\mathrm{grad} f(x_{k+1})\|^2_{x_{k+1}} \\
&\quad - \frac{2B_k}{\delta_{k+1}\beta_{k+1}}\langle w_{k+1}, \mathrm{grad} f(x_{k+1})\rangle_{x_{k+1}} \\
&= T_1 \underbrace{- \frac{B_k(\alpha_{k+1}+\beta_{k+1})}{\delta_{k+1}\beta_{k+1}}\|\mathrm{Log}_{x_{k+1}}(z_{k+1}) - w_{k+1}\|^2_{x_{k+1}}}_{\leq 0} + \frac{B_k}{\delta_{k+1}\beta_{k+1}}(\alpha_{k+1}-\mu) d_{x_{k+1}}(x_{k+1}, x^\star)^2 \\
&+ \frac{2B_k}{\delta_{k+1}\beta_{k+1}}(f(x^\star) - f(x_{k+1})) + \frac{2B_k}{\delta_{k+1}\beta_{k+1}(\alpha_{k+1}+\beta_{k+1})}\|\mathrm{grad} f(x_{k+1})\|^2_{x_{k+1}} \\
&\quad - \frac{2B_k}{\delta_{k+1}\beta_{k+1}}\langle w_{k+1}, \mathrm{grad} f(x_{k+1})\rangle_{x_{k+1}} \\
&\leq T_1 + \frac{B_k}{\delta_{k+1}\beta_{k+1}}(\alpha_{k+1}-\mu) d_{x_{k+1}}(x_{k+1}, x^\star)^2 + \frac{2B_k}{\delta_{k+1}\beta_{k+1}}(f(x^\star) - f(x_{k+1})) \\
&+ \frac{2B_k}{\delta_{k+1}\beta_{k+1}(\alpha_{k+1}+\beta_{k+1})}\|\mathrm{grad} f(x_{k+1})\|^2_{x_{k+1}} - \frac{2B_k}{\delta_{k+1}\beta_{k+1}}\langle w_{k+1}, \mathrm{grad} f(x_{k+1})\rangle_{x_{k+1}} \\
&\overset{(ix)}{=} \frac{B_k}{\delta_{k+1}\beta_{k+1}}(\alpha_{k+1}-\mu) d_{x_{k+1}}(x_{k+1}, x^\star)^2 + \frac{2B_k}{\delta_{k+1}\beta_{k+1}(\alpha_{k+1}+\beta_{k+1})}\|\mathrm{grad} f(x_{k+1})\|^2_{x_{k+1}} \\
&+ \frac{2B_k}{\delta_{k+1}\beta_{k+1}}(f(x^\star) - f(x_{k+1})) - \frac{2B_k}{\delta_{k+1}\beta_{k+1}}\langle w_{k+1}, \mathrm{grad} f(x_{k+1})\rangle_{x_{k+1}} \\
&\overset{(x)}{=} \frac{B_k}{\delta_{k+1}\beta_{k+1}}(\alpha_{k+1}-\mu) d_{x_{k+1}}(x_{k+1}, x^\star)^2 + \frac{2B_k}{\delta_{k+1}\beta_{k+1}(\alpha_{k+1}+\beta_{k+1})}\|\mathrm{grad} f(x_{k+1})\|^2_{x_{k+1}} \\
&+ \frac{2B_k}{\delta_{k+1}\beta_{k+1}}(f(x^\star) - f(x_{k+1})) - \frac{2B_k}{\delta_{k+1}(\alpha_{k+1}+\beta_{k+1})}\langle \mathrm{Log}_{x_{k+1}}(z_k), \mathrm{grad} f(x_{k+1})\rangle_{x_{k+1}} \\
&\overset{(xi)}{=} \frac{B_k}{\delta_{k+1}\beta_{k+1}}(\alpha_{k+1}-\mu) d_{x_{k+1}}(x_{k+1}, x^\star)^2 + \frac{2B_k}{\delta_{k+1}\beta_{k+1}(\alpha_{k+1}+\beta_{k+1})}\|\mathrm{grad} f(x_{k+1})\|^2_{x_{k+1}} \\
&+ \frac{2B_k}{\delta_{k+1}\beta_{k+1}}(f(x^\star) - f(x_{k+1})) + \frac{2B_k(1-\tau_{k+1})}{\delta_{k+1}(\alpha_{k+1}+\beta_{k+1})\tau_{k+1}}\langle \mathrm{Log}_{x_{k+1}}(y_k), \mathrm{grad} f(x_{k+1})\rangle_{x_{k+1}} \\
&\leq \frac{B_k}{\delta_{k+1}\beta_{k+1}}(\alpha_{k+1}-\mu) d_{x_{k+1}}(x_{k+1}, x^\star)^2 + \frac{2B_k}{\delta_{k+1}\beta_{k+1}(\alpha_{k+1}+\beta_{k+1})}\|\mathrm{grad} f(x_{k+1})\|^2_{x_{k+1}} \\
&+ \frac{2B_k}{\delta_{k+1}\beta_{k+1}}(f(x^\star) - f(x_{k+1})) \\
&\quad + \frac{2B_k(1-\tau_{k+1})}{\delta_{k+1}(\alpha_{k+1}+\beta_{k+1})\tau_{k+1}}(f(y_k) - f(x_{k+1}) \underbrace{- \mu/2 \cdot d_{x_{k+1}}(x_{k+1}, y_k)^2}_{\leq 0}) \\
&\overset{(xii)}{\leq} \overline{A}_k\left(\frac{\overline{B}_k}{\overline{A}_k} - \mu\right) d_{x_{k+1}}(x_{k+1}, x^\star)^2 + \frac{2\overline{A}_k^2}{B_{k+1}}\|\mathrm{grad} f(x_{k+1})\|^2_{x_{k+1}} + 2\overline{A}_k(f(x^\star) - f(x_{k+1})) \\
&+ \frac{2B_k(1-\tau_{k+1})\overline{A}_k}{\delta_{k+1}B_{k+1}\tau_{k+1}}(f(y_k) - f(x_{k+1})) \\
&= (\overline{B}_k - \overline{A}_k\mu) d_{x_{k+1}}(x_{k+1}, x^\star)^2 + \frac{2\overline{A}_k^2}{B_{k+1}}\|\mathrm{grad} f(x_{k+1})\|^2_{x_{k+1}} + 2\overline{A}_k(f(x^\star) - f(x_{k+1}))
\end{aligned}
$$

