Foundations of Computational Mathematics
https://doi.org/10.1007/s10208-022-09573-9

An Accelerated First-Order Method for Non-convex Optimization on Manifolds

Christopher Criscitiello[1] · **Nicolas Boumal**[1]

Received: 19 August 2020 / Revised: 3 February 2022 / Accepted: 1 March 2022

Abstract

We describe the first gradient methods on Riemannian manifolds to achieve accelerated rates in the non-convex case. Under Lipschitz assumptions on the Riemannian gradient and Hessian of the cost function, these methods find approximate first-order critical points faster than regular gradient descent. A randomized version also finds approximate second-order critical points. Both the algorithms and their analyses build extensively on existing work in the Euclidean case. The basic operation consists in running the Euclidean accelerated gradient descent method (appropriately safe-guarded against non-convexity) in the current tangent space, then moving back to the manifold and repeating. This requires lifting the cost function from the manifold to the tangent space, which can be done for example through the Riemannian exponential map. For this approach to succeed, the lifted cost function (called the pullback) must retain certain Lipschitz properties. As a contribution of independent interest, we prove precise claims to that effect, with explicit constants. Those claims are affected by the Riemannian curvature of the manifold, which in turn affects the worst-case complexity bounds for our optimization algorithms.

Keywords Optimization on manifolds · Accelerated gradient descent · Non-convex optimization · First-order method · Riemannian manifold · Jacobi field · Curvature

Mathematics Subject Classification 65K05 · 65J05 · 90C26 · 90C48 · 90C60 · 58C05

Communicated by James Renegar.

✉ Nicolas Boumal
nicolas.boumal@epfl.ch

Christopher Criscitiello
christopher.criscitiello@epfl.ch

[1] Ecole Polytechnique Fédérale de Lausanne (EPFL), Institute of Mathematics, EPFL FSB SMA, Station 8, 1015 Lausanne, Switzerland

Published online: 08 June 2022

1 Introduction

We consider optimization problems of the form

$$\min_{x \in \mathcal{M}} f(x) \tag{P}$$

where f is lower-bounded and twice continuously differentiable on a Riemannian manifold $\mathcal{M}$. For the special case where $\mathcal{M}$ is a Euclidean space, problem (P) amounts to smooth, unconstrained optimization. The more general case is important for applications notably in scientific computing, statistics, imaging, learning, communications and robotics: see for example [1, 27].

For a general non-convex objective f, computing a global minimizer of (P) is hard. Instead, our goal is to compute approximate first- and second-order critical points of (P). A number of non-convex problems of interest exhibit the property that second-order critical points are optimal [7, 11, 14, 24, 30, 36, 49]. Several of these are optimization problems on nonlinear manifolds. Therefore, theoretical guarantees for approximately finding second-order critical points can translate to guarantees for approximately solving these problems.

It is therefore natural to ask for fast algorithms which find approximate second-order critical points on manifolds, within a tolerance ϵ (see below). Existing algorithms include RTR [13], ARC [2] and perturbed RGD [20, 44]. Under some regularity conditions, ARC uses Hessian-vector products to achieve a rate of $O(\epsilon^{-7/4})$. In contrast, under the same regularity conditions, perturbed RGD uses only function value and gradient queries, but achieves a poorer rate of $O(\epsilon^{-2})$. Does there exist an algorithm which finds approximate second-order critical points with a rate of $O(\epsilon^{-7/4})$ using only function value and gradient queries? The answer was known to be yes in Euclidean space. Can it also be done on Riemannian manifolds, hence extending applicability to applications treated in the aforementioned references? We resolve that question positively with the algorithm `PTAGD` below.

From a different perspective, the recent success of momentum-based first-order methods in machine learning [42] has encouraged interest in momentum-based first-order algorithms for non-convex optimization which are provably faster than gradient descent [15, 28]. We show such provable guarantees can be extended to optimization under a manifold constraint. From this perspective, our paper is part of a body of work theoretically explaining the success of momentum methods in non-convex optimization.

There has been significant difficulty in accelerating *geodesically convex* optimization on Riemannian manifolds. See "Related literature" below for more details on best known bounds [3] as well as results proving that acceleration in certain settings is impossible on manifolds [26]. Given this difficulty, it is not at all clear a priori that it is possible to accelerate *non-convex* optimization on Riemannian manifolds. Our paper shows that it is in fact possible.

We design two new algorithms and establish worst-case complexity bounds under Lipschitz assumptions on the gradient and Hessian of f. Beyond a theoretical contribution, we hope that this work will provide an impetus to look for more practical fast first-order algorithms on manifolds.

More precisely, if the gradient of f is L-Lipschitz continuous (in the Riemannian sense defined below), it is known that Riemannian gradient descent can find an ϵ-approximate first-order critical point[1] in at most $O(\Delta_f L/\epsilon^2)$ queries, where Δ_f upper-bounds the gap between initial and optimal cost value [8, 13, 47]. Moreover, this rate is optimal in the special case where $\mathcal{M}$ is a Euclidean space [16], but it can be improved under the additional assumption that the Hessian of f is ρ-Lipschitz continuous.

Recently in Euclidean space, Carmon et al. [15] have proposed a deterministic algorithm for this setting (L-Lipschitz gradient, ρ-Lipschitz Hessian) which requires at most $\tilde{O}(\Delta_f L^{1/2}\rho^{1/4}/\epsilon^{7/4})$ queries (up to logarithmic factors), and is independent of dimension. This is a speed up of Riemannian gradient descent by a factor of $\tilde{\Theta}(\sqrt{\frac{L}{\sqrt{\rho\epsilon}}})$. For the Euclidean case, it has been shown under these assumptions that first-order methods require at least $\Omega(\Delta_f L^{3/7}\rho^{2/7}/\epsilon^{12/7})$ queries [17,Thm. 2]. This leaves a gap of merely $\tilde{O}(1/\epsilon^{1/28})$ in the ϵ-dependency.

Soon after, Jin et al. [28] showed how a related algorithm with randomization can find $(\epsilon, \sqrt{\rho\epsilon})$-approximate second-order critical points[2] with the same complexity, up to polylogarithmic factors in the dimension of the search space and in the (reciprocal of) the probability of failure.

Both the algorithm of Carmon et al. [15] and that of Jin et al. [28] fundamentally rely on Nesterov's accelerated gradient descent method (AGD) [40], with safe-guards against non-convexity. To achieve improved rates, AGD builds heavily on a notion of momentum which accumulates across several iterations. This makes it delicate to extend AGD to nonlinear manifolds, as it would seem that we need to transfer momentum from tangent space to tangent space, all the while keeping track of fine properties.

In this paper, we build heavily on the Euclidean work of Jin et al. [28] to show the following. Assume f has Lipschitz continuous gradient and Hessian on a complete Riemannian manifold satisfying some curvature conditions. With at most $\tilde{O}(\Delta_f L^{1/2}\hat{\rho}^{1/4}/\epsilon^{7/4})$ queries (where $\hat{\rho}$ is larger than ρ by an additive term affected by L and the manifold's curvature),

1. It is possible to compute an ϵ-approximate first-order critical point of f with a deterministic first-order method,
2. It is possible to compute an $(\epsilon, \sqrt{\hat{\rho}\epsilon})$-approximate second-order critical point of f with a randomized first-order method.

In the first case, the complexity is independent of the dimension of $\mathcal{M}$. In the second case, the complexity includes polylogarithmic factors in the dimension of $\mathcal{M}$ and in the probability of failure. This parallels the Euclidean setting. In both cases (and in contrast to the Euclidean setting), the Riemannian curvature of $\mathcal{M}$ affects the complexity in two ways: (a) because $\hat{\rho}$ is larger than ρ, and (b) because the results only apply when the target accuracy ϵ is small enough in comparison with some curvature-dependent

[1] That is, a point where the gradient of f has norm smaller than ϵ.

[2] That is, a point where the gradient of f has norm smaller than ϵ and the eigenvalues of the Hessian of f are at least $-\sqrt{\rho\epsilon}$.

thresholds. It is an interesting open question to determine whether such a curvature dependency is inescapable.

We call our first algorithm TAGD for *tangent accelerated gradient descent*,[3] and the second algorithm PTAGD for *perturbed tangent accelerated gradient descent*. Both algorithms and (even more so) their analyses closely mirror the perturbed accelerated gradient descent algorithm (PAGD) of Jin et al. [28], with one core design choice that facilitates the extension to manifolds: instead of transporting momentum from tangent space to tangent space, we run several iterations of AGD (safe-guarded against non-convexity) in individual tangent spaces. After an AGD run in the current tangent space, we "retract" back to a new point on the manifold and initiate another AGD run in the new tangent space. In so doing, we only need to understand the fine behavior of AGD in one tangent space at a time. Since tangent spaces are linear spaces, we can capitalize on existing Euclidean analyses. This general approach is in line with prior work in [20], and is an instance of the dynamic trivializations framework of Lezcano-Casado [33].

In order to run AGD on the tangent space $\mathrm{T}_x\mathcal{M}$ at x, we must "pullback" the cost function f from $\mathcal{M}$ to $\mathrm{T}_x\mathcal{M}$. A geometrically pleasing way to do so is via the exponential map[4] $\mathrm{Exp}_x\colon \mathrm{T}_x\mathcal{M} \to \mathcal{M}$, whose defining feature is that for each $v \in \mathrm{T}_x\mathcal{M}$ the curve $\gamma(t) = \mathrm{Exp}_x(tv)$ is the geodesic of $\mathcal{M}$ passing through $\gamma(0) = x$ with velocity $\gamma'(0) = v$. Then, $\hat{f}_x = f \circ \mathrm{Exp}_x$ is a real function on $\mathrm{T}_x\mathcal{M}$ called the *pullback* of f at x. To analyze the behavior of AGD applied to $\hat{f}_x$, the most pressing question is:

To what extent does $\hat{f}_x = f \circ \mathrm{Exp}_x$ *inherit the Lipschitz properties of* f?

In this paper, we show that if f has Lipschitz continuous gradient and Hessian *and if the gradient of* f *at* x *is sufficiently small*, then $\hat{f}_x$ *restricted to a ball around the origin of* $\mathrm{T}_x\mathcal{M}$ has Lipschitz continuous gradient and retains *partial Lipschitz-type properties for its Hessian*. The norm condition on the gradient and the radius of the ball are dictated by the Riemannian curvature of $\mathcal{M}$. These geometric results are of independent interest.

Because $\hat{f}_x$ retains only partial Lipschitzness, our algorithms depart from the Euclidean case in the following ways: (a) at points where the gradient is still large, we perform a simple gradient step; and (b) when running AGD in $\mathrm{T}_x\mathcal{M}$, we are careful not to leave the prescribed ball around the origin: if we ever do, we take appropriate action. For those reasons and also because we do not have full Lipschitzness but only radial Lipschitzness for the Hessian of $\hat{f}_x$, minute changes throughout the analysis of Jin et al. [28] are in order.

To be clear, in their current state, TAGD and PTAGD are theoretical constructs. As one can see from later sections, running them requires the user to know the value of several parameters that are seldom available (including the Lipschitz constants L and ρ); the target accuracy ϵ must be set ahead of time; and the tuning constants as dictated here by the theory are (in all likelihood) overly cautious.

---

[3] We refrain from calling our first algorithm "accelerated Riemannian gradient descent," thinking this name should be reserved for algorithms which emulate the momentum approach on the manifold directly.

[4] The exponential map is a *retraction*: our main optimization results are stated for general retractions.

Moreover, to compute the gradient of $\hat{f}_x$ we need to differentiate through the exponential map (or a retraction, as the case may be). This is sometimes easy to do in closed form (see [33] for families of examples), but it could be a practical hurdle. On the other hand, our algorithms do not require parallel transport. It remains an interesting open question to develop *practical* accelerated gradient methods for non-convex problems on manifolds.

In closing this introduction, we give simplified statements of our main results. These are all phrased under the following assumption (see Sect. 2 for geometric definitions):

A 1 The Riemannian manifold $\mathcal{M}$ and the cost function $f : \mathcal{M} \to \mathbb{R}$ have these properties:

- $\mathcal{M}$ is complete, its sectional curvatures are in the interval $[-K, K]$, and the covariant derivative of its Riemann curvature endomorphism is bounded by F in operator norm; and
- f is lower-bounded by f_{low}, has L-Lipschitz continuous gradient $\operatorname{grad} f$ and ρ-Lipschitz continuous Hessian $\operatorname{Hess} f$ on $\mathcal{M}$.

1.1 Main Geometry Results

As a geometric contribution, we show that pullbacks through the exponential map retain certain Lipschitz properties of f. Explicitly, at a point $x \in \mathcal{M}$ we have the following statement.

Theorem 1.1 *Let $x \in \mathcal{M}$. Under A1, let $B_x(b)$ be the closed ball of radius $b \leq \min\left(\frac{1}{4\sqrt{K}}, \frac{K}{4F}\right)$ around the origin in $\mathrm{T}_x\mathcal{M}$. If $\|\operatorname{grad} f(x)\| \leq Lb$, then*

1. *The pullback $\hat{f}_x = f \circ \mathrm{Exp}_x$ has $2L$-Lipschitz continuous gradient $\nabla \hat{f}_x$ on $B_x(b)$, and*
2. *For all $s \in B_x(b)$, we have $\|\nabla^2 \hat{f}_x(s) - \nabla^2 \hat{f}_x(0)\| \leq \hat{\rho}\|s\|$ with $\hat{\rho} = \rho + L\sqrt{K}$.*

(Above, $\|\cdot\|$ denotes both the Riemannian norm on $\mathrm{T}_x\mathcal{M}$ and the associated operator norm. Also, $\nabla \hat{f}_x$ and $\nabla^2 \hat{f}_x$ are the gradient and Hessian of $\hat{f}_x$ on the Euclidean space $\mathrm{T}_x\mathcal{M}$.)

We expect this result to be useful in several other contexts. Section 2 provides a more complete (and somewhat more general) statement. At the same time and independently, Lezcano-Casado [35] develops similar geometric bounds and applies them to study gradient descent in tangent spaces—see "Related literature" below for additional details.

1.2 Main Optimization Results

We aim to compute approximate first- and second-order critical points of f, as defined here:

Definition 1.2 A point $x \in \mathcal{M}$ is an *ϵ-FOCP* for (P) if $\|\operatorname{grad} f(x)\| \leq \epsilon$. A point $x \in \mathcal{M}$ is an *(ϵ_1, ϵ_2)-SOCP* for (P) if $\|\operatorname{grad} f(x)\| \leq \epsilon_1$ and $\lambda_{\min}(\operatorname{Hess} f(x)) \geq -\epsilon_2$, where $\lambda_{\min}(\cdot)$ extracts the smallest eigenvalue of a self-adjoint operator.

In Sect. 5, we define and analyze the algorithm TAGD. Resting on the geometric result above, that algorithm with the exponential retraction warrants the following claim about the computation of first-order points. The $O(\cdot)$ notation is with respect to scaling in ϵ.

Theorem 1.3 *If A1 holds, there exists an algorithm (*TAGD*) which, given any $x_0 \in \mathcal{M}$ and small enough tolerance $\epsilon > 0$, namely,*

$$\begin{aligned}\epsilon &\leq \frac{1}{144}\min\left(\frac{1}{K}\hat{\rho}, \frac{K^2}{F^2}\hat{\rho}, \frac{36\ell^2}{\hat{\rho}}\right)\\ &= \frac{1}{144}\min\left(\frac{1}{K}, \frac{K^2}{F^2}, \left(\frac{12L}{\rho + L\sqrt{K}}\right)^2\right)(\rho + L\sqrt{K}),\end{aligned} \tag{1}$$

produces an ϵ-FOCP for (P) using at most a constant multiple of T function and pullback gradient queries, and a similar number of evaluations of the exponential map, where

$$\begin{aligned}T &= (f(x_0) - f_{\text{low}})\frac{\hat{\rho}^{1/4}\ell^{1/2}}{\epsilon^{7/4}}\log\left(\frac{16\ell}{\sqrt{\hat{\rho}\epsilon}}\right)^6\\ &= O\left((f(x_0) - f_{\text{low}})(\rho + L\sqrt{K})^{1/4}L^{1/2}\cdot\frac{1}{\epsilon^{7/4}}\log\left(\frac{1}{\epsilon}\right)^6\right),\end{aligned}$$

with $\ell = 2L$ and $\hat{\rho} = \rho + L\sqrt{K}$. The algorithm uses no Hessian queries and is deterministic.

This result is dimension-free but not curvature-free because K and F constrain ϵ and affect $\hat{\rho}$.

Remark 1.4 In the statements of all theorems and lemmas, the notations $O(\cdot)$, $\Theta(\cdot)$ only hide universal constants, i.e., numbers like $\frac{1}{2}$ or 100. They do not hide any parameters. Moreover, $\tilde{O}(\cdot)$, $\tilde{\Theta}(\cdot)$ only hide universal constants and logarithmic factors in the parameters.

Remark 1.5 If ϵ is large enough (that is, if $\epsilon > \Theta(\frac{\ell^2}{\hat{\rho}})$), then TAGD reduces to vanilla Riemannian gradient descent with constant step-size. The latter is known to produce an ϵ-FOCP in $O(1/\epsilon^2)$ iterations, yet our result here announces this same outcome in $O(1/\epsilon^{7/4})$ iterations. This is not a contradiction: when ϵ is large, $1/\epsilon^{7/4}$ can be worse than $1/\epsilon^2$. In short: the rates are only meaningful for small ϵ, in which case TAGD does use accelerated gradient descent steps.

In Sect. 6 we define and analyze the algorithm PTAGD. With the exponential retraction, the latter warrants the following claim about the computation of second-order points.

Theorem 1.6 *If A1 holds, there exists an algorithm (*PTAGD*) which, given any $x_0 \in \mathcal{M}$, any $\delta \in (0, 1)$ and small enough tolerance $\epsilon > 0$ (same condition as in Theorem 1.3)* *produces an ϵ-FOCP for (P) using at most a constant multiple of T function and pullback gradient queries, and a similar number of evaluations of the exponential map, where*

$$T = (f(x_0) - f_{\text{low}})\frac{\hat{\rho}^{1/4}\ell^{1/2}}{\epsilon^{7/4}}\log\left(\frac{d^{1/2}\ell^{3/2}\Delta_f}{(\hat{\rho}\epsilon)^{1/4}\epsilon^2\delta}\right)^6 + \frac{\ell^{1/2}}{\hat{\rho}^{1/4}\epsilon^{1/4}}\log\left(\frac{d^{1/2}\ell^{3/2}\Delta_f}{(\hat{\rho}\epsilon)^{1/4}\epsilon^2\delta}\right)$$
$$= O\left((f(x_0) - f_{\text{low}})(\rho + L\sqrt{K})^{1/4}L^{1/2}\cdot\frac{1}{\epsilon^{7/4}}\log\left(\frac{d}{\epsilon\delta}\right)^6\right),$$

with $\ell = 2L$, $\hat{\rho} = \rho + L\sqrt{K}$, $d = \dim \mathcal{M}$ and any $\Delta_f \geq \max(f(x_0) - f_{\text{low}}, \sqrt{\epsilon^3/\hat{\rho}})$. With probability at least $1-2\delta$, that point is also $(\epsilon, \sqrt{\hat{\rho}\epsilon})$-SOCP. The algorithm uses no Hessian queries and is randomized.

This result is *almost* dimension-free, and still not curvature-free for the same reasons as above.

1.3 Related Literature

At the same time and independently, Lezcano-Casado [35] develops geometric bounds similar to our own. Both papers derive the same second-order inhomogenous linear ODE (ordinary differential equation) describing the behavior of the second derivative of the exponential map. Lezcano-Casado [35] then uses ODE comparison techniques to derive the geometric bounds, while the present work uses a bootstrapping technique. Lezcano-Casado [35] applies these bounds to study gradient descent in tangent spaces, whereas we study non-convex accelerated algorithms for finding first- and second-order critical points.

The technique of pulling back a function to a tangent space is frequently used in other settings within optimization on manifolds. See for example the recent papers of Bergmann et al. [9] and Lezcano-Casado [34]. Additionally, the use of Riemannian Lipschitz conditions in optimization as they appear in Section 2 can be traced back to [21,Def. 4.1] and [23,Def. 2.2].

Accelerating optimization algorithms on Riemannian manifolds has been well-studied in the context of *geodesically convex* optimization problems. Such problems can be solved globally, and usually the objective is to bound the suboptimality gap rather than finding approximate critical points. A number of papers have studied Riemannian versions of AGD; however, none of these papers have been able to achieve a fully accelerated rate for convex optimization. Zhang and Sra [48] show that if the initial iterate is sufficiently close to the minimizer, then acceleration is possible. Intuitively this makes sense, since manifolds are locally Euclidean. Ahn and Sra [3] pushed this further, developing an algorithm converging strictly faster than RGD, and which also achieves acceleration when sufficiently close to the minimizer.

Alimisis et al. [4–6] analyze the problem of acceleration on the class of non-strongly convex functions, as well as under weaker notions of convexity. Interestingly, they also show that in the continuous limit (using an ODE to model optimization algorithms) acceleration is possible. However, it is unclear whether the discretization of this ODE preserves a similar acceleration.

Recently, Hamilton and Moitra [26] have shown that full acceleration (in the geodesically convex case) is impossible in the hyperbolic plane, in the setting where function values and gradients are corrupted by a very small amount of noise. In contrast, in the analogous Euclidean setting, acceleration is possible even with noisy oracles [22].

2 Riemannian Tools and Regularity of Pullbacks

In this section, we build up to and state our main geometric result. As we do so, we provide a few reminders of Riemannian geometry. For more on this topic, we recommend the modern textbooks by Lee [31, 32]. For book-length, optimization-focused introductions see [1, 12]. Some proofs of this section appear in Appendices A and B.

We consider a manifold $\mathcal{M}$ with Riemannian metric $\langle \cdot, \cdot \rangle_x$ and associated norm $\| \cdot \|_x$ on the tangent spaces $\mathrm{T}_x\mathcal{M}$. (In other sections, we omit the subscript x.) The *tangent bundle*

$$\mathrm{T}\mathcal{M} = \{(x, s) : x \in \mathcal{M} \text{ and } s \in \mathrm{T}_x\mathcal{M}\}$$

is itself a smooth manifold. The Riemannian metric provides a notion of gradient.

Definition 2.1 The *Riemannian gradient* of a differentiable function $f : \mathcal{M} \to \mathbb{R}$ is the unique vector field $\mathrm{grad} f$ on $\mathcal{M}$ which satisfies:

$$\mathrm{D} f(x)[s] = \langle \mathrm{grad} f(x), s \rangle_x \qquad \text{for all } (x, s) \in \mathrm{T}\mathcal{M},$$

where $\mathrm{D} f(x)[s]$ is the directional derivative of f at x along s.

The Riemannian metric further induces a uniquely defined *Riemannian connection* ∇ (used to differentiate vector fields on $\mathcal{M}$) and an associated *covariant derivative* D_t (used to differentiate vector fields along curves on $\mathcal{M}$). (The symbol ∇ here is not to be confused with its use elsewhere to denote differentiation of scalar functions on Euclidean spaces.) Applying the connection to the gradient vector field, we obtain Hessians.

Definition 2.2 The *Riemannian Hessian* of a twice differentiable function $f : \mathcal{M} \to \mathbb{R}$ at x is the linear operator $\mathrm{Hess} f(x)$ to and from $\mathrm{T}_x\mathcal{M}$ defined by

$$\mathrm{Hess} f(x)[s] = \nabla_s \mathrm{grad} f = \mathrm{D}_t \mathrm{grad} f(c(t))|_{t=0},$$

where in the last equality c can be any smooth curve on $\mathcal{M}$ satisfying $c(0) = x$ and $c'(0) = s$. This operator is self-adjoint with respect to the metric $\langle \cdot, \cdot \rangle_x$.

We can also define the Riemmannian third derivative $\nabla^3 f$ (a tensor of order three), see [12,Ch. 10] for details. We write $\left\|\nabla^3 f(x)\right\| \leq \rho$ to mean $\left|\nabla^3 f(x)(u, v, w)\right| \leq \rho$ for all unit vectors $u, v, w \in \mathrm{T}_x\mathcal{M}$.

A *retraction* R is a smooth map from (a subset of) $\mathrm{T}\mathcal{M}$ to $\mathcal{M}$ with the following property: for all $(x, s) \in \mathrm{T}\mathcal{M}$, the smooth curve $c(t) = \mathrm{R}(x, ts) = \mathrm{R}_x(ts)$ on $\mathcal{M}$ passes through $c(0) = x$ with velocity $c'(0) = s$. Such maps are used frequently in Riemannian optimization in order to move on a manifold. For example, a key ingredient of Riemannian gradient descent is the curve $c(t) = \mathrm{R}_x(-t\,\mathrm{grad} f(x))$ which initially moves away from x along the negative gradient direction.

To a curve c, we naturally associate a velocity vector field c'. Using the covariant derivative D_t, we differentiate this vector field along c to define the *acceleration* $c'' = \mathrm{D}_t c'$ of c: this is also a vector field along c. In particular, the *geodesics* of $\mathcal{M}$ are the curves with zero acceleration.

The *exponential map* $\mathrm{Exp}\colon \mathcal{O} \to \mathcal{M}$—defined on an open subset $\mathcal{O}$ of the tangent bundle—is a special retraction whose curves are geodesics. Specifically, $\gamma(t) = \mathrm{Exp}(x, ts) = \mathrm{Exp}_x(ts)$ is the unique geodesic on $\mathcal{M}$ which passes through $\gamma(0) = x$ with velocity $\gamma'(0) = s$. If the domain of Exp is the whole tangent bundle, we say $\mathcal{M}$ is *complete*.

To compare tangent vectors in distinct tangent spaces, we use *parallel transports*. Explicitly, let c be a smooth curve connecting the points $c(0) = x$ and $c(1) = y$. We say a vector field Z along c is *parallel* if its covariant derivative $\mathrm{D}_t Z$ is zero. Conveniently, for any given $v \in \mathrm{T}_x\mathcal{M}$ there exists a unique parallel vector field along c whose value at $t = 0$ is v. Therefore, the value of that vector field at $t = 1$ is a well-defined vector in $\mathrm{T}_y\mathcal{M}$: we call it the parallel transport of v from x to y along c. We introduce the notation

$$P_t^c \colon \mathrm{T}_{c(0)}\mathcal{M} \to \mathrm{T}_{c(t)}\mathcal{M}$$

to denote parallel transport along a smooth curve c from $c(0)$ to $c(t)$. This is a linear isometry: $(P_t^c)^{-1} = (P_t^c)^*$, where the star denotes an adjoint with respect to the Riemannian metric. For the special case of parallel transport along the geodesic $\gamma(t) = \mathrm{Exp}_x(ts)$, we write

$$P_{ts} \colon \mathrm{T}_x\mathcal{M} \to \mathrm{T}_{\mathrm{Exp}_x(ts)}\mathcal{M} \tag{2}$$

with the meaning $P_{ts} = P_t^{\gamma}$.

Using these tools, we can define Lipschitz continuity of gradients and Hessians. Note that in the particular case where $\mathcal{M}$ is a Euclidean space we have $\mathrm{Exp}_x(s) = x + s$ and parallel transports are identities, so that this reduces to the usual definitions.

Definition 2.3 The gradient of $f\colon \mathcal{M} \to \mathbb{R}$ is *L-Lipschitz continuous* if

$$\|P_s^* \mathrm{grad} f(\mathrm{Exp}_x(s)) - \mathrm{grad} f(x)\|_x \leq L\|s\|_x \qquad \text{for all } (x, s) \in \mathcal{O}, \tag{3}$$

where P_s^* is the adjoint of P_s with respect to the Riemannian metric.

The Hessian of f is *ρ-Lipschitz continuous* if

$$\|P_s^* \circ \mathrm{Hess} f(\mathrm{Exp}_x(s)) \circ P_s - \mathrm{Hess} f(x)\|_x \leq \rho\|s\|_x \qquad \text{for all } (x, s) \in \mathcal{O}, \tag{4}$$

where $\|\cdot\|_x$ denotes both the Riemannian norm on $\mathrm{T}_x\mathcal{M}$ and the associated operator norm.

It is well known that these Lipschitz conditions are equivalent to convenient inequalities, often used to study the complexity of optimization algorithms. More details appear in [12,Ch. 10].

Proposition 2.4 *If a function $f : \mathcal{M} \to \mathbb{R}$ has L-Lipschitz continuous gradient, then*

$$\left| f(\mathrm{Exp}_x(s)) - f(x) - \langle \mathrm{grad} f(x), s \rangle_x \right| \leq \frac{L}{2}\|s\|_x^2 \qquad \textit{for all } (x, s) \in \mathcal{O}.$$

If in addition f is twice differentiable, then $\|\mathrm{Hess} f(x)\| \leq L$ for all $x \in \mathcal{M}$.

If f has ρ-Lipschitz continuous Hessian, then

$$\begin{aligned} &\left| f(\mathrm{Exp}_x(s)) - f(x) - \langle \mathrm{grad} f(x), s \rangle_x - \frac{1}{2}\langle s, \mathrm{Hess} f(x)[s] \rangle_x \right| \\ &\leq \frac{\rho}{6}\|s\|_x^3 \quad \textit{and} \\ &\left\| P_s^* \mathrm{grad} f(\mathrm{Exp}_x(s)) - \mathrm{grad} f(x) - \mathrm{Hess} f(x)[s] \right\|_x \\ &\leq \frac{\rho}{2}\|s\|_x^2 \quad \textit{for all } (x, s) \in \mathcal{O}. \end{aligned}$$

If in addition f is three times differentiable, then $\left\|\nabla^3 f(x)\right\| \leq \rho$ for all $x \in \mathcal{M}$.

The other way around, if f is three times continuously differentiable and the stated inequalities hold, then its gradient and Hessian are Lipschitz continuous with the stated constants.

For sufficiently simple algorithms, these inequalities may be all we need to track progress in a sharp way. As an example, the iterates of Riemannian gradient descent with constant step-size $1/L$ satisfy $x_{k+1} = \mathrm{Exp}_{x_k}(s_k)$ with $s_k = -\frac{1}{L}\mathrm{grad} f(x_k)$. It follows directly from the first inequality above that $f(x_k) - f(x_{k+1}) \geq \frac{1}{2L}\|\mathrm{grad} f(x_k)\|^2$. From there, it takes a brief argument to conclude that this method finds a point with gradient smaller than ϵ in at most $2L(f(x_0) - f_{\mathrm{low}})\frac{1}{\epsilon^2}$ steps. A similar (but longer) story applies to the analysis of Riemannian trust regions and adaptive cubic regularization [2, 13].

However, the inequalities in Proposition 2.4 fall short when finer properties of the algorithms are only visible at the scale of multiple combined iterations. This is notably the case for accelerated gradient methods. For such algorithms, individual iterations may not achieve spectacular cost decrease, but a long sequence of them may accumulate an advantage over time (using momentum). To capture this advantage in an analysis, it is not enough to apply inequalities above to individual iterations. As we turn to assessing a string of iterations jointly by relating the various gradients and step directions we encounter, the nonlinearity of $\mathcal{M}$ generates significant hurdles.

For these reasons, we study the pullbacks of the cost function, namely, the functions

$$\hat{f}_x = f \circ \mathrm{Exp}_x : \mathrm{T}_x\mathcal{M} \to \mathbb{R}. \tag{5}$$

Each pullback is defined on a linear space, hence we can in principle run any Euclidean optimization algorithm on $\hat{f}_x$ directly: our strategy is therefore to apply a momentum-based method on $\hat{f}_x$. To this end, we now work towards showing that if f has Lipschitz continuous gradient and Hessian then $\hat{f}_x$ also has certain Lipschitz-type properties.

The following formulas appear in [2,Lem. 5]: we are interested in the case $\mathrm{R} = \mathrm{Exp}$. (We use ∇ and ∇^2 to designate gradients and Hessians of functions on Euclidean spaces: not to be confused with the connection ∇.)

Lemma 2.5 *Given $f : \mathcal{M} \to \mathbb{R}$ twice continuously differentiable and (x, s) in the domain of a retraction R, the gradient and Hessian of the pullback $\hat{f}_x = f \circ \mathrm{R}_x$ at $s \in \mathrm{T}_x\mathcal{M}$ are given by*

$$\nabla \hat{f}_x(s) = T_s^* \mathrm{grad} f(\mathrm{R}_x(s)) \quad \textit{and} \quad \nabla^2 \hat{f}_x(s) = T_s^* \circ \mathrm{Hess} f(\mathrm{R}_x(s)) \circ T_s + W_s, \tag{6}$$

where T_s is the differential of R_x at s (a linear operator):

$$T_s = \mathrm{DR}_x(s) \colon \mathrm{T}_x\mathcal{M} \to \mathrm{T}_{\mathrm{R}_x(s)}\mathcal{M}, \tag{7}$$

and W_s is a self-adjoint linear operator on $\mathrm{T}_x\mathcal{M}$ defined through polarization by

$$\langle W_s[\dot{s}], \dot{s}\rangle_x = \left\langle \mathrm{grad} f(\mathrm{R}_x(s)), c''(0)\right\rangle_{\mathrm{R}_x(s)}, \tag{8}$$

with $c''(0) \in \mathrm{T}_{\mathrm{R}_x(s)}\mathcal{M}$ the (intrinsic) acceleration on $\mathcal{M}$ of $c(t) = \mathrm{R}_x(s + t\dot{s})$ at $t = 0$.

Remark 2.6 Throughout, $s, \dot{s}, \ddot{s}$ will simply denote tangent vectors.

