Distributed optimization on Riemannian manifolds for multi-agent networks

Suhail M. Shah
May 28, 2018

Abstract

We consider the consensual distributed optimization problem on a Riemannian manifold. Specifically, the minimization of a sum of functions is studied where each individual function in the sum is located at a node of a network. An algorithm to solve the problem is proposed. A detailed convergence analysis is carried out for smooth functions. The algorithm is demonstrated using some standard applications which fit the proposed framework.

1 Introduction

In the last few decades there has been a major effort to develop algorithms dealing with distributed optimization problems and consensus protocols. The majority of these aim at solving problems arising in wireless and sensor networks, machine learning, multi-vehicle coordination and internet transmission protocols. In such scenarios, the networks are typically spatially distributed over a large area and potentially have a large number of nodes. The absence of a central node for access to the complete system information lends them the name "decentralized algorithms". Such a distinction is made because the lack of a central entity makes the conventional centralized optimization techniques inapplicable. This has initiated the development of distributed computational models and algorithms to support efficient operations over such networks. We refer the interested reader to [1] for a succinct account of the theory of distributed optimization algorithms and the recent developments therein.

The problem of consensual distributed optimization for the Euclidean case be formally stated as:

\[
\begin{align*}
& \text{minimize} \quad \sum_{i=1}^{N} f_i(x^i), \\
& \text{subject to} \quad \sum_{j=1}^{N} q_{ij} \|x^i - x^j\|^2 = 0 \quad \forall i \in \{1, \ldots, N\},
\end{align*}
\]

where \( x^i \in \mathbb{R}^n, f^i : \mathbb{R}^n \to \mathbb{R} \) and scalars \( q_{ij} \) with \( q_{ij} > 0 \) if the nodes \( i \) and \( j \) can communicate and equal to zero if not. There are usually some additional assumptions on the network which we do not state here. Our objective here is to consider the above problem for Riemannian manifolds. The obvious generalization of the above would be:
\[
\text{minimize} \quad \sum_{i=1}^{N} f_i(x^i), \\
\text{subject to} \quad x^i \in \mathcal{M}, \\
\sum_{j \in \mathcal{N}_i} q_{ij} d^2(x^i, x^j) = 0 \quad \forall i \in \{1, \ldots, N\},
\]

where $\mathcal{M}$ is a Riemannian Manifold, $f^i : \mathcal{M} \rightarrow \mathbb{R}$ and $d^2(x^i, x^j)$ is the squared distance between $x^i$ and $x^j$ (see section II for a definition of distance for Riemannian manifolds). The algorithm we propose for solving (1) and its convergence properties are fundamentally different from the Euclidean case. The aim of the present work is to provide such an algorithm and study its convergence properties.

Relevant Literature: [2] traces the key ideas of gradient descent algorithms with line search on manifolds to [3]. As computationally viable alternatives to calculating geodesics became available, the literature dealing with optimization on Riemannian manifolds grew significantly. In the last few decades, these algorithms have been increasingly challenging various mainstream algorithms for the state of art in many popular applications. As a result, there has been considerable interest in optimization and related algorithms on Riemannian manifolds, particularly on matrix manifolds. [2] and [4] serve as standard references for the topic along with [5] for convex optimization on Riemannian manifolds. We refer the reader to the “notes and reference” section in [2] for a wonderful account of the history and development of this field.

The literature on distributed optimization is vast building upon the works of [6] for the unconstrained case and [7] for the constrained case. [8] was one of the earliest works to address the issue of achieving a consensus solution to an optimization problem in a network of computational agents. In [7], the same problem was considered subject to constraints on the solution. The works [9] and [10] extended this framework and studied different variants and extensions (asynchrony, noisy links, varying communication graphs, etc.). The non-convex version was considered in [11] and a nonlinear version was studied in [12].

Finally, we briefly mention some of the works which study distributed consensus algorithms on Riemannian manifolds. Some of the direct applications include distributed pose estimation, camera sensor network utilization among others. [13] lists the earliest works on Riemannian consensus algorithms as [14] and [15]. The former considers the spherical manifold while the latter deals with the $N$-torus. [16] considers a more general class of compact homogeneous manifolds. However, these deal with only embedded sub-manifolds of Euclidean space. In particular they depend on the specific embedding used and hence are extrinsic. Some other relevant works which deal with coordination on Lie groups include [17] and [18]. The most important work for our purposes is [13]. In it, the authors consider the consensus problem as an optimization problem and employ the standard Riemannian gradient descent to solve it. A primary advantage of using this approach is that it is intrinsic and hence is independent of the embedding map used.

Overview of the present work: We propose an algorithm which is a natural extension of the distributed optimization framework for Euclidean space. To do so, we leverage ideas from [6] and [13]. The present generalization to Riemannian manifolds faces two (surmountable) hurdles:

(i) The Riemannian distance function, which is the analog of the Euclidean distance for manifolds, is not convex. This introduces local minima in the cost function that is being minimized in order to achieve consensus. These local minima constitute non-consensus configurations, by which we mean that the limiting values at all the nodes need not be equal. However, this prob-
lem can be overcome by restricting our attention to local neighborhoods where convexity of the Riemannian distance function is guaranteed.

(ii) The gradient vectors of the cost function evaluated at different nodes are tangent vectors lying in different tangent planes. This renders the usual vector operations on them useless. This issue is addressed using the concept of vector transport.

The convergence properties of the proposed algorithm are studied and convergence is established under certain assumptions. The algorithm is presented for the noisy case, i.e. the gradient evaluations are assumed to contain noise. This is done in order to study the algorithm in greater generality and subsumes the noiseless case.

The paper is organized as follows: In Section 2 we provide basic definitions and recall concepts of Riemannian manifolds which we will be using here. Also, a quick review of Riemannian consensus and distributed optimization on Euclidean spaces is provided. Section 3 presents the algorithm along with relevant assumptions needed to establish convergence. Section 4 details the convergence analysis. We first show that the algorithm achieves consensus subject to certain assumptions. Then we compare the asymptotic behaviour to a related dynamical system. In section 6 we test the proposed algorithm on several different examples.

Some Notation: Since we are dealing with distributed computation, we use a stacked vector notation. In particular a boldface \( \mathbf{w}_k = (w^1_k, \ldots, w^N_k) \), where \( w^i_k \) is the value stored at the \( i \)’th node, so that the superscript indicates the node while the subscript the iteration count. The notation \( f(x) = o(g(x)) \) denotes

\[
\lim_{x \to \infty} \frac{f(x)}{g(x)} \to 0,
\]

while \( f(x) = O(g(x)) \) represents

\[
\limsup_{x \to \infty} \left| \frac{f(x)}{g(x)} \right| \leq M
\]

for some constant \( M < \infty \).

2 Additional Background

In this section we briefly recall the basic definitions and concepts regarding Riemannian manifolds while also establishing notation. For a detailed treatment we refer the reader to [19] or [20]. We also cover the basics of Riemannian consensus and Euclidean distributed optimization.

2.1 Basic Definitions

(i) Riemannian Manifolds: Throughout this paper we let \( \mathcal{M} \) denote a connected Riemannian manifold. A smooth \( n \)-dimensional manifold is a pair \( (\mathcal{M}, \mathcal{A}) \), where \( \mathcal{M} \) is a Hausdorff second countable topological space and \( \mathcal{A} \) is a collection of charts \( \{ \mathcal{U}_\alpha, \psi_\alpha \} \) of the set \( \mathcal{M} \) (for more details see [13]). A Riemannian manifold is a manifold whose tangent spaces are endowed with a smoothly varying inner product \( \langle \cdot, \cdot \rangle_x \) called the Riemannian metric. The tangent space at any point \( x \in \mathcal{M} \) is denoted by \( T_x \mathcal{M} \) (for definition see [2] or [19]). We recall that a tangent space admits a structure of a vector space and for a Riemannian manifold, it is a normed vector space. The tangent bundle \( \mathcal{T} \mathcal{M} \) is defined to be the disjoint union \( \bigcup_{x \in \mathcal{M}} \{ x \} \times T_x \mathcal{M} \). The normal space at the point \( x \) denoted by \( \mathcal{N}_\mathcal{M}(x) \) is the set of all vectors orthogonal (w.r.t to
\langle \cdot, \cdot \rangle \) to the tangent space at \( x \). Using the norm, one can also define the arc length of a curve \( \gamma : [a, b] \to M \) as

\[
L(\gamma) = \int_a^b \sqrt{\langle \dot{\gamma}(t), \dot{\gamma}(t) \rangle_{\gamma(t)}} \, dt.
\]

We let \( d(\cdot, \cdot) \) denote the Riemannian distance between any two points \( x, y \in M \), i.e.,

\[
d : M \times M \to \mathbb{R} : d(x, y) = \inf_{\gamma} L(\gamma),
\]

where \( \Gamma \) is the set of all curves in \( M \) joining \( x \) and \( y \). We recall that \( d(\cdot, \cdot) \) defines a metric on \( M \). We also assume that the sectional curvature of the manifold is bounded above by \( \kappa_{\text{max}} \) and below by \( \kappa_{\text{min}} \).