$$+ \frac{2B_k(1-\tau_{k+1})\overline{A}_k}{\delta_{k+1}B_{k+1}\tau_{k+1}}(f(y_k) - f(x_{k+1})).$$

We explain the specific marked steps in the calculations made above next.

(i)**:** This holds from the existence of a valid distortion rate.

(ii)**:** Since the tangent space $\mathcal{T}_{x_{k+1}}\mathcal{M}$ is Euclidean, and we have the canonical three-term lemma:

$$d_w(a,b)^2 + d_w(b,c)^2 - d_w(c,a)^2 = 2\langle \mathrm{Log}_w(b) - \mathrm{Log}_w(a), \mathrm{Log}_w(b) - \mathrm{Log}_w(c)\rangle_w$$

where $w = x_{k+1}, a = x^\star, b = z_{k+1}, c = z_k$.

(iii)**:** This holds due to the update step (Eq. 2c).

$$\frac{\alpha_{k+1}+\beta_{k+1}}{\beta_{k+1}}\mathrm{Log}_{x_{k+1}}(z_{k+1}) = \mathrm{Log}_{x_{k+1}}(z_k) - \frac{1}{\beta_{k+1}}\mathrm{grad}f(x_{k+1})$$
$$\Rightarrow \mathrm{Log}_{x_{k+1}}(z_{k+1}) - \mathrm{Log}_{x_{k+1}}(z_k) = -\frac{\alpha_{k+1}}{\beta_{k+1}}\mathrm{Log}_{x_{k+1}}(z_{k+1}) - \frac{1}{\beta_{k+1}}\mathrm{grad}f(x_{k+1}).$$

(iv)**:** $\mathrm{Log}_a(a) = 0$ for all $a \in \mathcal{M}$.

(v)**:** Reuses the three-term lemma with $w = x_{k+1}$, $a = x_{k+1}$, $b = z_{k+1}$, $c = x^\star$.

(vi)**:** Here, we use the convexity of the Euclidean norm. Specifically,

$$\frac{\lambda}{2}\|a-b\|^2 + \frac{1-\lambda}{2}\|a-c\|^2 \geq \frac{1}{2}\|a - \lambda b - (1-\lambda)c\|^2.$$

and we use this with $a = \mathrm{Log}_{x_{k+1}}(z_{k+1}), b = \mathrm{Log}_{x_{k+1}}(x_{k+1}), c = \mathrm{Log}_{x_{k+1}}(z_k), \lambda = \frac{\alpha_{k+1}}{\beta_{k+1}+\alpha_{k+1}}$ and $\lambda b + (1-\lambda)c = w_{k+1}$ for some $w_{k+1} \in \mathcal{T}_{x_{k+1}}\mathcal{M}$.