We turn to curvature. The *Lie bracket* of two smooth vector fields X, Y on $\mathcal{M}$ is itself a smooth vector field, conveniently expressed in terms of the Riemannian connection as $[X, Y] = \nabla_X Y - \nabla_Y X$. Using this notion, the *Riemann curvature endomorphism R* of $\mathcal{M}$ is an operator which maps three smooth vector fields X, Y, Z of $\mathcal{M}$ to a fourth smooth vector field as:

$$R(X, Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z. \tag{9}$$

Whenever R is identically zero, we say $\mathcal{M}$ is *flat*: this is the case notably when $\mathcal{M}$ is a Euclidean space and when $\mathcal{M}$ has dimension one (e.g., a circle is flat, while a sphere is not).

Though it is not obvious from the definition, the value of the vector field $R(X, Y)Z$ at $x \in \mathcal{M}$ depends on X, Y, Z only through their value at x. Therefore, given $u, v, w \in \mathrm{T}_x\mathcal{M}$ we can make sense of the notation $R(u, v)w$ as denoting the vector in $\mathrm{T}_x\mathcal{M}$ corresponding to $R(X, Y)Z$ at x, where X, Y, Z are arbitrary smooth vector fields whose values at x are u, v, w, respectively. The map $(u, v, w) \mapsto R(u, v)w$ is linear in each input.

Two linearly independent tangent vectors u, v at x span a two-dimensional plane of $\mathrm{T}_x\mathcal{M}$. The *sectional curvature* of $\mathcal{M}$ along that plane is a real number $K(u, v)$ defined as

$$K(u, v) = \frac{\langle R(u, v)v, u\rangle_x}{\|u\|_x^2\|v\|_x^2 - \langle u, v\rangle_x^2}. \tag{10}$$

Of course, all sectional curvatures of flat manifolds are zero. Also, all sectional curvatures of a sphere of radius r are $1/r^2$ and all sectional curvatures of the hyperbolic space with parameter r are $-1/r^2$—see [32,Thm. 8.34].

Using the connection ∇, we differentiate the curvature endomorphism R as follows. Given any smooth vector field U, we let $\nabla_U R$ be an operator of the same type as R itself, in the sense that it maps three smooth vector fields X, Y, Z to a fourth one denoted $(\nabla_U R)(X, Y)Z$ through

$$\begin{aligned}(\nabla_U R)(X, Y)Z = \nabla_U(R(X, Y)Z) - R(\nabla_U X, Y)Z \\ - R(X, \nabla_U Y)Z - R(X, Y)\nabla_U Z.\end{aligned} \tag{11}$$

Observe that this formula captures a convenient chain rule on $\nabla_U(R(X, Y)Z)$. As for R, the value of $\nabla R(X, Y, Z, U) \triangleq (\nabla_U R)(X, Y)Z$ at x depends on X, Y, Z, U only through their values at x. Therefore, ∇R unambiguously maps $u, v, w, z \in \mathrm{T}_x\mathcal{M}$ to $\nabla R(u, v, w, z) \in \mathrm{T}_x\mathcal{M}$, linearly in all inputs. We say the operator norm of ∇R at x is bounded by F if

$$\|\nabla R(u, v, w, z)\|_x \leq F\|u\|_x\|v\|_x\|w\|_x\|z\|_x$$

for all $u, v, w, z \in \mathrm{T}_x\mathcal{M}$. We say ∇R has operator norm bounded by F if this holds for all x. If $F = 0$ (that is, $\nabla R \equiv 0$), we say R is *parallel* and $\mathcal{M}$ is called *locally symmetric*. This is notably the case for manifolds with constant sectional curvature—Euclidean spaces, spheres and hyperbolic spaces—and (Riemannian) products thereof [41,pp. 219–221].

We are ready to state the main result of this section. Note that $\mathcal{M}$ need not be complete.

Theorem 2.7 *Let $\mathcal{M}$ be a Riemannian manifold whose sectional curvatures are in the interval $[K_{\mathrm{low}}, K_{\mathrm{up}}]$, and let $K = \max(|K_{\mathrm{low}}|, |K_{\mathrm{up}}|)$. Also assume ∇R—the covariant derivative of the Riemann curvature endomorphism R—is bounded by F in operator norm.*

Let $f : \mathcal{M} \to \mathbb{R}$ be twice continuously differentiable and select $b > 0$ such that

$$b \leq \min\left(\frac{1}{4\sqrt{K}}, \frac{K}{4F}\right).$$

Pick any point $x \in \mathcal{M}$ such that Exp_x is defined on the closed ball $B_x(b)$ of radius b around the origin in $\mathrm{T}_x\mathcal{M}$. We have the following three conclusions:

1. *If f has L-Lipschitz continuous gradient and $\|\mathrm{grad} f(x)\|_x \leq Lb$, then $\hat{f}_x = f \circ \mathrm{Exp}_x$ has $2L$-Lipschitz continuous gradient in $B_x(b)$, that is, for all $u, v \in B_x(b)$ it holds that $\|\nabla \hat{f}_x(u) - \nabla \hat{f}_x(v)\|_x \leq 2L\|u - v\|_x$.*
2. *If moreover f has ρ-Lipschitz continuous Hessian, then $\|\nabla^2 \hat{f}_x(s) - \nabla^2 \hat{f}_x(0)\|_x \leq \hat{\rho}\|s\|_x$ for all $s \in B_x(b)$, with $\hat{\rho} = \rho + L\sqrt{K}$.*
3. *For all $s \in B_x(b)$, the singular values of $T_s = \mathrm{DExp}_x(s)$ lie in the interval* $[2/3, 4/3]$.

A few comments are in order:

1. For locally symmetric spaces ($F = 0$), we interpret K/F as infinite (regardless of K).
2. If $\mathcal{M}$ is compact, then it is complete and there necessarily exist finite K and F. See work by Greene [25] for a discussion on non-compact manifolds.
3. If $\mathcal{M}$ is a homogeneous Riemannian manifold (not necessarily compact), then there exist finite K and F, and these can be assessed by studying a single point on the manifold. This follows directly from the definition of homogeneous Riemannian manifold [32,p. 55].
4. All symmetric spaces are homogeneous and locally symmetric [32,Exercise 6–19, Exercise 7–3 and p. 78] so there exists finite K and $F = 0$. Let $\mathrm{Sym}(d)$ be the set of real $d \times d$ symmetric matrices. The set of $d \times d$ positive definite matrices
$$\mathcal{P}_d = \{P \in \mathrm{Sym}(d) : P \succ 0\}$$
endowed with the so-called affine invariant metric
$$\langle X, Y \rangle_P = \mathrm{Tr}(P^{-1}XP^{-1}Y) \quad \text{for } P \in \mathcal{P}_d \text{ and } X, Y \in \mathrm{T}_P\mathcal{P}_d \cong \mathrm{Sym}(d)$$
is a non-compact symmetric space of non-constant curvature. It is commonly used in practice [10, 37, 38, 43]. One can show that $K = \frac{1}{2}$ and $F = 0$ are the right constants for this manifold.
5. The following statements are equivalent: (a) $\mathcal{M}$ is complete; (b) Exp is defined on the whole tangent bundle: $\mathcal{O} = \mathrm{T}\mathcal{M}$; and (c) for some $b > 0$, Exp_x is defined on $B_x(b)$ for all $x \in \mathcal{M}$. In later sections, we need to apply Theorem 2.7 at various points of $\mathcal{M}$ with constant b, which is why we then assume $\mathcal{M}$ is complete.
6. The properties of T_s are useful in combination with Lemma 2.5 to relate gradients and Hessians of the pullbacks to gradients and Hessians on the manifold. For example, if $\nabla \hat{f}_x(s)$ has norm ϵ, then $\mathrm{grad} f(\mathrm{Exp}_x(s))$ has norm somewhere between $\frac{3}{4}\epsilon$ and $\frac{3}{2}\epsilon$. Under the conditions of the theorem, W_s (8) is bounded as $\|W_s\|_x \leq \frac{9}{4}K\|\nabla \hat{f}_x(s)\|_x\|s\|_x$.
7. We only get satisfactory Lipschitzness at points where the gradient is bounded by Lb. Fortunately, for the algorithms we study, whenever we encounter a point with gradient larger than that threshold, it is sufficient to take a simple gradient descent step.

To prove Theorem 2.7, we must control $\nabla^2 \hat{f}_x(s)$. According to Lemma 2.5, this requires controlling both T_s (a differential of the exponential map) and $c''(0)$ (the intrinsic initial acceleration of a curve defined via the exponential map, but which is not itself a geodesic in general). On both counts, we must study differentials of exponentials. *Jacobi fields* are the tool of choice for such tasks. As a first step, we use Jacobi fields to investigate the difference between T_s and P_s: two linear operators from $\mathrm{T}_x\mathcal{M}$ to $\mathrm{T}_{\mathrm{Exp}_x(s)}\mathcal{M}$. We prove a general result in Appendix A (exact for constant sectional curvature) and state a sufficient particular case here. Control of T_s follows as a corollary because P_s (parallel transport) is an isometry.

Proposition 2.8 *Let $\mathcal{M}$ be a Riemannian manifold whose sectional curvatures are in the interval $[K_{\mathrm{low}}, K_{\mathrm{up}}]$, and let $K = \max(|K_{\mathrm{low}}|, |K_{\mathrm{up}}|)$. For any $(x, s) \in \mathcal{O}$ with $\|s\|_x \leq \frac{\pi}{\sqrt{K}}$,*

$$\|(T_s - P_s)[\dot{s}]\|_{\mathrm{Exp}_x(s)} \leq \frac{1}{3} K \|s\|_x^2 \|\dot{s}_\perp\|_x, \tag{12}$$

where $\dot{s}_\perp = \dot{s} - \frac{\langle s, \dot{s}\rangle_x}{\langle s, s\rangle_x} s$ is the component of $\dot{s}$ orthogonal to s.

Corollary 2.9 *Let $\mathcal{M}$ be a Riemannian manifold whose sectional curvatures are in the interval $[K_{\mathrm{low}}, K_{\mathrm{up}}]$, and let $K = \max(|K_{\mathrm{low}}|, |K_{\mathrm{up}}|)$. For any $(x, s) \in \mathcal{O}$ with $\|s\|_x \leq \frac{1}{\sqrt{K}}$,*

$$\sigma_{\min}(T_s) \geq \frac{2}{3} \qquad \text{and} \qquad \sigma_{\max}(T_s) \leq \frac{4}{3}. \tag{13}$$

Proof. By Proposition 2.8, the operator norm of $T_s - P_s$ is bounded above by $\frac{1}{3} K \|s\|_x^2 \leq \frac{1}{3}$. Furthermore, parallel transport P_s is an isometry: its singular values are equal to 1. Thus,

$$\sigma_{\max}(T_s) = \sigma_{\max}(P_s + T_s - P_s) \leq \sigma_{\max}(P_s) + \sigma_{\max}(T_s - P_s) \leq 1 + \frac{1}{3} = \frac{4}{3}.$$

Likewise, with min/max taken over unit-norm vectors $u \in \mathrm{T}_x\mathcal{M}$ and writing $y = \mathrm{Exp}_x(s)$,

$$\begin{aligned} \sigma_{\min}(T_s) &= \min_u \|T_s u\|_y \geq \min_u \|P_s u\|_y - \|(T_s - P_s)u\|_y \\ &= 1 - \max_u \|(T_s - P_s)u\|_y \geq \frac{2}{3}. \end{aligned}$$

□

We turn to controlling the term $c''(0)$ which appears in the definition of operator W_s in the expression for $\nabla^2 \hat{f}_x(s)$ provided by Lemma 2.5. We present a detailed proof in Appendix B for a general statement, and state a sufficient particular case here. The proof is fairly technical: it involves designing an appropriate nonlinear second-order ODE on the manifold and bounding its solutions. The ODE is related to the Jacobi equation, except we had to differentiate to the next order, and the equation is not homogeneous.

Proposition 2.10 *Let $\mathcal{M}$ be a Riemannian manifold whose sectional curvatures are in the interval $[K_{\mathrm{low}}, K_{\mathrm{up}}]$, and let $K = \max(|K_{\mathrm{low}}|, |K_{\mathrm{up}}|)$. Further assume ∇R is bounded by F in operator norm.*

Pick any $(x, s) \in \mathcal{O}$ *such that*

$$\|s\|_x \le \min\left(\frac{1}{4\sqrt{K}}, \frac{K}{4F}\right).$$

For any $\dot{s} \in \mathrm{T}_x\mathcal{M}$*, the curve* $c(t) = \mathrm{Exp}_x(s + t\dot{s})$ *has initial acceleration bounded as*

$$\|c''(0)\|_{\mathrm{Exp}_x(s)} \le \frac{3}{2} K \|s\|_x \|\dot{s}\|_x \|\dot{s}_\perp\|_x,$$

where $\dot{s}_\perp = \dot{s} - \frac{\langle s, \dot{s}\rangle_x}{\langle s, s\rangle_x} s$ *is the component of* $\dot{s}$ *orthogonal to* s.

Equipped with all of the above, it is now easy to prove the main theorem of this section.

Proof of Theorem 2.7. Consider the pullback $\hat{f}_x = f \circ \mathrm{Exp}_x$ defined on $\mathrm{T}_x\mathcal{M}$. Since $\mathrm{T}_x\mathcal{M}$ is linear, it is a classical exercise to verify that $\nabla \hat{f}_x$ is $2L$-Lipschitz continuous in $B_x(b)$ if and only if $\|\nabla^2 \hat{f}_x(s)\|_x \le 2L$ for all s in $B_x(b)$. Using Lemma 2.5, we start bounding the Hessian as follows:

$$\|\nabla^2 \hat{f}_x(s)\|_x \le \sigma_{\max}(T_s^*)\sigma_{\max}(T_s)\|\mathrm{Hess} f(\mathrm{Exp}_x(s))\|_{\mathrm{Exp}_x(s)} + \|W_s\|_x,$$

with operator W_s defined by (8). Since $\mathrm{grad} f$ is L-Lipschitz continuous, $\|\mathrm{Hess} f(y)\|_y \le L$ for all $y \in \mathcal{M}$ (this follows fairly directly from Proposition 2.4). To bound W_s, we start with a Cauchy–Schwarz inequality then we consider the worst case for the magnitude of $c''(0)$:

$$\|W_s\|_x \le \|\mathrm{grad} f(\mathrm{Exp}_x(s))\|_{\mathrm{Exp}_x(s)} \cdot \max_{\dot{s} \in \mathrm{T}_x\mathcal{M}, \|\dot{s}\|_x = 1} \|c''(0)\|_{\mathrm{Exp}_x(s)}.$$

Combining these steps yields a first bound of the form

$$\begin{aligned} \|\nabla^2 \hat{f}_x(s)\|_x \le\ & \sigma_{\max}(T_s)^2 L \\ & + \|\mathrm{grad} f(\mathrm{Exp}_x(s))\|_{\mathrm{Exp}_x(s)} \cdot \max_{\dot{s} \in \mathrm{T}_x\mathcal{M}, \|\dot{s}\|_x = 1} \|c''(0)\|_{\mathrm{Exp}_x(s)}. \end{aligned} \tag{14}$$

To proceed, we keep working on the W_s-terms: use Proposition 2.10, L-Lipschitz-continuity of the gradient, and our bounds on the norms of s and $\mathrm{grad} f(x)$ to see that:

$$\begin{aligned} \|W_s\|_x &\le \max_{\dot{s} \in \mathrm{T}_x\mathcal{M}, \|\dot{s}\|_x = 1} \|c''(0)\|_{\mathrm{Exp}_x(s)} \cdot \|\mathrm{grad} f(\mathrm{Exp}_x(s))\|_{\mathrm{Exp}_x(s)} \\ &\le \frac{3}{2} K \|s\|_x \cdot \|P_s^* \mathrm{grad} f(\mathrm{Exp}_x(s)) - \mathrm{grad} f(x) + \mathrm{grad} f(x)\|_x \\ &\le \frac{3}{2} K \|s\|_x \cdot (L\|s\|_x + \|\mathrm{grad} f(x)\|_x) \\ &\le 3KLb\|s\|_x \le \frac{3}{4} L\sqrt{K}\|s\|_x \le \frac{3}{16} L. \end{aligned} \tag{15}$$

Returning to (14) and using Corollary 2.9 to bound T_s confirms that

$$\|\nabla^2 \hat{f}_x(s)\|_x \leq \frac{16}{9}L + \frac{3}{16}L < 2L.$$

Thus, $\nabla \hat{f}_x$ is $2L$-Lipschitz continuous in the ball of radius b around the origin in $\mathrm{T}_x\mathcal{M}$.

To establish the second part of the claim, we use the same intermediate results and ρ-Lipschitz continuity of the Hessian. First, using Lemma 2.5 twice and noting that $W_0 = 0$ so that $\nabla^2 \hat{f}_x(0) = \operatorname{Hess} f(x)$, we have:

$$\begin{aligned}\nabla^2 \hat{f}_x(s) - \nabla^2 \hat{f}_x(0) = {} & P_s^* \circ \operatorname{Hess} f(\operatorname{Exp}_x(s)) \circ P_s - \operatorname{Hess} f(x) \\ & + (T_s - P_s)^* \circ \operatorname{Hess} f(\operatorname{Exp}_x(s)) \circ T_s \\ & + P_s^* \circ \operatorname{Hess} f(\operatorname{Exp}_x(s)) \circ (T_s - P_s) \\ & + W_s.\end{aligned}$$

We bound this line by line calling upon Proposition 2.8, Corollary 2.9 and (15) to get:

$$\begin{aligned}\|\nabla^2 \hat{f}_x(s) - \nabla^2 \hat{f}_x(0)\|_x &\leq \rho\|s\|_x + \frac{4}{9}LK\|s\|_x^2 + \frac{1}{3}LK\|s\|_x^2 + 3LKb\|s\|_x \\ &\leq \left(\rho + \frac{1}{9}L\sqrt{K} + \frac{1}{12}L\sqrt{K} + \frac{3}{4}L\sqrt{K}\right)\|s\|_x \\ &\leq \left(\rho + L\sqrt{K}\right)\|s\|_x.\end{aligned}$$

This shows a type of Lipschitz continuity of the Hessian of the pullback with respect to the origin, in the ball of radius b. □

3 Assumptions and parameters for TAGD and PTAGD

Our algorithms apply to the minimization of $f\colon \mathcal{M} \to \mathbb{R}$ on a Riemannian manifold $\mathcal{M}$ equipped with a retraction R defined on the whole tangent bundle $\mathrm{T}\mathcal{M}$. The pullback of f at $x \in \mathcal{M}$ is $\hat{f}_x = f \circ \mathrm{R}_x\colon \mathrm{T}_x\mathcal{M} \to \mathbb{R}$. In light of Sect. 2, we make the following assumptions.

A 2 There exists a constant f_{low} such that $f(x) \geq f_{\text{low}}$ for all $x \in \mathcal{M}$. Moreover, f is twice continuously differentiable and there exist constants ℓ, $\hat{\rho}$ and b such that, for all $x \in \mathcal{M}$ with $\|\operatorname{grad} f(x)\| \leq \frac{1}{2}\ell b$,

1. $\nabla \hat{f}_x$ is ℓ-Lipschitz continuous in $B_x(3b)$ (in particular, $\|\nabla^2 \hat{f}_x(0)\| \leq \ell$),
2. $\|\nabla^2 \hat{f}_x(s) - \nabla^2 \hat{f}_x(0)\| \leq \hat{\rho}\|s\|$ for all $s \in B_x(3b)$, and
3. $\sigma_{\min}(T_s) \geq \frac{1}{2}$ with $T_s = \mathrm{D}\mathrm{R}_x(s)$ for all $s \in B_x(3b)$,

where $B_x(3b) = \{u \in \mathrm{T}_x\mathcal{M} : \|u\| \leq 3b\}$. Finally, for all $(x, s) \in \mathrm{T}\mathcal{M}$ it holds that

4. $\hat{f}_x(s) \leq \hat{f}_x(0) + \langle \nabla \hat{f}_x(0), s \rangle + \frac{\ell}{2}\|s\|^2$.

The first three items in A2 confer Lipschitz properties to the derivatives of the pullbacks $\hat{f}_x$ restricted to balls around the origins of tangent spaces: these are the balls where we shall run accelerated gradient steps. We only need these guarantees at points where the gradient is below a threshold. For all other points, a regular gradient step provides ample progress: the last item in A2 serves that purpose only, see Proposition 5.2.

Section 2 tells us that A2 holds in particular when we use the exponential map as a retraction and f itself has appropriate (Riemannian) Lipschitz properties. This is the link between Theorems 1.3 and 1.6 in the introduction and Theorems 5.1 and 6.1 in later sections.

Corollary 3.1 *If we use the exponential retraction* $\mathrm{R} = \mathrm{Exp}$ *and A1 holds, then A2 holds with* f_{low}, *and*

$$\ell = 2L, \qquad \hat{\rho} = \rho + L\sqrt{K}, \qquad b = \frac{1}{12}\min\left(\frac{1}{\sqrt{K}}, \frac{K}{F}\right). \tag{16}$$

With constants as in A2, we further define a number of parameters. First, the user specifies a tolerance ϵ which must not be too loose.

A 3 The tolerance $\epsilon > 0$ satisfies $\sqrt{\hat{\rho}\epsilon} \leq \frac{1}{2}\ell$ and $\epsilon \leq b^2\hat{\rho}$.

Then, we fix a first set of parameters (see [28] for more context; in particular, κ plays the role of a condition number; under A3, we have $\kappa \geq 2$):

$$\eta = \frac{1}{4\ell}, \qquad \kappa = \frac{\ell}{\sqrt{\hat{\rho}\epsilon}}, \qquad \theta = \frac{1}{4\sqrt{\kappa}}, \qquad \gamma = \frac{\sqrt{\hat{\rho}\epsilon}}{4}, \qquad s = \frac{1}{32}\sqrt{\frac{\epsilon}{\hat{\rho}}}. \tag{17}$$

We define a second set of parameters based on some $\chi \geq 1$ (as set in some of the lemmas and theorems below) and a universal constant $c > 0$ (implicitly defined as the smallest real satisfying a finite number of lower-bounds required throughout the paper):

$$r = \eta\epsilon\chi^{-5}c^{-8}, \qquad \mathscr{T} = \sqrt{\kappa}\chi c, \qquad \mathscr{E} = \sqrt{\frac{\epsilon^3}{\hat{\rho}}}\chi^{-5}c^{-7},$$
$$\mathscr{L} = \sqrt{\frac{4\epsilon}{\hat{\rho}}}\chi^{-2}c^{-3}, \qquad \mathscr{M} = \frac{\epsilon\sqrt{\kappa}}{\ell}c^{-1}. \tag{18}$$

When we say "with $\chi \geq A \geq 1$" (for example, in Theorems 5.1 and 6.1), we mean: "with χ the smallest value larger than A such that $\mathscr{T}$ is a positive integer multiple of 4."

Lemma C.1 in Appendix C lists useful relations between the parameters.

Algorithm 1 $\texttt{TSS}(x, s_0)$ with $(x, s_0) \in \mathrm{T}\mathcal{M}$ and parameters $\epsilon, \eta, b, \theta, \gamma, s, \mathscr{T}$

1: If s_0 is not provided, set $s_0 = 0$ and $\texttt{perturbed} =$ **false**; otherwise, set $\texttt{perturbed} =$ **true**.
2: $v_0 = 0$
3: **for** j in $0, 1, \ldots, \mathscr{T} - 1$ **do**
4: $\quad u_j = s_j + (1 - \theta_j) v_j$ with ▷ AGD: capped momentum step

$$\theta_j = \begin{cases} \theta \text{ if } \|s_j + (1-\theta) v_j\| \le 2b, \\ \hat{\theta} \in [\theta, 1] \text{ such that } \|s_j + (1 - \hat{\theta}) v_j\| = 2b \text{ otherwise.} \end{cases} \tag{19}$$

5: $\quad$ **if** (NCC) triggers with (x, s_j, u_j) **then** ▷ Negative curvature detection
6: $\quad\quad$ **return** $\mathrm{R}_x(\texttt{NCE}(x, s_j, v_j))$ ▷ (Cases 2a, 3a)
7: $\quad$ **end if**
8: $\quad s_{j+1} = u_j - \eta \nabla \hat{f}_x(u_j)$ ▷ AGD: gradient step
9: $\quad v_{j+1} = s_{j+1} - s_j$ ▷ AGD: momentum update
10: $\quad$ **if** $\left(\|s_{j+1}\| > b\right)$ **or** $\left((\textbf{not } \texttt{perturbed}) \textbf{ and } \|\nabla \hat{f}_x(s_{j+1})\| \le \epsilon/2\right)$ **then**
11: $\quad\quad$ **return** $\mathrm{R}_x(s_{j+1})$ ▷ (Cases 2b, 2c, 3b)
12: $\quad$ **end if**
13: **end for**
14: **return** $\mathrm{R}_x(s_{\mathscr{T}})$ ▷ (Cases 2d, 3d)

4 Accelerated Gradient Descent in a Ball of a Tangent Space

The main ingredient of algorithms TAGD and PTAGD is TSS: the *tangent space steps* algorithm. Essentially, the latter runs the classical accelerated gradient descent algorithm (AGD) from convex optimization on $\hat{f}_x$ in a tangent space $\mathrm{T}_x\mathcal{M}$, with a few tweaks:

1. Because $\hat{f}_x$ need not be convex, TSS monitors the generated sequences for signs of non-convexity. If $\hat{f}_x$ happens to behave like a convex function along the sequence TSS generates, then we reap the benefits of convexity. Otherwise, the direction along which $\hat{f}_x$ behaves in a non-convex way can be used as a good descent direction. This is the idea behind the "convex until proven guilty" paradigm developed by Carmon et al. [15] and also exploited by Jin et al. [28]. Explicitly, given $x \in \mathcal{M}$ and $s, u \in \mathrm{T}_x\mathcal{M}$, for a specified parameter $\gamma > 0$, we check the *negative curvature condition* (one might also call it the *non-convexity condition)* (NCC):

$$\hat{f}_x(s) < \hat{f}_x(u) + \langle \nabla \hat{f}_x(u), s - u \rangle - \frac{\gamma}{2} \|s - u\|^2 . \tag{NCC}$$

   If (NCC) triggers with a triplet (x, s, u) and s is not too large, we can exploit that fact to generate substantial cost decrease using the *negative curvature exploitation* algorithm, NCE: see Lemma 4.4. (This is about curvature of the cost function, not the manifold.)
2. In contrast to the Euclidean case in [28], our assumption A2 provides Lipschitz-type guarantees only in a ball of radius $3b$ around the origin in $\mathrm{T}_x\mathcal{M}$. Therefore, we must act if iterates generated by TSS leave that ball. This is done in two places. First, the momentum step in step 4 of TSS is capped so that $\|u_j\|$ remains in the ball of radius $2b$ around the origin. Second, if s_{j+1} leaves the ball of radius b (as checked in step 10) then we terminate this run of TSS by returning to the manifold.

Algorithm 2 $\texttt{NCE}(x, s_j, v_j)$ with $x \in \mathcal{M}$, $s_j, v_j \in \mathrm{T}_x\mathcal{M}$ and parameter s

if $\|v_j\| \geq s$ **then**
  return s_j
else
  $\dot{v} = s\frac{v_j}{\|v_j\|}$
  return $\mathrm{argmin}_{\dot{s} \in \{s_j, s_j+\dot{v}, s_j-\dot{v}\}} \hat{f}_x(\dot{s})$
end if

Lemma 4.1 guarantees that the iterates indeed remain in appropriate balls, that θ_j (19) in the capped momentum step is uniquely defined, and that if a momentum step is capped, then immediately after that TSS terminates.

The initial momentum v_0 is always set to zero. By default, the AGD sequence is initialized at the origin: $s_0 = 0$. However, for PTAGD we sometimes want to initialize at a different point (a perturbation away from the origin): this is only relevant for Sect. 6.

In the remainder of this section, we provide four general purpose lemmas about TSS. Proofs are in Appendix D. We note that TAGD and PTAGD call TSS only at points x where $\|\mathrm{grad} f(x)\| \leq \frac{1}{2}\ell b$. The first lemma below notably guarantees that, for such runs, all iterates u_j, s_j generated by TSS remain (a fortiori) in balls of radius $3b$, so that the strongest provisions of A2 always apply: we use this fact often without mention.

Lemma 4.1 (TSS stays in balls) *Fix parameters and assumptions as laid out in Sect. 3. Let $x \in \mathcal{M}$ satisfy $\|\mathrm{grad} f(x)\| \leq \frac{1}{2}\ell b$. If* $\texttt{TSS}(x)$ *or* $\texttt{TSS}(x, s_0)$ *(with $\|s_0\| \leq b$) defines vectors $u_0, \ldots, u_q$ (and possibly more), then it also defines vectors $s_0, \ldots, s_q$, and we have:*

$$\|s_0\|, \ldots, \|s_q\| \leq b, \qquad \|u_0\|, \ldots, \|u_q\| \leq 2b, \qquad \text{and} \qquad 2\eta\gamma \leq \theta \leq \theta_j \leq 1.$$

If s_{q+1} is defined, then $\|s_{q+1}\| \leq 3b$ and, if $\|u_q\| = 2b$, then $\|s_{q+1}\| > b$ and u_{q+1} is undefined.

Along the iterates of AGD, the value of the cost function $\hat{f}_x$ may not monotonically decrease. Fortunately, there is a useful quantity which monotonically decreases along iterates: Jin et al. [28] call it the *Hamiltonian*. In several ways, it serves the purpose of a Lyapunov function. Importantly, the Hamiltonian decreases regardless of any special events that occur while running TSS. It is built as a combination of the cost function value and the momentum. The next lemma makes this precise: we use monotonic decrease of the Hamiltonian often without mention. This corresponds to [28,Lem. 9 and 20].

Lemma 4.2 (Hamiltonian decrease) *Fix parameters and assumptions as laid out in Sect. 3. Let $x \in \mathcal{M}$ satisfy $\|\mathrm{grad} f(x)\| \leq \frac{1}{2}\ell b$. For each pair (s_j, v_j) defined by* $\texttt{TSS}(x)$ *or* $\texttt{TSS}(x, s_0)$ *(with $\|s_0\| \leq b$), define the Hamiltonian*

$$E_j = \hat{f}_x(s_j) + \frac{1}{2\eta}\|v_j\|^2. \tag{20}$$

If E_{j+1} is defined, then E_j, θ_j and u_j are also defined and:

$$E_{j+1} \le E_j - \frac{\theta_j}{2\eta}\|v_j\|^2 - \frac{\eta}{4}\|\nabla \hat{f}_x(u_j)\|^2 \le E_j.$$

If moreover $\|v_j\| \ge \mathscr{M}$, then $E_j - E_{j+1} \ge \frac{4\mathscr{E}}{\mathscr{T}}$.

Jin et al. [28] formalize an important property of TSS sequences in the Euclidean case, namely, the fact that "either the algorithm makes significant progress or the iterates do not move much." They call this the *improve or localize* phenomenon. The next lemma states this precisely in our context. This corresponds to [28,Cor. 11].

Lemma 4.3 (Improve or localize) *Fix parameters and assumptions as laid out in Sect. 3. Let $x \in \mathcal{M}$ satisfy $\|\mathrm{grad} f(x)\| \le \frac{1}{2}\ell b$. If $\mathtt{TSS}(x)$ or $\mathtt{TSS}(x, s_0)$ (with $\|s_0\| \le b$) defines vectors $s_0, \ldots, s_q$ (and possibly more), then $E_0, \ldots, E_q$ are defined by (20) and, for all $0 \le q' \le q$,*

$$\|s_q - s_{q'}\|^2 \le (q - q') \sum_{j=q'}^{q-1} \|s_{j+1} - s_j\|^2 \le 16\sqrt{\kappa}\eta(q - q')(E_{q'} - E_q).$$

For $q' = 0$ in particular, using $E_0 = \hat{f}_x(s_0)$ we can write $E_q \le \hat{f}_x(s_0) - \frac{\|s_q - s_0\|^2}{16\sqrt{\kappa}\eta q}$.

As outlined earlier, in case the TSS sequence witnesses non-convexity in $\hat{f}_x$ through the (NCC) check, we call upon the NCE algorithm to exploit this event. The final lemma of this section formalizes the fact that this yields appropriate cost improvement. (Indeed, if $\|s_j\| > \mathscr{L}$ one can argue that sufficient progress was already achieved; otherwise, the lemma applies and we get a result from $E_j \le E_0 = \hat{f}_x(s_0)$.) This corresponds to [28,Lem. 10 and 17].

Lemma 4.4 (Negative curvature exploitation) *Fix parameters and assumptions as laid out in Sect. 3. Let $x \in \mathcal{M}$ satisfy $\|\mathrm{grad} f(x)\| \le \frac{1}{2}\ell b$. Assume $\mathtt{TSS}(x)$ or $\mathtt{TSS}(x, s_0)$ (with $\|s_0\| \le b$) defines u_j, so that s_j, v_j are also defined, and E_j is defined by (20). If (NCC) triggers with (x, s_j, u_j) and $\|s_j\| \le \mathscr{L}$, then $\hat{f}_x(\mathtt{NCE}(x, s_j, v_j)) \le E_j - 2\mathscr{E}$.*

5 First-Order Critical Points

Our algorithm to compute ϵ-approximate first-order critical points on Riemannian manifolds is TAGD: this is a deterministic algorithm which does not require access to the Hessian of the cost function. Our main result regarding TAGD, namely, Theorem 5.1, states that it does so in a bounded number of iterations. As worked out in Theorem 1.3, this bound scales as $\epsilon^{-7/4}$, up to polylogarithmic terms. The complexity is independent of the dimension of the manifold.