(ii) Geodesics and Exponential Map : A geodesic on \( M \) is a curve that locally minimizes the arc length (equivalently, these are the curves that satisfy \( \gamma'' \in \mathcal{N}_M(\gamma(t)) \) for all \( t \)). The exponential of a tangent vector \( u \) at \( x \), denoted by \( \exp_x(u) \), is defined to be \( \Gamma(1, x, u) \), where \( t \to \Gamma(t, x, u) \) is the geodesic that satisfies

\[
\Gamma(0, x, u) = x \quad \text{and} \quad \frac{d}{dt} \Gamma(0, x, u)|_{t=0} = u.
\]

Throughout the paper we assume a geodesically complete manifold, i.e. there is always a minimal length geodesic between any two points on the manifold. Let \( B(x, r) \) denote the open geodesic ball centered at \( x \) of radius \( r \), i.e.

\[
B(x, r) = \{ y \in M : d(x, y) < r \}.
\]

Let \( \hat{I}_x \) denote the maximal open set in \( T_xM \) on which \( \exp_x \) is a diffeomorphism. On the set \( I_x = \exp_x(\hat{I}_x) \), the exponential map is invertible. This inverse is called the logarithm map and is denoted as \( \log_x y = \exp_x^{-1} y \) for \( y \in I_x \).

(iii) Injectivity Radius : The radius of the maximal geodesic ball centered at \( x \) entirely contained in \( I_x \) is called the injectivity radius at \( x \) and is denoted as \( \text{inj}_x M \). Also,

\[
\text{inj}_x M = \inf_x \{ \text{inj}_x M \}.
\]

(iv) Retraction : A retraction on \( M \) is a smooth mapping \( \mathcal{R} : TM \to M \), where \( TM \) is the tangent bundle, with the following properties :

i) \( \mathcal{R}_x(0_x) = x \), where \( \mathcal{R}_x \) is the restriction of the retraction to \( T_xM \) and \( 0_x \) denotes the zero element of \( T_xM \).

ii) [Local rigidity] With the canonical identification \( T_0 T_xM \approx T_xM \), \( \mathcal{R}_x \) satisfies

\[
D\mathcal{R}_x(0_x) = \text{id}_{T_xM},
\]

where \( \text{id}_{T_xM} \) denotes the identity mapping on \( T_xM \). We assume in addition that the following property holds :

\[
d(x, R_x(\xi_x)) \leq c \| \xi_x \|. \tag{2}
\]

This obviously holds for the exponential with an equality for \( c = 1 \). We let \( c = 1 \) in \( (2) \) for a retraction without any loss of generality.
(v) Convexity Radius: The convexity radius \( r_c > 0 \) is defined as:

\[
r_c = \frac{1}{2} \left\{ \text{inj } \mathcal{M}, \frac{\pi}{\sqrt{\kappa_{\text{max}}}} \right\},
\]

where, if \( \kappa_{\text{max}} \leq 0 \), we set \( 1/\sqrt{\kappa_{\text{max}}} = +\infty \). This quantity plays an important role in studying the convergence properties since any open ball with radius \( r \leq r_c \) is convex. A subset \( \mathcal{X} \) of \( \mathcal{M} \) is said to be a geodesically convex set if, given any two points in \( \mathcal{X} \), there is a minimizing geodesic contained within \( \mathcal{X} \) that joins those two points. Additionally, the function \( x \mapsto d^2(y, x) \) for any fixed \( y \) is (strongly) convex when restricted to \( B(y, r_c) \).

(vi) Riemannian Gradient: Let \( f: \mathcal{M} \to \mathbb{R} \) be a smooth function and \( v \in T_x\mathcal{M} \) be a tangent vector. If \( Df(x)[v] \) is the directional derivative of \( f \) at \( x \) in the direction \( v \), then the gradient of \( f \) at \( x \), denoted by \( \text{grad} f(x) \), is defined as the unique element of \( T_x\mathcal{M} \) that satisfies,

\[
Df(x)[v] = \langle \text{grad} f(x), v \rangle.
\]

(vii) Parallel/Vector Transport: A parallel transport \( T_y^x : T_x\mathcal{M} \to T_y\mathcal{M} \) translates the vector \( \xi_x \in T_x\mathcal{M} \) along the geodesic to \( T_y^x(\xi_x) \in T_y\mathcal{M} \), while preserving norm and in some sense the direction. Generalizing the notion of parallel transport, a vector transport associated with a retraction \( \mathcal{R} \) (or a geodesic in general) is defined as a smooth mapping (\( \oplus \) denotes the Whitney sum)

\[
T\mathcal{M} \oplus T\mathcal{M} \to T\mathcal{M} : (\eta_x, \xi_x) \mapsto T_{\eta_x}(\xi_x) \in T\mathcal{M}
\]

satisfying:

i) If \( \pi(\cdot) \) denotes the foot of a tangent vector, then \( \pi(\mathcal{R}_x(\eta_x)) = \pi(T_{\eta_x}(\xi_x)) \).

ii) \( T_{\eta_x} : T_x\mathcal{M} \to T_{\mathcal{R}_x(\eta_x)}\mathcal{M} \) is a linear map,

iii) \( T_0(\xi_x) = \xi_x \).

Vector transports are in general more computationally appealing than parallel transport. We use the more intuitive notation (the same one as used for parallel transport) \( T_y^y := T_{\eta_y}(\xi_y) \) for a vector transport with \( y = \mathcal{R}_x(\eta_x) \). The vector transport is well defined as long as we stay within the injectivity radius. Note that we don’t explicitly specify the retraction or the geodesic in the notation.

2.2 Riemannian Consensus

The behavior of consensus algorithms on Riemannian Manifolds is radically different from the Euclidean case. The fundamental differences seem to arise from the fact that the squared Riemannian distance function is not globally convex in general. Also, the effect of the curvature of the manifolds has to be accounted for (although for non-positive curvature, the behavior is almost the same as for the Euclidean case because of an infinite convexity radius, see Corollary 14, [13]). Because of this, only local convergence can be claimed. Although a number of algorithms have been suggested to achieve consensus on manifolds, the best suited for our purposes is the one suggested in [13]. This seems to be the most natural and effective generalization of the Euclidean case and involves minimizing the following potential function:

\[
\varphi(w) = \frac{1}{2} \sum_{\{i,j\} \in \mathcal{E}} d^2(w^i, w^j).
\]

Doing a Riemannian gradient descent on the above yields,

\[
w^i_{k+1} = \text{Exp}_{w^i_k}(-\epsilon \text{grad}_{w^i} \varphi(w_k)),
\]

5
where \( \epsilon \) is an admissible time step depending upon the upper bound \( \mu_{\text{max}} \) on the Hessian of the function \( \varphi(\cdot) \) along the geodesic on some open subset \( X \subset M \) (see Proposition 8, 9 in [13] for the exact formula), i.e. for any \( x \in X \) and \( v \in T_xM \), the second derivative of \( \varphi(\cdot) \) along \( \gamma_{x_0}(t) = \text{Exp}_{x_0}(tv) \) satisfies
\[
\frac{d}{dt} \frac{d}{dt} \varphi(\gamma_{x_0}(t)) \bigg|_{t=0} = \langle v, \text{Hess} \varphi(x_0) v \rangle \leq \mu_{\text{max}} \|v\|^2
\]
(3)

Alternatively, one could use a time step decaying to zero without worrying about the Hessian bound. The downside will be a possibly slower convergence rate.