(vii)**:** Adding and subtracting w_{k+1} in the inner products.

$(viii)$**:** Holds due to the fact that f is μ-strongly g-convex. Additionally, we use the fact that

$$w_{k+1} = \lambda b + (1-\lambda)c = \frac{\beta_{k+1}}{\alpha_{k+1}+\beta_{k+1}}\mathrm{Log}_{x_{k+1}}(z_k)$$
$$= \mathrm{Log}_{x_{k+1}}(z_{k+1}) + \frac{1}{\alpha_{k+1}+\beta_{k+1}}\mathrm{grad}f(x_{k+1})$$

(ix)**:** The choice of α_{k+1} and β_{k+1} satisfies $\overline{B}_k = \frac{B_k\alpha_{k+1}}{\delta_{k+1}\beta_{k+1}}$ and hence $T_1 = 0$.

(x)**:** Reuses the midpoint fact in $(viii)$.

(xi)**:** Uses the fact that if $z = \mathrm{Exp}_x(\alpha\mathrm{Log}_x(y))$ then $-\mathrm{Log}_z(x) = \frac{\alpha}{1-\alpha}\mathrm{Log}_z(y)$ with $\alpha = \tau_{k+1}, z = x_{k+1}, x = y_k, y = z_k$.

(xii)**:** Holds due to the form of the constants α_{k+1} and β_{k+1}.

Next, we look at ΔE_k^F.

$$\Delta E_k^F = A_{k+1}(f(y_{k+1}) - f(x_{k+1})) + A_k(f(x_{k+1}) - f(y_k)) + \overline{A}_k(f(x_{k+1}) - f(x^\star)).$$

As a result of these computations:

$$\Delta E_k \leq A_{k+1}(f(y_{k+1}) - f(x_{k+1})) + \left(A_{k+1} - \frac{B_k\overline{A}_k(1-\tau_{k+1})}{\delta_{k+1}B_{k+1}\tau_{k+1}}\right)(f(x_{k+1}) - f(y_k)) + \frac{2\overline{A}_k^2}{B_{k+1}}\|\mathrm{grad} f(x_{k+1})\|_{x_{k+1}}^2 + (\overline{B}_k - \mu\overline{A}_k)d_{x_{k+1}}(x_{k+1}, x^\star)^2.$$

The choice of τ_{k+1} ensure that the coefficient of the $f(x_{k+1}) - f(y_k)$ term is 0. This leads to, along with $\mu = 0$,

$$\Delta E_k \leq A_{k+1}(f(y_{k+1}) - f(x_{k+1})) + \frac{2\overline{A}_k^2}{B_{k+1}}\|\mathrm{grad} f(x_{k+1})\|_{x_{k+1}}^2 + \overline{B}_k d_{x_{k+1}}(x_{k+1}, x^\star)^2.$$

When, G_c is a 2-backward descent method, $f(y_{k+1}) - f(x_{k+1}) \leq -c\|\mathrm{grad} f(x_{k+1})\|_{x_{k+1}}^2$. This gives

$$E_{k+1} - E_k \leq -\left(cA_{k+1} - \frac{2\overline{A}_k^2}{B_{k+1}}\right)\|\mathrm{grad} f(x_{k+1})\|_{x_{k+1}}^2 + \overline{B}_k R^2.$$

Choose $B_{k+1} = \frac{4}{c}$ and $A_{k+1} = \frac{(k+1)(k+2)}{2}$. Note that for this choice $\overline{A}_k = A_{k+1} - A_k = (k+1)$ and therefore, $cA_{k+1} > \frac{c\overline{A}_k^2}{2}$. Due to this,

$$E_{k+1} - E_k \leq \frac{4}{c}\left(1 - \frac{1}{\delta_{k+1}}\right)\mathrm{diam}(A)^2 \Rightarrow E_T - E_0 \leq \frac{4T}{c}\left(1 - \frac{1}{\delta_{\max}}\right)\mathrm{diam}(A)^2.$$

This gives us a rate

$$f(y_T) - f(x^\star) \leq \frac{E_0}{A_T} + \frac{\frac{4}{c}\left(1 - \frac{1}{\delta_{\max}}\right)T}{A_T} \leq \frac{E_0}{T^2} + \frac{\frac{4}{c}\left(1 - \frac{1}{\delta_{\max}}\right)\mathrm{diam}(A)^2}{T}.$$