The proof of Theorem 5.1 rests on two propositions introduced hereafter in this section. Interestingly, it is only in the proof of Theorem 5.1 that we track the behavior

Algorithm 3 $\texttt{TAGD}(x_0)$ with $x_0 \in \mathcal{M}$ and parameters $\epsilon, \ell, \eta, b, \theta, \gamma, s, \mathscr{T}, \mathscr{M}$

$t \leftarrow 0$
while true do
  if $\|\mathrm{grad} f(x_t)\| > 2\ell\mathscr{M}$ **then**
    $x_{t+1} = \mathrm{R}_{x_t}(-\eta \mathrm{grad} f(x_t))$ ▷ Case 1: one Riemannian gradient step
    $t \leftarrow t + 1$
  else if $\|\mathrm{grad} f(x_t)\| > \epsilon$ **then**
    $x_{t+\mathscr{T}} = \texttt{TSS}(x_t)$ ▷ Case 2: accelerated gradient in $\mathrm{T}_{x_t}\mathcal{M}$
    $t \leftarrow t + \mathscr{T}$
  else
    return x_t ▷ Approximate FOCP
  end if
end while

of iterates of TAGD across multiple points on the manifold. This is done by tracking decrease of the value of the cost function f. All supporting results (lemmas and propositions) handle a single tangent space at a time. As a result, lemmas and propositions fully benefit from the linear structure of tangent spaces. This is why we can salvage most of the Euclidean proofs of Jin et al. [28], up to mostly minor (but numerous and necessary) changes.

Theorem 5.1 *Fix parameters and assumptions as laid out in Sect. 3, with*

$$\chi \geq \log_2(\theta^{-1}) \geq 1. \tag{21}$$

Given $x_0 \in \mathcal{M}$, $\texttt{TAGD}(x_0)$ *returns* $x_t \in \mathcal{M}$ *satisfying* $f(x_t) \leq f(x_0)$ *and* $\|\mathrm{grad} f(x_t)\| \leq \epsilon$ *with*

$$t \leq T_1 \triangleq \frac{f(x_0) - f_{\mathrm{low}}}{\mathscr{E}} \mathscr{T}. \tag{22}$$

Running the algorithm requires at most $2T_1$ *pullback gradient queries and* $3T_1$ *function queries (but no Hessian queries), and a similar number of calls to the retraction.*

Proof of Theorem 5.1 The call to $\texttt{TAGD}(x_0)$ generates a sequence of points $x_{t_0}, x_{t_1}, x_{t_2}, \ldots$ on $\mathcal{M}$, with $t_0 = 0$. A priori, this sequence may be finite or infinite. Considering two consecutive indices t_i and t_{i+1}, we either have $t_{i+1} = t_i + 1$ (if the step from x_{t_i} to $x_{t_{i+1}}$ is a single gradient step (Case 1)) or $t_{i+1} = t_i + \mathscr{T}$ (if that same step is obtained through a call to TSS (Case 2)). Moreover:

- In Case 1, Proposition 5.2 applies and guarantees

$$f(x_{t_i}) - f(x_{t_{i+1}}) \geq \frac{\mathscr{E}}{\mathscr{T}} = \frac{\mathscr{E}}{\mathscr{T}}(t_{i+1} - t_i).$$

- In Case 2, Proposition 5.3 applies and guarantees that if $\|\mathrm{grad} f(x_{t_{i+1}})\| > \epsilon$ then

$$f(x_{t_i}) - f(x_{t_{i+1}}) \geq \mathscr{E} = \frac{\mathscr{E}}{\mathscr{T}}(t_{i+1} - t_i).$$

It is now clear that TAGD(x_0) terminates after a finite number of steps. Indeed, if it does not, then the above reasoning shows that the algorithm produces an amortized decrease in the cost function f of $\frac{\mathscr{E}}{\mathscr{T}}$ per unit increment of the counter t, yet the value of f cannot decrease by more than $f(x_0) - f_{\mathrm{low}}$ because f is globally lower-bounded by f_{low}.

Accordingly, assume TAGD(x_0) generates $x_{t_0}, \ldots, x_{t_k}$ and terminates there, returning x_{t_k}. We know that $f(x_{t_k}) \leq f(x_0)$ and $\|\mathrm{grad} f(x_{t_k})\| \leq \epsilon$. Moreover, from the discussion above and $t_0 = 0$, we know that

$$f(x_0) - f_{\mathrm{low}} \geq f(x_0) - f(x_{t_k}) = \sum_{i=0}^{k-1} f(x_{t_i}) - f(x_{t_{i+1}})$$
$$\geq \frac{\mathscr{E}}{\mathscr{T}} \sum_{i=0}^{k-1} t_{i+1} - t_i = \frac{\mathscr{E}}{\mathscr{T}} t_k.$$

Thus, $t_k \leq \frac{f(x_0) - f_{\mathrm{low}}}{\mathscr{E}} \mathscr{T} \triangleq T_1$.

How much work does it take to run the algorithm? Each (regular) gradient step requires one gradient query and increases the counter by one. Each run of TSS requires at most $2\mathscr{T}$ gradient queries and $2\mathscr{T} + 3 \leq 3\mathscr{T}$ function queries ($3 \leq \mathscr{T}$ because $\mathscr{T}$ is a positive integer multiple of 4) and increases the counter by $\mathscr{T}$. Therefore, by the time TAGD produces x_t it has used at most $2t$ gradient queries and $3t$ function queries. □

The two following propositions form the backbone of the proof of Theorem 5.1. Each handles one of the two possible cases in one (outer) iteration of TAGD, namely: Case 1 is a "vanilla" Riemannian gradient descent step, while Case 2 is a call to TSS to run (modified) AGD in the current tangent space. The former has a trivial and standard proof. The latter relies on all lemmas from Sect. 4 and on two additional lemmas introduced in Appendix F, all following Jin et al. [28].

Proposition 5.2 (Case 1) *Fix parameters and assumptions as laid out in Sect. 3. Assume $x \in \mathcal{M}$ satisfies $\|\mathrm{grad} f(x)\| > 2\ell\mathscr{M}$. Then, $x_+ = \mathrm{R}_x(-\eta \mathrm{grad} f(x))$ satisfies $f(x) - f(x_+) \geq \frac{\mathscr{E}}{\mathscr{T}}$.*

Proof of Proposition 5.2 This follows directly by property 4 in A2 with $\hat{f}_x = f \circ \mathrm{R}_x$ since $\hat{f}_x(0) = f(x)$ and $\nabla \hat{f}_x(0) = \mathrm{grad} f(x)$ by properties of retractions, and also using $\ell\eta = 1/4$:

$$f(x_+) = \hat{f}_x(-\eta \mathrm{grad} f(x)) \leq \hat{f}_x(0) - \eta \|\mathrm{grad} f(x)\|^2$$
$$+ \frac{\ell}{2} \|\eta \mathrm{grad} f(x)\|^2 \leq f(x) - (7/8)\ell\mathscr{M}^2.$$

To conclude, it remains to use that $(7/8)\ell\mathscr{M}^2 \geq \frac{\mathscr{E}}{\mathscr{T}}$, as shown in Lemma C.1. □

The next proposition corresponds mostly to [28,Lem. 12]. A proof is in Appendix F.

Algorithm 4 PTAGD(x_0) with $x_0 \in \mathcal{M}$ and parameters $\epsilon, \ell, \eta, b, \theta, \gamma, s, r, \mathscr{T}, \mathscr{E}, \mathscr{M}$

$t \leftarrow 0$
while true do
  if $\|\mathrm{grad} f(x_t)\| > 2\ell\mathscr{M}$ **then**
    $x_{t+1} = \mathrm{R}_{x_t}(-\eta \mathrm{grad} f(x_t))$ ▷ Case 1: one Riemannian gradient step
    $t \leftarrow t+1$
  else if $\|\mathrm{grad} f(x_t)\| > \epsilon$ **then**
    $x_{t+\mathscr{T}} = \mathtt{TSS}(x_t)$ ▷ Case 2: accelerated gradient in $\mathrm{T}_{x_t}\mathcal{M}$
    $t \leftarrow t + \mathscr{T}$
  else
    $\xi \sim \mathrm{Uniform}(B_{x_t}(r))$ ▷ Random perturbation
    $x_{t+\mathscr{T}} = \mathtt{TSS}(x_t, \xi)$ ▷ Case 3: Perturbed accelerated gradient in $\mathrm{T}_{x_t}\mathcal{M}$
    if $f(x_t) - f(x_{t+\mathscr{T}}) < \frac{1}{2}\mathscr{E}$ **then**
      return x_t ▷ Approximate FOCP, likely an approximate SOCP
    end if
    $t \leftarrow t + \mathscr{T}$
  end if
end while

Proposition 5.3 (Case 2) *Fix parameters and assumptions as laid out in Sect. 3, with*

$$\chi \geq \log_2(\theta^{-1}) \geq 1. \tag{23}$$

If $x \in \mathcal{M}$ satisfies $\|\mathrm{grad} f(x)\| \leq 2\ell\mathscr{M}$, then $x_{\mathscr{T}} = \mathtt{TSS}(x)$ falls in one of two cases:

1. *Either $\|\mathrm{grad} f(x_{\mathscr{T}})\| \leq \epsilon$ and $f(x) - f(x_{\mathscr{T}}) \geq 0$,*
2. *Or $\|\mathrm{grad} f(x_{\mathscr{T}})\| > \epsilon$ and $f(x) - f(x_{\mathscr{T}}) \geq \mathscr{E}$.*

6 Second-Order Critical Points

As discussed in the previous section, TAGD produces ϵ-approximate first-order critical points at an accelerated rate, deterministically. Such a point might happen to be an approximate second-order critical point, or it might not. In order to produce approximate second-order critical points, PTAGD builds on top of TAGD as follows.

Whenever TAGD produces a point with gradient smaller than ϵ, PTAGD generates a random vector ξ close to the origin in the current tangent space and runs TSS starting from that perturbation. The run of TSS itself is deterministic. However, the randomized initialization has the following effect: if the current point is not an approximate second-order critical point, then with high probability the sequence generated by TSS produces significant cost decrease. Intuitively, this is because the current point is a saddle point, and gradient descent-type methods slowly but likely escape saddles. If this happens, we simply proceed with the algorithm. Otherwise, we can be reasonably confident that the point from which we ran the perturbed TSS is an approximate second-order critical point, and we terminate there.

Our main result regarding PTAGD, namely, Theorem 6.1, states that it computes approximate second-order critical points with high probability in a bounded number of iterations. As worked out in Theorem 1.6, this bound scales as $\epsilon^{-7/4}$, up to polylogarithmic terms which include a dependency in the dimension of the manifold and the probability of success.

Mirroring Sect. 5, the proof of Theorem 6.1 rests on the two propositions of that section and on an additional proposition introduced hereafter in this section.

Theorem 6.1 *Pick any $x_0 \in \mathcal{M}$. Fix parameters and assumptions as laid out in Sect. 3, with $d = \dim \mathcal{M}$, $\delta \in (0,1)$, any $\Delta_f \geq \max\left(f(x_0) - f_{\mathrm{low}}, \sqrt{\frac{\epsilon^3}{\hat{\rho}}}\right)$ and*

$$\chi \geq \log_2\left(\frac{d^{1/2}\ell^{3/2}\Delta_f}{(\hat{\rho}\epsilon)^{1/4}\epsilon^2\delta}\right) \geq \log_2(\theta^{-1}) \geq 1.$$

The call to PTAGD(x_0) *returns $x_t \in \mathcal{M}$ satisfying $f(x_t) \leq f(x_0)$, $\|\mathrm{grad} f(x_t)\| \leq \epsilon$ and (with probability at least $1 - 2\delta$) also $\lambda_{\min}(\nabla^2 \hat{f}_{x_t}(0)) \geq -\sqrt{\hat{\rho}\epsilon}$ with*

$$t + \mathscr{T} \leq T_2 \triangleq \left(2 + 4\frac{f(x_0) - f_{\mathrm{low}}}{\mathscr{E}}\right)\mathscr{T}. \tag{24}$$

To reach termination, the algorithm requires at most $2T_2$ pullback gradient queries and $4T_2$ function queries (but no Hessian queries), and a similar number of calls to the retraction.

Notice how this result gives a (probabilistic) guarantee about the smallest eigenvalue of the Hessian of the pullback $\hat{f}_x$ at 0 rather than about the Hessian of f itself at x. Owing to Lemma 2.5, the two are equal in particular when we use the exponential retraction (more generally, when we use a *second-order retraction*): see also [13,§3.5].

Proof of Theorem 6.1 The proof starts the same way as that of Theorem 5.1. The call to PTAGD(x_0) generates a sequence of points $x_{t_0}, x_{t_1}, x_{t_2}, \ldots$ on $\mathcal{M}$, with $t_0 = 0$. A priori, this sequence may be finite or infinite. Considering two consecutive indices t_i and t_{i+1}, we either have $t_{i+1} = t_i + 1$ (if the step from x_{t_i} to $x_{t_{i+1}}$ is a single gradient step (Case 1)) or $t_{i+1} = t_i + \mathscr{T}$ (if that same step is obtained through a call to TSS, with or without perturbation (Cases 3 and 2, respectively)). Moreover:

- In Case 1, Proposition 5.2 applies and guarantees

  $$f(x_{t_i}) - f(x_{t_{i+1}}) \geq \frac{\mathscr{E}}{\mathscr{T}} = \frac{\mathscr{E}}{\mathscr{T}}(t_{i+1} - t_i).$$

  The algorithm does not terminate here.
- In Case 2, Proposition 5.3 applies and guarantees that if $\|\mathrm{grad} f(x_{t_{i+1}})\| > \epsilon$ then

  $$f(x_{t_i}) - f(x_{t_{i+1}}) \geq \mathscr{E} = \frac{\mathscr{E}}{\mathscr{T}}(t_{i+1} - t_i),$$

  and the algorithm does not terminate here.
  If, however, $\|\mathrm{grad} f(x_{t_{i+1}})\| \leq \epsilon$, then $f(x_{t_i}) - f(x_{t_{i+1}}) \geq 0$ and the step from $x_{t_{i+1}}$ to $x_{t_{i+2}}$ does not fall in Case 2: it must fall in Case 3. (Indeed, it cannot fall

in Case 1 because the fact that a Case 2 step occurred tells us $\epsilon < 2\ell\mathscr{M}$.) The algorithm terminates with $x_{t_{i+1}}$ unless $f(x_{t_{i+1}}) - f(x_{t_{i+2}}) \geq \frac{1}{2}\mathscr{E}$. In other words, if the algorithm does not terminate with $x_{t_{i+1}}$, then

$$f(x_{t_i}) - f(x_{t_{i+2}}) = f(x_{t_i}) - f(x_{t_{i+1}}) + f(x_{t_{i+1}}) - f(x_{t_{i+2}}) \geq \frac{1}{2}\mathscr{E} = \frac{\mathscr{E}}{4\mathscr{T}}(t_{i+2} - t_i).$$

- In Case 3, the algorithm terminates with x_{t_i} unless

$$f(x_{t_i}) - f(x_{t_{i+1}}) \geq \frac{1}{2}\mathscr{E} = \frac{\mathscr{E}}{2\mathscr{T}}(t_{i+1} - t_i).$$

Clearly, $\texttt{PTAGD}(x_0)$ must terminate after a finite number of steps. Indeed, if it does not, then the above reasoning shows that the algorithm produces an amortized decrease in the cost function f of $\frac{\mathscr{E}}{4\mathscr{T}}$ per unit increment of the counter t, yet the value of f cannot decrease by more than $f(x_0) - f_{\text{low}}$.

Accordingly, assume $\texttt{PTAGD}(x_0)$ generates $x_{t_0}, \ldots, x_{t_{k+1}}$ and terminates there (returning x_{t_k}). The step from x_{t_k} to $x_{t_{k+1}}$ necessarily falls in Case 3: $t_{k+1} - t_k = \mathscr{T}$. The step from $x_{t_{k-1}}$ to x_{t_k} could be of any type. If it falls in Case 2, it could be that $f(x_{t_{k-1}}) - f(x_{t_k})$ is as small as zero and that $t_k - t_{k-1} = \mathscr{T}$. (All other scenarios are better, in that the cost function decreases more, and the counter increases as much or less.) Moreover, for all steps prior to that, each unit increment of t brings about an amortized decrease in f of $\frac{\mathscr{E}}{4\mathscr{T}}$. Thus, $t_{k+1} \leq t_{k-1} + 2\mathscr{T}$ and

$$f(x_0) - f_{\text{low}} \geq f(x_0) - f(x_{t_{k-1}}) \geq \frac{\mathscr{E}}{4\mathscr{T}} t_{k-1}.$$

Combining, we find

$$t_k + \mathscr{T} = t_{k+1} \leq \left(2 + 4\frac{f(x_0) - f_{\text{low}}}{\mathscr{E}}\right)\mathscr{T} \triangleq T_2.$$

What can we say about the point that is returned, x_{t_k}? Deterministically, $f(x_{t_k}) \leq f(x_0)$ and $\|\text{grad} f(x_{t_k})\| \leq \epsilon$ (notice that we cannot guarantee the same about $x_{t_{k+1}}$). Let us now discuss the role of randomness.

In any run of $\texttt{PTAGD}(x_0)$, there are at most $T_2/\mathscr{T}$ perturbations, that is, "Case 3" steps. By Proposition 6.2, the probability of any single one of those steps failing to prevent termination at a point where the smallest eigenvalue of the Hessian of the pullback at the origin is strictly less than $-\sqrt{\hat{\rho}\epsilon}$ is at most $\frac{\delta\mathscr{E}}{3\Delta_f}$. Thus, by a union bound, the probability of failure in any given run of $\texttt{PTAGD}(x_0)$ is at most (we use $\Delta_f \geq \max\left(f(x_0) - f_{\text{low}}, \sqrt{\frac{\epsilon^3}{\hat{\rho}}}\right) \geq \max\left(f(x_0) - f_{\text{low}}, 2^7\mathscr{E}\right)$ because $\chi \geq 1$ and $c \geq 2$):

$$\frac{T_2}{\mathscr{T}} \cdot \frac{\delta\mathscr{E}}{3\Delta_f} = \left(2 + 4\frac{f(x_0) - f_{\text{low}}}{\mathscr{E}}\right)\frac{\delta\mathscr{E}}{3\Delta_f} \leq \left(\frac{2\mathscr{E}}{3\Delta_f} + \frac{4}{3}\right)\delta \leq 2\delta.$$

In all other events, we have $\lambda_{\min}(\nabla^2 \hat{f}_{x_{t_k}}(0)) \geq -\sqrt{\hat{\rho}\epsilon}$.

For accounting of the maximal amount of work needed to run $\texttt{PTAGD}(x_0)$, use reasoning similar to that at the end of the proof of Theorem 5.1, adding the cost of checking the condition "$f(x_t) - f(x_{t+\mathscr{T}}) < \frac{1}{2}\mathscr{E}$" after each perturbed call to TSS.

Note: the inequality $\frac{d^{1/2}\ell^{3/2}\sqrt{\epsilon^3/\hat{\rho}}}{(\hat{\rho}\epsilon)^{1/4}\epsilon^2\delta} \geq \theta^{-1}$ holds for all $d \geq 1$ and $\delta \in (0, 1)$ with $c \geq 4$. □

The next proposition corresponds mostly to [28,Lem. 13]. A proof is in Appendix G.

Proposition 6.2 (Case 3) *Fix parameters and assumptions as laid out in Sect. 3, with* $d = \dim \mathcal{M}$, $\delta \in (0, 1)$, *any* $\Delta_f > 0$ *and*

$$\chi \geq \max\left(\log_2(\theta^{-1}), \log_2\left(\frac{d^{1/2}\ell^{3/2}\Delta_f}{(\hat{\rho}\epsilon)^{1/4}\epsilon^2\delta}\right)\right) \geq 1.$$

If $x \in \mathcal{M}$ *satisfies* $\|\mathrm{grad} f(x)\| \leq \min(\epsilon, 2\ell\mathscr{M})$ *and* $\lambda_{\min}(\nabla^2 \hat{f}_x(0)) \leq -\sqrt{\hat{\rho}\epsilon}$, *and* ξ *is sampled uniformly at random from the ball of radius* r *around the origin in* $\mathrm{T}_x\mathcal{M}$, *then* $x_{\mathscr{T}} = \texttt{TSS}(x, \xi)$ *satisfies* $f(x) - f(x_{\mathscr{T}}) \geq \mathscr{E}/2$ *with probability at least* $1 - \frac{\delta\mathscr{E}}{3\Delta_f}$ *over the choice of* ξ.

7 Conclusions and Perspectives

Our main complexity results for TAGD and PTAGD (Theorems 1.3 and 1.6) recover known Euclidean results when $\mathcal{M}$ is a Euclidean space. In particular, they retain the important properties of scaling essentially with $\epsilon^{-7/4}$ and of being either dimension-free (for TAGD) or almost dimension-free (for PTAGD). Those properties extend as is to the Riemannian case.

However, our Riemannian results are negatively impacted by the Riemannian curvature of $\mathcal{M}$, and also by the covariant derivative of the Riemann curvature endomorphism. We do not know whether such a dependency on curvature is necessary to achieve acceleration. In particular, the non-accelerated rates for Riemannian gradient descent, Riemannian trust-regions and Riemannian adaptive regularization with cubics under Lipschitz assumptions do not suffer from curvature [2, 13].

Curvature enters our complexity bounds through our geometric results (Theorem 2.7). For the latter, we do believe that curvature must play a role. Thus, it is natural to ask:

> *Can we achieve acceleration for first-order methods on Riemannian manifolds with weaker (or without) dependency on the curvature of the manifold?*

For the geodesically convex case, all algorithms we know of are affected by curvature [3–5, 48]. Additionally, Hamilton and Moitra [26] show that curvature can significantly slow down convergence rates in the geodesically convex case with noisy gradients.

Adaptive regularization with cubics (ARC) may offer insights in that regard. ARC is a cubically regularized approximate Newton method with optimal iteration complexity on the class of cost functions with Lipschitz continuous Hessian, assuming access to gradients and Hessians [19, 39]. Specifically, assuming f has ρ-Lipschitz continuous Hessian, ARC finds an $(\epsilon, \sqrt{\rho\epsilon})$-approximate second-order critical point in at most $\tilde{O}(\Delta_f \rho^{1/2}/\epsilon^{3/2})$ iterations, omitting logarithmic factors. This also holds on complete Riemannian manifolds [2,Cor. 3, eqs (16), (26)]. Note that this is dimension-free *and* curvature-free. Each iteration, however, requires solving a separate subproblem more costly than a gradient evaluation. Carmon and Duchi [18,§3] argue that it is possible to solve the subproblems accurately enough so as to find ϵ-approximate first-order critical points with $\sim 1/\epsilon^{7/4}$ Hessian-vector products overall, with randomization and a logarithmic dependency in dimension. Compared to TAGD, this has the benefit of being curvature-free, at the cost of randomization, a logarithmic dimension dependency, and of requiring Hessian-vector products. The latter could conceivably be approximated with finite differences of the gradients. Perhaps that operation leads to losses tied to curvature? If not, as it is unclear why there ought to be a trade-off between curvature dependency and randomization, this may be the indication that the curvature dependency is not necessary for acceleration.

On a distinct note and as pointed out in the introduction, TAGD and PTAGD are theoretical constructs. Despite having the theoretical upper-hand in worst-case scenarios, we do not expect them to be competitive against time-tested algorithms such as Riemannian versions of nonlinear conjugate gradients or the trust-region methods. It remains an interesting open problem to devise a truly practical accelerated first-order method on manifolds.

In the Euclidean case, Carmon et al. [15] showed that if one assumes not only the gradient and the Hessian of f but also the third derivative of f are Lipschitz continuous, then it is possible to find ϵ-approximate first-order critical points in just $\tilde{O}(\epsilon^{-5/3})$ iterations. We suspect that our proof technique could be used to prove a similar result on manifolds, possibly at the cost of also assuming a bound on the second covariant derivative of the Riemann curvature endomorphism.

Funding Open access funding provided by EPFL Lausanne.

A Parallel Transport vs Differential of Exponential Map

In this section, we give a proof for Proposition 2.8 regarding the difference between parallel transport along a geodesic and the differential of the exponential map. We use these families of functions parameterized by $K_{\text{low}} \in \mathbb{R}$:

$$h_{K_{\text{low}}}(t) = \begin{cases} t & \text{if } K_{\text{low}} = 0, \\ r\sin(t/r) & \text{if } K_{\text{low}} = 1/r^2 > 0, \\ r\sinh(t/r) & \text{if } K_{\text{low}} = -1/r^2 < 0. \end{cases} \tag{25}$$

$$g_{K_{\text{low}}}(t) = \int_0^t h_{K_{\text{low}}}(\tau)\mathrm{d}\tau = \begin{cases} \frac{t^2}{2} & \text{if } K_{\text{low}} = 0, \\ r^2\left(1-\cos(t/r)\right) & \text{if } K_{\text{low}} = 1/r^2 > 0, \\ r^2\left(\cosh(t/r)-1\right) & \text{if } K_{\text{low}} = -1/r^2 < 0. \end{cases} \tag{26}$$

$$f_{K_{\text{low}}}(t) = \frac{1}{t}\int_0^t g_{K_{\text{low}}}(\tau)\mathrm{d}\tau = \begin{cases} \frac{t^2}{6} & \text{if } K_{\text{low}} = 0, \\ r^2\left(1-\frac{\sin(t/r)}{t/r}\right) & \text{if } K_{\text{low}} = 1/r^2 > 0, \\ r^2\left(\frac{\sinh(t/r)}{t/r}-1\right) & \text{if } K_{\text{low}} = -1/r^2 < 0. \end{cases} \tag{27}$$

Under the assumptions we make below, these functions are only ever evaluated at points where they are nonnegative. In all cases, functions are dominated by the case $K_{\text{low}} < 0$; formally, for all $K_{\text{low}} \in \mathbb{R}$, all $K \geq |K_{\text{low}}|$ and all $t \geq 0$:

$$h_{K_{\text{low}}}(t) \leq h_{-K}(t), \qquad g_{K_{\text{low}}}(t) \leq g_{-K}(t), \qquad f_{K_{\text{low}}}(t) \leq f_{-K}(t). \tag{28}$$

If $K_{\text{low}} \geq 0$ and $t \geq 0$, then

$$h_{K_{\text{low}}}(t) \leq t, \qquad g_{K_{\text{low}}}(t) \leq \frac{1}{2}t^2, \qquad f_{K_{\text{low}}}(t) \leq \frac{1}{6}t^2. \tag{29}$$

Independently of the sign of K_{low}, if $0 \leq t \leq \pi/\sqrt{|K_{\text{low}}|}$, then

$$h_{K_{\text{low}}}(t) \leq t + 0.2712 \cdot \quad K_{\text{low}} t^3 \leq 3.6761 \cdot t,$$
$$g_{K_{\text{low}}}(t) \leq 1.0732 \cdot t^2, \quad f_{K_{\text{low}}}(t) \leq 0.2712 \cdot t^2.$$

For t bounded as indicated, this last line shows that up to constants the sign of K_{low} does not substantially affect bounds.

To state our result, we need the notion of conjugate points along geodesics on Riemannian manifolds. The following definition is equivalent to the standard one [32,Prop. 10.20 and p. 303]. We are particularly interested in situations where there are no conjugate points on some interval: we discuss that event in a remark.

Definition A.1 Let $\mathcal{M}$ be a Riemannian manifold. Consider $(x, s) \in \mathrm{T}\mathcal{M}$ and the geodesic $\gamma(t) = \mathrm{Exp}_x(ts)$ defined on an open interval I around zero. For $t \in I$, we say $\gamma(t)$ is *conjugate to x along γ* if $\mathrm{DExp}_x(ts)$ is rank deficient. We say γ *has an interior conjugate point on* $[0, \bar{t}] \subset I$ if $\gamma(t)$ is conjugate to x along γ for some $t \in (0, \bar{t})$.

Remark A.2 Let γ be a geodesic on a Riemannian manifold $\mathcal{M}$. If γ is minimizing on the interval $[0, \bar{t}]$, then it has no interior conjugate point on that interval

[32,Thm. 10.26]. Assume the sectional curvatures of $\mathcal{M}$ are in the interval $[K_{\text{low}}, K_{\text{up}}]$. Then:

1. If $K_{\text{up}} \le 0$, γ has no conjugate points at all [32,Pb. 10-7];
2. If $K_{\text{up}} > 0$, γ has no interior conjugate points on $[0, \pi/\sqrt{K_{\text{up}}}]$ [32,Thm. 11.9a]; and
3. If $K_{\text{low}} > 0$ and γ has no interior conjugate point on $[0, \bar{t}]$, then $\bar{t} \le \pi/\sqrt{K_{\text{low}}}$ [32,p. 298 and Thm. 11.9b]. This will be why, under our assumptions, $h_{K_{\text{low}}}$ (25) is only ever evaluated at points where it is nonnegative.

We now state and prove the main result of this section. A similar result appears in [45,Lem. 6] for general retractions. Constants there are not explicit (they are absorbed in $O(\cdot)$ notation). Their proof is based on Taylor expansions of the differential of the exponential map as they appear in [46,Thm. A.2.9], namely, for $s \mapsto \mathrm{DExp}_x(s)$ around $s = 0$. In the next section, we investigate a situation around $s \neq 0$. In appendices, we typically omit subscripts for inner products and norms.

Proposition A.3 *Let $\mathcal{M}$ be a Riemannian manifold whose sectional curvatures are in the interval $[K_{\text{low}}, K_{\text{up}}]$, and let $K = \max(|K_{\text{low}}|, |K_{\text{up}}|)$. Consider $(x, s) \in \mathrm{T}\mathcal{M}$ and the geodesic $\gamma(t) = \mathrm{Exp}_x(ts)$. If γ is defined and has no interior conjugate point on the interval $[0, 1]$, then*

$$\forall \dot{s} \in \mathrm{T}_x\mathcal{M}, \qquad \|(T_s - P_s)[\dot{s}]\| \le K \cdot f_{K_{\text{low}}}(\|s\|) \cdot \|\dot{s}_\perp\|, \tag{30}$$

where $\dot{s}_\perp = \dot{s} - \frac{\langle s, \dot{s}\rangle}{\langle s, s\rangle} s$ is the component of $\dot{s}$ orthogonal to s, $T_s = \mathrm{DExp}_x(s)$ and P_{ts} denotes parallel transport along γ from $\gamma(0)$ to $\gamma(t)$. (The inequality holds with equality if $K_{\text{low}} = K_{\text{up}}$.)

If it also holds that $\|s\| \le \pi/\sqrt{|K_{\text{low}}|}$, then

$$\forall \dot{s} \in \mathrm{T}_x\mathcal{M}, \qquad \|(T_s - P_s)[\dot{s}]\| \le \frac{1}{3} K \|s\|^2 \|\dot{s}_\perp\|. \tag{31}$$

Proof For convenience, we consider $\|s\| = 1$: the result follows by a simple rescaling of t. Given any tangent vector $\dot{s} \in \mathrm{T}_x\mathcal{M}$, consider the following smooth vector field along γ:

$$J(t) = \mathrm{DExp}_x(ts)[t\dot{s}]. \tag{32}$$

By [32,Prop. 10.10], this is the unique Jacobi field satisfying the initial conditions

$$J(0) = 0 \qquad \text{and} \qquad \mathrm{D}_t J(0) = \dot{s}, \tag{33}$$

where D_t is the covariant derivative along curves induced by the Riemannian connection. Thus, J is smooth and obeys the ordinary differential equation (ODE) known as the Jacobi equation:

$$\mathrm{D}_t^2 J(t) + R(J(t), \gamma'(t))\gamma'(t) = 0, \tag{34}$$

where R denotes Riemannian curvature. Fix $e_d = s$ and pick $e_1, \ldots, e_{d-1}$ so that $e_1, \ldots, e_d$ form an orthonormal basis for $\mathrm{T}_x\mathcal{M}$. Parallel transport this basis along γ as

$$E_i(t) = P_{ts}(e_i), \qquad i = 1, \ldots, d, \tag{35}$$

so that $E_1(t), \ldots, E_d(t)$ form an orthonormal basis for $\mathrm{T}_{\gamma(t)}\mathcal{M}$. Expand J as

$$J(t) = \sum_{i=1}^{d} a_i(t) E_i(t) \tag{36}$$

with uniquely defined smooth, real functions $a_1, \ldots, a_d$. Plugging this expansion into the Jacobi equation yields the ODE

$$\sum_{i=1}^{d} a_i''(t) E_i(t) + \sum_{i=1}^{d} a_i(t) R(E_i(t), E_d(t)) E_d(t) = 0, \tag{37}$$

where we used the Leibniz rule on D_t, the fact that $\mathrm{D}_t E_i = 0$, linearity of the Riemann curvature endomorphism in its inputs, and the fact that

$$\gamma'(t) = P_{ts}(\gamma'(0)) = E_d(t).$$

Taking an inner product of this ODE against each one of the fields $E_j(t)$ yields d ODEs:

$$a_j''(t) = -\sum_{i=1}^{d} a_i(t) \left\langle R(E_i(t), E_d(t)) E_d(t), E_j(t) \right\rangle, \qquad j = 1, \ldots, d. \tag{38}$$

Furthermore, the initial conditions fix $a_i(0) = 0$ and $a_i'(0) = \langle \dot{s}, e_i \rangle$ for $i = 1, \ldots, d$.