The behavior of the above algorithm has been thoroughly studied in [13] and the discussion that follows is directly based on it. The underlying idea of [13] to overcome the non-convexity (and hence local minima) of \( d^2(\cdot, \cdot) \) is to restrict the initial measurements to a convex neighborhood. By definition, the global minima of \( \varphi(\cdot) \) belong to the consensus submanifold \( D \) for an undirected connected graph, where \( D \) is defined as
\[
D = \{(w, ..., w) \in M^N : w \in M\}.
\]
(4)

This set is the diagonal space of \( M^N \) and it represents all possible consensus configurations of the network. Let \( \mathcal{G}_{\text{diam}} \) denote the diameter of the network graph. Then using a simple calculation, we can bound the distance between any two node estimates \( w^p \) and \( w^q \), connected by the path \( \{i_k\}_{k=0}^K \), as,
\[
d^2(w^p, w^q) \leq \left( \sum_{k=0}^{K-1} d(x^{i_k}, x^{i_{k+1}}) \right)^2 \leq K \sum_{k=0}^{K-1} d^2(x^{i_k}, x^{i_{k+1}})
\]
\[
\leq K \left( \sum_{\{i,j\} \in E} d^2(w^i, w^j) \right) \leq 2\mathcal{G}_{\text{diam}} \varphi(w),
\]
(5)

where in the last inequality we use \( K \leq \mathcal{G}_{\text{diam}} \). So, as long as \( \varphi(w) < (r_c)^2/2\mathcal{G}_{\text{diam}} \), the node estimates remain within a convex neighbourhood and eventually converge to \( D \) since \( \varphi(w_k) \) is decreasing on the sequence \( w_k \). This sub-level set of \( \varphi(\cdot) \) plays an important role for the case of manifolds with finite convexity radius and we denote it as :
\[
S_{\text{conv}} = \left\{ w \in M^N : \varphi(w) \leq \frac{(r_c)^2}{2\mathcal{G}_{\text{diam}}} \right\}.
\]
(6)

Note that above set is convex. We also point out that the estimate in [3] is highly conservative and usually the convergence set will in fact might be larger. In fact, if \( r_c = \infty \) (e.g. PSD matrices, hyperbolic space), then global convergence to \( D \) is guaranteed.

### 2.3 Euclidean Distributed Optimization

Suppose we have a network of \( N \) agents indexed by 1, ..., \( N \). We associate with each agent \( i \), a function \( f^i : \mathbb{R}^n \rightarrow \mathbb{R} \) and a global convex constraint set \( X \). Also, let \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) denote
\[
f(\cdot) := \frac{1}{N} \sum_{i=1}^{N} f^i(\cdot).
\]
(7)

Let the communication network be modeled by a static undirected connected graph \( \mathcal{G} = \{\mathcal{V}, \mathcal{E}\} \) where \( \mathcal{V} = \{1, ..., N\} \) is the node set and \( \mathcal{E} \subset \mathcal{V} \times \mathcal{V} \) is the set of links \( (i, j) \) indicating that agent
j can send information to agent i.

We associate with the network a non-negative weight matrix $Q = [(q_{ij})]_{i,j \in V}$ such that

$$q_{ij} > 0 \iff (i,j) \in E.$$

In addition, the following assumptions are made on the matrix $Q$:

(N1) [Double Stochasticity] $1^T Q = 1^T$ and $Q 1 = 1$.

(N2) [Irreducibility and aperiodicity] We assume that the underlying graph is irreducible, i.e., there is a directed path from any node to any other node, and aperiodic, i.e., the g.c.d. of lengths of all paths from a node to itself is one. It is known that the choice of node in this definition is immaterial. This property can be guaranteed, e.g., by making $q_{ii} > 0$ for some $i$.

This implies that the spectral norm $\gamma$ of $Q - \frac{11^T}{N}$ satisfies $\gamma < 1$. This guarantees in particular that

$$\| (Q^k - Q^*) u \| \leq \kappa \beta^{-k} \| u \|$$

for some $\kappa > 0$, $\beta > 1$, with $Q^*$ denoting the matrix $\frac{11^T}{N}$.

The objective of distributed optimization is to minimize (7) subject to staying in the constraint $\mathcal{X}$ while simultaneously achieve consensus, i.e.

$$\text{minimize} \sum_{i=1}^{N} f^i(x^i)$$

subject to

$$x^i \in \mathcal{X}$$

$$\sum_{j \in \mathcal{N}(i)} q_{ij} ||x^i - x^j||^2 = 0 \ \forall i.$$

The most popular way to solve this is studied in [8] for the unconstrained case and subsequently in [7] for the constrained case. It involves the following two steps:

(S1) [Consensus Step] This step involves local averaging at each node and is aimed at achieving consensus,

$$v_k^i = \sum_{j \in \mathcal{N}(i)} q_{ij} x_k^j.$$

(S2) [Gradient Descent Step] This step is the gradient descent part aimed at minimizing $f^i$ at each node:

$$x_{k+1}^i = P_\mathcal{X}(v_k^i - a_k \nabla f^i(v_k^i)),$$

where $P_\mathcal{X}(\cdot)$ is the projection on the set $\mathcal{X}$ and $a_k$ is a positive scalar decaying at a suitable rate to zero.

The convergence properties of the above algorithm have been extremely well studied for convex as well as the non-convex case. In the next section we modify the above scheme to give an analogous version for Riemannian Manifolds.
2.4 Gradient Flow

We analyze the proposed algorithm using the ODE method. This involves comparing a suitably interpolated trajectory of the iterates obtained from the algorithm with the corresponding gradient flow which we define next. Let \( h : \mathcal{M} \to \mathbb{R} \) be a smooth function on \( \mathcal{M} \). Then by the gradient flow, we mean the dynamics

\[
\dot{w}(t) = -\nabla h(w(t)).
\] (9)

Note that by definition of the gradient vector, the above ODE stays on the manifold. The existence and uniqueness theorem for ordinary differential equations guarantees that there exists a unique smooth function \( \Phi : \mathbb{R} \times \mathcal{M} \to \mathcal{M} \) such that

i) \( \Phi_\cdot(t) : \mathcal{M} \to \mathcal{M} \) is a diffeomorphism for each \( t \in \mathbb{R} \);

ii) \( \Phi_w(t + s) = \Phi_{\Phi_w(t)}(s) \); and

iii) for each \( w \in \mathcal{M} \),

\[
\left. \frac{d}{dt} \Phi_w(t) \right|_{t=0} = -\nabla h(w).
\]

Also, by the definition of the gradient vector field, the equilibria of the differential equation (9) are precisely the critical points of \( h : \mathcal{M} \to \mathbb{R} \). We note here that \( h \) itself serves as a Lyapunov function:

\[
\frac{d}{dt} h(w(t)) = \langle \nabla h(w(t)), \dot{w}(t) \rangle = -\|\nabla h(w(t))\|^2 \leq 0.
\]

We recall here a result (Proposition 12.1, [4]) which we use later:

**Proposition 1.** Let \( h : \mathcal{M} \to \mathbb{R} \) be a smooth function on a Riemannian manifold with compact sub-level sets, i.e. for all \( c \in \mathbb{R} \) the sub-level set \( \{ w \in \mathcal{M} \mid h(w) \leq c \} \) is a compact subset of \( \mathcal{M} \). Then every solution \( w(t) \) of (9) converges to a connected component of the set of critical points of \( h(\cdot) \) as \( t \to +\infty \).

3 A Distributed Optimization algorithm on Riemannian Manifolds

The proposed algorithm to solve (1) on Riemannian manifolds is shown in Algorithm 1.
Algorithm 1 Distributed Optimization on Riemannian Manifolds

**Input**: Manifold $\mathcal{M}$, cost functions $f^i(\cdot)$ for all $i$, retraction $\mathcal{R}(\cdot)$; exponential map $\text{Exp}(\cdot)$; injectivity radius $\iota(\mathcal{M})$; vector transport $\mathcal{T}(\cdot)$; time steps $\epsilon$ and $a_k$; network graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with weight matrix $Q$; consensus steps sequence $\{n_k\}$.

**Assumptions**:
(A1)-(A6) given below and (N1)-(N2) of Section 2.

**Initial Conditions**:
Initialize $w_0 \in S_{\text{conv}}$.

For $k = 0, 1, 2, \ldots$ do:

At each node $i \in \{1, \ldots, N\}$ do:

(S1) **[Riemannian Consensus Step]**:
- $v_k^i = \mathcal{R}_{w_k^i}(−\epsilon \nabla w_k^i \varphi(w_k))$.