■

Proof [Proof of Proposition 5] The proof of this proposition is directly given by Ahn and Sra [2, Theorem 3.1] where we make the substitution $\Delta_\gamma \to c$. By definition $2\mu\Delta_\gamma = 2\mu \cdot \gamma(1 - {}^{L\gamma}/_{2})$ is strictly less than 0 under the preconditions of their theorem, whereas due to our generality, we will have to enforce it as a property of G_c. We also find that their theorem holds more generally when Exp and Log is well-defined at every $x \in A$, hence the additional assumptions ($\mathbf{A}_1$), ($\mathbf{A}_2$) and ($\mathbf{A}_3$). As noted earlier, when $\mathcal{M}$ is a Hadamard manifold, these assumptions hold. ■

D.2. Proofs for the sufficient conditions

Proof [Proof of Lemma 6] We carefully follow the proof of Lemma 4.2 in Ahn and Sra [2]. By our assumption, $\xi_0 \leq \sqrt{2\mu c}$ and $2\mu c < 1$. For convenience, we use $\lambda_{k+1} = \frac{\beta_{k+1}}{\beta_{k+1}+\alpha_{k+1}}$ and $\eta_{k+1} = \frac{1}{\beta_{k+1}+\alpha_{k+1}}$.

$$
\begin{aligned}
d(x_{k+1}, z_{k+1}) &\overset{(i)}{=} \|\mathrm{Log}_{x_{k+1}}(z_{k+1})\|_{x_{k+1}} \\
&\overset{(ii)}{=} \|\lambda_{k+1}\mathrm{Log}_{x_{k+1}}(z_k) - \eta_{k+1}\mathrm{grad}f(x_{k+1})\|_{x_{k+1}} \\
&\overset{(iii)}{\leq} \lambda_{k+1}\|\mathrm{Log}_{x_{k+1}}(z_k)\|_{x_{k+1}} + \eta_{k+1}\|\mathrm{grad}f(x_{k+1})\|_{x_{k+1}} \\
&\overset{(iv)}{\leq} \lambda_{k+1}d(x_{k+1}, z_k) + \eta_{k+1}Ld(x_{k+1}, x^\star) \\
&\overset{(v)}{\leq} \lambda_{k+1}d(x_{k+1}, z_k) + \eta_{k+1}Ld(x_{k+1}, y_k) + \eta_{k+1}Ld(y_k, x^\star) \\
&\overset{(vi)}{=} \lambda_{k+1}(1-\tau_{k+1})d(y_k, z_k) + \eta_{k+1}L\tau_{k+1}d(y_k, z_k) + \eta_{k+1}Ld(y_k, x^\star) \\
&\overset{(vii)}{=} d(y_k, z_k)(\lambda_{k+1}(1-\tau_{k+1}) + \eta_{k+1}L\tau_{k+1}) + \eta_{k+1}Ld(y_k, x^\star).
\end{aligned}
$$

We provide justifications for the steps above next.

(i)**:** This holds by definition.

(ii)**:** This is due to Eq. 2c with the new notation for parameters.

(iii)**:** Uses the triangle inequality in the tangent sapce at x_{k+1} which is Euclidean.

(iv)**:** Uses the fact that f is L-g-smooth.

(v)**:** Employs a triangle inequality over $\mathcal{M}$.

(vi)**:** This holds since x_{k+1} lies between y_k and z_k.

To get a bound on $d(x_{k+1}, z_{k+1})$, we need to have a bound on $d(y_k, z_k)$ and $d(y_k, x^\star)$. We can use the energy inequality from Prop. 5 with μ-strong g-convexity to get the following statements

$$
\frac{\mu}{2} \cdot d(y_k, x^\star)^2 \leq \prod_{j=1}^{k}(1-\xi_j)D_0 \Leftrightarrow d(y_k, x^\star) \leq \sqrt{\prod_{j=1}^{k}(1-\xi_j)D_0}\sqrt{\frac{2}{\mu}}, \tag{11}
$$

$$
\mu^2 c \cdot d_{x_k}(z_k, x^\star)^2 \leq \prod_{j=1}^{k}(1-\xi_j)D_0 \Leftrightarrow d_{x_k}(z_k, x^\star) \leq \sqrt{\prod_{j=1}^{k}(1-\xi_j)D_0}\sqrt{\frac{1}{\mu^2 c}}. \tag{12}
$$