Owing to symmetries of Riemannian curvature, the summation above can be restricted to the range $1, \ldots, d-1$. For the same reason, $a_d''(t) = 0$, so that

$$a_d(t) = a_d(0) + t a_d'(0) = t \langle \dot{s}, s \rangle. \tag{39}$$

It remains to solve for the first $d-1$ coefficients (they are decoupled from a_d). This effectively splits the solution J into two fields: one tangent (aligned with γ'), and one normal (orthogonal to γ'):

$$J(t) = t \langle \dot{s}, s \rangle \gamma'(t) + J_\perp(t), \qquad J_\perp(t) = \sum_{i=1}^{d-1} a_i(t) E_i(t). \tag{40}$$

The normal part is the Jacobi field with initial conditions $J_\perp(0) = 0$ and $\mathrm{D}_t J_\perp(0) = \dot{s}_\perp$, where $\dot{s}_\perp = \dot{s} - \langle \dot{s}, s \rangle s$ is the component of $\dot{s}$ orthogonal to s.

Introducing vector notation for the first $d-1$ ODEs, let $a(t) \in \mathbb{R}^{d-1}$ have components $a_1(t), \ldots, a_{d-1}(t)$, and let $M(t) \in \mathbb{R}^{(d-1)\times(d-1)}$ have entries

$$M_{ji}(t) = \left\langle R(E_i(t), E_d(t))E_d(t), E_j(t)\right\rangle. \tag{41}$$

Then, equations in (38) for $j = 1, \ldots, d-1$ can be written succinctly as

$$a''(t) = -M(t)a(t). \tag{42}$$

Since $a(t)$ is smooth, it holds that

$$a(t) = a(0) + \int_0^t a'(\tau)\mathrm{d}\tau = a(0) + ta'(0) + \int_0^t \int_0^\tau a''(\theta)\mathrm{d}\theta\mathrm{d}\tau. \tag{43}$$

Initial conditions specify $a(0) = 0$, so that (with $\|\cdot\|$ also denoting the standard Euclidean norm and associated operator norm in real space):

$$\|a(t) - ta'(0)\| \le \int_0^t \int_0^\tau \|M(\theta)\|\|a(\theta)\|\mathrm{d}\theta\mathrm{d}\tau. \tag{44}$$

The left-hand side is exactly what we seek to control. Indeed, initial conditions ensure $\dot{s} = a_1'(0)e_1 + \cdots + a_d'(0)e_d$, and:

$$\begin{aligned}\|(\mathrm{DExp}_x(ts) - P_{ts})[t\dot{s}]\| &= \|J(t) - P_{ts}(t\dot{s})\| \\ &= \left\|\sum_{i=1}^{d}\left[a_i(t)E_i(t) - ta_i'(0)E_i(t)\right]\right\| \\ &= \sqrt{\|a(t) - ta'(0)\|^2 + |a_d(t) - ta_d'(0)|^2} \\ &= \|a(t) - ta'(0)\|.\end{aligned}$$

For the right-hand side of (44), first note that $M(t)$ is a symmetric matrix owing to the symmetries of R.

Additionally, for any unit-norm $z \in \mathbb{R}^{d-1}$,

$$z^\top M(t)z = \sum_{i,j=1}^{d-1} z_i z_j \left\langle R(E_i(t), E_d(t))E_d(t), E_j(t)\right\rangle = \left\langle R(v, \gamma'(t))\gamma'(t), v\right\rangle, \tag{45}$$

where $v = z_1E_1(t) + \cdots + z_{d-1}E_{d-1}(t)$ is a tangent vector at $\gamma(t)$: it is orthogonal to $\gamma'(t)$ and also has unit norm. By definition of sectional curvature $K(\cdot,\cdot)$ (10), it follows that

$$z^\top M(t)z = K(v, \gamma'(t)). \tag{46}$$

By symmetry of $M(t)$, we conclude that

$$\|M(t)\| = \max_{z \in \mathbb{R}^{d-1}, \|z\|=1} |z^\top M(t) z| \le K, \tag{47}$$

where $K \ge 0$ is such that all sectional curvatures of $\mathcal{M}$ along γ are in the interval $[-K, K]$. Going back to (44), we have so far shown that

$$\|(\mathrm{DExp}_x(ts) - P_{ts})[t\dot{s}]\| \le K \int_0^t \int_0^\tau \|a(\theta)\| \mathrm{d}\theta \mathrm{d}\tau. \tag{48}$$

It remains to bound $\|a(\theta)\|$. By (40), we see that $\|a(t)\| = \|J_\perp(t)\|$. By the Jacobi field comparison theorem [32,Thm. 11.9b] and our assumed lower-bound on sectional curvature, we can now claim that, for $t \ge 0$, with $h_{K_{\mathrm{low}}}(t)$ as defined by (25),

$$\|a(t)\| = \|J_\perp(t)\| \le h_{K_{\mathrm{low}}}(t)\|\dot{s}_\perp\|, \tag{49}$$

provided γ has no interior conjugate point on $[0, t]$. Combining with (48) and with the definitions of $h_{K_{\mathrm{low}}}$ (25), $g_{K_{\mathrm{low}}}$ (26) and $f_{K_{\mathrm{low}}}$ (27), we find

$$\begin{aligned}\|(\mathrm{DExp}_x(ts) - P_{ts})[t\dot{s}]\| &\le K\|\dot{s}_\perp\| \int_0^t \int_0^\tau h_{K_{\mathrm{low}}}(\theta) \mathrm{d}\theta \mathrm{d}\tau \\ &= K\|\dot{s}_\perp\| \int_0^t g_{K_{\mathrm{low}}}(\tau) \mathrm{d}\tau \\ &= K\|\dot{s}_\perp\| \cdot t f_{K_{\mathrm{low}}}(t).\end{aligned} \tag{50}$$

It only remains to divide through by t, and to rescale s so that t plays the role of $\|s\|$.

For the special case where $K_{\mathrm{up}} = K_{\mathrm{low}} = \pm K$ (constant sectional curvature), one can show (for example by polarization) that $M(t) = \pm K I_{d-1}$, that is, $M(t)$ is a multiple of the identity matrix. As a result, the ODEs separate and are easily solved (see also [32,Prop. 10.12]). Explicitly, with $\|s\| = 1$,

$$\mathrm{DExp}_x(ts)[t\dot{s}] = J(t) = h_{\pm K}(t) P_{ts}(\dot{s}_\perp) + t P_{ts}(\dot{s}_\parallel), \tag{51}$$

where $\dot{s}_\parallel = \langle \dot{s}, s\rangle\, s$ is the component of $\dot{s}$ parallel to s. Hence,

$$\mathrm{DExp}_x(ts)[t\dot{s}] - P_{ts}(t\dot{s}) = (h_{\pm K}(t) - t) P_{ts}(\dot{s}_\perp), \tag{52}$$

and the claim follows easily after dividing through by t and rescaling. □

As a continuation of the previous proof and in anticipation of our needs in Appendix B, we provide a lemma controlling the Jacobi field J and its covariant derivative, assessing both the full field and its normal component.

Lemma A.4 *Let $\mathcal{M}$ be a Riemannian manifold whose sectional curvatures are in the interval $[K_{\mathrm{low}}, K_{\mathrm{up}}]$, and let $K = \max(|K_{\mathrm{low}}|, |K_{\mathrm{up}}|)$. Consider $(x, s) \in \mathrm{T}\mathcal{M}$ with*

$\|s\| = 1$ *and the geodesic* $\gamma(t) = \mathrm{Exp}_x(ts)$. *Given a tangent vector* $\dot{s} \in \mathrm{T}_x\mathcal{M}$, *consider the Jacobi field* J *defined by* (40)*:*

$$J(t) = t\,\langle \dot{s}, s\rangle\,\gamma'(t) + J_\perp(t),$$

where $J_\perp$ *is the Jacobi field along* γ *with initial conditions* $J_\perp(0) = 0$ *and* $\mathrm{D}_t J_\perp(0) = \dot{s}_\perp$, *and* $\dot{s}_\perp = \dot{s} - \langle \dot{s}, s\rangle\, s$ *is the component of* $\dot{s}$ *orthogonal to* s. *For* $t \geq 0$ *such that* γ *is defined and has no interior conjugate point on the interval* $[0, t]$, *the following inequalities hold:*

$$\|J(t)\| \leq \max(t, h_{K_{\mathrm{low}}}(t))\|\dot{s}\|, \qquad \|\mathrm{D}_t J(t)\| \leq \big(1 + K g_{K_{\mathrm{low}}}(t)\big)\,\|\dot{s}\|,$$
$$\|J_\perp(t)\| \leq h_{K_{\mathrm{low}}}(t)\|\dot{s}_\perp\|, \qquad \|\mathrm{D}_t J_\perp(t)\| \leq \big(1 + K g_{K_{\mathrm{low}}}(t)\big)\,\|\dot{s}_\perp\|,$$

where $h_{K_{\mathrm{low}}}(t)$ *and* $g_{K_{\mathrm{low}}}(t)$ *are defined by* (25) *and* (26).

Proof The proof is a continuation of that of Proposition A.3. Using notation as in there,

$$J_\perp(t) = \sum_{i=1}^{d-1} a_i(t) E_i(t).$$

Since $J_\perp$ and $\mathrm{D}_t J_\perp$ are orthogonal to $\gamma' = E_d$, we know that

$$\|J\|^2 = t^2\,\langle \dot{s}, s\rangle^2 + \|J_\perp\|^2 \qquad \text{and} \qquad \|\mathrm{D}_t J\|^2 = \langle \dot{s}, s\rangle^2 + \|\mathrm{D}_t J_\perp\|^2.$$

The bound $\|J_\perp(t)\| \leq h_{K_{\mathrm{low}}}(t)\|\dot{s}_\perp\|$ appears explicitly as (49). With α denoting the angle between s and $\dot{s}$, we may write $\langle \dot{s}, s\rangle^2 = \cos(\alpha)^2\|\dot{s}\|^2$ and $\|\dot{s}_\perp\|^2 = \sin(\alpha)^2\|\dot{s}\|^2$, so that

$$\|J\|^2 \leq t^2\left(\cos(\alpha)^2 + \left(\frac{h_{K_{\mathrm{low}}}(t)}{t}\right)^2 \sin(\alpha)^2\right)\|\dot{s}\|^2.$$

Since the maximum of $\alpha \mapsto \cos(\alpha)^2 + q\sin(\alpha)^2$ with $q \in \mathbb{R}$ is $\max(1, q)$, we find for $t \geq 0$ that

$$\|J\| \leq \max(t, h_{K_{\mathrm{low}}}(t))\|\dot{s}\|.$$

With the same tools, we may also bound $\mathrm{D}_t J = \langle \dot{s}, s\rangle\,\gamma' + \mathrm{D}_t J_\perp$. Indeed, its coordinates in the frame $E_1, \ldots, E_d$ are given by $a_1', \ldots, a_d'$ with $a_d'(t) = \langle \dot{s}, s\rangle$, so that

$$\|\mathrm{D}_t J(t)\|^2 = \langle \dot{s}, s\rangle^2 + \|\mathrm{D}_t J_\perp(t)\|^2 = \langle \dot{s}, s\rangle^2 + \|a'(t)\|^2,$$

where $a(t) \in \mathbb{R}^{d-1}$ collects the $d-1$ first coordinates. Moreover,

$$a'(t) = a'(0) + \int_0^t a''(\tau)\mathrm{d}\tau = a'(0) - \int_0^t M(\tau)a(\tau)\mathrm{d}\tau.$$

Note that

$$\left\| \int_0^t M(\tau)a(\tau)\mathrm{d}\tau \right\| \leq K\|\dot{s}_\perp\| \int_0^t h_{K_{\mathrm{low}}}(\tau)\mathrm{d}\tau = K\|\dot{s}_\perp\| g_{K_{\mathrm{low}}}(t).$$

Combining with the fact that $\|a'(0)\| = \|\dot{s}_\perp\|$, we get

$$\|\mathrm{D}_t J_\perp(t)\| \leq \left(1 + K g_{K_{\mathrm{low}}}(t)\right) \|\dot{s}_\perp\|,$$

as announced. We now conclude along the same lines as above with

$$\|\mathrm{D}_t J(t)\|^2 \leq \left(\cos(\alpha)^2 + \left(1 + K g_{K_{\mathrm{low}}}(t)\right)^2 \sin(\alpha)^2\right) \|\dot{s}\|^2.$$

Since $\max(1, 1 + K g_{K_{\mathrm{low}}}(t)) = 1 + K g_{K_{\mathrm{low}}}(t)$, we reach the desired conclusion. □

B Controlling the Initial Acceleration $c''(0)$

In this section, we build a proof for Proposition 2.10, whose aim is to control the initial intrinsic acceleration $c''(0)$ of the curve $c(t) = \mathrm{Exp}_x(s + t\dot{s})$. Since $c'(t) = \mathrm{DExp}_x(s + t\dot{s})[\dot{s}]$, we can think of this result as giving us access to a second derivative of the exponential map Exp_x away from the origin. As a first step, we build an ODE whose solution encodes $c''(0)$.

Proposition B.1 *Let $\mathcal{M}$ be a Riemannian manifold with Riemannian connection ∇ and Riemann curvature endomorphism R. Consider $(x, s) \in \mathrm{T}\mathcal{M}$ with $\|s\| = 1$ and the geodesic $\gamma(t) = \mathrm{Exp}_x(ts)$. Furthermore, consider a tangent vector $\dot{s} \in \mathrm{T}_x\mathcal{M}$ and the curve*

$$c_{ts,\dot{s}}(q) = \mathrm{Exp}_x(ts + q\dot{s})$$

defined for some fixed t. Let J be the Jacobi field along γ with initial conditions $J(0) = 0$ and $\mathrm{D}_t J(0) = \dot{s}$. We use it to define a new vector field H along γ:

$$H = 4R(\gamma', J)\mathrm{D}_t J + (\nabla_J R)(\gamma', J)\gamma' + \left(\nabla_{\gamma'} R\right)(\gamma', J)J.$$

The smooth vector field W along γ defined by the linear ODE

$$\mathrm{D}_t^2 W + R(W, \gamma')\gamma' = H$$

with initial conditions $W(0) = 0$ and $\mathrm{D}_t W(0) = 0$ is also defined on the same domain as γ. This vector field is related to the initial intrinsic acceleration of the curve $c_{ts,\dot{s}}$ as follows:

$$W(t) = t^2 c''_{ts,\dot{s}}(0).$$

Furthermore, the vector field H is equivalently defined as

$$H = 4R(\gamma', J_\perp)\mathrm{D}_t J + (\nabla_J R)(\gamma', J_\perp)\gamma' + \left(\nabla_{\gamma'} R\right)(\gamma', J_\perp)J,$$

where $J_\perp$ the Jacobi field along γ with initial conditions $J_\perp(0) = 0$ and $\mathrm{D}_t J_\perp(0) = \dot{s}_\perp = \dot{s} - \langle \dot{s}, s\rangle s$.

Proof Define

$$\Gamma(q, t) = \mathrm{Exp}_x(t(s + q\dot{s})),$$

a variation through geodesics of the geodesic

$$\gamma(t) = \Gamma(0, t) = \mathrm{Exp}_x(ts).$$

Then,

$$J(t) = \partial_q \Gamma(0, t) = \mathrm{DExp}_x(t(s + q\dot{s}))[t\dot{s}]\Big|_{q=0} = \mathrm{DExp}_x(ts)[t\dot{s}]$$

is the Jacobi field along γ with initial conditions $J(0) = 0$ and $\mathrm{D}_t J(0) = \dot{s}$: the same field we considered in the proof of Proposition A.3. Further consider

$$W(t) = \left(\mathrm{D}_q \partial_q \Gamma\right)(0, t), \tag{53}$$

another smooth vector field along γ. This field is related to acceleration of curves of the form

$$c_{s,\dot{s}}(q) = \mathrm{Exp}_x(s + q\dot{s}),$$

because $c_{ts,t\dot{s}}(q) = \Gamma(q, t)$. Specifically,

$$W(t) = \left(\mathrm{D}_q \partial_q \Gamma\right)(0, t) = c''_{ts,t\dot{s}}(0) = t^2 c''_{ts,\dot{s}}(0). \tag{54}$$

To verify the last equality, differentiate the identity $c_{ts,t\dot{s}}(q) = c_{ts,\dot{s}}(tq)$ twice with respect to q, with the chain rule. This shows in particular that

$$W(0) = 0 \qquad \text{and} \qquad \mathrm{D}_t W(0) = 0. \tag{55}$$

Our goal is to derive a second-order ODE for W. In so doing, we repeatedly use the two following results from Riemannian geometry which allow us to commute certain derivatives:

- [32,Prop. 7.5] For every smooth vector field V along Γ (meaning $V(q,t)$ is tangent to $\mathcal{M}$ at $\Gamma(q,t)$),

$$\mathrm{D}_t\mathrm{D}_q V - \mathrm{D}_q\mathrm{D}_t V = R(\partial_t\Gamma, \partial_q\Gamma)V, \tag{56}$$

where R is the Riemann curvature endomorphism.

- [32,Lem. 6.2] The *symmetry lemma* states

$$\mathrm{D}_q\partial_t\Gamma = \mathrm{D}_t\partial_q\Gamma. \tag{57}$$

With the link between W and $\mathrm{D}_q\partial_q\Gamma$ in mind, we compute a first derivative with respect to t:

$$\mathrm{D}_t\mathrm{D}_q\partial_q\Gamma = \mathrm{D}_q\mathrm{D}_t\partial_q\Gamma + R(\partial_t\Gamma, \partial_q\Gamma)\partial_q\Gamma,$$

then a second derivative:

$$\mathrm{D}_t\mathrm{D}_t\mathrm{D}_q\partial_q\Gamma = \mathrm{D}_t\mathrm{D}_q\mathrm{D}_t\partial_q\Gamma + \mathrm{D}_t\left\{R(\partial_t\Gamma, \partial_q\Gamma)\partial_q\Gamma\right\}.$$

Our goal is to evaluate this expression for $q = 0$, in which case the left-hand side yields $\mathrm{D}_t^2 W$. However, it is unclear how to evaluate the first term on the right-hand side at $q = 0$. Focusing on that term for now, apply the commutation rule on the first two derivatives:

$$\mathrm{D}_t\mathrm{D}_q\mathrm{D}_t\partial_q\Gamma = \mathrm{D}_q\mathrm{D}_t\mathrm{D}_t\partial_q\Gamma + R(\partial_t\Gamma, \partial_q\Gamma)\mathrm{D}_t\partial_q\Gamma.$$

Focusing on the first term once more, apply the symmetry lemma then the commutation rule:

$$\begin{aligned}\mathrm{D}_q\mathrm{D}_t\mathrm{D}_t\partial_q\Gamma &= \mathrm{D}_q\left\{\mathrm{D}_t\mathrm{D}_q\partial_t\Gamma\right\} = \mathrm{D}_q\left\{\mathrm{D}_q\mathrm{D}_t\partial_t\Gamma + R(\partial_t\Gamma, \partial_q\Gamma)\partial_t\Gamma\right\}\\ &= \mathrm{D}_q\left\{R(\partial_t\Gamma, \partial_q\Gamma)\partial_t\Gamma\right\}.\end{aligned}$$

To reach the last equality, we used that $\mathrm{D}_t\partial_t\Gamma$ vanishes identically since $t \mapsto \Gamma(q,t)$ is a geodesic for every fixed q. Combining, we find

$$\begin{aligned}\mathrm{D}_t\mathrm{D}_t\mathrm{D}_q\partial_q\Gamma &= R(\partial_t\Gamma, \partial_q\Gamma)\mathrm{D}_t\partial_q\Gamma\\ &\quad + \mathrm{D}_t\left\{R(\partial_t\Gamma, \partial_q\Gamma)\partial_q\Gamma\right\} + \mathrm{D}_q\left\{R(\partial_t\Gamma, \partial_q\Gamma)\partial_t\Gamma\right\}.\end{aligned} \tag{58}$$

Using the chain rule for tensors as in (11) (see also [32,pp. 95–103] or [41,Def. 3.17]), we can further expand the right-most term:

$$\begin{aligned}\mathrm{D}_q\left\{R(\partial_t\Gamma, \partial_q\Gamma)\partial_t\Gamma\right\} &= \left(\nabla_{\partial_q\Gamma}R\right)(\partial_t\Gamma, \partial_q\Gamma)\partial_t\Gamma + R\left(\mathrm{D}_q\partial_t\Gamma, \partial_q\Gamma\right)\partial_t\Gamma\\ &\quad + R\left(\partial_t\Gamma, \mathrm{D}_q\partial_q\Gamma\right)\partial_t\Gamma + R\left(\partial_t\Gamma, \partial_q\Gamma\right)\mathrm{D}_q\partial_t\Gamma.\end{aligned}$$

It is now easier to evaluate the whole expression at $q = 0$: using

$$\partial_q \Gamma(0,t) = J(t), \qquad \partial_t \Gamma(0,t) = \gamma'(t) \qquad \text{and} \qquad \left(\mathrm{D}_q \partial_q \Gamma\right)(0,t) = W(t)$$

repeatedly, and also $\mathrm{D}_q \partial_t \Gamma = \mathrm{D}_t \partial_q \Gamma$ twice so that it evaluates to $\mathrm{D}_t J$ at $q = 0$, we find

$$\begin{aligned} \mathrm{D}_t^2 W = {} & 2R(\gamma', J)\mathrm{D}_t J + \mathrm{D}_t \left\{ R(\gamma', J)J \right\} \\ & + (\nabla_J R)(\gamma', J)\gamma' + R\left(\mathrm{D}_t J, J\right)\gamma' + R(\gamma', W)\gamma'. \end{aligned}$$

This is now an ODE in the single variable t, involving smooth vector fields J, W and γ' along the geodesic γ. We may apply the chain rule for tensors again (we could just as well have done this earlier too):

$$\mathrm{D}_t \left\{ R(\gamma', J)J \right\} = \left(\nabla_{\gamma'} R\right)(\gamma', J)J + R\left(\gamma', \mathrm{D}_t J\right)J + R(\gamma', J)\mathrm{D}_t J,$$

here too simplifying one term since γ'' vanishes. The algebraic Bianchi identity [32,p. 203] states $R(X,Y)Z + R(Y,Z)X + R(Z,X)Y = 0$, so that in particular

$$R\left(\mathrm{D}_t J, J\right)\gamma' + R\left(\gamma', \mathrm{D}_t J\right)J = -R(J, \gamma')\mathrm{D}_t J = R(\gamma', J)\mathrm{D}_t J.$$

(We also used anti-symmetry of R). Overall, $\mathrm{D}_t^2 W + R(W, \gamma')\gamma' = H$ with

$$H = 4R(\gamma', J)\mathrm{D}_t J + (\nabla_J R)(\gamma', J)\gamma' + \left(\nabla_{\gamma'} R\right)(\gamma', J)J.$$

The Jacobi field J splits into its tangent and normal parts (40):

$$J(t) = t \left\langle \dot{s}, s \right\rangle \gamma'(t) + J_\perp(t).$$

Since $R(\gamma', \gamma') = 0$ by anti-symmetry of R, and since for the same reason $(\nabla_. R)(\gamma', \gamma') = 0$ as well, by linearity, we may simplify H to:

$$H = 4R(\gamma', J_\perp)\mathrm{D}_t J + (\nabla_J R)(\gamma', J_\perp)\gamma' + \left(\nabla_{\gamma'} R\right)(\gamma', J_\perp)J.$$

This concludes the proof. □

To reach our main result, it remains to bound the solutions of the ODE in W. In order to do so, we notably need to bound the inhomogeneous term H. For that reason, we require a bound on the covariant derivative of Riemannian curvature.

Theorem B.2 *Let $\mathcal{M}$ be a Riemannian manifold whose sectional curvatures are in the interval $[K_{\mathrm{low}}, K_{\mathrm{up}}]$, and let $K = \max(|K_{\mathrm{low}}|, |K_{\mathrm{up}}|)$. Also assume ∇R—the covariant derivative of the Riemann curvature endomorphism—is bounded by F in* *operator norm. Pick any* $(x, s) \in \mathrm{T}\mathcal{M}$ *such that the geodesic* $\gamma(t) = \mathrm{Exp}_x(ts)$ *is defined for all* $t \in [0, 1]$*, and such that*

$$\|s\| \le \min\left(C\frac{1}{\sqrt{K}}, C'\frac{K}{F}\right)$$

with some constants $C \le \pi$ *and* C'*. For any* $\dot{s} \in \mathrm{T}_x\mathcal{M}$*, the curve*

$$c(t) = \mathrm{Exp}_x(s + t\dot{s})$$

has initial acceleration bounded as

$$\|c''(0)\| \le \bar{\bar{\bar{W}}} K \|s\| \|\dot{s}\| \|\dot{s}_\perp\|,$$

where $\dot{s}_\perp = \dot{s} - \frac{\langle s, \dot{s}\rangle}{\langle s, s\rangle} s$ *is the component of* $\dot{s}$ *orthogonal to* s *and* $\bar{\bar{\bar{W}}} \in \mathbb{R}$ *is only a function of* C *and* C'*. In particular, for* $C, C' \le \frac{1}{4}$*, we have* $\bar{\bar{\bar{W}}} \le \frac{3}{2}$*.*

Proof By Remark A.2, since $C \le \pi$ we know that γ has no interior conjugate point on $[0, 1]$. Since the claim is clear for either $s = 0$ or $\dot{s} = 0$, assume $\|s\| = 1$ for now—we rescale at the end—and $\dot{s} \ne 0$. We also assume $K > 0$: the case $K = 0$ follows easily by inspection of the proof below.

Following Proposition B.1, the goal is to bound W: the solution of an ODE with right-hand side given by the vector field H. As we did in earlier proofs, pick an orthonormal basis $e_1, \ldots, e_d$ for $\mathrm{T}_x\mathcal{M}$ with $e_d = s$ and transport it along γ as $E_i(t) = P_{ts}(e_i)$. We expand W and H as

$$W(t) = \sum_{i=1}^{d} w_i(t)E_i(t), \qquad H(t) = \sum_{i=1}^{d} h_i(t)E_i(t). \tag{59}$$

This allows us to write the ODE in coordinates:

$$w''(t) + M(t)w(t) = h(t), \tag{60}$$

where $M(t)$ is as in (41) but defined in $\mathbb{R}^{d\times d}$ (thus, it has an extra row and column of zeros), and $w(t), h(t) \in \mathbb{R}^d$ are vectors containing the coordinates of $W(t)$ and $H(t)$. Since $W(0) = \mathrm{D}_t W(0) = 0$, we have $w(0) = w'(0) = 0$ and we deduce

$$w(t) = w(0) + tw'(0) + \int_0^t \int_0^\tau w''(\theta)\mathrm{d}\theta\mathrm{d}\tau = \int_0^t \int_0^\tau -M(\theta)w(\theta) + h(\theta)\,\mathrm{d}\theta\mathrm{d}\tau.$$

Thus,

$$\|W(t)\| = \|w(t)\| \le \int_0^t \int_0^\tau K\|w(\theta)\| + \|h(\theta)\|\,\mathrm{d}\theta\mathrm{d}\tau. \tag{61}$$

To proceed, we need a bound on $\|H(t)\| = \|h(t)\|$ and a first bound on $\|W(t)\|$. We will then improve the latter by bootstrapping.

Let us first bound H.

Following [29,eq. (9)], we know that R is bounded (as an operator) as follows:

$$\|R(X,Y)Z\| \leq K_0\|X\|\|Y\|\|Z\| \quad \text{with} \\ K_0 = \sqrt{K^2 + (25/36)(K_{\text{up}} - K_{\text{low}})^2} \leq 2K, \tag{62}$$

where X, Y, Z are arbitrary vector fields along γ. We further assume that

$$\|(\nabla_U R)(X,Y)Z\| \leq F\|U\|\|X\|\|Y\|\|Z\| \tag{63}$$

for some finite $F \geq 0$. Then,

$$\|H\| \leq 4K_0\|\gamma'\|\,\|\mathrm{D}_t J\|\,\|J_\perp\| + 2F\|\gamma'\|^2\|J\|\|J_\perp\|. \tag{64}$$

Since $\|\gamma'(t)\| = \|s\| = 1$ for all t, this expression simplifies somewhat. Using Lemma A.4, we can also bound all terms involving J and $J_\perp$, so that, also using $K_0 \leq 2K$,

$$\|H(t)\| \leq h_{K_{\text{low}}}(t)\Big(8K\left(1 + Kg_{K_{\text{low}}}(t)\right) + 2F\max(t, h_{K_{\text{low}}}(t))\Big)\|\dot{s}\|\|\dot{s}_\perp\|. \tag{65}$$

Since $h_{K_{\text{low}}}(t) \leq h_{-K}(t) = t\frac{\sinh(\sqrt{K}t)}{\sqrt{K}t}$ and likewise $Kg_{K_{\text{low}}}(t) \leq Kg_{-K}(t) = \cosh(\sqrt{K}t) - 1$, and since $h_{-K}(t) \geq t$, we find

$$\|H(t)\| \leq t\frac{\sinh(\sqrt{K}t)}{\sqrt{K}t}\left(8K\cosh(\sqrt{K}t) + 2Ft\frac{\sinh(\sqrt{K}t)}{\sqrt{K}t}\right)\|\dot{s}\|\|\dot{s}_\perp\| \tag{66}$$

for all $t \geq 0$. Assuming $0 \leq \sqrt{K}t \leq C$ for some $C > 0$, we find

$$\|H(t)\| \leq (aK + bFt)\,t\|\dot{s}\|\|\dot{s}_\perp\| \tag{67}$$

with $a = 8\frac{\sinh(C)\cosh(C)}{C}$ and $b = 2\frac{\sinh(C)^2}{C^2}$. Let us further assume that $0 \leq t \leq C'\frac{K}{F}$. Then, $Ft \leq C'K$ and we write:

$$\frac{\|H(t)\|}{\|\dot{s}\|\|\dot{s}_\perp\|} \leq \left(a + bC'\right)Kt \triangleq \bar{H}Kt. \tag{68}$$

Let us now obtain a first crude bound on $\|W(t)\|$. To this end, introduce

$$u(t) = w'(t)/\sqrt{K}, \qquad y(t) = \|\dot{s}\|\|\dot{s}_\perp\|/\sqrt{K}, \qquad z(t) = \begin{bmatrix} u(t) \\ w(t) \\ y(t) \end{bmatrix}.$$

Then,

$$z'(t) = A(t)z(t), \qquad \text{with} \qquad A(t) = \begin{bmatrix} 0 & -M(t)/\sqrt{K} & h(t)/(\|\dot{s}\|\|\dot{s}_\perp\|) \\ \sqrt{K}I & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Let $g(t) = \|z(t)\|^2$. Then, $g(0) = \|\dot{s}\|^2\|\dot{s}_\perp\|^2/K$ and

$$g'(t) = 2\left\langle z(t), z'(t)\right\rangle = 2\left\langle z(t), A(t)z(t)\right\rangle \leq 2\|A(t)\|\|z(t)\|^2 = 2\|A(t)\|g(t).$$

Grönwall's inequality states that

$$g(t) \leq g(0)\exp\left(2\int_0^t \|A(\tau)\|\mathrm{d}\tau\right).$$

By triangle inequality and using $\|M(t)\| \leq K$, we have $\|A(t)\| \leq 2\sqrt{K} + \|h(t)\|/(\|\dot{s}\|\|\dot{s}_\perp\|)$. Thus, $\|z(t)\|^2$ can be bounded above and below:

$$\|w(t)\|^2 + \frac{\|\dot{s}\|^2\|\dot{s}_\perp\|^2}{K} \leq \|z(t)\|^2 \leq \frac{\|\dot{s}\|^2\|\dot{s}_\perp\|^2}{K} \times \exp\left(4\sqrt{K}t + 2\int_0^t \|h(\tau)\|/(\|\dot{s}\|\|\dot{s}_\perp\|)\mathrm{d}\tau\right). \tag{69}$$

Using our bound on $H(t)$ (68), we find

$$\exp\left(4\sqrt{K}t + 2\int_0^t \|h(\tau)\|/(\|\dot{s}\|\|\dot{s}_\perp\|)\mathrm{d}\tau\right) \leq \exp\left(4\sqrt{K}t + \bar{H}Kt^2\right).$$

Using $\sqrt{K}t \leq C$ again we deduce this crude bound:

$$\frac{\|W(t)\|}{\|\dot{s}\|\|\dot{s}_\perp\|} \leq \frac{1}{\sqrt{K}}\sqrt{\exp\left(4C + \bar{H}C^2\right) - 1} \triangleq \frac{1}{\sqrt{K}}\bar{W}. \tag{70}$$

We now return to (61) and plug in our bounds for H (68) and W (70) to get an improved bound on W: assuming t satisfies the stated conditions,

$$\frac{\|W(t)\|}{\|\dot{s}\|\|\dot{s}_\perp\|} \leq \int_0^t\int_0^\tau \bar{W}\sqrt{K} + \bar{H}K\theta\,\mathrm{d}\theta\mathrm{d}\tau = \frac{1}{2}\bar{W}\sqrt{K}t^2 + \frac{1}{6}\bar{H}Kt^3.$$

Plug this new and improved bound on W in (61) once again to get:

$$\begin{aligned}\frac{\|W(t)\|}{\|\dot{s}\|\|\dot{s}_\perp\|} &\leq \int_0^t\int_0^\tau K\left(\frac{1}{2}\bar{W}\sqrt{K}\theta^2 + \frac{1}{6}\bar{H}K\theta^3\right) + \bar{H}K\theta\,\mathrm{d}\theta\mathrm{d}\tau \\ &= \frac{1}{24}\bar{W}K^{3/2}t^4 + \frac{1}{120}\bar{H}K^2t^5 + \frac{1}{6}\bar{H}Kt^3\end{aligned}$$

$$= \left(\frac{1}{6}\bar{H} + \frac{1}{24}\bar{W}\sqrt{K}t + \frac{1}{120}\bar{H}Kt^2 \right) Kt^3.$$

We could now bound $\sqrt{K}t$ and Kt^2 by C and C^2, respectively, and stop here. However, this yields a constant which depends on $\bar{W}$: this can be quite large. Instead, we plug our new bound in (61) again, repeatedly. Doing so infinitely many times, we obtain a sequence of upper bounds, all of them valid. The limit of these bounds exists, and is hence also a valid bound. It is tedious but not difficult to check that this reasoning leads to the following:

$$\frac{\|W(t)\|}{\|\dot{s}\|\|\dot{s}_\perp\|} \leq \frac{1}{6}\bar{H}\left(1 + \frac{C^2}{6\cdot 7} + \frac{C^4}{6\cdot 7\cdot 8\cdot 9} + \frac{C^6}{6\cdots 11} + \cdots\right) Kt^3.$$

It is clear that the series converges. Let z be the value it converges to; then:

$$\begin{aligned} z &= 1 + \frac{C^2}{6\cdot 7} + \frac{C^4}{6\cdot 7\cdot 8\cdot 9} + \frac{C^6}{6\cdots 11} + \cdots \\ &= 1 + \frac{C^2}{6\cdot 7}\left(1 + \frac{C^2}{8\cdot 9} + \frac{C^4}{8\cdots 11} + \cdots\right) \leq 1 + \frac{C^2}{42}z. \end{aligned}$$

Thus, $z \leq \frac{1}{1-\frac{C^2}{42}}$. All in all, we conclude that

$$\frac{\|W(t)\|}{\|\dot{s}\|\|\dot{s}_\perp\|} \leq \bar{\bar{W}}Kt^3 \quad \text{with} \quad \bar{\bar{W}} = \frac{1}{6}\bar{H}\frac{1}{1-\frac{C^2}{42}} \text{ and}$$

$$\bar{H} = 8\frac{\sinh(C)\cosh(C)}{C} + 2\frac{\sinh(C)^2}{C^2}C'.$$

For example, with $C, C' \leq \frac{1}{4}$, we have $\bar{\bar{W}} \leq \frac{3}{2}$.