(S2) **[Gradient Consensus Step]**:
- (a) Select the number of consensus steps $n_k$. Let $q_{nj}^{nk}$ denotes the $ij$'th entry of $Q^{nk} = \underbrace{Q \times \cdots \times Q}_{n_k \text{ times}}$.
- (b) Derive a noisy sample of the gradient of the function, $\overline{\nabla f^i(v_k^i)} = \nabla f^i(v_k^i) + \bar{M}_{k+1}^i$.
- (c) Compute, $\xi_k^i = \sum_{j=1}^{N} q_{ij}^{nk} \mathcal{T}_{v_k^i}^{v_k^i} \{\nabla f^j(v_k^j)\}$.

(S3) **[Gradient Descent Step]**:
- (a) $w_{k+1}^i = \mathcal{R}_{v_k^i}(−a_k \xi_k^i)$.

$k \leftarrow k + 1$

end

We make the following important assumptions:

(A1) The cost functions $f^i : \mathcal{M} \to \mathbb{R}$ are smooth for all $i$.

(A2) Step-sizes $\{a(n)\}$ are positive scalars satisfying:
\[
\sum_{n} a_n = \infty, \sum_{n} a_n^2 < \infty. \quad (10)
\]
and ε is a constant satisfying $\epsilon \in (0, 2\mu_{\text{max}}^{-1})$, where $\mu_{\text{max}}$ is as in (3).

(A3) $\{\bar{M}_n\}$ is a martingale difference sequence with respect to the increasing σ-fields

$$\mathcal{F}_n = \sigma(w_0, \bar{M}_m^k, m \leq n, k \in \{1, \ldots, N\}), n \geq 0$$

(11)

so that

$$\mathbb{E}[\bar{M}_{n+1}^i|\mathcal{F}_n] = 0 \text{ a.s.}$$

for all $i$. Furthermore, we assume that

$$\sup_n \mathbb{E}[\|\bar{M}_{n+1}^i\|^2|\mathcal{F}_n] < \infty.$$  (12)

(A4) The iterates $w_k$ and $v_k$ of Algorithm 1 remain in a (possibly sample point dependent) compact subset $\mathcal{X}$ of $\mathcal{M}$. We let $R$ denote the diameter of $\mathcal{X}$.

(A5) There exists a sequence of positive integers $\{n_k\}$, denoting the number of consensus steps taken during the gradient consensus step (Step 2) of the $k$'th iteration, such that the following holds:

$$|q_{pq}^{n_k} - \frac{1}{N}| \leq \delta_k \to 0.$$  (13)

(A6) For manifolds with a finite convexity radius $r_c$, $w_k \in \mathcal{S}_{\text{conv}}$

for all $k$, with $\mathcal{S}_{\text{conv}}$ defined in (6).

Remark 2. Alternatively, one could use the exponential map $\text{Exp}_w(\cdot)$ instead of a retraction in Step 1 and 3 of Algorithm 1 with no change in the behavior of the algorithm. The main advantage of a retraction is a lower computational overhead.

Remark 3. Note that the norm $\|\cdot\|$ used in (A3) is derived from the Riemannian metric $\langle \cdot, \cdot \rangle$. Condition (A4) is a stability condition that needs to be separately verified, see [27], Chapter 3 for some tests for stability in the Euclidean case. On the plus side, many of the manifolds of interest in optimization, such as the sphere and the Stiefel manifolds, enjoy compactness making this assumption redundant.

Remark 4. In theory, an infinite number of consensus steps are needed to achieve full consensus in Step 2. This makes the condition $\delta_k \to 0$ impossible to achieve. But from a practical viewpoint, a very few number of iterations seem enough for convergence within a reasonable tolerance. This because the operation performed in this step enjoys a geometric convergence rate (see equation (8)), it converges very quickly and the computational overhead involved is manageable. We provide an alternative to Step 2 requiring only one consensus round per iteration in Section 5.

We discuss the steps of Algorithm 1 to gain some intuition and compare them with their Euclidean counterpart:

(S1) This constitutes the consensus part of the algorithm. At each point $w_k^i$, the gradient of the function $\varphi(\cdot)$ is evaluated. The estimate at node $i$ moves from the current estimate $w_k^i$ to a new estimate, designated as $v_k^i$, along the retraction in the direction of the negative gradient with an admissible step size $\epsilon$.  

\[^{1}\text{i.e., } R = \sup_{(w_p, w_q) \in \mathcal{X}} d(w_p, w_q)\]
(S2) This step is taken to ensure consensus in the tangent space at each estimate $v_k^i$. We remark here that this step is unnecessary in the Euclidean case, or when each node is optimizing the same function or has access to every $f^i(\cdot)$. Also, for stochastic gradient descent where we minimize an expectation, skipping this step does not pose a problem as we can lump the error into the noise term. For Riemannian manifolds, unlike the Euclidean case, the method employed to achieve consensus is a nonlinear operation and does not preserve the Fréchet mean at each iteration. Consider the unconstrained version of the Euclidean distributed optimization algorithm discussed in Section IIC,

$$x_{k+1}^i = \sum_{j \in \mathcal{N}(i)} q_{ij} x_j^i - a_k \nabla f^i \left( \sum_{j \in \mathcal{N}(i)} q_{ij} x_j^i \right)$$

Summing the above over $i = 1$ to $N$ and using the double stochasticity of $Q$ we get,

$$\bar{x}_{k+1} = \bar{x}_k - a_k (\nabla f(\bar{x}_k) + \epsilon_k)$$

with $\bar{x}_k = \sum_{j=1}^n x_k^j$ and $\epsilon_k \to 0$. Studying the above system is the same as studying the individual iterations at each node because of the fact $x_k^i \to \sum_{j=1}^n x_k^j$ for all $i$, which is the empirical mean of all states. The failure of this latter fact for the case of Riemannian consensus necessitates this step.

Moreover, since $\nabla f^i(\bar{x}_k^i) \in T_{v_k^i} \mathcal{M}$ the gradients evaluated at various nodes belong to different tangent planes and as such vector operations between them are undefined. To overcome this obstacle, we use the concept of vector transport. The linear mapping $x \to (Q \otimes I_n)x$ has the property,

$$\lim_{\eta \to \infty} (W \otimes I_n)^\eta x = \lim_{\eta \to \infty} (W^n \otimes I_n)x \to \frac{1}{n} 1_n^T N x,$$

so that the result of its repeated application to $X$ converges to the mean of $X = \sum_{i=1}^N X^i$. For algorithm 1, we will have $N$ parallel consensus steps for

$$X_k^i = (T_{v_k^i} \{ \nabla f^1(v_k^i) \}, ..., \nabla f^i(v_k^i), ..., T_{v_k^i} \{ \nabla f^N (v_k^N) \})$$

for each $i$. We note here that this step involves only local communication. The number of consensus steps $\eta \in \mathbb{N}$ are decided upon beforehand.

(S3) This step constitutes the optimization part of the algorithm. It is the standard Riemannian gradient descent algorithm performed using a retraction with a suitably fast decaying time step. The slow decaying time step is unavoidable here because of the need to achieve consensus as well as combat the noise in the gradient evaluations. The direction along which the new estimate $w_{k+1}$ is obtained is computed from the Step 2. Note that although $\varphi(w_k) - \varphi(v_k) < 0$, there is no reason that $\varphi(w_{k+1}) - \varphi(w_k) < 0$. But as in the Euclidean case, since the consensus part is executed on a faster time scale, the convergence to the set

$$\bar{w} = \arg\min_{w \in \mathcal{M}} \sum_{i=1}^N d^2(w^i, w)$$

In fact the consensus value may not even lie in the convex hull of the initial measurements for non-constant curvature manifolds (Section IVC, [13]).

4We could also use parallel transport.
D happens faster and is guaranteed as long as the iterates belong to $S_{\text{conv}}$ at each step $k$. This point will be proved later in the convergence analysis and ways to ensure the latter fact will also be discussed (see Remark 6). We refer the reader to Chapter 6, [21] for a more detailed treatment of the "two time scale" phenomenon.

Before concluding this section we discuss assumption (A6) which is again redundant for the Euclidean case. This assumption is not as restrictive as it may seem. However, if $w_k$ for any $k$ violates this, the algorithm will need to be restarted as consensus will be impossible to achieve in such a situation in general. One could preferably reset the values at all the nodes to the estimate of the node having the lowest cost function value. A possible remedy to this predicament is discussed in Remark 6.