With these we also have

$$
\begin{aligned}
d_{x_k}(y_k, z_k) &\leq d_{x_k}(y_k, x^\star) + d_{x_k}(z_k, x^\star) && \because \triangle \text{ inequality} \\
&\leq d(y_k, x^\star) + d_{x_k}(z_k, x^\star) && \because d_a(b,c) \leq d(b,c) \text{ for Hadamard manifolds} \\
&\leq \sqrt{\prod_{j=1}^{k}(1-\xi_j)D_0}\left(\sqrt{\frac{2}{\mu}} + \sqrt{\frac{1}{\mu^2 c}}\right) && \because \text{Eqs. } 11, 12.
\end{aligned}
$$

However, this doesn't quite help us yet, since $d_{x_k}(y_k, z_k) \leq d(y_k, z_k)$, and we need the quantity on the RHS for the upper bound on $d(x_{k+1}, z_{k+1})$. Following the proof of Ahn and Sra [2, Prop. C.7], we will analyse the quantity $d_{x_{k+1}}(y_{k+1}, z_{k+1})$.

$$\begin{aligned}
d_{x_{k+1}}(y_{k+1}, z_{k+1}) &\overset{(i)}{\geq} -d_{x_{k+1}}(y_{k+1}, x_{k+1}) + d_{x_{k+1}}(x_{k+1}, z_{k+1}) \\
&\overset{(ii)}{=} -d_{x_{k+1}}(y_{k+1}, x_{k+1}) + \|\mathrm{Log}_{x_{k+1}}(z_{k+1})\|_{x_{k+1}} \\
&\overset{(iii)}{=} -d_{x_{k+1}}(y_{k+1}, x_{k+1}) + \|\lambda_{k+1}\mathrm{Log}_{x_{k+1}}(z_k) - \eta_{k+1}\mathrm{grad}f(x_{k+1})\|_{x_{k+1}} \\
&\overset{(iv)}{\geq} -d_{x_{k+1}}(y_{k+1}, x_{k+1}) + \lambda_{k+1}d(z_k, x_{k+1}) - \eta_{k+1}\|\mathrm{grad}f(x_{k+1})\|_{x_{k+1}} \\
&\overset{(v)}{\geq} -d(y_{k+1}, x_{k+1}) + \lambda_{k+1}(1-\tau_{k+1})d(y_k, z_k) - \eta_{k+1}\|\mathrm{grad}f(x_{k+1})\|_{x_{k+1}} \\
&\overset{(vi)}{\geq} -d(y_{k+1}, x_{k+1}) + \lambda_{k+1}(1-\tau_{k+1})d(y_k, z_k) - \eta_{k+1}Ld(x_{k+1}, x^\star) \\
&\overset{(vii)}{\geq} -d(y_{k+1}, x_{k+1}) + \lambda_{k+1}(1-\tau_{k+1})d(y_k, z_k) - \eta_{k+1}Ld(x_{k+1}, y_k) - \eta_{k+1}Ld(y_k, x^\star) \\
&\overset{(viii)}{=} -d(y_{k+1}, x_{k+1}) + \lambda_{k+1}(1-\tau_{k+1})d(y_k, z_k) - \eta_{k+1}L\tau_{k+1}d(y_k, z_k) - \eta_{k+1}Ld(y_k, x^\star).
\end{aligned}$$

The steps above are justified below.

(i) Uses the triangle inequality in the tangent space at x_{k+1} which is Euclidean.

(ii) Holds by definition.

(iii) Uses the form of the update Eq. 2c.

(iv) Again uses the triangle inequality in the tangent space at x_{k+1}.

(v) x_{k+1} lies between y_k and z_k.

(vi) Uses the fact that f is L-g-smooth.

(vii) Uses the triangle inequality over $\mathcal{M}$.

$(viii)$ Again uses the fact that x_{k+1} lies between y_k and z_k.