From Proposition B.1, we know that for the curve

$$c_{ts,\dot{s}}(q) = \mathrm{Exp}_x(ts + q\dot{s})$$

(recall that s has unit norm) it holds that $W(t) = t^2 c''_{ts,\dot{s}}(0)$. Thus,

$$\|c''_{ts,\dot{s}}(0)\| \leq \bar{\bar{W}}K\|\dot{s}\|\|\dot{s}_\perp\|t.$$

Allowing s to have norm different from one and rescaling t, we conclude that for the curve

$$c(t) = \mathrm{Exp}_x(s + t\dot{s})$$

we have

$$\|c''(0)\| \leq \bar{\bar{W}}K\|s\|\|\dot{s}\|\|\dot{s}_\perp\|,$$

provided $\|s\| \leq C\frac{1}{\sqrt{K}}$ with $C \leq \pi$ and $\|s\| \leq C'\frac{K}{F}$ and $\gamma(t) = \mathrm{Exp}_x(ts)$ is defined $[0, 1]$. □

C Lemma About Parameter Relations

As a general comments: here and throughout, constants are not optimized at all. In part, this is so that there is leeway in the precise definition of parameters. For example, the step-size η does not need to be exactly equal to $1/4\ell$, but it is convenient to assume equality to simplify many tedious computations.

Lemma C.1 *With parameters and assumptions as laid out in Sect. 3, the following hold:*

1. $\kappa \geq 2 and \log_2(\theta^{-1}) \geq \frac{5}{2}$,
2. $\epsilon \leq \frac{1}{2}\ell b and 2\ell\mathscr{M} < \frac{1}{2}\ell b$,
3. $r \leq \frac{1}{64}\mathscr{L} and \mathscr{L} \leq s \leq \frac{1}{32}b$,
4. $\ell\mathscr{M}^2 \geq \frac{64\mathscr{E}}{\mathscr{T}} and \theta\ell\mathscr{M}^2 \geq \frac{4\mathscr{E}}{\mathscr{T}}$,
5. $\epsilon r + \frac{\ell}{2}r^2 \leq \frac{1}{4}\mathscr{E}$,
6. $\frac{\mathscr{L}^2}{16\sqrt{\kappa}\eta\mathscr{T}} = \mathscr{E}$,
7. $\frac{s^2}{2\eta} \geq 2\mathscr{E} and \frac{(\gamma-4\hat{\rho}s)s^2}{2} \geq 2\mathscr{E}$,
8. $\hat{\rho}(\mathscr{L} + \mathscr{M}) \leq \sqrt{\hat{\rho}\epsilon}$.

D Proofs from Sect. 4 About AGD in a Ball of a Tangent Space

We give a proof of the lemma which states that iterates generated by `TSS` remain in certain balls. Such a lemma is not necessary in the Euclidean case.

Proof of Lemma 4.1 Because of how `TSS` works, if it defines u_j for some j, then s_j must have already been defined. Moreover, if $\|s_{j+1}\| > b$, then the algorithm terminates before defining u_{j+1}. It follows that if $u_0, \ldots, u_q$ are defined then $\|s_0\|, \ldots, \|s_q\|$ are all at most b. Also, `TSS` ensures $\|u_0\|, \ldots, \|u_q\|$ are all at most $2b$ by construction.

Recall that $\theta = \frac{1}{4\sqrt{\kappa}}$. From Lemma C.1 we know $\kappa \geq 2$ so that $\theta \leq 1$. Moreover, $2\eta\gamma = \frac{1}{8\kappa} = \frac{1}{2\sqrt{\kappa}}\theta \leq \theta$. It follows that θ_j as presented in (19) is well defined in the interval $[\theta, 1]$. Indeed, either $\|s_j + (1-\theta)v_j\| \leq 2b$, in which case $\theta_j = \theta$; or the line segment connecting s_j to $s_j + (1-\theta)v_j$ intersects the boundary of the sphere of radius $2b$ at exactly one point. By definition, this happens at $s_j + (1-\theta_j)v_j$ with $1-\theta_j$ chosen in the interval $[0, 1-\theta]$, that is, $\theta_j \in [\theta, 1]$.

Now assume that $\|\mathrm{grad} f(x)\| \leq \frac{1}{2}\ell b$. Then, for all $0 \leq j \leq q$ we have

$$\begin{aligned} \|\eta\nabla\hat{f}_x(u_j)\| &\leq \eta\left(\|\nabla\hat{f}_x(u_j) - \nabla\hat{f}_x(0)\| + \|\nabla\hat{f}_x(0)\|\right) \\ &\leq \eta\left(\ell\|u_j\| + \frac{1}{2}\ell b\right) \leq \frac{5}{2}\eta\ell b = \frac{5}{8}b < b, \end{aligned}$$

where we used the fact that $\|u_j\| \leq 2b$ and that $\nabla\hat{f}_x$ is ℓ-Lipschitz continuous in the ball of radius $3b$ around the origin (by A2), the fact that $\mathrm{grad} f(x) = \nabla\hat{f}_x(0)$, and the fact that $\eta\ell = \frac{1}{4}$ by definition of η. Consequently, if s_{q+1} is defined, then

$$\|s_{q+1}\| = \|u_q - \eta\nabla\hat{f}_x(u_q)\| \le \|u_q\| + \|\eta\nabla\hat{f}_x(u_q)\| \le 3b.$$

If additionally it holds that $\|u_q\| = 2b$, then

$$\|s_{q+1}\| = \|u_q - \eta\nabla\hat{f}_x(u_q)\| \ge \|u_q\| - \|\eta\nabla\hat{f}_x(u_q)\| > b.$$

(Mind the strict inequality: this one will matter.) □

Lemma 4.1 applies under the assumptions of Lemmas 4.2, 4.3 and 4.4. This ensures all vectors u_j, s_j remain in $B_x(3b)$, and hence the strongest provisions of A2 apply: we use this often in the proofs below.

We give a proof of the lemma which states that the Hamiltonian is monotonically decreasing along iterations.

Proof of Lemma 4.2 This follows almost exactly [28,Lem. 9 and 20], with one modification to allow θ_j (19) to be larger than $1/2$: this is necessary in our setup because we need to cap u_j to the ball of radius $2b$, requiring values of θ_j which can be arbitrarily close to 1.

Since $\nabla\hat{f}_x$ is ℓ-Lipschitz continuous in $B_x(3b)$ and $u_j, s_{j+1} \in B_x(3b)$, standard calculus and the identity $s_{j+1} = u_j - \eta\nabla\hat{f}_x(u_j)$ show that

$$\begin{aligned}\hat{f}_x(s_{j+1}) &\le \hat{f}_x(u_j) + \langle s_{j+1} - u_j, \nabla\hat{f}_x(u_j)\rangle + \frac{\ell}{2}\|s_{j+1} - u_j\|^2 \\ &= \hat{f}_x(u_j) - \eta\left(1 - \frac{\ell\eta}{2}\right)\|\nabla\hat{f}_x(u_j)\|^2.\end{aligned}$$

Since $\ell\eta = \frac{1}{4} \le \frac{1}{2}$, it follows that

$$\hat{f}_x(s_{j+1}) \le \hat{f}_x(u_j) - \frac{3\eta}{4}\|\nabla\hat{f}_x(u_j)\|^2.$$

Turning to E_{j+1} as defined by (20) and with the identity $v_{j+1} = s_{j+1} - s_j$, we compute:

$$E_{j+1} = \hat{f}_x(s_{j+1}) + \frac{1}{2\eta}\|v_{j+1}\|^2 \le \hat{f}_x(u_j) - \frac{3\eta}{4}\|\nabla\hat{f}_x(u_j)\|^2 + \frac{1}{2\eta}\|s_{j+1} - s_j\|^2.$$

Notice that

$$\begin{aligned}\|s_{j+1} - s_j\|^2 &= \|u_j - \eta\nabla\hat{f}_x(u_j) - s_j\|^2 \\ &= \|u_j - s_j\|^2 - 2\eta\langle u_j - s_j, \nabla\hat{f}_x(u_j)\rangle + \eta^2\|\nabla\hat{f}_x(u_j)\|^2.\end{aligned}$$

Moreover, the fact that s_{j+1} is defined means that (NCC) does not trigger with (x, s_j, u_j); in other words:

$$\hat{f}_x(s_j) \geq \hat{f}_x(u_j) - \langle u_j - s_j, \nabla \hat{f}_x(u_j)\rangle - \frac{\gamma}{2}\left\|u_j - s_j\right\|^2.$$

Combining, we find that

$$\begin{aligned} E_{j+1} &\leq \hat{f}_x(u_j) - \frac{3\eta}{4}\|\nabla \hat{f}_x(u_j)\|^2 - \langle u_j - s_j, \nabla \hat{f}_x(u_j)\rangle \\ &\quad + \frac{1}{2\eta}\|u_j - s_j\|^2 + \frac{\eta}{2}\|\nabla \hat{f}_x(u_j)\|^2 \\ &\leq \hat{f}_x(s_j) + \left(\frac{\gamma}{2} + \frac{1}{2\eta}\right)\left\|u_j - s_j\right\|^2 - \frac{\eta}{4}\|\nabla \hat{f}_x(u_j)\|^2. \end{aligned}$$

Using the identities $u_j - s_j = (1-\theta_j)v_j$ and $E_j = \hat{f}_x(s_j) + \frac{1}{2\eta}\|v_j\|^2$, we can further write:

$$\begin{aligned} E_{j+1} &\leq \hat{f}_x(s_j) + \left(\frac{\gamma}{2} + \frac{1}{2\eta}\right)(1-\theta_j)^2\left\|v_j\right\|^2 - \frac{\eta}{4}\|\nabla \hat{f}_x(u_j)\|^2 \\ &= E_j + \left(\frac{\gamma(1-\theta_j)^2}{2} + \frac{(1-\theta_j)^2 - 1}{2\eta}\right)\|v_j\|^2 - \frac{\eta}{4}\|\nabla \hat{f}_x(u_j)\|^2 \\ &= E_j + \frac{1}{2\eta}\left(\eta\gamma(1-\theta_j)^2 + (1-\theta_j)^2 - 1\right)\|v_j\|^2 - \frac{\eta}{4}\|\nabla \hat{f}_x(u_j)\|^2. \end{aligned}$$

From Lemma 4.1 we know that $\eta\gamma \leq \frac{1}{2}\theta_j$ and that θ_j is in the interval $[0, 1]$. It is easy to check that the function $\theta_j \mapsto \frac{1}{2}\theta_j(1-\theta_j)^2 + (1-\theta_j)^2 - 1$ is upper-bounded by $-\theta_j$ over the interval $[0, 1]$.

Thus,

$$E_{j+1} \leq E_j - \frac{\theta_j}{2\eta}\|v_j\|^2 - \frac{\eta}{4}\|\nabla \hat{f}_x(u_j)\|^2 \leq E_j,$$

as announced.

In closing, note that if $\|v_j\| \geq \mathscr{M}$ then Lemma C.1 shows

$$E_j - E_{j+1} \geq \frac{\theta_j}{2\eta}\|v_j\|^2 \geq \frac{\theta}{2\eta}\mathscr{M}^2 = 2\theta\ell\mathscr{M}^2 \geq \frac{4\mathscr{E}}{\mathscr{T}},$$

which concludes the proof. □

We give a proof of the improve-or-localize lemma.

Proof of Lemma 4.3 This follows from [28,Cor. 11], with some modifications for variable θ_j and because we allow $\theta_j > \frac{1}{2}$. By triangular inequality then Cauchy–Schwarz, we have

$$\|s_q - s_{q'}\|^2 = \left\|\sum_{j=q'}^{q-1} s_{j+1} - s_j\right\|^2 \leq \left(\sum_{j=q'}^{q-1}\|s_{j+1} - s_j\|\right)^2$$

$$\leq (q - q') \sum_{j=q'}^{q-1} \|s_{j+1} - s_j\|^2.$$

Now use the inequality $\|a + b\|^2 \leq (1 + C)\|a\|^2 + \frac{1+C}{C}\|b\|^2$ (valid for all vectors a, b and reals $C > 0$) with $C = 2\sqrt{\kappa} - 1$ (positive owing to $\kappa \geq 1$ by Lemma C.1) to see that

$$\|s_{j+1} - s_j\|^2 = \|(s_{j+1} - u_j) + (u_j - s_j)\|^2 \leq 2\sqrt{\kappa}\|s_{j+1} - u_j\|^2 \\ + \frac{2\sqrt{\kappa}}{2\sqrt{\kappa} - 1}\|u_j - s_j\|^2.$$

By construction, we have $s_{j+1} = u_j - \eta\nabla \hat{f}_x(u_j)$ and $u_j = s_j + (1 - \theta_j)v_j$. Thus:

$$\|s_{j+1} - s_j\|^2 \leq 2\sqrt{\kappa}\eta^2\|\nabla \hat{f}_x(u_j)\|^2 + \frac{2\sqrt{\kappa}(1 - \theta_j)^2}{2\sqrt{\kappa} - 1}\|v_j\|^2 \\ = 16\sqrt{\kappa}\eta \left(\frac{\eta}{8}\|\nabla \hat{f}_x(u_j)\|^2 + \frac{1}{2\eta}\frac{(1 - \theta_j)^2}{4(2\sqrt{\kappa} - 1)}\|v_j\|^2\right).$$

We focus on the second term: recall from Lemma 4.1 that $\theta_j \in [\theta, 1]$ with $\theta = \frac{1}{4\sqrt{\kappa}}$, and notice that $(1-t)^2 \leq 4(2\sqrt{\kappa} - 1)t$ for all t in the interval defined by $\frac{1-\theta\pm\sqrt{1-2\theta}}{\theta}$.
This holds a fortiori for all t in $[\theta, 1]$ because $\theta \leq \frac{1}{4}$ owing to $\kappa \geq 1$.
It follows that

$$\|s_{j+1} - s_j\|^2 \leq 16\sqrt{\kappa}\eta \left(\frac{\eta}{8}\|\nabla \hat{f}_x(u_j)\|^2 + \frac{\theta_j}{2\eta}\|v_j\|^2\right).$$

Apply Lemma 4.2 to the parenthesized expression to deduce that

$$\|s_{j+1} - s_j\|^2 \leq 16\sqrt{\kappa}\eta \left(E_j - E_{j+1}\right).$$

Plug this into the first inequality of this proof to conclude with a telescoping sum. □

We give a proof of the lemma which states that, upon witnessing significant non-convexity, it is possible to exploit that observation to drive significant decrease in the cost function value.

Proof of Lemma 4.4 This follows almost exactly [28,Lem. 10 and 17]. We need a slight modification because the Hessian $\nabla^2 \hat{f}_x$ may not be Lipschitz continuous in all of $B_x(3b)$: our assumptions only guarantee a type of Lipschitz continuity with respect to the origin of $\mathrm{T}_x\mathcal{M}$. Interestingly, even if the last momentum step was capped (that is, if $\theta_j \neq \theta$)—something which does not happen in the Euclidean case—the result goes through.

First, consider the case $\|v_j\| \geq s$, where s is a parameter set in Sect. 3. Then, $\mathtt{NCE}(x, s_j, v_j) = s_j$. It follows from the definition of E_j (20) that

$$\hat{f}_x(\mathtt{NCE}(x, s_j, v_j)) = \hat{f}_x(s_j) = E_j - \frac{1}{2\eta}\|v_j\|^2 \leq E_j - \frac{s^2}{2\eta}.$$

Second, consider the case $\|v_j\| < s$. We know that $v_j \neq 0$ as otherwise $u_j = s_j + (1-\theta_j)v_j = s_j$: this would contradict the assumption that (NCC) triggers with (x, s_j, u_j). Expand $\hat{f}_x$ around u_j in a truncated Taylor series with Lagrange remainder to see that

$$\hat{f}_x(s_j) = \hat{f}_x(u_j) + \langle \nabla \hat{f}_x(u_j), s_j - u_j\rangle + \frac{1}{2}\langle \nabla^2 \hat{f}_x(\zeta_j)[s_j - u_j], s_j - u_j\rangle$$

with $\zeta_j = ts_j + (1-t)u_j$ for some $t \in [0, 1]$. Since (NCC) triggers with (x, s_j, u_j), we also know that

$$\hat{f}_x(s_j) < \hat{f}_x(u_j) + \langle \nabla \hat{f}_x(u_j), s_j - u_j\rangle - \frac{\gamma}{2}\left\|s_j - u_j\right\|^2.$$

The last two claims combined yield:

$$\langle \nabla^2 \hat{f}_x(\zeta_j)[s_j - u_j], s_j - u_j\rangle < -\gamma\left\|s_j - u_j\right\|^2. \tag{71}$$

Consider $\dot{v} = s\frac{v_j}{\|v_j\|}$ as defined in the call to NCE. Let $\tilde{v}$ be either $\dot{v}$ or $-\dot{v}$, chosen so that $\langle \nabla \hat{f}_x(s_j), \tilde{v}\rangle \leq 0$ (at least one of the two choices satisfies this condition). By construction, $\mathtt{NCE}(x, s_j, v_j)$ is the element of the triplet $\{s_j, s_j + \dot{v}, s_j - \dot{v}\}$ where $\hat{f}_x$ is minimized. Since $s_j + \tilde{v}$ belongs to this triplet, it follows through another truncated Taylor series with Lagrange remainder (this time around s_j) that

$$\begin{aligned}\hat{f}_x(\mathtt{NCE}(x, s_j, v_j)) \leq \hat{f}_x(s_j + \tilde{v}) &= \hat{f}_x(s_j) + \langle \nabla \hat{f}_x(s_j), \tilde{v}\rangle + \frac{1}{2}\langle \nabla^2 \hat{f}_x(\zeta_j')[\tilde{v}], \tilde{v}\rangle \\ &\leq \hat{f}_x(s_j) + \frac{1}{2}\langle \nabla^2 \hat{f}_x(\zeta_j')[\tilde{v}], \tilde{v}\rangle \end{aligned} \tag{72}$$

with $\zeta_j' = s_j + t'\tilde{v}$ for some $t' \in [0, 1]$. Since $\tilde{v}$ is parallel to v_j which itself is parallel to $s_j - u_j$ (by definition of u_j), we deduce from (71) that

$$\langle \nabla^2 \hat{f}_x(\zeta_j)[\tilde{v}], \tilde{v}\rangle < -\gamma\|\tilde{v}\|^2 = -\gamma s^2.$$

We aim to use this to work on (72), but notice that $\nabla^2 \hat{f}_x$ is evaluated at two possibly distinct points, namely, ζ_j and ζ_j': we need to use the Lipschitz properties of the Hessian to relate them. To this end, notice that ζ_j and ζ_j' both live in $B_x(3b)$. Indeed, $\|\tilde{v}\| = \|\dot{v}\| = s \leq b$ by Lemma C.1 and $\|s_j\| \leq b$, $\|u_j\| \leq 2b$ by Lemma 4.1. Thus, $\|\zeta_j\| \leq \|s_j\| + \|u_j\| \leq b + 2b = 3b$ and $\|\zeta_j'\| \leq \|s_j\| + \|\tilde{v}\| \leq b + b = 2b$.

In contrast to the proof in [28], we have no Lipschitz guarantee for $\nabla^2 \hat{f}_x$ along the line segment connecting ζ_j and ζ_j', but A2 still offers such guarantees along the line segments connecting the origin of $\mathrm{T}_x\mathcal{M}$ to each of ζ_j and ζ_j'. Thus, we can write:

$$\begin{aligned}
\langle \nabla^2 \hat{f}_x(\zeta_j')[\tilde{v}], \tilde{v}\rangle &= \langle \nabla^2 \hat{f}_x(\zeta_j)[\tilde{v}], \tilde{v}\rangle + \langle (\nabla^2 \hat{f}_x(\zeta_j') - \nabla^2 \hat{f}_x(0))[\tilde{v}], \tilde{v}\rangle \\
&\quad - \langle (\nabla^2 \hat{f}_x(\zeta_j) - \nabla^2 \hat{f}_x(0))[\tilde{v}], \tilde{v}\rangle \\
&\leq -\gamma s^2 + \Big(\|\nabla^2 \hat{f}_x(\zeta_j') - \nabla^2 \hat{f}_x(0)\| \\
&\quad + \|\nabla^2 \hat{f}_x(\zeta_j) - \nabla^2 \hat{f}_x(0)\| \Big) \|\tilde{v}\|^2 \\
&\leq \Big(-\gamma + \hat{\rho}(\|\zeta_j'\| + \|\zeta_j\|) \Big) s^2 \\
&\leq \big(-\gamma + 2\hat{\rho}(s + \|s_j\|) \big) s^2,
\end{aligned}$$

where on the last line we used $\zeta_j = ts_j + (1-t)u_j$, $u_j = s_j + (1-\theta_j)v_j$, $\theta_j \in [0, 1]$ and $\|v_j\| \leq s$ to claim that $\|\zeta_j\| = \|s_j + (1-t)(1-\theta_j)v_j\| \leq \|s_j\| + \|v_j\| \leq \|s_j\| + s$, and also (more directly) that $\|\zeta_j'\| \leq \|s_j\| + \|\tilde{v}\| = \|s_j\| + s$. Plugging our findings into (72), it follows that

$$\hat{f}_x(\mathtt{NCE}(x, s_j, v_j)) \leq \hat{f}_x(s_j) - \frac{1}{2}\big(\gamma - 2\hat{\rho}(s + \|s_j\|)\big) s^2. \tag{73}$$

Since $\hat{f}_x(s_j) \leq E_j$ by definition (20), the main part of the lemma's claim is now proved.

We now turn to the last part of the lemma's claim, for which we further assume $\|s_j\| \leq \mathscr{L}$. Recall from Lemma C.1 that $\mathscr{L} \leq s$. We deduce from the main claim that

$$\hat{f}_x(\mathtt{NCE}(x, s_j, v_j)) \leq E_j - \min\left(\frac{s^2}{2\eta}, \frac{(\gamma - 4\hat{\rho}s)s^2}{2} \right).$$

To conclude, use Lemma C.1 anew to bound the right-most term. □

E Supporting Lemmas

In this section, we state and prove three additional lemmas about accelerated gradient descent in balls of tangent spaces that are useful for proofs in subsequent sections. The statements apply more broadly than the setup of parameters and assumptions in Sect. 3, but of course it is under those provisions that the conclusions are useful to us.

Throughout this section, we use the following notation. For some $x \in \mathcal{M}$, let $\mathcal{H} = \nabla^2 \hat{f}_x(0)$. Given $s_0 \in \mathrm{T}_x\mathcal{M}$, set $v_0 = 0$ and define for $j = 0, 1, 2, \ldots$:

$$u_j = s_j + (1-\theta)v_j, \quad s_{j+1} = u_j - \eta \nabla \hat{f}_x(u_j) \quad \text{and} \quad v_{j+1} = s_{j+1} - s_j \tag{74}$$

with some arbitrary $\theta \in [0, 1]$ and $\eta > 0$. Also define $s_{-1} = s_0 - v_0$ for convenience and

$$\delta_k = \nabla \hat{f}_x(u_k) - \nabla \hat{f}_x(0) - \mathcal{H}u_k,$$
$$\delta'_k = \nabla \hat{f}_x(u_k) - \nabla \hat{f}_x(s_\tau) - \mathcal{H}(u_k - s_\tau), \tag{75}$$

where $\tau \geq 0$ is a fixed index. Notice that iterates generated by $\mathtt{TSS}(x, s_0)$ with parameters and assumptions as laid out in Sect. 3 conform to this notation so long as $\theta_j = \theta$. Owing to Lemma 4.1, the latter condition holds in particular if TSS runs all its iterations in full because if at any point $\theta_j \neq \theta$ then $\|s_{j+1}\| > b$ and TSS terminates early. This is the setting in which we call upon lemmas from this section.

The first lemma is a variation on [28,Lem. 18].

Lemma E.1 *With notation as above, for all $j \geq 0$ we can write*

$$\begin{pmatrix} s_{\tau+j} \\ s_{\tau+j-1} \end{pmatrix} = A^j \begin{pmatrix} s_\tau \\ s_{\tau-1} \end{pmatrix} - \eta \sum_{k=0}^{j-1} A^{j-1-k} \begin{pmatrix} \nabla \hat{f}_x(0) + \delta_{\tau+k} \\ 0 \end{pmatrix} \tag{76}$$

and

$$\begin{pmatrix} s_{\tau+j} - s_\tau \\ s_{\tau+j-1} - s_\tau \end{pmatrix} = A^j \begin{pmatrix} 0 \\ -v_\tau \end{pmatrix} - \eta \sum_{k=0}^{j-1} A^{j-1-k} \begin{pmatrix} \nabla \hat{f}_x(s_\tau) + \delta'_{\tau+k} \\ 0 \end{pmatrix} \tag{77}$$

where

$$A = \begin{pmatrix} (2-\theta)(I - \eta\mathcal{H}) & -(1-\theta)(I - \eta\mathcal{H}) \\ I & 0 \end{pmatrix}. \tag{78}$$

Proof By definition of $\delta_{\tau+j-1}$, we have $\nabla \hat{f}_x(u_{\tau+j-1}) = \nabla \hat{f}_x(0) + \mathcal{H}u_{\tau+j-1} + \delta_{\tau+j-1}$. Thus,

$$\begin{aligned} s_{\tau+j} &= u_{\tau+j-1} - \eta \nabla \hat{f}_x(u_{\tau+j-1}) \\ &= u_{\tau+j-1} - \eta \nabla \hat{f}_x(0) - \eta \mathcal{H} u_{\tau+j-1} - \eta \delta_{\tau+j-1} \\ &= (I - \eta\mathcal{H}) u_{\tau+j-1} - \eta(\nabla \hat{f}_x(0) + \delta_{\tau+j-1}). \end{aligned}$$

Use the definitions of u_k and v_k to verify that $u_k = (2-\theta)s_k - (1-\theta)s_{k-1}$ (we use this several times in subsequent proofs). Plug this in the previous identity to see that

$$s_{\tau+j} = (2-\theta)(I - \eta\mathcal{H})s_{\tau+j-1} - (1-\theta)(I - \eta\mathcal{H})s_{\tau+j-2} - \eta(\nabla \hat{f}_x(0) + \delta_{\tau+j-1}).$$

Equivalently in matrix form, then reasoning by induction, it follows that

$$\begin{aligned}\begin{pmatrix} s_{\tau+j} \\ s_{\tau+j-1} \end{pmatrix} &= \begin{pmatrix} (2-\theta)(I-\eta\mathcal{H}) & -(1-\theta)(I-\eta\mathcal{H}) \\ I & 0 \end{pmatrix} \begin{pmatrix} s_{\tau+j-1} \\ s_{\tau+j-2} \end{pmatrix} \\ &\quad - \eta \begin{pmatrix} \nabla \hat{f}_x(0) + \delta_{\tau+j-1} \\ 0 \end{pmatrix} \\ &= A^j \begin{pmatrix} s_\tau \\ s_{\tau-1} \end{pmatrix} - \eta \sum_{k=0}^{j-1} A^{j-1-k} \begin{pmatrix} \nabla \hat{f}_x(0) + \delta_{\tau+k} \\ 0 \end{pmatrix}.\end{aligned}$$

This verifies Eq. (76). To prove Eq. (77), observe that (76) together with

$$\delta_{\tau+k} = \delta'_{\tau+k} + \nabla \hat{f}_x(s_\tau) - \nabla \hat{f}_x(0) - \mathcal{H} s_\tau$$

and $s_{\tau-1} = s_\tau - v_\tau$ imply

$$\begin{aligned}\begin{pmatrix} s_{\tau+j} - s_\tau \\ s_{\tau+j-1} - s_\tau \end{pmatrix} &= A^j \begin{pmatrix} 0 \\ -v_\tau \end{pmatrix} - \eta \sum_{k=0}^{j-1} A^{j-1-k} \begin{pmatrix} \nabla \hat{f}_x(s_\tau) + \delta'_{\tau+k} \\ 0 \end{pmatrix} \\ &\quad + (A^j - I) \begin{pmatrix} s_\tau \\ s_\tau \end{pmatrix} + \sum_{k=0}^{j-1} A^{j-1-k} \begin{pmatrix} \eta \mathcal{H} s_\tau \\ 0 \end{pmatrix}.\end{aligned}$$

The last two terms cancel. Indeed, let $M \triangleq \sum_{k=0}^{j-1} A^{j-1-k} = A^0 + \cdots + A^{j-1}$. Notice that $M(A-I) = MA - M = A^j - I$. Thus,

$$\begin{aligned}&\sum_{k=0}^{j-1} A^{j-1-k} \begin{pmatrix} \eta \mathcal{H} s_\tau \\ 0 \end{pmatrix} + (A^j - I) \begin{pmatrix} s_\tau \\ s_\tau \end{pmatrix} \\ &= M \left[\begin{pmatrix} \eta\mathcal{H} & 0 \\ 0 & \eta\mathcal{H} \end{pmatrix} \begin{pmatrix} s_\tau \\ 0 \end{pmatrix} + (A - I) \begin{pmatrix} s_\tau \\ s_\tau \end{pmatrix} \right] \\ &= M \left[\begin{pmatrix} 0 & -\eta\mathcal{H} \\ 0 & \eta\mathcal{H} \end{pmatrix} \begin{pmatrix} s_\tau \\ 0 \end{pmatrix} + (A - I) \begin{pmatrix} -I & -I \\ -I & -I \end{pmatrix} \begin{pmatrix} s_\tau \\ 0 \end{pmatrix} + (A - I) \begin{pmatrix} s_\tau \\ s_\tau \end{pmatrix} \right] \\ &= M \left[0 \right] = 0.\end{aligned}$$

To reach the second-to-last line, verify that $(A - I)\begin{pmatrix} -I & -I \\ -I & -I \end{pmatrix} = \begin{pmatrix} \eta\mathcal{H} & \eta\mathcal{H} \\ 0 & 0 \end{pmatrix}$ using (78). The last line follows by direct calculation. □

The lemma below is a direct continuation from the lemma above. We use it only for the proof of Lemma G.1.

Lemma E.2 *Use notation from Lemma E.1. Given $s_0, s_0' \in \mathrm{T}_x\mathcal{M}$, define two sequences $\{s_j, u_j, v_j\}$ and $\{s_j', u_j', v_j'\}$ by the update equations (74). Let $w_j = s_j - s_j'$. Then,*

$$\begin{pmatrix} w_j \\ w_{j-1} \end{pmatrix} = A^j \begin{pmatrix} w_0 \\ w_{-1} \end{pmatrix} - \eta \sum_{k=0}^{j-1} A^{j-1-k} \begin{pmatrix} \delta_k'' \\ 0 \end{pmatrix}$$

where $\delta_k'' = \nabla \hat{f}_x(u_k) - \nabla \hat{f}_x(u_k') - \mathcal{H}(u_k - u_k')$.

Proof By Lemma E.1 with $\tau = 0$, both of these identities hold:

$$\begin{pmatrix} s_j \\ s_{j-1} \end{pmatrix} = A^j \begin{pmatrix} s_0 \\ s_{-1} \end{pmatrix} - \eta \sum_{k=0}^{j-1} A^{j-1-k} \begin{pmatrix} \nabla \hat{f}_x(u_k) - \mathcal{H}u_k \\ 0 \end{pmatrix},$$

$$\begin{pmatrix} s_j' \\ s_{j-1}' \end{pmatrix} = A^j \begin{pmatrix} s_0' \\ s_{-1}' \end{pmatrix} - \eta \sum_{k=0}^{j-1} A^{j-1-k} \begin{pmatrix} \nabla \hat{f}_x(u_k') - \mathcal{H}u_k' \\ 0 \end{pmatrix}.$$

Taking the difference of these two equations reveals that

$$\begin{pmatrix} w_j \\ w_{j-1} \end{pmatrix} = A^j \begin{pmatrix} w_0 \\ w_{-1} \end{pmatrix} - \eta \sum_{k=0}^{j-1} A^{j-1-k} \begin{pmatrix} \nabla \hat{f}_x(u_k) - \nabla \hat{f}_x(u_k') - \mathcal{H}(u_k - u_k') \\ 0 \end{pmatrix}.$$

Conclude with the definition of δ_k''. □

The next lemma corresponds to [28,Prop. 19]. The claim applies in particular to iterates generated by TSS with parameters and assumptions as laid out in Sect. 3 and $R \le b$, so long as $\theta_j = \theta$ and the s_j remain in the appropriate balls. There are a few changes related to indexing and to the fact that our Lipschitz assumptions are limited to balls.