4 Convergence Analysis

In this section we analyze the convergence properties under assumption (A1)-(A6) and (N1)-(N2). For reference, we summarize Algorithm 1 here in the following two equations:

\begin{align}
    v_k^i &= \mathcal{R}_{w_k^i} (-\epsilon \nabla_{w_k} \varphi(w_k)), \\
    w_{k+1}^i &= \mathcal{R}_{v_k^i} (-a_k (\nabla f^\delta(v_k, v_k^i) + M_{k+1}^i)),
\end{align}

where

\begin{align}
    \nabla f^\delta(v_k, v_k^i) &= \sum_{j=1}^{N} q_{ij}^{nk} T_{v_k^i} \{ \nabla f^j(v_k^i) \}
\end{align}

and

\begin{align}
    M_{k+1}^i &= \sum_{j=1}^{N} q_{ij}^{nk} T_{v_k^i} \{ M_{k+1}^j \}.
\end{align}

Note that the second argument in $\nabla f^\delta(v_k, v_k^i)$ indicates that $\nabla f^\delta(v_k, v_k^i) \in T_{v_k^i} \mathcal{M}$. Also, note that we drop the subscript $k$ in $\delta_k$ in (14) as it clutters up the notation. It is understood that $\delta = \delta_k \to 0$ (see (A5)). When the superscript $\delta$ is missing in $\nabla f^\delta(v_k, v_k^i)$, we mean

\begin{align}
    \nabla f(v_k, v_k^i) &= \frac{1}{N} \sum_{j=1}^{N} T_{v_k^i} \{ \nabla f^j(v_k^i) \}.
\end{align}

Also, if the argument belongs to $\mathcal{M}$ (hence not in boldface), we mean

\begin{align}
    \nabla f(v_k^i) &= \frac{1}{N} \sum_{j=1}^{N} \nabla f^j(v_k^i).
\end{align}

The second argument is unnecessary in the above, since no parallel transport is required to evaluate it. When dealing with $\mathcal{M}^N$, let $\nabla_{w_k} \varphi(w_k)$ denote the following stacked vector:

\begin{align}
    \nabla_{w_k} \varphi(w_k) &= [\nabla_{w_k^1} \varphi(w_k), ..., \nabla_{w_k^N} \varphi(w_k)].
\end{align}

In this notation, (13) can be written as

\begin{align}
    v_k &= \mathcal{R}_{w_k} (-\epsilon \nabla_{w_k} \varphi(w_k))).
\end{align}
Along the same lines, let $\nabla^2 f^\delta(v_k, v_k) \doteq [\nabla^2 f^\delta(v_k, v^1_k), ..., \nabla^2 f^\delta(v_k, v^N_k)]$ and $M_k = [M^1_k, ..., M^N_k]$. Then (13) can be written as:

$$w_{k+1} = R_{v_k}(-a_k(\nabla^2 f^\delta(v_k, v_k) + M_{k+1}))$$

(16) with $R_{v_k}(\cdot) = [R_{v^1_k}(\cdot), ..., R_{v^N_k}(\cdot)]$. We have the following bound,

$$\|\nabla^2 f^\delta(v_k, v_k) - \nabla^2 f(v_k, v_k)\| = \| \sum_{j=1}^{N} \{ q_{ij}^\delta T_{v_j}^\delta \{ \nabla^2 f^\delta(v_j^i) \} - \frac{1}{N} T_{v_k}^\delta \{ \nabla^2 f^\delta(v_k^i) \} \} \|

\leq \sum_{j=1}^{N} \| q_{ij}^\delta - \frac{1}{N} \| \nabla^2 f^\delta(v_k) \| \leq C \delta$$

(17) for some constant $C$. In the second inequality we use the fact that $T^\delta_y$ is an isometric transformation preserving the inner product and in the last inequality we use (A4) and (A5).

### 4.1 Consensus

The following lemma shows that Algorithm 1 asymptotically achieves consensus.

**Lemma 5.** Suppose that Assumptions (A1)-(A6) and (N1)-(N2) are satisfied. Then we have,

$$\mathbb{E}[\|\nabla w_i^\delta \varphi(w_k)\|^2] \to 0, \ \forall i.$$ 

(18) so that,

$$w_k \to D \text{ as } k \to \infty,$$

almost surely, i.e.,

$$d(w_i^k, w_j^k) \to 0 \text{ as } k \to \infty, \ \forall i, j \text{ a.s.}$$

**Proof.** We establish the lemma using a routine calculation. Doing a second order Taylor expansion of $\varphi \circ R(\cdot)$ using (15), we get the following inequality

$$\varphi(v_k) \leq \varphi(w_k) - \|\nabla w_k \varphi(w_k)\|^2 \epsilon + \frac{\mu_{\text{max}} \|\nabla w_k \varphi(w_k)\|^2 \epsilon^2}{2}.$$ 

Rearranging terms, we have

$$\varphi(w_k) - \varphi(v_k) \geq \|\nabla w_k \varphi(w_k)\|^2 \epsilon \left( 1 - \frac{\mu_{\text{max}} \epsilon}{2} \right).$$ 

(19)

Note that since $\epsilon \in (0, 2\mu_{\text{max}}^{-1})$ from (A2), the RHS of the above inequality is strictly positive with equality throughout if and only if $\|\nabla w_k \varphi(w_k)\|^2 = 0$. Similarly, from (16) we get

$$\varphi(w_{k+1}) \leq \varphi(v_k) - \langle \nabla v_k \varphi(v_k), \nabla f^\delta(v_k, v_k) + M_{k+1} \rangle a_k + \frac{\mu_{\text{max}} \|\nabla f^\delta(v_k, v_k) + M_{k+1}\|^2}{2} a_k^2.$$ 

Taking conditional expectation of the above w.r.t $F_k$, we get almost surely

$$\mathbb{E}[\varphi(w_{k+1}) | F_k] \leq \varphi(v_k) - \langle \nabla v_k \varphi(v_k), \nabla f^\delta(v_k, v_k) \rangle a_k + \mathbb{E} \left[ \frac{\mu_{\text{max}} \|\nabla f^\delta(v_k, v_k) + M_{k+1}\|^2}{2} | F_k \right] a_k^2.$$ 

(20)
Note that $M_{k+1}$ is still a martingale owing to the linearity of the vector transport. We also have from (A1), (12) and (A4),

$$E[\|\text{grad}^k(v_k, v_k) + M_{k+1}\|^2] \leq 2E[\|\text{grad}^k(v_k, v_k)\|^2] + 2E[\|M_{k+1}\|^2] \leq \bar{C}. \quad (21)$$

some constant $\bar{C} > 0$. Taking expectation in (20), we get

$$E[\varphi(w_{k+1})] \leq E[\varphi(v_k)] + C E[\|\text{grad}_{v_k} \varphi(v_k)\|] a_k + \bar{C} a_k^2. \quad (22)$$

where $C$ is a suitable bound on $\|\text{grad}^k(v_k)\|$. Rearranging we have,

$$E[\varphi(v_k)] - E[\varphi(w_{k+1})] \geq -C E[\|\text{grad}_{v_k} \varphi(v_k)\|] a_k - \bar{C} a_k^2.$$

Taking expectation in (19) and adding it the above equation we get

$$E[\varphi(w_k)] - E[\varphi(w_{k+1})] \geq E[\|\text{grad}_{w_k} \varphi(w_k)\|^2] \epsilon \left(1 - \frac{\mu_{\text{max}}\epsilon}{2}\right) - \bar{C} a_k^2 - C E[\|\text{grad}_{v_k} \varphi(v_k)\|] a_k.$$

Summing the above equation over $k$ we get

$$\sum_{k=0}^{\infty} \left\{ E[\|\text{grad}_{w_k} \varphi(w_k)\|^2] \epsilon \left(1 - \frac{\mu_{\text{max}}\epsilon}{2}\right) - C E[\|\text{grad}_{v_k} \varphi(v_k)\|] a_k \right\} \leq \varphi(w_0) - \varphi(w^*) + C \sum_{k=0}^{\infty} a_k^2 < \infty, \quad (23)$$

where $w^*$ is any limit point of the algorithm. We have used the facts that $\varphi(\cdot)$ is bounded from below and $\sum_k a_k^2 < \infty$. To conclude the proof, we show that

$$\lim_k \left\{ E[\|\text{grad}_{w_k} \varphi(w_k)\|^2] \epsilon \left(1 - \frac{\mu_{\text{max}}\epsilon}{2}\right) - C E[\|\text{grad}_{v_k} \varphi(v_k)\|] a_k \right\} = 0, \quad (24)$$

which implies that

$$E[\|\text{grad}_{w_k} \varphi(w_k)\|^2] \to 0 \quad (25)$$

for all $i$, since $a_k \to 0$. Assume the contrary, then

$$\lim_k \left\{ E[\|\text{grad}_{w_k} \varphi(w_k)\|^2] \epsilon \left(1 - \frac{\mu_{\text{max}}\epsilon}{2}\right) - C E[\|\text{grad}_{v_k} \varphi(v_k)\|] a_k \right\} = \lim_k E[\|\text{grad}_{w_k} \varphi(w_k)\|^2] \epsilon \left(1 - \frac{\mu_{\text{max}}\epsilon}{2}\right) := D > 0.$$