This gives us

$$\begin{aligned}
d(y_k, z_k)(\lambda_{k+1}(1-\tau_{k+1}) - \eta_{k+1}L\tau_{k+1}) &\leq d_{x_{k+1}}(y_{k+1}, z_{k+1}) + d(y_{k+1}, x_{k+1}) + \eta_{k+1}Ld(y_k, x^\star) \\
&\leq \sqrt{\prod_{j=1}^{k}(1-\xi_j)D_0}\left(\sqrt{\frac{2}{\mu}} + \sqrt{\frac{1}{\mu^2 c}}\right) + d(y_{k+1}, x_{k+1}) \\
&\quad + \eta_{k+1}L\sqrt{\prod_{j=1}^{k}(1-\xi_j)D_0}\sqrt{\frac{2}{\mu}}.
\end{aligned}$$

We make note of the fact that $\xi_{k+1} \leq 1$ and use the bound from Eq. 11 and Eq. 12. The final piece is to bound $d(y_{k+1}, x_{k+1})$ and to show that $\lambda_{k+1}(1-\tau_{k+1}) - \eta_{k+1}L\tau_{k+1}$ can be bounded in terms

of a constant involving L, μ, c alone. The first part is given by the statement of the lemma, which states

$$d(y_{k+1}, x_{k+1}) \leq \mathcal{C}'_{L,\mu,c} \sqrt{\prod_{j=1}^{k} (1 - \xi_j) \cdot D_0}.$$

For the second part, we make use of global properties of the recurrence relation governing the sequence $\{\xi_k\}$. From Ahn and Sra [2, Proposition C.9], we have that if $\xi_0 \leq \sqrt{a}$, then $\xi_k \leq \sqrt{a}$ for all $k \geq 0$, where

$$\frac{\xi_{k+1}(\xi_{k+1} - a)}{1 - \xi_{k+1}} = \frac{\xi_k^2}{\delta}$$

for any $\delta \geq 1$ and $a \in (0, 1)$. We use this statement with $a = 2\mu c$, and δ being the valid distortion rate at iteration k which is ≥ 1. Therefore,

$$\begin{aligned} \lambda_{k+1}(1 - \tau_{k+1}) - \eta_{k+1} L \tau_{k+1} &= \left[\frac{1 - 2\mu c \xi_{k+1}^{-1}}{1 - 2\mu c} \right] (1 - \xi_{k+1} - 2Lc) \\ &\geq \left[\frac{1 - 2\mu c \xi_{k+1}^{-1}}{1 - 2\mu c} \right] (1 - \sqrt{2\mu c} - 2Lc). \end{aligned}$$

The quantity $1 - \sqrt{2\mu c} - 2Lc$ strictly positive when $c < 1/6L$. Therefore, we have the bound on $d(y_k, z_k)$ as

$$d(y_k, z_k) \leq \left[\frac{1 - \sqrt{2\mu c \xi_{k+1}^{-1}}}{1 - 2\mu c} \right]^{-1} \frac{1}{1 - \sqrt{2\mu c} - 2Lc} \mathcal{C}''_{L,\mu,c} \sqrt{\prod_{j=1}^{k} (1 - \xi_j) \cdot D_0}.$$

Using this to bound $d(x_{k+1}, z_{k+1})$ we obtain

$$d(x_{k+1}, z_{k+1}) \leq \underbrace{\left(\frac{1 - 2\mu c + 2Lc}{1 - \sqrt{2\mu c} - 2Lc} \mathcal{C}''_{L,\mu,c} + \frac{L\sqrt{2}}{\mu\sqrt{\mu}} \right)}_{\mathcal{C}_{L,\mu,c}} \sqrt{\prod_{j=1}^{k} (1 - \xi_j) \cdot D_0}.$$

■

Proof [Proof of Proposition 7] This proposition can be proven using the analysis of the recurrence relation as presented in Ahn and Sra [2, Section C.7]. The key tool of the analysis is the distance shrinking lemma, which we have proven for 2-backward descent methods in general when c is sufficiently small. ■

Appendix E. Auxiliary lemmas

Lemma 11 (Conjugate lemma [6]) *Let s, u be vectors, and α be a scalar. Then*

$$\langle s, \alpha \cdot u \rangle - \frac{1}{q} \|s\|^q \leq \frac{q-1}{q} |\alpha|^{q/q-1} \|u\|^{q/q-1}.$$

Proof Let u be the maximum of the LHS (taken with respect to s). By the first order optimality, we have

$$\alpha \cdot u - \|s\|^{q-2} s = 0$$

Consequently,

$$\langle s, \alpha \cdot u \rangle - \frac{1}{q}\|s\|^q = \|s\|^q - \frac{1}{q}\|s\|^q = \frac{q-1}{q}\|s\|^q.$$

Also,

$$|\alpha|\|u\| = \|s\|^{q-1}.$$

Hence,

$$\langle s, \alpha \cdot u \rangle - \frac{1}{q}\|s\|^q = \frac{q-1}{q}|\alpha|^{q/q-1}\|u\|^{q/q-1}.$$

■