Lemma E.3 *Use notation from Lemma E.1. Assume $\|\nabla^2 \hat{f}_x(s) - \nabla^2 \hat{f}_x(0)\| \le \hat{\rho}\|s\|$ for all $s \in B_x(3R)$ with some $R > 0$, $\hat{\rho} > 0$. Also assume $\|s_k\| \le R$ for all $k = q' - 1, \ldots, q$. Then for all $k = q', \ldots, q$ we have $\|\delta_k\| \le 5\hat{\rho}R^2$. Moreover, for all $k = q' + 1, \ldots, q$ we have*

$$\|\delta_k - \delta_{k-1}\| \le 6\hat{\rho}R\big(\|s_k - s_{k-1}\| + \|s_{k-1} - s_{k-2}\|\big).$$

Additionally, we can bound their sum as:

$$\sum_{k=q'+1}^{q} \|\delta_k - \delta_{k-1}\|^2 \le 144\hat{\rho}^2 R^2 \sum_{k=q'}^{q} \|s_k - s_{k-1}\|^2.$$

(Mind the different ranges of summation.)

Proof Recall that $u_k = (2-\theta)s_k - (1-\theta)s_{k-1}$. In particular,

$$\|u_k\| \leq |2-\theta|\|s_k\| + |1-\theta|\|s_{k-1}\| \leq 3R \qquad \text{for} \qquad k = q', \ldots, q.$$

We use this to establish each of the three inequalities.

First, by definition of $\mathcal{H} = \nabla^2 \hat{f}_x(0)$ and of δ_k, we know that

$$\delta_k = \nabla \hat{f}_x(u_k) - \nabla \hat{f}_x(0) - \mathcal{H}u_k = \int_0^1 \nabla^2 \hat{f}_x(\phi u_k)[u_k] - \nabla^2 \hat{f}_x(0)[u_k]\mathrm{d}\phi.$$

Owing to $\|u_k\| \leq 3R$, we can use the Lipschitz properties of $\nabla^2 \hat{f}_x$ to find

$$\|\delta_k\| \leq \int_0^1 \left\|\nabla^2 \hat{f}_x(\phi u_k) - \nabla^2 \hat{f}_x(0)\right\| \mathrm{d}\phi \,\|u_k\| \leq \frac{1}{2}\hat{\rho}\,\|u_k\|^2 \leq \frac{9}{2}\hat{\rho}R^2.$$

This shows the first inequality for $k = q', \ldots, q$.

For the second inequality, first verify that

$$\begin{aligned}\|\delta_k - \delta_{k-1}\| &= \left\|\nabla \hat{f}_x(u_k) - \nabla \hat{f}_x(u_{k-1}) - \nabla^2 \hat{f}_x(0)[u_k - u_{k-1}]\right\| \\ &= \left\|\left(\int_0^1 \nabla^2 \hat{f}_x((1-\phi)u_{k-1} + \phi u_k) - \nabla^2 \hat{f}_x(0)\mathrm{d}\phi\right)[u_k - u_{k-1}]\right\|.\end{aligned}$$

Note that the distance between $(1-\phi)u_{k-1} + \phi u_k$ and the origin is at most $\max\{\|u_k\|, \|u_{k-1}\|\}$ for all $\phi \in [0,1]$. Since for $k = q'+1, \ldots q$ we have both $\|u_k\| \leq 3R$ and $\|u_{k-1}\| \leq 3R$, it follows that $\|(1-\phi)u_{k-1} + \phi u_k\| \leq 3R$ for all $\phi \in [0,1]$. As a result, we can use the Lipschitz-like properties of $\nabla^2 \hat{f}_x$ and write:

$$\|\delta_k - \delta_{k-1}\| \leq 3\hat{\rho}R\,\|u_k - u_{k-1}\|.$$

Combine $u_k = (2-\theta)s_k - (1-\theta)s_{k-1}$ and $u_{k-1} = (2-\theta)s_{k-1} - (1-\theta)s_{k-2}$ to find $u_k - u_{k-1} = (2-\theta)(s_k - s_{k-1}) - (1-\theta)(s_{k-1} - s_{k-2})$. From there, it follows that

$$\begin{aligned}\|\delta_k - \delta_{k-1}\| &\leq 3\hat{\rho}R\,\|(2-\theta)(s_k - s_{k-1}) - (1-\theta)(s_{k-1} - s_{k-2})\| \\ &\leq 3\hat{\rho}R\,(2\|s_k - s_{k-1}\| + \|s_{k-1} - s_{k-2}\|) \\ &\leq 6\hat{\rho}R\,(\|s_k - s_{k-1}\| + \|s_{k-1} - s_{k-2}\|).\end{aligned}$$

This establishes the second inequality for $k = q'+1, \ldots q$.

The third inequality follows from the second one through squaring and a sum, notably using $(a+b)^2 \leq 2(a^2+b^2)$ for $a, b \geq 0$:

$$\sum_{k=q'+1}^{q} \|\delta_k - \delta_{k-1}\|^2 \leq 36\hat{\rho}^2R^2 \sum_{k=q'+1}^{q} (\|s_k - s_{k-1}\| + \|s_{k-1} - s_{k-2}\|)^2$$

$$\leq 72\hat{\rho}^2 R^2 \sum_{k=q'+1}^{q} \left(\|s_k - s_{k-1}\|^2 + \|s_{k-1} - s_{k-2}\|^2 \right)$$

$$= 72\hat{\rho}^2 R^2 \left(\sum_{k=q'+1}^{q} \|s_k - s_{k-1}\|^2 + \sum_{k=q'}^{q-1} \|s_k - s_{k-1}\|^2 \right).$$

To conclude, extend the ranges of both sums to $q', \ldots, q$. □

We close this supporting section with important remarks about the matrix A (78), still following [28]. Recall the notation $\mathcal{H} = \nabla^2 \hat{f}_x(0)$: this is an operator on $\mathrm{T}_x\mathcal{M}$, self-adjoint with respect to the Riemannian inner product on $\mathrm{T}_x\mathcal{M}$. Let $e_1, \ldots, e_d \in \mathrm{T}_x\mathcal{M}$ form an orthonormal basis of eigenvectors of $\mathcal{H}$ associated to ordered eigenvalues $\lambda_1 \leq \cdots \leq \lambda_d$. We think of A as a linear operator to and from $\mathrm{T}_x\mathcal{M} \times \mathrm{T}_x\mathcal{M}$. Conveniently, the eigenvectors of $\mathcal{H}$ reveal how to block-diagonalize A. Indeed, from

$$A \begin{pmatrix} e_m \\ 0 \end{pmatrix} = \begin{pmatrix} (2-\theta)(I - \eta\mathcal{H}) & -(1-\theta)(I - \eta\mathcal{H}) \\ I & 0 \end{pmatrix} \begin{pmatrix} e_m \\ 0 \end{pmatrix}$$
$$= \begin{pmatrix} (2-\theta)(1 - \eta\lambda_m)e_m \\ e_m \end{pmatrix}$$

and

$$A \begin{pmatrix} 0 \\ e_m \end{pmatrix} = \begin{pmatrix} (2-\theta)(I - \eta\mathcal{H}) & -(1-\theta)(I - \eta\mathcal{H}) \\ I & 0 \end{pmatrix} \begin{pmatrix} 0 \\ e_m \end{pmatrix}$$
$$= \begin{pmatrix} -(1-\theta)(1 - \eta\lambda_m)e_m \\ 0 \end{pmatrix}$$

it is a simple exercise to check that

$$J^* A J = \mathrm{diag}(A_1, \ldots, A_d) \quad \text{with } J = \begin{pmatrix} e_1 & 0 & e_2 & 0 & \cdots & e_d & 0 \\ 0 & e_1 & 0 & e_2 & \cdots & 0 & e_d \end{pmatrix} \quad \text{and}$$
$$A_m = \begin{pmatrix} (2-\theta)(1 - \eta\lambda_m) & -(1-\theta)(1 - \eta\lambda_m) \\ 1 & 0 \end{pmatrix}. \tag{79}$$

Here, J is a unitary operator from $\mathbb{R}^{2d}$ (equipped with the standard Euclidean metric) to $\mathrm{T}_x\mathcal{M} \times \mathrm{T}_x\mathcal{M}$, and J^* denotes its adjoint (which is also its inverse). In particular, it becomes straightforward to investigate powers of A:

$$A^k = \left(J \mathrm{diag}(A_1, \ldots, A_d)\, J^*\right)^k = J \mathrm{diag}\left(A_1^k, \ldots, A_d^k\right) J^*. \tag{80}$$

For m, m' in $\{1, \ldots, d\}$ we have the useful identities

$$\left\langle \begin{pmatrix} e_{m'} \\ 0 \end{pmatrix}, A^k \begin{pmatrix} e_m \\ 0 \end{pmatrix} \right\rangle = \begin{cases} (A_m^k)_{11} & \text{if } m = m', \\ 0 & \text{if } m \neq m', \end{cases} \tag{81}$$

where $(A_m^k)_{11}$ is the top-left entry of the 2×2 matrix $(A_m)^k$. Likewise,

$$\left\langle \begin{pmatrix} e_{m'} \\ 0 \end{pmatrix}, A^k \begin{pmatrix} 0 \\ e_m \end{pmatrix} \right\rangle = \begin{cases} (A_m^k)_{12} & \text{if } m = m', \\ 0 & \text{if } m \neq m'. \end{cases} \tag{82}$$

Additionally, one can also check that [28,Lem. 24]:

$$\left\langle \begin{pmatrix} 0 \\ e_{m'} \end{pmatrix}, A^k \begin{pmatrix} e_m \\ 0 \end{pmatrix} \right\rangle = \begin{cases} (A_m^{k-1})_{11} & \text{if } m = m', \\ 0 & \text{if } m \neq m'. \end{cases} \tag{83}$$

F Proofs from Section 5 About TAGD

F.1 Proof of Proposition 5.3

The next two lemmas support Proposition 5.3. Proofs are in Appendix F.2. They correspond to [28,Lem. 21 and 22]. Notice that it is in Lemma F.2 that the condition on χ originates, then finds its way into the conditions of Theorem 5.1 through Proposition 5.3. Ultimately, this causes the polylogarithmic factor in the complexity of Theorem 1.3.

Lemma F.1 *Fix parameters and assumptions as laid out in Sect. 3. Let $x \in \mathcal{M}$ satisfy $\|\mathrm{grad} f(x)\| \leq \frac{1}{2}\ell b$. Let $\mathcal{S}$ denote the linear subspace of $\mathrm{T}_x\mathcal{M}$ spanned by the eigenvectors of $\nabla^2 \hat{f}_x(0)$ associated to eigenvalues strictly larger than $\frac{\theta^2}{\eta(2-\theta)^2}$. Let $P_{\mathcal{S}}$ denote orthogonal projection to $\mathcal{S}$. Assume $\mathtt{TSS}(x)$ runs its course in full.*

If there exists $\tau \in \{\mathscr{T}/4, \ldots, \mathscr{T}/2\}$ such that

$$\|s_\tau\| \leq \mathscr{L}, \qquad \|\nabla \hat{f}_x(s_\tau) - P_{\mathcal{S}} \nabla \hat{f}_x(s_\tau)\| \geq \epsilon/6,$$
$$\|v_\tau\| \leq \mathscr{M}, \text{ and} \qquad \left\langle P_{\mathcal{S}} v_\tau, \nabla^2 \hat{f}_x(0)[P_{\mathcal{S}} v_\tau] \right\rangle \leq \sqrt{\hat{\rho}\epsilon}\mathscr{M}^2,$$

then $E_{\tau-1} - E_{\tau+\mathscr{T}/4} \geq \mathscr{E}$.

Lemma F.2 *Fix parameters and assumptions as laid out in Sect. 3, with*

$$\chi \geq \log_2(\theta^{-1}) \geq 1.$$

Let $x \in \mathcal{M}$ satisfy $\|\mathrm{grad} f(x)\| \leq 2\ell\mathscr{M}$. Let $\mathcal{S}$ denote the linear subspace of $\mathrm{T}_x\mathcal{M}$ spanned by the eigenvectors of $\nabla^2 \hat{f}_x(0)$ associated to eigenvalues strictly larger than $\frac{\theta^2}{\eta(2-\theta)^2}$. Let $P_{\mathcal{S}}$ denote orthogonal projection to $\mathcal{S}$. Assume $\mathtt{TSS}(x)$ runs its course in full.

If $E_0 - E_{\mathscr{T}/2} \leq \mathscr{E}$, then for each j in $\{\mathscr{T}/4, \ldots, \mathscr{T}/2\}$ we have

$$\|P_{\mathcal{S}} \nabla \hat{f}_x(s_j)\| \leq \epsilon/6 \qquad \text{and} \qquad \left\langle P_{\mathcal{S}} v_j, \nabla^2 \hat{f}_x(0)[P_{\mathcal{S}} v_j] \right\rangle \leq \sqrt{\hat{\rho}\epsilon}\mathscr{M}^2.$$

In Lemmas F.1 and F.2, if $\mathcal{S}$ is empty then $P_{\mathcal{S}}$ maps all vectors to the zero vector, and the statements still hold.

Proof of Proposition 5.3 By Lemma C.1, $\|\mathrm{grad} f(x)\| \leq 2\ell\mathscr{M} < \frac{1}{2}\ell b$. Thus, the strongest provisions of A2 apply at x, as do Lemmas 4.1, 4.2, 4.3 and 4.4. Let u_j, s_j, v_j for $j = 0, 1, \ldots$ be the vectors generated by the computation of $x_{\mathscr{T}} = \mathtt{TSS}(x)$. Note that $s_0 = v_0 = 0$. There are several cases to consider, based on how TSS terminates:

- (Case 2a) The negative curvature condition (NCC) triggers with (x, s_j, u_j). There are two cases to check. Either $\|s_j\| \leq \mathscr{L}$, in which case Lemma 4.2 tells us $E_j \leq E_0 = f(x)$ and Lemma 4.4 further tells us that

$$f(x_{\mathscr{T}}) = \hat{f}_x(\mathtt{NCE}(x, s_j, v_j)) \leq E_j - 2\mathscr{E} \leq f(x) - 2\mathscr{E}.$$

  Or $\|s_j\| > \mathscr{L}$, in which case Lemma 4.3 used with $q = j < \mathscr{T}$ and $s_0 = 0$ implies

$$E_j \leq f(x) - \frac{\mathscr{L}^2}{16\sqrt{\kappa}\eta\mathscr{T}} = f(x) - \mathscr{E}.$$

  (See Lemma C.1 for that last equality.) Owing to how NCE works, we always have $f(x_{\mathscr{T}}) = \hat{f}_x(\mathtt{NCE}(x, s_j, v_j)) \leq \hat{f}_x(s_j) \leq E_j$ (the last inequality is by definition of E_j (20)). Thus, we conclude that $f(x_{\mathscr{T}}) \leq f(x) - \mathscr{E}$.
- (Case 2b) The iterate s_{j+1} leaves the ball of radius b, that is, $\|s_{j+1}\| > b$. In this case, apply Lemma 4.3 with $q = j + 1 \leq \mathscr{T}$ and $s_0 = 0$ to claim

$$\begin{aligned} f(x_{\mathscr{T}}) = \hat{f}_x(s_{j+1}) \leq E_{j+1} &\leq f(x) - \frac{\|s_{j+1}\|^2}{16\sqrt{\kappa}\eta\mathscr{T}} \\ &\leq f(x) - \frac{\mathscr{L}^2}{16\sqrt{\kappa}\eta\mathscr{T}} = f(x) - \mathscr{E}. \end{aligned}$$

  (The first inequality is by definition of E_{j+1} (20); subsequently, we use $\|s_{j+1}\| > b > \mathscr{L}$ as in Lemma C.1.)
- (Case 2c) The iterate s_{j+1} satisfies $\|\nabla \hat{f}_x(s_{j+1})\| \leq \epsilon/2$. Recall the chain rule identity relating gradients of f and gradients of the pullback $\hat{f}_x = f \circ \mathrm{R}_x$ with $T_s = \mathrm{DR}_x(s)$:

$$\nabla \hat{f}_x(s) = T_s^* \mathrm{grad} f(\mathrm{R}_x(s)).$$

  In our situation, $x_{\mathscr{T}} = \mathrm{R}_x(s_{j+1})$ and $\|s_{j+1}\| \leq b$ (otherwise, Case 2b applies). Thus, A2 ensures $\sigma_{\min}(T_{s_{j+1}}) \geq \frac{1}{2}$ and we deduce that

$$\begin{aligned} \|\mathrm{grad} f(x_{\mathscr{T}})\| &= \|(T_{s_{j+1}}^*)^{-1} \nabla \hat{f}_x(s_{j+1})\| \\ &\leq \|(T_{s_{j+1}}^*)^{-1}\| \|\nabla \hat{f}_x(s_{j+1})\| \leq 2 \cdot \frac{\epsilon}{2} = \epsilon. \end{aligned}$$

- (Case 2d) None of the other events occur: $\mathtt{TSS}(x)$ runs its $\mathscr{T}$ iterations in full. In this case, we apply the logic in the proof of [28,Lem. 12], as follows. We consider two cases. In the first case, $E_0 - E_{\mathscr{T}/2} > \mathscr{E}$. Then, we apply Lemma 4.2 to claim that $E_0 - E_{\mathscr{T}} \geq E_0 - E_{\mathscr{T}/2} \geq \mathscr{E}$. Moreover, $E_0 = f(x)$ and $E_{\mathscr{T}} \geq \hat{f}_x(s_{\mathscr{T}}) = f(x_{\mathscr{T}})$. Thus, in this case, $f(x) - f(x_{\mathscr{T}}) \geq \mathscr{E}$. In the second case, $E_0 - E_{\mathscr{T}/2} \leq \mathscr{E}$. Then, Lemma F.2 applies and we learn the following: Let $\mathcal{S}$ denote the linear subspace of $\mathrm{T}_x\mathcal{M}$ spanned by the eigenvectors of $\nabla^2 \hat{f}_x(0)$ associated to eigenvalues strictly larger than $\frac{\theta^2}{\eta(2-\theta)^2}$. Let $P_{\mathcal{S}}$ denote orthogonal projection to $\mathcal{S}$. For each j in $\{\mathscr{T}/4, \ldots, \mathscr{T}/2\}$ we have

$$\|P_{\mathcal{S}} \nabla \hat{f}_x(s_j)\| \leq \epsilon/6 \qquad \text{and} \qquad \left\langle P_{\mathcal{S}} v_j, \nabla^2 \hat{f}_x(0)[P_{\mathcal{S}} v_j] \right\rangle \leq \sqrt{\hat{\rho}\epsilon}\mathscr{M}^2.$$

Let τ be the first index in the range $\{\mathscr{T}/4, \ldots, \mathscr{T}\}$ for which $\|v_\tau\| \leq \mathscr{M}$. Again, there are two possibilities. In the first case, $\tau > \mathscr{T}/2$. Then, $\|v_j\| > \mathscr{M}$ for all j in $\{\mathscr{T}/4, \ldots, \mathscr{T}/2\}$. The last part of Lemma 4.2 implies that, for each such j, $E_j - E_{j+1} \geq \frac{4\mathscr{E}}{\mathscr{T}}$. It follows that $E_{\mathscr{T}/4} - E_{\mathscr{T}/2} \geq \mathscr{E}$. Conclude this case with Lemma 4.2 which justifies these statements: $f(x) = E_0$, $f(x_{\mathscr{T}}) = \hat{f}_x(s_{\mathscr{T}}) \leq E_{\mathscr{T}}$, and:

$$f(x) - f(x_{\mathscr{T}}) \geq E_0 - E_{\mathscr{T}} \geq E_{\mathscr{T}/4} - E_{\mathscr{T}/2} \geq \mathscr{E}.$$

In the second case, $\tau \in \{\mathscr{T}/4, \ldots, \mathscr{T}/2\}$. We aim to apply Lemma F.1: there are a few preconditions to check. Here is what we already know:

$$\|v_\tau\| \leq \mathscr{M}, \qquad \left\langle P_{\mathcal{S}} v_\tau, \nabla^2 \hat{f}_x(0)[P_{\mathcal{S}} v_\tau] \right\rangle \leq \sqrt{\hat{\rho}\epsilon}\mathscr{M}^2, \qquad \text{and}$$
$$\|P_{\mathcal{S}} \nabla \hat{f}_x(s_\tau)\| \leq \epsilon/6.$$

Regarding the third one above: we know that $\|\nabla \hat{f}_x(s_\tau)\| > \epsilon/2$ because $\mathtt{TSS}(x)$ did not terminate with s_τ. We deduce that

$$\|\nabla \hat{f}_x(s_\tau) - P_{\mathcal{S}} \nabla \hat{f}_x(s_\tau)\| \geq \|\nabla \hat{f}_x(s_\tau)\| - \|P_{\mathcal{S}} \nabla \hat{f}_x(s_\tau)\| \geq \frac{\epsilon}{2} - \frac{\epsilon}{6} > \frac{\epsilon}{6}.$$

We now have a final pair of cases to check. Either $\|s_\tau\| \leq \mathscr{L}$, in which case Lemma F.1 applies: it follows that $E_{\tau-1} - E_{\tau+\mathscr{T}/4} \geq \mathscr{E}$, and by arguments similar as above we conclude that $f(x) - f(x_{\mathscr{T}}) \geq \mathscr{E}$. Or $\|s_\tau\| > \mathscr{L}$, in which case Lemma 4.3 implies (using $s_0 = 0$):

$$f(x_{\mathscr{T}}) \leq E_{\mathscr{T}} \leq E_\tau \leq f(x) - \frac{\mathscr{L}^2}{16\sqrt{\kappa}\eta\tau} \leq f(x) - \mathscr{E}.$$

(For the second and last inequalities, we use $\tau < \mathscr{T}$ and Lemmas 4.2 and C.1.)

This covers all possibilities. □

F.2 Proofs of Lemmas F.1 and F.2

We include fulls proofs for the analogues of [28,Lem. 21 and 22] because we need small but important changes for our setting (as is the case for the other similar results we prove in full), and because of (ultimately inconsequential) small issues with some arguments pertaining to the subspace $\mathcal{S}$ in the original proofs. (Specifically, the subspace $\mathcal{S}$ is defined with respect to the Hessian of the cost function at a specific reference point, which for notational convenience in Jin et al. [28] is denoted by 0; however, this same convention is used in several lemmas, on at least one occasion referring to distinct reference points; the authors easily proposed a fix, and we use a different fix below; to avoid ambiguities, we keep all iterate references explicit.) Up to those minor changes, the proofs of the next two lemmas are due to Jin et al.

As a general heads-up for this and the next section: we call upon several lemmas from [28] which are purely algebraic facts about the entries of powers of the 2×2 matrices A_m (79): they do not change at all for our context, hence we do not include their proofs. We only note that Lemma 33 in [28] may not hold for all $x \in \left(\frac{\theta^2}{(2-\theta)^2}, \frac{1}{4}\right]$ as stated (there are some issues surrounding their eq. (17)), but it is only used twice, both times with $x \in \left(\frac{2\theta^2}{(2-\theta)^2}, \frac{1}{4}\right]$: in that interval the lemma does hold.

Proof of Lemma F.1 For contradiction, assume $E_{\tau-1} - E_{\tau+\mathscr{T}/4} < \mathscr{E}$. Then, Lemma 4.2 implies that $E_{\tau-1} - E_{\tau+j} < \mathscr{E}$ for all $-1 \leq j \leq \mathscr{T}/4$. Over that range, Lemmas 4.3 and C.1 tell us that

$$\|s_{\tau+j} - s_\tau\|^2 \leq 16\sqrt{\kappa}\eta|j||E_\tau - E_{\tau+j}| < 4\sqrt{\kappa}\eta\mathscr{T}\mathscr{E} = \frac{1}{4}\mathscr{L}^2. \tag{84}$$

The remainder of the proof consists in showing that $\|s_{\tau+\mathscr{T}/4} - s_\tau\|$ is in fact larger than $\frac{1}{2}\mathscr{L}$.

Starting now, consider $j = \mathscr{T}/4$. From (77) in Lemma E.1, we know that

$$\begin{pmatrix} s_{\tau+j} - s_\tau \\ s_{\tau+j-1} - s_\tau \end{pmatrix} = A^j \begin{pmatrix} 0 \\ -v_\tau \end{pmatrix} - \eta \sum_{k=0}^{j-1} A^{j-1-k} \begin{pmatrix} \nabla \hat{f}_x(s_\tau) + \delta'_{\tau+k} \\ 0 \end{pmatrix}.$$

Let $e_1, \ldots, e_d$ form an orthonormal basis of eigenvectors for $\mathcal{H} = \nabla^2 \hat{f}_x(0)$ with eigenvalues $\lambda_1 \leq \cdots \leq \lambda_d$. Expand v_τ, $\nabla \hat{f}_x(s_\tau)$ and $\delta'_{\tau+k}$ in that basis as:

$$v_\tau = \sum_{m=1}^{d} v^{(m)} e_m, \qquad \nabla \hat{f}_x(s_\tau) = \sum_{m=1}^{d} g^{(m)} e_m, \qquad \delta'_{\tau+k} = \sum_{m=1}^{d} (\delta'_{\tau+k})^{(m)} e_m. \tag{85}$$

Then,

$$\begin{pmatrix} s_{\tau+j} - s_\tau \\ s_{\tau+j-1} - s_\tau \end{pmatrix} = \sum_{m=1}^{d} \left[-v^{(m)} A^j \begin{pmatrix} 0 \\ e_m \end{pmatrix} - \eta \sum_{k=0}^{j-1} (g^{(m)} + (\delta'_{\tau+k})^{(m)}) A^{j-1-k} \begin{pmatrix} e_m \\ 0 \end{pmatrix} \right].$$

Owing to (81) and (82) which reveal how A block-diagonalizes in the basis e, we can further write

$$\left\langle \begin{pmatrix} e_m \\ 0 \end{pmatrix}, \begin{pmatrix} s_{\tau+j} - s_\tau \\ s_{\tau+j-1} - s_\tau \end{pmatrix} \right\rangle = -v^{(m)}(A_m^j)_{12} - \eta \sum_{k=0}^{j-1} \left(g^{(m)} + (\delta'_{\tau+k})^{(m)} \right) (A_m^{j-1-k})_{11}.$$

This reveals the expansion coefficients of $s_{\tau+j} - s_\tau$ in the basis $e_1, \ldots, e_d$, which is enough to study the norm of $s_{\tau+j} - s_\tau$. Explicitly,

$$\|s_{\tau+j} - s_\tau\|^2 = \sum_{m=1}^{d} \left(v^{(m)} b_{m,j} - \eta \sum_{k=0}^{j-1} \left(g^{(m)} + (\delta'_{\tau+k})^{(m)} \right) a_{m,j-1-k} \right)^2, \tag{86}$$

where we introduce the notation

$$a_{m,t} = (A_m^t)_{11}, \qquad\qquad b_{m,t} = -(A_m^t)_{12}. \tag{87}$$

To proceed, we need control over the coefficients $a_{m,t}$ and $b_{m,t}$, as provided by [28,Lem. 30]. We explore this for m in the set

$$S^c = \left\{ m : \eta\lambda_m \leq \frac{\theta^2}{(2-\theta)^2} \right\},$$

that is, for the eigenvectors orthogonal to $\mathcal{S}$. Under our general assumptions it holds that $\|\nabla^2 \hat{f}_x(0)\| \leq \ell$, so that $|\lambda_m| \leq \ell$ for all m. This ensures $\eta\lambda_m \in [-1/4, \theta^2/(2-\theta)^2]$ for $m \in S^c$. Recall that A_m (79) is a 2×2 matrix which depends on θ and $\eta\lambda_m$. It is reasonably straightforward to diagonalize A_m (or rather, to put it in Jordan normal form), and from there to get an explicit expression for any entry of A_m^k. The quantity $\sum_{k=0}^{j-1} a_{m,k}$ is a sum of such entries over a range of powers: this can be controlled as one would a geometric series. In [28,Lem. 30], it is shown that, for $m \in S^c$, if $j \geq 1 + 2/\theta$ and $\theta \in (0, 1/4]$, then

$$\sum_{k=0}^{j-1} a_{m,k} \geq \frac{1}{c_4\theta^2} \qquad \text{and} \qquad \frac{b_{m,j}}{\sum_{k'=0}^{j-1} a_{m,k'}} \leq c_5^{1/2} \max\left(\theta, \sqrt{|\eta\lambda_m|}\right), \tag{88}$$

with some universal constants c_4, c_5. The lemma applies because $\theta \in (0, 1/4]$ by Lemma C.1 and also $j = \mathscr{T}/4 = \chi(c/48) \cdot 3/\theta \geq 3/\theta \geq 4\sqrt{\kappa} + 2/\theta \geq 1 + 2/\theta$, with $c \geq 48$.

Building on the latter comments, we can define the following scalars for $m \in S^c$:

$$p_{m,k,j} = \frac{a_{m,j-1-k}}{\sum_{k'=0}^{j-1} a_{m,k'}}, \qquad q_{m,j} = -\frac{b_{m,j}}{\eta \sum_{k'=0}^{j-1} a_{m,k'}},$$

$$\tilde{\delta}_j'^{(m)} = \sum_{k=0}^{j-1} p_{m,k,j} (\delta'_{\tau+k})^{(m)} \qquad \tilde{v}_j^{(m)} = q_{m,j} v^{(m)}.$$

In analogy with notation in (85), we also consider vectors $\tilde{\delta}'_j$ and $\tilde{v}_j$ with expansion coefficients as above. These definitions are crafted specifically so that (86) yields:

$$\|s_{\tau+j} - s_\tau\|^2 \geq \sum_{m \in S^c} \left(\eta \left(\sum_{k=0}^{j-1} a_{m,k} \right) \left(g^{(m)} + \tilde{\delta}_j'^{(m)} + \tilde{v}_j^{(m)} \right) \right)^2 .$$

We deduce from (88) that

$$\begin{aligned} \|s_{\tau+j} - s_\tau\| &\geq \frac{\eta}{c_4 \theta^2} \sqrt{\sum_{m \in S^c} \left(g^{(m)} + \tilde{\delta}_j'^{(m)} + \tilde{v}_j^{(m)} \right)^2} \\ &= \frac{\eta}{c_4 \theta^2} \left\| P_{\mathcal{S}^c} \left(\nabla \hat{f}_x(s_\tau) + \tilde{\delta}'_j + \tilde{v}_j \right) \right\| \\ &\geq \frac{\eta}{c_4 \theta^2} \left(\frac{\epsilon}{6} - \|P_{\mathcal{S}^c}(\tilde{\delta}'_j)\| - \|P_{\mathcal{S}^c}(\tilde{v}_j)\| \right), \end{aligned} \tag{89}$$

where $\mathcal{S}^c$ is the orthogonal complement of $\mathcal{S}$, that is, it is the subspace of $\mathrm{T}_x\mathcal{M}$ spanned by eigenvectors $\{e_m\}_{m \in S^c}$, and $P_{\mathcal{S}^c}$ is the orthogonal projector to $\mathcal{S}^c$. In the last line, we used a triangular inequality and the assumption that $\|P_{\mathcal{S}^c}(\nabla \hat{f}_x(s_\tau))\| \geq \epsilon/6$. Our goal now is to show that $\|P_{\mathcal{S}^c}(\tilde{\delta}'_j)\|$ and $\|P_{\mathcal{S}^c}(\tilde{v}_j)\|$ are suitably small.

Consider the following vector with notation as in (75):

$$\begin{aligned} \Delta =& \delta_{\tau+k} - \delta'_{\tau+k} = \nabla \hat{f}_x(s_\tau) - \nabla \hat{f}_x(0) - \nabla^2 \hat{f}_x(0)[s_\tau] \\ =& \left(\int_0^1 \nabla^2 \hat{f}_x(\phi s_\tau) - \nabla^2 \hat{f}_x(0) \mathrm{d}\phi \right) [s_\tau]. \end{aligned}$$

By the Lipschitz-like properties of $\nabla^2 \hat{f}_x$ and the assumption $\|s_\tau\| \leq \mathscr{L} < b$, we deduce that

$$\|\Delta\| \leq \frac{1}{2} \hat{\rho} \|s_\tau\|^2 \leq \frac{1}{2} \hat{\rho} \mathscr{L}^2.$$

Note that $\sum_{k=0}^{j-1} p_{m,k,j} = 1$. This and the fact that Δ is independent of k justify that:

$$\begin{aligned} \|P_{\mathcal{S}^c}(\tilde{\delta}'_j)\|^2 = \sum_{m \in S^c} \left(\tilde{\delta}_j'^{(m)} \right)^2 &= \sum_{m \in S^c} \left(\sum_{k=0}^{j-1} p_{m,k,j} (\delta'_{\tau+k})^{(m)} \right)^2 \\ &= \sum_{m \in S^c} \left(\sum_{k=0}^{j-1} p_{m,k,j} \left((\delta_{\tau+k})^{(m)} - \Delta^{(m)} \right) \right)^2 \\ &= \sum_{m \in S^c} \left(\sum_{k=0}^{j-1} p_{m,k,j} (\delta_{\tau+k})^{(m)} - \Delta^{(m)} \right)^2, \end{aligned}$$

where $\Delta^{(m)}$ denotes the expansion coefficients of Δ in the basis e. Define the vector $\tilde{\delta}_j$ (without "prime") with expansion coefficients $\tilde{\delta}_j^{(m)} = \sum_{k=0}^{j-1} p_{m,k,j} (\delta_{\tau+k})^{(m)}$. Then, by construction,

$$\|P_{\mathcal{S}^c}(\tilde{\delta}'_j)\| = \|P_{\mathcal{S}^c}(\tilde{\delta}_j - \Delta)\| \leq \|P_{\mathcal{S}^c}(\tilde{\delta}_j)\| + \|\Delta\| \leq \|P_{\mathcal{S}^c}(\tilde{\delta}_j)\| + \hat{\rho}\mathscr{L}^2.$$

Through a simple reasoning using [28,Lem. 24, 26] one can conclude that, under our setting, both eigenvalues of A_m (for $m \in S^c$) are positive, and as a result that the coefficients $a_{m,k}$ (hence also $p_{m,k,j}$) are positive.