We have for a large enough $k'$,

$$\sum_{k=k'}^{\infty} \left\{ E[\|\text{grad}_{w_k} \varphi(w_k)\|^2] \epsilon \left(1 - \frac{\mu_{\text{max}}\epsilon}{2}\right) - C E[\|\text{grad}_{v_k} \varphi(v_k)\|] a_k \right\} > \frac{D}{2} \sum_{k=k'}^{\infty} 1 > \infty$$

which contradicts (23). \qed

**Remark 6.** To satisfy assumption (A6) we need to separate the time steps properly. We have

$$\varphi(v_k) \leq \varphi(w_k) - \|\text{grad}_{w_k} \varphi(w_k)\|^2 \epsilon \left(1 - \frac{\mu_{\text{max}}\epsilon}{2}\right).$$
A suitable approximation for $\varphi(\cdot)$ around $v_k$ obtained by neglecting the $O(a_k^2)$ term can be written as,

$$\varphi(w_{k+1}) \approx \varphi(v_k) - (\nabla \varphi(v_k), \nabla f^k(v_k, v_k) + M_{k+1}) a_k.$$  

Combining the above, we have,

$$\varphi(w_{k+1}) - \varphi(w_k) \leq \epsilon \left( 1 - \frac{\mu_{\text{max}} \epsilon}{2} \right) (\nabla \varphi(v_k), \nabla f^k(v_k, v_k) + M_{k+1}) a_k.$$  

We want the above approximation to be negative for $w_{k+1} \in S_{\text{conv}}$ to hold true (assuming that $w_k \in S_{\text{conv}}$). Note that the RHS can be negative only when the dot product in the second term of the above approximation is strictly negative with a value greater than the first term. Since $a_k \rightarrow 0$, this will not be a problem asymptotically. During the initial phases we can keep the ratio $a_k/\epsilon(1-\mu_{\text{max}} \epsilon/2) \ll 1$ as a precaution (which incidentally is what the two time scale argument itself suggests in the asymptotic sense). Another possible remedy would be to employ a line search for the Riemannian consensus and adjust the time step appropriately to get enough decrease in $\varphi(\cdot)$ in Step 1. In fact it is well known that the accelerated line search methods enjoy a linear convergence rate for manifolds (Section 4.5.2. [2]).

Before proceeding further, we prove the following fact which we use in the next section,

$$\sum_{k=0}^{\infty} E[\|\nabla v_k \varphi(w_k)\|^2] < \infty. \hspace{1cm} (26)$$  

We prove this by a contradiction. If (26) does not hold, then it is obvious from (23) that

$$\sum_{k=0}^{\infty} E[\|\nabla v_k \varphi(v_k)\|^2] a_k \rightarrow \infty. \hspace{1cm} (27)$$  

We also have

$$E[d(w_k, v_k)] = E[d(w_k, R_{w_k}(-\epsilon \nabla \varphi(w_k)))] - \epsilon E[\|\nabla \varphi(w_k)\|], \hspace{1cm} (28)$$

where the inequality follows from [2]. The smoothness of $\varphi(\cdot)$ implies [3]

$$E[\|\nabla v_k \varphi(v)\| - \|T_{w_k} \nabla \varphi(w)\|] \leq E[\|\nabla v_k \varphi(v) - T_{w_k} \nabla \varphi(w)\|] \leq LE[d(v, w)],$$

where $L$ is some constant [5] so that combining with the previous equation we have the bound,

$$E[\|\nabla v_k \varphi(v_k)\| - \|T_{w_k} \nabla \varphi(w_k)\|] \leq \epsilon L E[\|\nabla w_k \varphi(w_k)\|]. \hspace{1cm} (29)$$

It can be easily seen from (27) and (29) that

$$\sum_{k=0}^{\infty} E[\|\nabla w_k \varphi(w_k)\|^2] a_k \rightarrow \infty. \hspace{1cm} (30)$$

Now, consider (23) after adding and subtracting $a_k C \|\nabla w_k \varphi(w_k)\| = a_k C \|T_{w_k} \nabla \varphi(w_k)\|$ in the LHS,

$$\sum_{k=0}^{\infty} E[\|\nabla w_k \varphi(w_k)\|^2] a_k \left\{ \frac{E[\|\nabla w_k \varphi(w_k)\|^2]}{E[\|\nabla w_k \varphi(w_k)\|^2] a_k} \left( 1 - \frac{\mu_{\text{max}} \epsilon}{2} \right) \epsilon - K_k \right\} < \infty, \hspace{1cm} (31)$$

$^5T_{w_k} \nabla \varphi(w) = [T_{w_k} \nabla \varphi(w), \ldots, T_{w_k} \nabla \varphi(w)]$.

$^6$By (A6), we are constrained to stay inside the injectivity radius. The exponential (and hence $\varphi(\cdot)$) is guaranteed to be diffeomorphic here. See (Corollary 7.4.6. [2]) how smoothness on a compact set $\mathcal{X}$ gives the inequality used which looks similar to the Lipschitz property in the Euclidean case.
where \( K_k \) is given by the expression
\[
K_k = \left\{ \frac{\mathbb{E}[\|\nabla v_k \varphi(v_k)\| - \|T_{w_k}^* \nabla w_k \varphi(w_k)\|]}{\mathbb{E}[\|\nabla w_k \varphi(w_k)\|]} + 1 \right\} C.
\]
Note that \( K_k \leq (L\epsilon + 1)C := K \) by using (29) in the above expression and \( c_k \geq 0 \) for large \( k \), just as in lemma 5. Since \( c_k \geq 0 \) and \( \sum_{k=0}^{\infty} \mathbb{E}[\|\nabla w_k \varphi(w_k)\|] a_k \to \infty \) (see (30)), the finiteness of the sum in (31) implies \( \lim_k c_k = 0 \).
\[
\implies \limsup_k \frac{\mathbb{E}[\|\nabla w_k \varphi(w_k)\|]}{a_k} \leq \frac{K}{\epsilon(1 - \frac{\mu_{\text{max}}}{2})} = \tilde{K}.
\]
Using Jensen’s inequality, we have
\[
\limsup_k \mathbb{E}[\|\nabla w_k \varphi(w_k)\|] a_k \leq \tilde{K}.
\]
This gives a contradiction to (30):
\[
\sum_{k=0}^{\infty} \mathbb{E}[\|\nabla w_k \varphi(w_k)\|] a_k < \tilde{K} \sum_{k=0}^{\infty} a_k^2 < \infty.
\]

### 4.2 ODE analysis

In this section we analyze the algorithm via the gradient flow. The next lemma compares the iterates obtained from Algorithm 1 to the gradient flow in expectation. The proof builds upon the ideas of [23] in which a generalization of the ODE method to Riemannian manifolds in the context of stochastic approximation was studied. For the Euclidean case, we refer the reader to [21].

For the rest of the subsection, fix \( i \). Studying the behavior of any one particular agent is suffice because of Lemma 5. Let \( t_n = \sum_{k=0}^{n-1} a_k \) with \( t_0 = 0 \). Let \( w_i(t) \) be a continuous time trajectory evolving on the manifold defined by setting \( w_i(t_n) = w_i^n \), where \( w_i^n \) is the iterate produced by (14), and then joined by a geodesic on the interval \([t_n, t_{n+1}]\). Also, let \( w_i(s), t \geq s \), denote the solution to the following ODE evolving in \( \mathcal{M} \) starting at the point \( w(s) \).

\[
\dot{w}_i(t) = -\nabla f(w_i(t))
\]
with \( w_i(s) = w(s) \). This means that \( w_{t_n}(t), t \geq t_n \), is the solution to the ODE (32) with \( w_{t_n}(t_n) = w_i(t_n) = w_i^n \).