Therefore,

$$\begin{aligned}\|P_{\mathcal{S}^c}(\tilde{\delta}_j)\|^2 &= \sum_{m \in S^c} \left(\sum_{k=0}^{j-1} p_{m,k,j} (\delta_{\tau+k})^{(m)} \right)^2 \\ &\leq \sum_{m \in S^c} \left(\sum_{k=0}^{j-1} p_{m,k,j} \left(|(\delta_\tau)^{(m)}| + |(\delta_{\tau+k})^{(m)} - (\delta_\tau)^{(m)}| \right) \right)^2 .\end{aligned}$$

Notice that for all $0 \leq k \leq j-1$ we have

$$\begin{aligned}|(\delta_{\tau+k})^{(m)} - (\delta_\tau)^{(m)}| &\leq \sum_{k'=1}^{k} |(\delta_{\tau+k'})^{(m)} - (\delta_{\tau+k'-1})^{(m)}| \\ &\leq \sum_{k'=1}^{j-1} |(\delta_{\tau+k'})^{(m)} - (\delta_{\tau+k'-1})^{(m)}|,\end{aligned}$$

and this right-hand side is independent of k. Thus, we can factor out $\sum_{k=0}^{j-1} p_{m,k,j} = 1$ in the expression above to get:

$$\|P_{\mathcal{S}^c}(\tilde{\delta}_j)\|^2 \leq \sum_{m \in S^c} \left(|(\delta_\tau)^{(m)}| + \sum_{k=1}^{j-1} |(\delta_{\tau+k})^{(m)} - (\delta_{\tau+k-1})^{(m)}| \right)^2 .$$

Use first $(a+b)^2 \leq 2a^2 + 2b^2$ then (another) Cauchy–Schwarz to deduce

$$\begin{aligned}\|P_{\mathcal{S}^c}(\tilde{\delta}_j)\|^2 &\leq 2 \sum_{m \in S^c} |(\delta_\tau)^{(m)}|^2 + 2(j-1) \sum_{m \in S^c} \sum_{k=1}^{j-1} |(\delta_{\tau+k})^{(m)} - (\delta_{\tau+k-1})^{(m)}|^2 \\ &\leq 2\|\delta_\tau\|^2 + 2j \sum_{k=1}^{j-1} \|\delta_{\tau+k} - \delta_{\tau+k-1}\|^2.\end{aligned}$$

To bound this further, we call upon Lemma E.3 with $R = \frac{3}{2}\mathscr{L} \leq \frac{1}{3}b$, $q' = \tau$ and $q = \tau + \frac{\mathscr{T}}{4} - 1$. To this end, we must first verify that $\|s_{\tau+k}\| \leq R$ for $k = -1, \ldots, \frac{\mathscr{T}}{4} - 1$.

This is indeed the case owing to (84) and the assumption $\|s_\tau\| \leq \mathscr{L}$:

$$\|s_{\tau+k}\| \leq \|s_{\tau+k} - s_\tau\| + \|s_\tau\| \leq \frac{1}{2}\mathscr{L} + \mathscr{L} = R \qquad \text{for} \qquad k = -1, \ldots, \mathscr{T}/4.$$

This confirms that we can use the conclusions of Lemma E.3, reaching:

$$\begin{aligned}
\|P_{\mathcal{S}^c}(\tilde{\delta}_j)\|^2 &\leq 50\hat{\rho}^2 R^4 + 288\hat{\rho}^2 R^2 \cdot j \sum_{k=0}^{j-1} \|s_{\tau+k} - s_{\tau+k-1}\|^2 \\
&= \frac{4050}{16}\hat{\rho}^2\mathscr{L}^4 + 648\hat{\rho}^2\mathscr{L}^2 \cdot j \sum_{k=\tau-1}^{\tau+j-2} \|s_{k+1} - s_k\|^2 \\
&\leq 256\hat{\rho}^2\mathscr{L}^4 + 648\hat{\rho}^2\mathscr{L}^2 \cdot 16\sqrt{\kappa}\eta j (E_{\tau-1} - E_{\tau+j-2}),
\end{aligned}$$

where the first and last lines follow from the definition of R and from Lemma 4.3, respectively. Recall that we assume $E_{\tau-1} - E_{\tau+\mathscr{T}/4} < \mathscr{E}$ for contradiction. Then, monotonic decrease of the Hamiltonian (Lemma 4.2) tells us that $E_{\tau-1} - E_{\tau+j-2} < \mathscr{E}$ for $0 \leq j \leq \mathscr{T}/4$. Combining with $16\sqrt{\kappa}\eta\mathscr{T}\mathscr{E} = \mathscr{L}^2$ (Lemma C.1), we find:

$$\|P_{\mathcal{S}^c}(\tilde{\delta}_j)\|^2 \leq 256\hat{\rho}^2\mathscr{L}^4 + 162\hat{\rho}^2\mathscr{L}^4 = 418\hat{\rho}^2\mathscr{L}^4.$$

Thus, $\|P_{\mathcal{S}^c}(\tilde{\delta}_j)\| \leq 21\hat{\rho}\mathscr{L}^2 = 84\epsilon\chi^{-4}c^{-6} \leq \epsilon/24$ with $c \geq 4$ and $\chi \geq 1$, for $0 \leq j \leq \mathscr{T}/4$.

Recall that we aim to make progress from bound (89). The bound $\|P_{\mathcal{S}^c}(\tilde{\delta}_j)\| \leq \epsilon/24$ we just established is a first step. We now turn to bounding $\|P_{\mathcal{S}^c}(\tilde{v}_j)\|$. Owing to (88), we have this first bound assuming $j = \mathscr{T}/4$:

$$\begin{aligned}
\|P_{\mathcal{S}^c}(\tilde{v}_j)\|^2 &= \sum_{m\in S^c} q_{m,j}^2 (v^{(m)})^2 \\
&= \sum_{m\in S^c} \left(\frac{b_{m,j}}{\eta\sum_{k'=0}^{j-1} a_{m,k'}}\right)^2 (v^{(m)})^2 \leq \frac{c_5}{\eta^2} \sum_{m\in S^c} (v^{(m)})^2 \max\Big(\theta^2, |\eta\lambda_m|\Big).
\end{aligned} \tag{90}$$

(Recall from (85) that $v^{(m)}$ denotes the coefficients of v_τ in the basis $e_1, \ldots, e_d$.) We split the sum in order to resolve the max. To this end, note that $\theta \in [0, 1]$ implies $\theta^2 \geq \frac{\theta^2}{(2-\theta)^2}$, so that the max evaluates to θ^2 exactly when $-\theta^2 \leq \eta\lambda_m \leq \frac{\theta^2}{(2-\theta)^2}$ (remembering that $\eta\lambda_m \leq \frac{\theta^2}{(2-\theta)^2}$ because $m \in S^c$). Thus,

$$\begin{aligned}
&\sum_{m\in S^c} (v^{(m)})^2 \max\Big(\theta^2, |\eta\lambda_m|\Big) \\
&\quad= \sum_{m:-\theta^2\leq\eta\lambda_m\leq\frac{\theta^2}{(2-\theta)^2}} (v^{(m)})^2\theta^2 - \sum_{m:\eta\lambda_m<-\theta^2} (v^{(m)})^2\eta\lambda_m.
\end{aligned}$$

Let us rework the last sum (we get a first bound by extending the summation range, exploiting that the summands are non-positive):

$$
\begin{aligned}
-\sum_{m:\eta\lambda_m<-\theta^2}(v^{(m)})^2\eta\lambda_m &\le -\sum_{m:\eta\lambda_m\le 0}(v^{(m)})^2\eta\lambda_m \\
&= \sum_{m:\eta\lambda_m>0}(v^{(m)})^2\eta\lambda_m - \sum_{m=1}^{d}(v^{(m)})^2\eta\lambda_m \\
&= \sum_{m:\eta\lambda_m>0}(v^{(m)})^2\eta\lambda_m - \eta\left\langle v_\tau, \mathcal{H}v_\tau\right\rangle \\
&= \sum_{m:0<\eta\lambda_m\le\frac{\theta^2}{(2-\theta)^2}}(v^{(m)})^2\eta\lambda_m \\
&\quad + \eta\left\langle P_{\mathcal{S}}v_\tau, \mathcal{H}P_{\mathcal{S}}v_\tau\right\rangle - \eta\left\langle v_\tau, \mathcal{H}v_\tau\right\rangle \\
&\le \theta^2\|v_\tau\|^2 + \eta\left\langle P_{\mathcal{S}}v_\tau, \mathcal{H}P_{\mathcal{S}}v_\tau\right\rangle - \eta\left\langle v_\tau, \mathcal{H}v_\tau\right\rangle.
\end{aligned}
$$

(Recall that $P_{\mathcal{S}}$ projects to the subspace spanned by eigenvectors with eigenvalues strictly above $\frac{\theta^2}{\eta(2-\theta)^2}$.) Combining all work done since (90), it follows that

$$
\|P_{\mathcal{S}^c}(\tilde{v}_j)\|^2 \le \frac{c_5}{\eta^2}\left(2\theta^2\|v_\tau\|^2 + \eta\left\langle P_{\mathcal{S}}v_\tau, \mathcal{H}P_{\mathcal{S}}v_\tau\right\rangle - \eta\left\langle v_\tau, \mathcal{H}v_\tau\right\rangle\right).
$$

Use assumptions $\|v_\tau\| \le \mathscr{M}$ and $\left\langle P_{\mathcal{S}}v_\tau, \mathcal{H}P_{\mathcal{S}}v_\tau\right\rangle \le \sqrt{\hat{\rho}\epsilon}\mathscr{M}^2$ to see that

$$
\begin{aligned}
\|P_{\mathcal{S}^c}(\tilde{v}_j)\|^2 &\le \frac{c_5}{\eta^2}\left(2\theta^2\mathscr{M}^2 + \eta\sqrt{\hat{\rho}\epsilon}\mathscr{M}^2 - \eta\left\langle v_\tau, \mathcal{H}v_\tau\right\rangle\right) \\
&= 4\ell c_5\left(\frac{3}{2}\sqrt{\hat{\rho}\epsilon}\mathscr{M}^2 - \left\langle v_\tau, \mathcal{H}v_\tau\right\rangle\right).
\end{aligned} \tag{91}
$$

(For the last equality, use $2\theta^2 = \frac{\sqrt{\hat{\rho}\epsilon}}{2}\eta$ and $\eta = 1/4\ell$.) To proceed, we must bound $\left\langle v_\tau, \mathcal{H}v_\tau\right\rangle$. To this end, notice that by assumption the (NCC) condition did not trigger for (x, s_τ, u_τ). Therefore, we know that

$$
\hat{f}_x(s_\tau) \ge \hat{f}_x(u_\tau) + \left\langle\nabla\hat{f}_x(u_\tau), s_\tau - u_\tau\right\rangle - \frac{\gamma}{2}\|s_\tau - u_\tau\|^2.
$$

Moreover, it always holds that

$$
\begin{aligned}
\hat{f}_x(s_\tau) = \hat{f}_x(u_\tau) &+ \left\langle\nabla\hat{f}_x(u_\tau), s_\tau - u_\tau\right\rangle \\
&+ \frac{1}{2}\left\langle s_\tau - u_\tau, \nabla^2\hat{f}_x(\phi s_\tau + (1-\phi)u_\tau)[s_\tau - u_\tau]\right\rangle
\end{aligned}
$$

for some $\phi \in [0, 1]$. Also using $u_\tau = s_\tau + (1-\theta)v_\tau$, we deduce that

$$\langle v_\tau, \nabla^2 \hat{f}_x(\phi s_\tau + (1-\phi)u_\tau)[v_\tau]\rangle \geq -\gamma \|v_\tau\|^2.$$

With the help of Lemma C.1, note that

$$\|\phi s_\tau + (1-\phi)u_\tau\| = \|s_\tau + (1-\phi)(1-\theta)v_\tau\| \leq \|s_\tau\| + \|v_\tau\| \leq \mathscr{L} + \mathscr{M} \leq b.$$

Thus, the Lipschitz-type properties of $\nabla^2 \hat{f}_x$ apply up to that point and we get

$$\|\nabla^2 \hat{f}_x(\phi s_\tau + (1-\phi)u_\tau) - \mathcal{H}\| \leq \hat{\rho}(\mathscr{L} + \mathscr{M}) \leq \sqrt{\hat{\rho}\epsilon}.$$

Since $\gamma = \frac{\sqrt{\hat{\rho}\epsilon}}{4}$, it follows overall that

$$\langle v_\tau, \mathcal{H}v_\tau\rangle \geq -\frac{5}{4}\sqrt{\hat{\rho}\epsilon}\|v_\tau\|^2 \geq -\frac{5}{4}\sqrt{\hat{\rho}\epsilon}\mathscr{M}^2.$$

Plugging this back into (91) with $c \geq 80\sqrt{c_5}$ reveals that

$$\|P_{\mathcal{S}^c}(\tilde{v}_j)\|^2 \leq 11\ell c_5\sqrt{\hat{\rho}\epsilon}\mathscr{M}^2 = 11c_5\epsilon^2 c^{-2} \leq \epsilon^2/24^2.$$

This shows that $\|P_{\mathcal{S}^c}(\tilde{v}_j)\| \leq \epsilon/24$ for $j = \mathscr{T}/4$.

We plug $\|P_{\mathcal{S}^c}(\tilde{\delta}_j)\| \leq \epsilon/24$ and $\|P_{\mathcal{S}^c}(\tilde{v}_j)\| \leq \epsilon/24$ into (89) to state that, with $j = \mathscr{T}/4$,

$$\begin{aligned}\|s_{\tau+j} - s_\tau\| &\geq \frac{\eta}{c_4\theta^2}\left(\frac{\epsilon}{6} - \frac{\epsilon}{24} - \frac{\epsilon}{24}\right) = \frac{\eta\epsilon}{12c_4\theta^2} \\ &= \frac{1}{3c_4}\sqrt{\frac{\epsilon}{\hat{\rho}}} > \sqrt{\frac{\epsilon}{\hat{\rho}}}\chi^{-2}c^{-3} = \mathscr{L}/2.\end{aligned}$$

(We used $4\theta^2 = \sqrt{\hat{\rho}\epsilon}\eta$, then we also set $c > (3c_4)^{1/3}$.) This last inequality contradicts (84). Thus, the proof by contradiction is complete and we conclude that $E_{\tau-1} - E_{\tau+\mathscr{T}/4} \geq \mathscr{E}$. □

What follows is the equivalent of the proof of [28,Lem. 22], with the small changes needed for our purpose.

Proof of Lemma F.2 Since $E_0 - E_{\mathscr{T}/2} \leq \mathscr{E}$ and $s_0 = 0$, Lemmas 4.2, 4.3 and C.1 yield:

$$\forall j \leq \mathscr{T}/2, \qquad \|s_j\| = \|s_j - s_0\| \leq \sqrt{8\sqrt{\kappa}\eta\mathscr{T}\mathscr{E}} = \frac{\mathscr{L}}{\sqrt{2}} \leq \mathscr{L} \leq b. \tag{92}$$

By Lemma E.1 with $\tau = 0$ and noting that $s_0 = 0$, $s_{-1} = s_0 - v_0 = 0$, we know that, for all j,

$$\begin{pmatrix} s_j \\ s_{j-1} \end{pmatrix} = -\eta \sum_{k=0}^{j-1} A^{j-1-k} \begin{pmatrix} \nabla \hat{f}_x(0) + \delta_k \\ 0 \end{pmatrix}. \tag{93}$$

Define the operator $\Delta_j = \int_0^1 \nabla^2 \hat{f}_x(\phi s_j) - \mathcal{H} \mathrm{d}\phi$ with $\mathcal{H} = \nabla^2 \hat{f}_x(0)$. We can write:

$$P_{\mathcal{S}} \nabla \hat{f}_x(s_j) = P_{\mathcal{S}} \Big(\nabla \hat{f}_x(0) + \mathcal{H} s_j + \Delta_j s_j \Big). \tag{94}$$

We shall bound this term by term.

The third term is straightforward, so let us start with this one. Owing to (92), the Lipschitz-like properties of the Hessian apply to claim $\|\Delta_j\| \leq \frac{1}{2}\hat{\rho}\|s_j\|$. Therefore,

$$\|P_{\mathcal{S}} \Delta_j s_j\| \leq \|\Delta_j\| \|s_j\| \leq \frac{1}{2}\hat{\rho}\|s_j\|^2 \leq \frac{1}{2}\hat{\rho}\mathscr{L}^2 = 2\epsilon\chi^{-4}c^{-6} \leq \epsilon/18 \tag{95}$$

with $c \geq 2$ and $\chi \geq 1$. Below, we work toward bounding the other two terms.

As we did in the proof of Lemma F.1, let $e_1, \ldots, e_d$ form an orthonormal basis of eigenvectors for $\mathcal{H}$ with eigenvalues $\lambda_1 \leq \cdots \leq \lambda_d$. Expand $\nabla \hat{f}_x(0)$ and δ_k in that basis as

$$\nabla \hat{f}_x(0) = \sum_{m=1}^{d} g^{(m)} e_m, \qquad \delta_k = \sum_{m=1}^{d} \delta_k^{(m)} e_m.$$

From (93) and (81) it follows that

$$\begin{aligned} s_j &= \sum_{m'=1}^{d} \langle e_{m'}, s_j \rangle e_{m'} = -\eta \sum_{m'=1}^{d} \sum_{k=0}^{j-1} \sum_{m=1}^{d} \left\langle \begin{pmatrix} e_{m'} \\ 0 \end{pmatrix}, A^{j-1-k} \begin{pmatrix} e_m \\ 0 \end{pmatrix} \right\rangle (g^{(m)} + \delta_k^{(m)}) e_{m'} \\ &= -\eta \sum_{k=0}^{j-1} \sum_{m=1}^{d} (A_m^{j-1-k})_{11} (g^{(m)} + \delta_k^{(m)}) e_m. \end{aligned}$$

Motivated by (94) and reusing notation $a_{m,j-1-k} = (A_m^{j-1-k})_{11}$ as in (87), we further write

$$\begin{aligned} P_{\mathcal{S}} \Big(\nabla \hat{f}_x(0) + \mathcal{H} s_j \Big) &= \sum_{m \in S} \left[g^{(m)} - \eta \lambda_m \sum_{k=0}^{j-1} a_{m,j-1-k} (g^{(m)} + \delta_k^{(m)}) \right] e_m \\ &= \sum_{m \in S} \left[\left(1 - \eta \lambda_m \sum_{k=0}^{j-1} a_{m,k} \right) g^{(m)} \right. \end{aligned}$$

$$-\eta\lambda_m \sum_{k=0}^{j-1} a_{m,j-1-k}\delta_k^{(m)} \Bigg] e_m, \tag{96}$$

where $S = \left\{m : \eta\lambda_m > \frac{\theta^2}{(2-\theta)^2}\right\}$ indexes the eigenvalues of the eigenvectors which span $\mathcal{S}$. This identity splits in two parts, each of which we now aim to bound.

In the spirit of the comments surrounding (88), here too it is possible to control the coefficients $a_{m,k}$ and $b_{m,k}$ (both defined as in (87)), this time for $m \in S$. Specifically, combining [28,Lem. 25] with an identity in the proof of [28,Lem. 29], we see that

$$1 - \eta\lambda_m \sum_{k=0}^{j-1} a_{m,k} = a_{m,j} - b_{m,j}. \tag{97}$$

Moreover, owing to [28,Lem. 32] we know that

$$\forall j \geq 0, \forall m \in S, \qquad \max(|a_{m,j}|, |b_{m,j}|) \leq (j+1)(1-\theta)^{j/2}. \tag{98}$$

Thus, the first part of (96) is bounded as:

$$\left\| \sum_{m\in S} \left(1 - \eta\lambda_m \sum_{k=0}^{j-1} a_{m,k}\right) g^{(m)} e_m \right\|^2 = \sum_{m\in S} (a_{m,j} - b_{m,j})^2 (g^{(m)})^2 \leq 4(j+1)^2(1-\theta)^j \|\nabla \hat{f}_x(0)\|^2.$$

One can show using $\theta \in (0, 1/4]$, $\chi \geq \log_2(\theta^{-1})$ and $c \geq 256$ (which we all assume) that

$$\forall j \geq \mathscr{T}/4, \qquad (j+1)^2 \leq (1-\theta)^{-j/2}. \tag{99}$$

Then use the assumption $\|\nabla \hat{f}_x(0)\| \leq 2\ell\mathscr{M}$ and $j \geq \mathscr{T}/4$ again to replace the power with $j/2 \geq \sqrt{\kappa}\chi c/8 \geq 4\sqrt{\kappa} \cdot 2\chi$ (with $c \geq 64$) and see that

$$\left\| \sum_{m\in S} \left(1 - \eta\lambda_m \sum_{k=0}^{j-1} a_{m,k}\right) g^{(m)} e_m \right\|^2 \leq 16\ell^2\mathscr{M}^2(1-\theta)^{j/2}$$
$$\leq 16\epsilon^2\kappa c^{-2}\left(1 - \frac{1}{4\sqrt{\kappa}}\right)^{4\sqrt{\kappa}\cdot 2\chi}.$$

Use the fact that $0 < (1-t^{-1})^t < e^{-1} \leq 2^{-1}$ for $t \geq 4$ together with $\kappa \geq 1$ to bound the right-hand side by $16\epsilon^2\kappa c^{-2}2^{-2\chi}$. This itself is bounded by $16\epsilon^2\kappa c^{-2}\theta^2 = \epsilon^2 c^{-2}$

using $\chi \geq \log_2(\theta^{-1})$. Overall, we have shown that

$$\left\| \sum_{m \in S} \left(1 - \eta\lambda_m \sum_{k=0}^{j-1} a_{m,k}\right) g^{(m)} e_m \right\| \leq \epsilon/18, \tag{100}$$

with $c \geq 18$. This covers the first term in (96).

We turn to bounding the second term in (96). For this one, we need [28,Lem. 34] which states that, for $m \in S$ and $j \geq \mathscr{T}/4$, for any sequence $\{\epsilon_k\}$, we have

$$\sum_{k=0}^{j-1} a_{m,k}\epsilon_k \leq \frac{\sqrt{c_2}}{\eta\lambda_m} \left(|\epsilon_0| + \sum_{k=1}^{j-1} |\epsilon_k - \epsilon_{k-1}|\right), \text{ and} \tag{101}$$

$$\sum_{k=0}^{j-1} (a_{m,k} - a_{m,k-1})\epsilon_k \leq \frac{\sqrt{c_3}}{\sqrt{\eta\lambda_m}} \left(|\epsilon_0| + \sum_{k=1}^{j-1} |\epsilon_k - \epsilon_{k-1}|\right), \tag{102}$$

with some positive constants c_1, c_2, c_3 and $c \geq c_1$. Thus, to bound the remaining term in (96) we start with:

$$\begin{aligned}
\left\| \sum_{m \in S} \eta\lambda_m \sum_{k=0}^{j-1} a_{m,j-1-k} \delta_k^{(m)} e_m \right\|^2 &\leq c_2 \sum_{m \in S} \left(|\delta_{j-1}^{(m)}| + \sum_{k=1}^{j-1} |\delta_k^{(m)} - \delta_{k-1}^{(m)}|\right)^2 \\
&\leq 2c_2 \sum_{m \in S} \left[|\delta_{j-1}^{(m)}|^2 + \left(\sum_{k=1}^{j-1} |\delta_k^{(m)} - \delta_{k-1}^{(m)}|\right)^2 \right] \\
&\leq 2c_2 \sum_{m \in S} \left[|\delta_{j-1}^{(m)}|^2 + (j-1) \sum_{k=1}^{j-1} |\delta_k^{(m)} - \delta_{k-1}^{(m)}|^2 \right] \\
&\leq 2c_2 \|\delta_{j-1}\|^2 + 2c_2 j \sum_{k=1}^{j-1} \|\delta_k - \delta_{k-1}\|^2.
\end{aligned} \tag{103}$$

(We used $(a+b)^2 \leq 2a^2 + 2b^2$ again, and another Cauchy–Schwarz on the remaining sum.) In order to proceed, we call upon Lemma E.3 with $R = \mathscr{L}, q' = 0$ and $q = j-1$, which is justified by (92) (recall that $s_{-1} = 0$). This yields the first inequality in:

$$\begin{aligned}
\left\| \sum_{m \in S} \eta\lambda_m \sum_{k=0}^{j-1} a_{m,j-1-k} \delta_k^{(m)} e_m \right\|^2 &\leq 50c_2\hat{\rho}^2\mathscr{L}^4 + 2c_2 j \cdot 144\hat{\rho}^2\mathscr{L}^2 \sum_{k=0}^{j-1} \|s_k - s_{k-1}\|^2 \\
&\leq 50c_2\hat{\rho}^2\mathscr{L}^4 + 144c_2\hat{\rho}^2\mathscr{L}^4.
\end{aligned} \tag{104}$$

The second inequality above is supported by Lemmas 4.2, 4.3 and C.1 as well as $j \le \mathscr{T}/2$ and the assumption $E_0 - E_{\mathscr{T}/2} \le \mathscr{E}$, through:

$$j \sum_{k=0}^{j-1} \|s_k - s_{k-1}\|^2 \le 16\sqrt{\kappa}\eta j (E_0 - E_j) \le 8\sqrt{\kappa}\eta \mathscr{T}\mathscr{E} = \mathscr{L}^2/2. \tag{105}$$

Continuing from (104), we see that the right- (hence also left-) hand side is upper-bounded by

$$194c_2 \cdot \hat{\rho}^2 \mathscr{L}^4 = 194c_2 \cdot 16\epsilon^2 \chi^{-8} c^{-12} \le \epsilon^2/18^2,$$

with $c \ge 4c_2^{1/12}$ and $\chi \ge 1$. Combine this result with (94), (95), (96) and (100) to conclude that

$$\left\| P_{\mathcal{S}} \nabla \hat{f}_x(s_j) \right\| \le \frac{\epsilon}{18} + \frac{\epsilon}{18} + \frac{\epsilon}{18} = \frac{\epsilon}{6}$$

for all $\mathscr{T}/4 \le j \le \mathscr{T}/2$. This proves the first part of the lemma.

For the second part of the result, consider (93) anew then (81) and (83) to see that:

$$\begin{aligned} v_j = s_j - s_{j-1} &= \sum_{m'=1}^{d} \left\langle \begin{pmatrix} s_j \\ s_{j-1} \end{pmatrix}, \begin{pmatrix} e_{m'} \\ -e_{m'} \end{pmatrix} \right\rangle e_{m'} \\ &= -\eta \sum_{m'=1}^{d} \sum_{k=0}^{j-1} \sum_{m=1}^{d} \left(g^{(m)} - \delta_k^{(m)} \right) \left\langle A^{j-1-k} \begin{pmatrix} e_m \\ 0 \end{pmatrix}, \begin{pmatrix} e_{m'} \\ -e_{m'} \end{pmatrix} \right\rangle e_{m'} \\ &= -\eta \sum_{k=0}^{j-1} \sum_{m=1}^{d} \left(g^{(m)} - \delta_k^{(m)} \right) \left((A_m^{j-1-k})_{11} - (A_m^{j-2-k})_{11} \right) e_m. \end{aligned}$$

Using notation as in (87) for $a_{m,t}$, it follows that

$$P_{\mathcal{S}} v_j = -\eta \sum_{m \in S} \sum_{k=0}^{j-1} \left(g^{(m)} - \delta_k^{(m)} \right) \left(a_{m,j-1-k} - a_{m,j-2-k} \right) e_m.$$

We aim to upper-bound $\langle P_{\mathcal{S}} v_j, \mathcal{H} P_{\mathcal{S}} v_j \rangle$. Compute, then use (102) to bound the sum in k:

$$\begin{aligned} \langle P_{\mathcal{S}} v_j, \mathcal{H} P_{\mathcal{S}} v_j \rangle &= \eta^2 \sum_{m \in S} \lambda_m \left(\sum_{k=0}^{j-1} \left(g^{(m)} - \delta_k^{(m)} \right) \left(a_{m,j-1-k} - a_{m,j-2-k} \right) \right)^2 \\ &= \eta^2 \sum_{m \in S} \lambda_m \left(g^{(m)} \sum_{k=0}^{j-1} \left(a_{m,k} - a_{m,k-1} \right) \right. \end{aligned}$$

$$
\begin{aligned}
&\quad - \sum_{k=0}^{j-1} \delta_k^{(m)} \left(a_{m,j-1-k} - a_{m,j-2-k}\right)\Bigg)^2 \\
&\leq 2\eta^2 \sum_{m\in S} \lambda_m \left(g^{(m)} \sum_{k=0}^{j-1} \left(a_{m,k} - a_{m,k-1}\right)\right)^2 \\
&\quad + 2\eta^2 \sum_{m\in S} \lambda_m \left(\sum_{k=0}^{j-1} \delta_k^{(m)} \left(a_{m,j-1-k} - a_{m,j-2-k}\right)\right)^2 . \qquad (106)
\end{aligned}
$$

(We used $(a+b)^2 \leq 2a^2 + 2b^2$ again.)

Focusing on the first term of (106), use (97) twice to see that

$$
\begin{aligned}
\sum_{k=0}^{j-1} \left(a_{m,k} - a_{m,k-1}\right) &= \frac{1}{\eta\lambda_m}(1 - a_{m,j} + b_{m,j}) \\
&\quad - \frac{1}{\eta\lambda_m}(1 - a_{m,j-1} + b_{m,j-1}) - a_{m,-1} \\
&= \frac{1}{\eta\lambda_m}(a_{m,j-1} - b_{m,j-1} - a_{m,j} + b_{m,j}).
\end{aligned}
$$

(Indeed, $a_{m,-1} = 0$ as it is the top-left entry of a matrix of the form $\left(\begin{smallmatrix} a & b \\ 1 & 0 \end{smallmatrix}\right)^{-1}$: that is zero regardless of a and $b \neq 0$.)

Hence, the first term in (106) is equal to the right-hand side below; the first bound follows from $(a+b+c+d)^2 \leq 4(a^2+b^2+c^2+d^2)$ (Cauchy–Schwarz) and (98), while the second bound follows from (99) for $j \geq \mathscr{T}/4$:

$$
\begin{aligned}
&\sum_{m\in S} \frac{2}{\lambda_m} \left|g^{(m)}\right|^2 \left(a_{m,j-1} - b_{m,j-1} - a_{m,j} + b_{m,j}\right)^2 \\
&\quad \leq \sum_{m\in S} \frac{16}{\lambda_m} \left|g^{(m)}\right|^2 \left((j+1)^2(1-\theta)^j + j^2(1-\theta)^{j-1}\right) \\
&\quad \leq \sum_{m\in S} \frac{16}{\lambda_m} \left|g^{(m)}\right|^2 \left((1-\theta)^{j/2} + (1-\theta)^{j/2-1}\right) \\
&\quad \leq \sum_{m\in S} \frac{128}{3\lambda_m} \left|g^{(m)}\right|^2 (1-\theta)^{j/2}.
\end{aligned}
$$

(The last inequality uses $\theta \in (0, 1/4]$ so that $(1-\theta)^{-1} \leq 4/3$.) Moreover, for $m \in S$ we have $\lambda_m > \frac{\theta^2}{\eta(2-\theta)^2} \geq \frac{1}{4\eta}\theta^2 = \frac{1}{4\eta}\frac{1}{16}\frac{\sqrt{\hat{\rho}\epsilon}}{\ell} = \frac{\sqrt{\hat{\rho}\epsilon}}{16}$. Therefore, in light of the latest considerations and using the assumption $\|\nabla \hat{f}_x(0)\| \leq 2\ell\mathscr{M}$ and also $j/2 \geq \sqrt{\kappa}\chi c/8$

owing to $j \geq \mathscr{T}/4$, the first term in (106) is upper-bounded by:

$$\begin{aligned}\sum_{m\in S} \frac{128}{3}\frac{16}{\sqrt{\hat{\rho}\epsilon}}\left|g^{(m)}\right|^2 (1-\theta)^{j/2} &\leq 3000\frac{\ell^2\mathscr{M}^2}{\sqrt{\hat{\rho}\epsilon}}(1-\theta)^{\sqrt{\kappa}\chi c/8}\\ &= 3000\mathscr{M}^2\sqrt{\hat{\rho}\epsilon}\kappa^2\left(1-\frac{1}{4\sqrt{\kappa}}\right)^{4\sqrt{\kappa}\cdot 4\chi\cdot c/128}\\ &\leq 3000\mathscr{M}^2\sqrt{\hat{\rho}\epsilon}\kappa^2\cdot 2^{-4\chi}2^{-c/128} \leq \frac{1}{4}\mathscr{M}^2\sqrt{\hat{\rho}\epsilon},\end{aligned}$$

where the second-to-last inequality uses again that $0 < (1-t^{-1})^t < 2^{-1}$ for $t \geq 4$, as well as $4\chi \cdot c/128 \geq 4\chi + c/128$ with $c \geq 128$; and the last inequality uses $\chi \geq \log_2(\theta^{-1}) = \log_2(4\sqrt{\kappa})$ to see that $\kappa^2 2^{-4\chi} \leq 4^{-4}$, and also $3000\cdot 4^{-4}\cdot 2^{-c/128} \leq 1/4$ with $c \geq 720$. (With care, one could improve the constant, here and in many other places.)

Now focusing on the second term of (106), we start with (102) to see that

$$\begin{aligned}&2\eta^2\sum_{m\in S}\lambda_m\left(\sum_{k=0}^{j-1}\delta_k^{(m)}\left(a_{m,j-1-k}-a_{m,j-2-k}\right)\right)^2\\ &\leq 2c_3\eta\sum_{m\in S}\left(|\delta_{j-1}^{(m)}| + \sum_{k=1}^{j-1}|\delta_k^{(m)} - \delta_{k-1}^{(m)}|\right)^2\\ &\leq 4c_3\eta\|\delta_{j-1}\|^2 + 4c_3\eta j\sum_{k=1}^{j-1}\|\delta_k - \delta_{k-1}\|^2\\ &\leq 388c_3\eta\cdot\hat{\rho}^2\mathscr{L}^4.\end{aligned}$$

The last inequality follows through the same reasoning that was applied to go from (103) to (104). Through simple parameter manipulation we find

$$\begin{aligned}388c_3\eta\cdot\hat{\rho}^2\mathscr{L}^4 &= \frac{97c_3}{\ell}\cdot 16\epsilon^2\chi^{-8}c^{-12}\cdot\frac{\ell^2}{\epsilon^2\kappa}c^2\cdot\mathscr{M}^2\\ &= 97c_3\cdot 16\chi^{-8}c^{-10}\cdot\sqrt{\hat{\rho}\epsilon}\mathscr{M}^2 \leq \frac{1}{4}\mathscr{M}^2\sqrt{\hat{\rho}\epsilon},\end{aligned}$$

with $c \geq 3c_3^{1/10}$ and $\chi \geq 1$.

To conclude, we combine the two main results about (106) to confirm that $\left\langle P_{\mathcal{S}}v_j, \mathcal{H}P_{\mathcal{S}}v_j\right\rangle \leq \frac{1}{4}\mathscr{M}^2\sqrt{\hat{\rho}\epsilon} + \frac{1}{4}\mathscr{M}^2\sqrt{\hat{\rho}\epsilon} = \frac{1}{2}\mathscr{M}^2\sqrt{\hat{\rho}\epsilon} \leq \mathscr{M}^2\sqrt{\hat{\rho}\epsilon}$ for all $\mathscr{T}/4 \leq j \leq \mathscr{T}/2$. This proves the second part of the lemma. □

G Proofs from Sect. 6 About PTAGD

G.1 Proof of Proposition 6.2

The following lemma supports Proposition 6.2. The proof is in Appendix G.2. It corresponds to [28,Lem. 23]. The condition on χ originates in this lemma, and from here appears in Theorem 6.1 through Proposition 6.2. It causes the occurrence of dimension in the polylogarithmic factor in the complexity of Theorem 1.6, but note that the real reason why d appears in the condition on χ here is so that dimension can be canceled out in the probabilistic argument in the proof of Proposition 6.2.

Lemma G.1 *Fix parameters and assumptions as laid out in Sect. 3, with $d = \dim \mathcal{M}$, $\delta \in (0, 1)$, any $\Delta_f > 0$ and*

$$\chi \geq \max\left(\log_2(\theta^{-1}), \log_2\left(\frac{d^{1/2}\ell^{3/2}\Delta_f}{(\hat{\rho}\epsilon)^{1/4}\epsilon^2\delta}\right)\right) \geq 1.$$

Let $s_0, s_0' \in B_x(r)$ be such that

1. *$s_0 - s_0' = r_0 e_1$ where e_1 is an eigenvector of $\nabla^2 \hat{f}_x(0)$ associated to the smallest eigenvalue and $r_0 \geq \frac{\delta\mathscr{E}}{2\Delta_f}\frac{r}{\sqrt{d}}$, and*
2. *$\mathtt{TSS}(x, s_0)$ and $\mathtt{TSS}(x, s_0')$ both run their $\mathscr{T}$ iterations in full, respectively, generating vectors u_j, s_j, v_j and u_j', s_j', v_j', with corresponding Hamiltonians E_j, E_j'.*

If $\|\mathrm{grad} f(x)\| \leq \frac{1}{2}\ell b$ and $\lambda_{\min}(\nabla^2 \hat{f}_x(0)) \leq -\sqrt{\hat{\rho}\epsilon}$, then $\max\left(E_0 - E_{\mathscr{T}}, E_0' - E_{\mathscr{T}}'\right) \geq 2\mathscr{E}$.

Proof of Proposition 6.2 By Lemma C.1, $\|\mathrm{grad} f(x)\| \leq 2\ell\mathscr{M} < \frac{1}{2}\ell b$ and $\|\xi\| \leq r < b$. Thus, the strongest provisions of A2 apply at x, as do Lemmas 4.1, 4.2, 4.3 and 4.4. Let u_j, s_j, v_j for $j = 0, 1, \ldots$ be the vectors generated by the computation of $x_{\mathscr{T}} = \mathtt{TSS}(x, \xi)$. Note that $s_0 = \xi$ and $v_0 = 0$. Owing to how $\mathtt{TSS}$ works, there are several cases to consider, based on how it terminates. We remark that cases 3a and 3b are deterministic (they only use the fact that $\|s_0\| \leq r$), that there is no case 3c, and that case 3d is the only place where probabilities are involved. Throughout, it is useful to observe that, since $f(x) = \hat{f}_x(0)$, $\|\mathrm{grad} f(x)\| \leq \epsilon$ and $\mathrm{grad} f(x) = \nabla \hat{f}_x(0)$, the first property of A2 ensures:

$$\hat{f}_x(s_0) - f(x) \leq \langle \mathrm{grad} f(x), s_0 \rangle + \frac{\ell}{2}\|s_0\|^2 \leq \epsilon r + \frac{\ell}{2}r^2 \leq \frac{1}{4}\mathscr{E}. \tag{107}$$

(Use Lemma C.1 to relate parameters.) Compare details below with Proposition 5.3.

- (Case 3a) The negative curvature condition (NCC) triggers with (x, s_j, u_j). Either $\|s_j\| \leq \mathscr{L}$, in which case Lemma 4.2 tells us $E_j \leq E_0 = \hat{f}_x(s_0)$ and, by Lemma 4.4,

$$f(x_{\mathscr{T}}) = \hat{f}_x(\mathtt{NCE}(x, s_j, v_j)) \leq E_j - 2\mathscr{E} \leq f(x) - 2\mathscr{E} + \hat{f}_x(s_0) - f(x).$$

Or $\|s_j\| > \mathscr{L}$, in which case Lemma 4.3 used with $q = j < \mathscr{T}$ and $\|s_q - s_0\| \geq \|s_q\| - \|s_0\| \geq \mathscr{L} - r \geq \frac{63}{64}\mathscr{L}$ implies

$$f(x_{\mathscr{T}}) \leq E_j \leq \hat{f}_x(s_0) - \frac{63^2}{64^2}\frac{\mathscr{L}^2}{16\sqrt{\kappa}\eta\mathscr{T}} = f(x) - \frac{63^2}{64^2}\mathscr{E} + \hat{f}_x(s_0) - f(x).$$

(We used Lemma C.1 to relate parameters.) Either way, bound $\hat{f}_x(s_0) - f(x)$ with (107). Overall, we conclude that $f(x) - f(x_{\mathscr{T}}) \geq \frac{1}{2}\mathscr{E}$ (deterministically).

- (Case 3b) The iterate s_{j+1} leaves the ball of radius b, that is, $\|s_{j+1}\| > b$. In this case, apply Lemma 4.3 with $q = j + 1 \leq \mathscr{T}$ and

$$\|s_{j+1} - s_0\| \geq \|s_{j+1}\| - \|s_0\| \geq b - r \geq 4\mathscr{L} - \frac{1}{64}\mathscr{L} \geq \mathscr{L}$$

to claim (as always, we use Lemma C.1 repeatedly to relate parameters)

$$\begin{aligned} f(x_{\mathscr{T}}) =& \hat{f}_x(s_{j+1}) \leq E_{j+1} \leq \hat{f}_x(s_0) - \frac{\|s_{j+1} - s_0\|^2}{16\sqrt{\kappa}\eta\mathscr{T}} \\ &\leq \hat{f}_x(s_0) - \frac{\mathscr{L}^2}{16\sqrt{\kappa}\eta\mathscr{T}} = \hat{f}_x(s_0) - \mathscr{E}. \end{aligned}$$

By (107), it follows that $f(x) - f(x_{\mathscr{T}}) \geq \frac{3}{4}\mathscr{E}$ (deterministically).

- (Case 3d) None of the other events occur: $\texttt{TSS}(x, s_0)$ runs its $\mathscr{T}$ iterations in full. In this case, we apply the logic in the proof of [28,Lem. 13], as follows. Define the set $\mathcal{X}_x^{(\text{stuck})}$ as containing exactly all tangent vectors $s^* \in B_x(r)$ such that

1. $\texttt{TSS}(x, s^*)$ runs its $\mathscr{T}$ iterations in full, and
2. $E_0^* - E_{\mathscr{T}}^* \leq 2\mathscr{E}$, where E_j^* denotes the Hamiltonians associated to $\texttt{TSS}(x, s^*)$.

There are two cases. Either s_0 is not in $\mathcal{X}_x^{(\text{stuck})}$, in which case $E_0 - E_{\mathscr{T}} > 2\mathscr{E}$: it is then easy to conclude (using (107)) that $f(x) - f(x_{\mathscr{T}}) > \frac{7}{4}\mathscr{E}$. Or s_0 is in $\mathcal{X}_x^{(\text{stuck})}$, in which case we do not lower-bound $f(x) - f(x_{\mathscr{T}})$. The probability of this happening is

$$\text{Prob}\left\{\xi \in \mathcal{X}_x^{(\text{stuck})}\right\} = \frac{\text{Vol}\left(\mathcal{X}_x^{(\text{stuck})}\right)}{\text{Vol}\left(\mathbb{B}_r^d\right)},$$

where $\text{Vol}(\cdot)$ denotes the volume of a set, and $\text{Vol}\left(\mathbb{B}_r^d\right)$ is the volume of a Euclidean ball of radius r in a d-dimensional vector space. In order to upper-bound the volume of $\mathcal{X}_x^{(\text{stuck})}$, we resort to Lemma G.1: this is where we use the assumption $\lambda_{\min}(\nabla^2 \hat{f}_x(0)) \leq -\sqrt{\hat{\rho}\epsilon}$.

Let e_1 denote an eigenvector of $\nabla^2 \hat{f}_x(0)$ with minimal eigenvalue, and let s_0, s_0' be two arbitrary vectors in $\mathcal{X}_x^{(\text{stuck})}$ such that $s_0 - s_0'$ is parallel to e_1. Lemma G.1 implies that $\|s_0 - s_0'\| \leq \frac{\delta\mathscr{E}}{2\Delta_f}\frac{r}{\sqrt{d}}$. Now consider a point $a \in B_x(r)$ orthogonal to e_1, and let ℓ_a

denote the line parallel to e_1 passing through a. The previous reasoning tells us that the intersection of ℓ_a with $\mathcal{X}_x^{(\text{stuck})}$ is contained in a segment of ℓ_a of length at most $\frac{\delta\mathscr{E}}{2\Delta_f}\frac{r}{\sqrt{d}}$. Thus, with $\mathbf{1}$ denoting the indicator function,

$$\begin{aligned}\mathrm{Vol}\left(\mathcal{X}_x^{(\text{stuck})}\right) &= \int_{B_x(r)} \mathbf{1}_{\mathcal{X}_x^{(\text{stuck})}}(y)\mathrm{d}y \\ &= \int_{a\in B_x(r):a\perp e_1}\left[\int_{\ell_a}\mathbf{1}_{\mathcal{X}_x^{(\text{stuck})}}(z)\mathrm{d}z\right]\mathrm{d}a \\ &\leq \frac{\delta\mathscr{E}}{2\Delta_f}\frac{r}{\sqrt{d}}\mathrm{Vol}\left(\mathbb{B}_r^{d-1}\right).\end{aligned}$$

With Γ denoting the Gamma function, it follows that

$$\begin{aligned}\mathrm{Prob}\left\{\xi\in\mathcal{X}_x^{(\text{stuck})}\right\} &\leq \frac{\delta\mathscr{E}}{2\Delta_f}\frac{r}{\sqrt{d}}\cdot\frac{\mathrm{Vol}\left(\mathbb{B}_r^{d-1}\right)}{\mathrm{Vol}\left(\mathbb{B}_r^{d}\right)} \\ &= \frac{\delta\mathscr{E}}{2\Delta_f}\frac{r}{\sqrt{d}}\cdot\frac{1}{r\sqrt{\pi}}\frac{\Gamma(1+d/2)}{\Gamma(1+(d-1)/2)}.\end{aligned}$$

One can check (for example, using Gautschi's inequality) that the last fraction is upper-bounded by $\sqrt{d}$ for all $d\geq 1$. Thus,

$$\mathrm{Prob}\left\{\xi\in\mathcal{X}_x^{(\text{stuck})}\right\}\leq\frac{\delta\mathscr{E}}{2\sqrt{\pi}\Delta_f}\leq\frac{\delta\mathscr{E}}{3\Delta_f}.$$

This limits the probability of the only bad event.

This covers all possibilities. □

G.2 Proof of Lemma G.1

Proof of Lemma G.1 For contradiction, assume $E_0-E_{\mathscr{T}}$ and $E_0'-E_{\mathscr{T}}'$ are both strictly less than $2\mathscr{E}$. Then, by Lemmas 4.2, 4.3 and C.1 and the assumption $\|s_0\|, \|s_0'\|\leq r$, we have

$$\begin{aligned}\forall j\leq\mathscr{T},\ \|s_j\| &\leq r+\|s_j-s_0\|\leq\mathscr{L}/64+\sqrt{32\sqrt{\kappa}\eta\mathscr{T}\mathscr{E}} \\ &= (1/64+\sqrt{2})\mathscr{L}\leq 2\mathscr{L},\ \|s_j'\|\leq 2\mathscr{L}.\end{aligned} \tag{108}$$

The aim is to show that this cannot hold for $j=\mathscr{T}$.

Define $w_j=s_j-s_j'$ for all j. Observe $w_{-1}=s_{-1}-s_{-1}'=(s_0-v_0)-(s_0'-v_0')=s_0-s_0'=w_0$ since $v_0=v_0'=0$. Then, Lemma E.2 provides that

$$\begin{pmatrix}w_j\\ w_{j-1}\end{pmatrix}=A^j\begin{pmatrix}w_0\\ w_0\end{pmatrix}-\eta\sum_{k=0}^{j-1}A^{j-1-k}\begin{pmatrix}\delta_k''\\ 0\end{pmatrix}, \tag{109}$$

where A is as defined and discussed in Appendix E, and

$$\begin{aligned}\delta_k'' &\triangleq \nabla \hat{f}_x(u_k) - \nabla \hat{f}_x(u_k') - \mathcal{H}(u_k - u_k') \\ &= \left(\int_0^1 \left(\nabla^2 \hat{f}_x(\phi u_k + (1-\phi)u_k') - \nabla^2 \hat{f}_x(0)\right) \mathrm{d}\phi\right)[u_k - u_k'].\end{aligned}$$

Recall that $u_k = (2-\theta)s_k - (1-\theta)s_{k-1}$. In particular, using (108) and Lemma C.1 we have:

$$\|u_k\| \leq |2-\theta|\|s_k\| + |1-\theta|\|s_{k-1}\| \leq 6\mathscr{L} \leq b.$$

The same holds for $\|u_k'\|$, and $\|\phi u_k + (1-\phi)u_k'\| \leq \max\left(\|u_k\|, \|u_k'\|\right) \leq 6\mathscr{L} \leq b$ for $\phi \in [0, 1]$. It follows that the Lipschitz-type properties of $\nabla^2 \hat{f}_x$ apply along rays from the origin of $\mathrm{T}_x\mathcal{M}$ to any point of the form $\phi u_k + (1-\phi)u_k'$ for $\phi \in [0, 1]$. Therefore,

$$\begin{aligned}\|\delta_k''\| &\leq 6\hat{\rho}\mathscr{L}\|u_k - u_k'\| = 6\hat{\rho}\mathscr{L}\|(2-\theta)w_k - (1-\theta)w_{k-1}\| \\ &\leq 12\hat{\rho}\mathscr{L}\left(\|w_k\| + \|w_{k-1}\|\right). \qquad (110)\end{aligned}$$

This will come in handy momentarily.

As we did in previous proofs, let $e_1, \ldots, e_d$ form an orthonormal basis of eigenvectors for $\mathcal{H}$ with eigenvalues $\lambda_1 \leq \cdots \leq \lambda_d$. Expand the vectors w_j and δ_k'' in this basis as:

$$w_j = \sum_{m=1}^{d} w_j^{(m)} e_m, \qquad \delta_k'' = \sum_{m=1}^{d} (\delta_k'')^{(m)} e_m.$$

Going back to (109), we can write

$$\begin{aligned}w_j &= \sum_{m'=1}^{d} \left\langle \begin{pmatrix} e_{m'} \\ 0 \end{pmatrix}, \begin{pmatrix} w_j \\ w_{j-1} \end{pmatrix} \right\rangle e_{m'} \\ &= \sum_{m'=1}^{d} \sum_{m=1}^{d} \Bigg[\left\langle \begin{pmatrix} e_{m'} \\ 0 \end{pmatrix}, A^j \begin{pmatrix} e_m \\ e_m \end{pmatrix} \right\rangle w_0^{(m)} \\ &\quad - \eta \sum_{k=0}^{j-1} \left\langle \begin{pmatrix} e_{m'} \\ 0 \end{pmatrix}, A^{j-1-k} \begin{pmatrix} e_m \\ 0 \end{pmatrix} \right\rangle (\delta_k'')^{(m)} \Bigg] e_{m'}.\end{aligned}$$

Owing to (81) and (82), only the terms with $m = m'$ survive. Also, recalling that $w_0 = r_0 e_1$ by assumption, we have

$$w_j = \left(a_{1,j} - b_{1,j}\right) r_0 e_1 - \eta \sum_{m=1}^{d} \sum_{k=0}^{j-1} a_{m,j-1-k} (\delta_k'')^{(m)} e_m, \qquad (111)$$

where $a_{m,j}, b_{m,j}$ are defined by (87).

We aim to show that $w_{\mathscr{T}} = s_{\mathscr{T}} - s'_{\mathscr{T}}$ is larger than $4\mathscr{L}$, as this will contradict the claim that both $\|s_{\mathscr{T}}\|$ and $\|s'_{\mathscr{T}}\|$ are smaller than $2\mathscr{L}$: in view of (108), this is sufficient to prove the lemma. To this end, we introduce two new sequences of vectors to split w_j according to (111):

$$w_j = y_j - z_j, \quad y_j = \left(a_{1,j} - b_{1,j}\right) r_0 e_1, \quad z_j = \eta \sum_{m=1}^{d} \sum_{k=0}^{j-1} a_{m,j-1-k} (\delta''_k)^{(m)} e_m.$$

First, we show by induction that $\|z_j\| \leq \frac{1}{2}\|y_j\|$ for all j. The base case holds since $z_0 = 0$. Now assuming the claim holds for $z_0, \ldots, z_j$, we must prove that $\|z_{j+1}\| \leq \frac{1}{2}\|y_{j+1}\|$. Owing to the induction hypothesis, we know that

$$\forall j' \leq j, \qquad \|w_{j'}\| \leq \|y_{j'}\| + \|z_{j'}\| \leq \frac{3}{2}\|y_{j'}\|. \tag{112}$$

By assumption, λ_1 (the smallest eigenvalue of $\nabla^2 \hat{f}_x(0)$) is less than $-\sqrt{\hat{\rho}\epsilon}$. In particular, it is non-positive. Hence [28,Lem. 37] asserts that $\max_{m=1,\ldots,d} |a_{m,j-k}| = |a_{1,j-k}|$, so that, also using (110) then (112):

$$\begin{aligned}
\|z_{j+1}\| &\leq \eta \sum_{k=0}^{j} \left\| \sum_{m=1}^{d} a_{m,j-k} (\delta''_k)^{(m)} e_m \right\| \leq \eta \sum_{k=0}^{j} |a_{1,j-k}| \|\delta''_k\| \\
&\leq 12\eta\hat{\rho}\mathscr{L} \sum_{k=0}^{j} |a_{1,j-k}| \left(\|w_k\| + \|w_{k-1}\|\right) \\
&\leq 18\eta\hat{\rho}\mathscr{L} \sum_{k=0}^{j} |a_{1,j-k}| \left(\|y_k\| + \|y_{k-1}\|\right).
\end{aligned}$$

Moreover, [28,Lem. 38] applies and tells us that

$$\forall j', \qquad \|y_{j'+1}\| \geq \|y_{j'}\| \geq \frac{\theta r_0}{2} \left(1 + \frac{1}{2} \min\left(\frac{|\eta\lambda_1|}{\theta}, \sqrt{|\eta\lambda_1|}\right)\right)^{j'}. \tag{113}$$

In particular, $\|y_j\|$ is non-decreasing with j. Thus, continuing from above, we find that

$$\|z_{j+1}\| \leq 36\eta\hat{\rho}\mathscr{L} \sum_{k=0}^{j} |a_{1,j-k}| \|y_k\| = 36\eta\hat{\rho}\mathscr{L} r_0 \sum_{k=0}^{j} |a_{1,j-k}| |a_{1,k} - b_{1,k}|,$$

where the last equality follows from the definition of y_k. Owing to [28,Lem. 36], the fact that λ_1 is non-positive implies that

$$\forall 0 \leq k \leq j, \qquad |a_{1,j-k}||a_{1,k} - b_{1,k}| \leq \left(\frac{2}{\theta} + (j+1)\right) |a_{1,k+1} - b_{1,k+1}|.$$

Moreover, $j+1 \le \mathscr{T}$ (as otherwise we are done with the proof by induction), and $\frac{2}{\theta} \le 2\mathscr{T}$ with $c \ge 4$. Hence,

$$\|z_{j+1}\| \le 108\eta\hat{\rho}\mathscr{L}\mathscr{T} r_0 \sum_{k=0}^{j} |a_{1,k+1} - b_{1,k+1}| = 108\eta\hat{\rho}\mathscr{L}\mathscr{T} \sum_{k=0}^{j} \|y_{k+1}\|.$$

Recall that $\|y_k\|$ is non-decreasing with k to see that, using $j+1 \le \mathscr{T}$ once more:

$$\|z_{j+1}\| \le 108\eta\hat{\rho}\mathscr{L}\mathscr{T}^2 \|y_{j+1}\| \le \frac{1}{2}\|y_{j+1}\|.$$

(The last inequality holds with $c \ge 108$ because $108\eta\hat{\rho}\mathscr{L}\mathscr{T}^2 = 54c^{-1}$.) This concludes the induction, from which we learn that $\|w_j\| \ge \|y_j\| - \|z_j\| \ge \frac{1}{2}\|y_j\|$ for all $j \le \mathscr{T}$. In particular, it holds owing to (113) that

$$\|w_{\mathscr{T}}\| \ge \frac{1}{2}\|y_{\mathscr{T}}\| \ge \frac{\theta r_0}{4}\left(1 + \frac{1}{2}\min\left(\frac{|\eta\lambda_1|}{\theta}, \sqrt{|\eta\lambda_1|}\right)\right)^{\mathscr{T}}.$$

As per our assumptions, $\lambda_1 \le -\sqrt{\hat{\rho}\epsilon}$. Therefore, using the definitions of θ, η and κ,

$$\min\left(\frac{|\eta\lambda_1|}{\theta}, \sqrt{|\eta\lambda_1|}\right) \ge \min\left(\frac{\sqrt{\hat{\rho}\epsilon}\sqrt{\kappa}}{\ell}, \sqrt{\frac{\sqrt{\hat{\rho}\epsilon}}{4\ell}}\right) = \min\left(\frac{1}{\sqrt{\kappa}}, \frac{1}{2}\frac{1}{\sqrt{\kappa}}\right) = \frac{1}{2}\frac{1}{\sqrt{\kappa}}.$$

Moreover, $\mathscr{T} = \sqrt{\kappa}\chi c = 4\sqrt{\kappa}\chi c/4$, so that, using $(1+1/t)^t \ge 2$ for $t \ge 4$ and $\kappa \ge 1$, $\chi c \ge 4$:

$$\|w_{\mathscr{T}}\| \ge \frac{\theta r_0}{4}\left(1 + \frac{1}{4\sqrt{\kappa}}\right)^{4\sqrt{\kappa}\cdot\chi c/4} \ge \frac{\theta r_0}{4} 2^{\chi c/4} \ge \frac{\theta}{4}\frac{\delta\mathscr{E}}{2\Delta_f}\frac{r}{\sqrt{d}} 2^{\chi(c/4-1)} 2^{\chi}.$$

At this point, we finally use the assumption $\chi \ge \log_2\left(\frac{d^{1/2}\ell^{3/2}\Delta_f}{(\hat{\rho}\epsilon)^{1/4}\epsilon^2\delta}\right)$ on the 2^{χ} factor:

$$\|w_{\mathscr{T}}\| \ge \frac{\theta}{4}\frac{\delta\mathscr{E}}{2\Delta_f}\frac{r}{\sqrt{d}} 2^{\chi(c/4-1)} \frac{d^{1/2}\ell^{3/2}\Delta_f}{(\hat{\rho}\epsilon)^{1/4}\epsilon^2\delta} = \frac{1}{1024}\chi^{-8}c^{-12}2^{\chi(c/4-1)} \cdot 4\mathscr{L} > 4\mathscr{L}.$$

(The last inequality holds with $c \ge 500$ and $\chi \ge 1$: this fact is straightforward to show by taking derivatives of $\frac{2^{\chi(c/4-1)}}{\chi^8 c^{12}}$ with respect to χ and c, and showing those derivatives are positive.) This concludes the proof by contradiction, from which we deduce that at least one of $E_0 - E_{\mathscr{T}}$ or $E'_0 - E'_{\mathscr{T}}$ must be larger than or equal to $2\mathscr{E}$. □

References

1. P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.
2. N. Agarwal, N. Boumal, B. Bullins, and C. Cartis. Adaptive regularization with cubics on manifolds. *Mathematical Programming*, 188(1):85–134, 2020.
3. Kwangjun Ahn and Suvrit Sra. From nesterov's estimate sequence to riemannian acceleration. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 84–118. PMLR, 09–12 Jul 2020.
4. F. Alimisis, A. Orvieto, G. Bécigneul, and A. Lucchi. Practical accelerated optimization on Riemannian manifolds. arXiv:2002.04144, 2020.
5. Foivos Alimisis, Antonio Orvieto, Gary Becigneul, and Aurelien Lucchi. A continuous-time perspective for modeling acceleration in riemannian optimization. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1297–1307. PMLR, 26–28 Aug 2020.
6. Foivos Alimisis, Antonio Orvieto, Gary Becigneul, and Aurelien Lucchi. Momentum improves optimization on riemannian manifolds. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1351–1359. PMLR, 13–15 Apr 2021.
7. A.S. Bandeira, N. Boumal, and V. Voroninski. On the low-rank approach for semidefinite programs arising in synchronization and community detection. In *Proceedings of The 29th Conference on Learning Theory, COLT 2016, New York, NY, June 23–26*, 2016.
8. 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.
9. Ronny Bergmann, Roland Herzog, Maurício Silva Louzeiro, Daniel Tenbrinck, and Jose Vidal-Nunez. Fenchel duality theory and a primal-dual algorithm on riemannian manifolds. *Foundations of Computational Mathematics*, 2021.
10. R. Bhatia. *Positive definite matrices*. Princeton University Press, 2007.
11. S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3873–3881. Curran Associates, Inc., 2016.
12. N. Boumal. An introduction to optimization on smooth manifolds. Available online, 2020.
13. N. Boumal, P.-A. Absil, and C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2018.
14. N. Boumal, V. Voroninski, and A.S. Bandeira. The non-convex Burer–Monteiro approach works on smooth semidefinite programs. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2757–2765. Curran Associates, Inc., 2016.
15. J.C. Carmon, Y nd Duchi, O. Hinder, and A. Sidford. "convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 654–663. JMLR.org, 2017.
16. Y. Carmon, J.C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 2019.
17. Y. Carmon, J.C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points II: first-order methods. *Mathematical Programming*, September 2019.
18. Yair Carmon and John C Duchi. Analysis of Krylov subspace solutions of regularized nonconvex quadratic problems. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 10728–10738. Curran Associates, Inc., 2018.
19. C. Cartis, N.I.M. Gould, and P. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function- and derivative-evaluation complexity. *Mathematical Programming*, 130:295–319, 2011.
20. C. Criscitiello and N. Boumal. Efficiently escaping saddle points on manifolds. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5985–5995. Curran Associates, Inc., 2019.

21. J.X. da Cruz Neto, L.L. de Lima, and P.R. Oliveira. Geodesic algorithms in Riemannian geometry. *Balkan Journal of Geometry and Its Applications*, 3(2):89–100, 1998.
22. Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods with inexact oracle: the strongly convex case. LIDAM Discussion Papers CORE 2013016, Universite catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2013.
23. O.P. Ferreira and B.F. Svaiter. Kantorovich's theorem on Newton's method in Riemannian manifolds. *Journal of Complexity*, 18(1):304–329, 2002.
24. R. Ge, J.D. Lee, and T. Ma. Matrix completion has no spurious local minimum. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2973–2981. Curran Associates, Inc., 2016.
25. R.E. Greene. Complete metrics of bounded curvature on noncompact manifolds. *Archiv der Mathematik*, 31(1):89–95, 1978.
26. Linus Hamilton and Ankur Moitra. No-go theorem for acceleration in the hyperbolic plane. arXiv: 2101.05657, 2021.
27. J. Hu, X. Liu, Z.-W. Wen, and Y.-X. Yuan. A brief introduction to manifold optimization. *Journal of the Operations Research Society of China*, 8(2):199–248, 2020.
28. C. Jin, P. Netrapalli, and M.I. Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In S. Bubeck, V. Perchet, and P. Rigollet, editors, *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1042–1085. PMLR, 06–09 Jul 2018.
29. H. Karcher. A short proof of Berger's curvature tensor estimates. *Proceedings of the American Mathematical Society*, 26(4):642–642, 1970.
30. Kenji Kawaguchi. Deep learning without poor local minima. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.
31. J.M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer-Verlag New York, 2nd edition, 2012.
32. J.M. Lee. *Introduction to Riemannian Manifolds*, volume 176 of *Graduate Texts in Mathematics*. Springer, 2nd edition, 2018.
33. M. Lezcano-Casado. Trivializations for gradient-based optimization on manifolds. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9157–9168, 2019.
34. Mario Lezcano-Casado. Adaptive and momentum methods on manifolds through trivializations. arXiv: 2010.04617, 2020.
35. Mario Lezcano-Casado. Curvature-dependant global convergence rates for optimization on manifolds of bounded geometry. arXiv: 2008.02517, 2020.
36. Song Mei, Theodor Misiakiewicz, Andrea Montanari, and Roberto Imbuzeiro Oliveira. Solving sdps for synchronization and maxcut problems via the grothendieck inequality. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1476–1515. PMLR, 07–10 Jul 2017.
37. M. Moakher. A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM J. Matrix Anal. Appl.*, 26(3):735–747, 2005.
38. M. Moakher and P.G. Batchelor. *Symmetric Positive-Definite Matrices: From Geometry to Applications and Visualization*, pages 285–298. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.
39. Y. Nesterov and B.T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.
40. Y. E. Nesterov. A method of solving a convex programming with convergence rate $o(1/k^2)$. *Soviet Mathematics Doklady*, 2(27):372–376, 1983.
41. B. O'Neill. *Semi-Riemannian geometry: with applications to relativity*, volume 103. Academic Press, 1983.
42. Sebastian Ruder. An overview of gradient descent optimization algorithms. arXiv:1609.04747, 2016.
43. S. Sra and R. Hosseini. Conic geometric optimization on the manifold of positive definite matrices. *SIAM Journal on Optimization*, 25(1):713–739, 2015.
44. Y. Sun, N. Flammarion, and M. Fazel. Escaping from saddle points on Riemannian manifolds. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7276–7286. Curran Associates, Inc., 2019.
45. N. Tripuraneni, N. Flammarion, F. Bach, and M.I. Jordan. Averaging stochastic gradient descent on Riemannian manifolds. In *Proceedings of The 31st Conference on Learning Theory, COLT*, 2018.

46. S. Waldmann. Geometric wave equations. arXiv:1208.4706, 2012.
47. H. Zhang and S. Sra. First-order methods for geodesically convex optimization. In *Conference on Learning Theory*, pages 1617–1638, 2016.
48. H. Zhang and S. Sra. An estimate sequence for geodesically convex optimization. In S. Bubeck, V. Perchet, and P. Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1703–1723. PMLR, 06–09 Jul 2018.
49. Y. Zhang, Q. Qu, and J. Wright. From symmetry to geometry: Tractable nonconvex problems. arXiv:2007.06753, 2020.