**Lemma 7.** For any \( T > 0 \),
\[
\lim_{s \to \infty} \mathbb{E}\left[ \sup_{t \in [s, s+T]} d^2(w^i(t), w_i(t)) \right] = \mathcal{O}(\epsilon), \ \forall i.
\]

**Proof.** Let \( \mathcal{U}_n, \psi_n \) be a neighborhood of the point \( w_i^n \). The following notation is used to denote the coordinate expressions:
\[
\dot{w}_i^n = \psi_n(w_i^n),
\]

\footnote{We recall that \( f(w_i(t)) = \sum_{j=1}^{N} f^j(w_i(t)) \), so that argument of \( f(\cdot) \) belongs to \( \mathcal{M} \) here instead of \( \mathcal{M}^N \) (in order to be consistent with the established notation).}
\[
\text{grad} f^\delta(\tilde{w}_n) = D\psi_n(u^i_n)[\text{grad} f^\delta(w_n, w^i_n)],
\]
\[
M^i_n = D\psi_n(u^i_n)[M^i_n],
\]
where \(D\psi_n(w^i_n): T_{w^i_n}M \to \mathbb{R}^n\) is the differential of the coordinate mapping \(\psi_n\) and is a linear map. Note that we suppress the second argument of \(\text{grad} f^\delta(w_n, w^i_n)\) in the coordinate notation since the vectors are mapped to the Euclidean space under the coordinate transformation. We assume without loss of generality that \(w(t_k) \in U_n\) for all \(t_n < t_k < t_n + T\) and \(\psi(\cdot), D\psi_n(\cdot)\) are bounded maps. This is not a trivial statement and is justified by selecting a suitable set of neighborhoods which cover \(X\) and preserve the aforementioned properties (see pg. 8, [23]). We have from [14]:
\[
w^i_{n+1} = R_{\psi_n} \{ -a_n(\text{grad} f^\delta(v_n, v^i_n) + M^i_{n+1}) \}. \tag{33}
\]
The coordinate expression of the above can be written as:
\[
\psi_n(w^i_{n+1}) = \psi_n(R_{\psi_n} \{ -a_n(\text{grad} f^\delta(v_n, v^i_n) + M^i_{n+1}) \}). \tag{34}
\]
A Taylor expansion of the term on the RHS of \((34)\) gives,
\[
\psi_n(R_{\psi_n} \{ -a_n(\text{grad} f^\delta(v_n, v^i_n) + M^i_{n+1}) \}) = \psi_n(R_{\psi_n}(0_{v^i_n})) + D\psi_n(R_{\psi_n}(0_{v^i_n})) \times D\psi_n(-a_n(\text{grad} f^\delta(v_n, v^i_n) + M^i_{n+1}))) + \mathcal{O}(a^2_n), \tag{35}
\]
Here the remainder \(\mathcal{R}(v^i_n) = \mathcal{O}(a^2_n)\) because of (A1),(A3) and (A4). Thus using the local rigidity property of a retraction \((DR_w(0_w) = id_{T_wM})\) and the fact that \(R_w(0_w) = w\), we can write \((35)\) as
\[
\psi_n(R_{\psi_n} \{ -a_n(\text{grad} f^\delta(v_n, v^i_n) + M^i_{n+1}) \}) = \psi_n(v^i_n) + D\psi_n(v^i_n) \times (-a_n(\text{grad} f^\delta(v_n, v^i_n) + M^i_{n+1}))
+ \mathcal{O}(a^2_n),
\]
which gives
\[
\psi_n(R_{\psi_n} \{ -a_n(\text{grad} f^\delta(v_n, v^i_n) + M^i_{n+1}) \}) = \hat{v}^i_n + a_n(-\text{grad} f^\delta(\hat{v}_n) + \hat{M}^i_{n+1}) + \mathcal{O}(a^2_n), \tag{36}
\]
where we used the linearity of the map \(D\psi_n(\cdot)\) and absorb the negative sign into \(\hat{M}^i_{n+1}\) without loss of generality. We recall that \(\delta = \delta_n \to 0\) is as in (A5) reflecting the number of consensus steps. Also for any \(k\),
\[
||\text{grad} f^\delta(\hat{v}) - \text{grad} f(\hat{v})|| = ||D\psi_n(v^i_n)\left(\sum_{j=1}^N q^j_{ij} T_{v^i_n} \{\text{grad} f^j(v^i_n)\} - \frac{1}{N} \sum_{j=1}^N T_{v^i_n} \{\text{grad} f^j(v^i_n)\} \right)||
\leq \sum_{j=1}^N q^j_{ij} - \frac{1}{N} ||\text{grad} f(\hat{v}_n)|| \leq C \delta_n,
\]
for some constant \(C\), where we have have used the linearity and the boundedness of \(D\psi(\cdot)\) in the second inequality and (A5) in the last inequality. Using this in \((36)\) and absorbing the constant \(C\) into \(\delta_n\), we get
\[
\tilde{w}^i_{n+1} = \tilde{v}^i_n + a_n(-\text{grad} f(\hat{v}_n) + \delta_n) + a_n\hat{M}^i_{n+1} + \mathcal{O}(a^2_n). \tag{37}
\]
A similar coordinate expression for \(\hat{v}_{n+1}\) can be obtained using \((13)\),
\[
\tilde{v}^i_n = \tilde{w}^i_n + \epsilon(-\text{grad}_{w^i_n} \phi(\tilde{w}_n) + \epsilon^2 \mathcal{O}(||\text{grad}_{w^i_n} \phi(w^i_n)||^2)).
\]

17
Substituting the above equation in (37) we get,
\[
\hat{w}_{n+1}^i = \hat{w}_n^i + \epsilon (-\hat{\text{grad}}_{\hat{w}_n^i} \varphi(\hat{w}_n) + \epsilon^2 \mathcal{O}(\|\text{grad}_{\hat{w}_n^i} \varphi(\hat{w}_n)\|^2) + \frac{1}{a_n} (\text{grad}f(\hat{x}_n) + \delta_n) + a_n \hat{M}_{n+1} + \mathcal{O}(a_n^2).
\]
Doing a recursion on the above for any \(1 \leq m \leq m_n\) with \(m_n = \inf\{m : \sum_{k=0}^m a_{n+k} \geq T\}\), we have
\[
\hat{w}_{n+m}^i = \hat{w}_n^i - \sum_{k=0}^{m-1} a_{n+k} \text{grad}f(\hat{w}_{n+k}^i) + \beta_{n,n+m} + \gamma_{n,n+m} + \kappa_{n,n+m},
\] (38)
where
\[
\beta_{n,n+m} = \sum_{k=0}^{m-1} a_{n+k} \{ \hat{M}_{n+k+1}^i + \mathcal{O}(a_{n+k}) + \delta_{n+k} \},
\]
\[
\gamma_{n,n+m} = \sum_{k=0}^{m-1} \epsilon \{ -\text{grad}w_{n+k}^i \varphi(\hat{w}_{n+k}) + \epsilon \mathcal{O}(\|\text{grad}_{w_{n+k}} \varphi(\hat{w}_{n+k})\|^2) \},
\] (39)
and
\[
\kappa_{n,n+m} = -\sum_{k=0}^{m-1} a_{n+k} \{ \text{grad}f(\hat{v}_{n+k}) - \text{grad}f(\hat{w}_{n+k}) \}.
\]
We consider the the expectation norm of all of the above terms, as \(n \to \infty\). Consider \(\beta_{n,n+m}\),
\[
\mathbb{E}[\|\beta_{n,n+m}\|^2]^{\frac{1}{2}} \leq \mathbb{E}[\|\sum_{k=0}^{m-1} a_{n+k} \hat{M}_{n+k+1}^i \|^2]^{\frac{1}{2}} + \sum_{k=0}^{m-1} \{ \mathcal{O}(a_{n+k}^2) + a_{n+k} \delta_{n+k} \},
\]
where we have used the triangle inequality for the expectation norm. Since vector transport is an isometric transformation we have \(\|T_{v_k}^i \{ \hat{M}_{k+1}^j \} \| = \hat{M}_{k+1}^j\). Using this we can bound the expectation of \(\hat{M}_{k+1}^j\),
\[
\mathbb{E}[\|\hat{M}_{n+1}^i\|^2 | \mathcal{F}_n] = \mathbb{E}[\|\sum_{j=1}^N D\psi(w_n) a_{ij}^n T_{v_k}^i \{ \hat{M}_{k+1}^j \} \|^2 | \mathcal{F}_n] \\
\leq \mathbb{E}[\sum_{j=1}^N (\|D\psi(w_n)\|^2 \times \|a_{ij}^n T_{v_k}^i \| \hat{M}_{k+1}^j \|^2) | \mathcal{F}_n] \\
< \infty,
\]
where \(\|D\psi(x_n)\|\) denotes the norm of \(D\psi(x_n)\) viewed as a linear operator and is assumed to be bounded as stated earlier. Note that we have used (12) in the above to bound \(\hat{M}_{k+1}^j\). So, the first term in the expression for \(\beta_{n,n+m}\) converges to 0 from the martingale convergence theorem (Appendix C, [21]). The second one converges from (A2) and third from (A5).

We next consider \(\gamma_{n,n+m}\). We first prove
\[
\limsup_k \frac{\mathbb{E}[\|\text{grad}_{v_k} \varphi(\mathcal{V}_k)\|]}{a_k} = C < \infty.
\] (40)
by a contradiction argument. Suppose no sub-sequence exists for which the above holds (if it
does, we can use that sub-sequence for the rest of the proof without any loss of generality). This
implies \( \liminf_{k} \frac{E[\|\text{grad}_{\mathbf{v}} \varphi(\mathbf{v}_k)\|]}{a_k} = \infty \). Then, we have
\[
\sum_{k=0}^{\infty} E[\|\text{grad}_{\mathbf{v}} \varphi(\mathbf{v}_k)\|]a_k > \liminf_{k} \frac{E[\|\text{grad}_{\mathbf{v}} \varphi(\mathbf{v}_k)\|]}{a_k} \sum_{k} a_k^2 = \infty.
\]
But this is is a contradiction since we have \( \sum_{k=0}^{\infty} E[\|\text{grad}_{\mathbf{v}} \varphi(\mathbf{v}_k)\|]a_k < \infty \) by using (26) in [23]. Then we have
\[
\frac{E[\|\text{grad}_{\mathbf{w}_k} \varphi(\mathbf{w}_k)\|]}{a_k} = \frac{E[\|\text{grad}_{\mathbf{v}_k} \varphi(\mathbf{v}_k)\|]}{a_k} + \frac{E[\|\text{grad}_{\mathbf{w}_k} \varphi(\mathbf{w}_k)\| - \|\text{grad}_{\mathbf{v}_k} \varphi(\mathbf{v}_k)\|]}{a_k} \\
\leq \frac{E[\|\text{grad}_{\mathbf{v}_k} \varphi(\mathbf{v}_k)\|]}{a_k} + \frac{L^{\mathbf{v}} (\mathbf{v}_k, \mathbf{w}_k)}{a_k},
\]
so that from (28),
\[
\limsup_{k} \frac{E[\|\text{grad}_{\mathbf{w}_k} \varphi(\mathbf{w}_k)\|]}{a_k} \leq \bar{C}.
\]
We use the above to bound the first term in \( \gamma_{n,n+m} \),
\[
\| \sum_{k=0}^{m-1} \text{grad}_{\mathbf{w}_n+k} \varphi(\mathbf{w}_n+k) \| \leq \epsilon^2 \limsup_{k} \left\{ \frac{\| \text{grad}_{\mathbf{w}_n+k} \varphi(\mathbf{w}_n+k) \|}{a_{n+k}} \right\} (\sum_{k=0}^{m-1} a_n)^2 \\
\leq \epsilon^2 \limsup_{k} \left\{ \frac{\| \text{grad}_{\mathbf{w}_n+k} \varphi(\mathbf{w}_n+k) \|}{a_{n+k}} \right\} T^2 \leq \epsilon^2 \bar{C}^2 T^2,
\]
so that
\[
\lim_{n} E \left[ \| \sum_{k=0}^{m-1} \text{grad}_{\mathbf{w}_n+k} \varphi(\mathbf{w}_n+k) \| \right]^{\frac{1}{2}} = O(\epsilon).
\]
The second term of \( \gamma_{n,n+m} \) can be shown to converge to zero from (26) implying that
\( E[\|\gamma_{n,n+m}\|^{\frac{3}{2}}] \to 0 \) by the triangle inequality for the expectation norm. Since \( f(\cdot) \) is smooth
we have:
\[
\| \text{grad} f^i(\hat{x}^i) - \text{grad} f^i(\hat{y}^i) \| \leq \hat{L} \| \hat{x}^i - \hat{y}^i \|,
\]
for some constant \( \hat{L} \). Using this we have
\[
E[\|\kappa_{n,n+m}\|^{\frac{3}{2}}] \leq \sum_{k=0}^{m-1} \left\{ E[a_{n+k}^2 \| \text{grad} f(\hat{w}_{n+k}) - \text{grad} f(\hat{w}_{n+k}) \|^{\frac{1}{2}}} \right\} \\
+ E[u_{n+k}^2 \| \text{grad} f(\hat{w}_{n+k}) - \text{grad} f(\hat{w}_{n+k}) \|^{\frac{1}{2}}} \\
\leq \sum_{k=0}^{m-1} a_{n+k} \hat{L} \left\{ \sum_{j=0}^{N} \| \hat{v}^j_{n+k} - \hat{w}^j_{n+k} \|^{\frac{1}{2}} \right\} + \sum_{j=0}^{N} \| \hat{w}^j_{n+k} - \hat{w}^j_{n+k} \|^{\frac{1}{2}}
\]
so that
\[
E[\|\kappa_{n,n+m}\|^{\frac{3}{2}}] \leq T \hat{L} \sup_{0 \leq k \leq m_n} \left\{ \sum_{j=1}^{N} \| \hat{v}^j_{n+k} - \hat{w}^j_{n+k} \|^{\frac{1}{2}} + \sum_{j=1}^{N} \| \hat{w}^j_{n+k} - \hat{w}^j_{n+k} \|^{\frac{1}{2}} \right\}.
\]
We also have for all $i$,
\[
\mathbb{E}[d^2(w^i_k, v^i_k)] = \mathbb{E}[d^2(w^i_k, \mathcal{R}_{w^i_k}(-c\nabla w^i_k \varphi(w^i_k)))] \leq \epsilon^2 \mathbb{E}[\|\nabla w^i_k \varphi(w^i_k)\|^2] \rightarrow 0,
\]
so that $\mathbb{E}[\|\hat{v}^i_{n+k} - \hat{w}^i_{n+k}\|^2] \rightarrow 0$ for all $i$. Similarly, from Lemma 5, the second term in (42) goes to $0$. So to recap, if we set $K_{T,n} = \mathbb{E}[\beta \|n,n+m\|/2] + \mathbb{E}[\gamma \|n,n+m\|/2] + \mathbb{E}[\kappa \|n,n+m\|/2]$, we have
\[
\lim_{n \to \infty} K_{T,n} = \mathcal{O}(\epsilon).
\]

Next we establish similar bounds for the ODE (32) considered in the coordinate expression. Integrating the co-ordinate expression of (32) between the limits $t_n$ to $t_{n+m}$ we have:
\[
\dot{\hat{w}}_{t_n}(t) = \hat{w}_{t_n}(t_n) - \int_{t_n}^{t_{n+m}} \nabla f(\hat{w}_{t_n}(t))dt,
\]
\[
= \hat{w}_{t_n}(t_n) - \sum_{k=0}^{m-1} a_{n+k} \nabla f(\hat{w}_{t_n}(t_{n+k})) + \int_{t_n}^{t_{n+m}} \{\nabla f(\hat{w}_{t_n}(t)) - \nabla f(\hat{w}_{t_n}(t))\}dt,
\]
where $[t] = \max\{t_n : t_n \leq t\}$. We now establish a bound on the last integral on the right hand side. Integrating (32) again between $t_n$ to $t \in [t_n, t_{n+T}]$,
\[
\|\dot{\hat{w}}_{t_n}(t)\| \leq \|\hat{w}_{t_n}(t_n)\| + \int_{t_n}^{t} \|\nabla f(\hat{w}_{t_n}(s))\|ds,
\]
\[
\leq \|\hat{w}_{t_n}(t_n)\| + \|\nabla f(0)\| + \int_{t_n}^{t} \|\nabla f(\hat{w}_{t_n}(s)) - \nabla f(0)\|ds,
\]
\[
\leq c_T + L \int_{t_n}^{t} \|\hat{w}_{t_n}(s)\|ds, \quad \text{(from (41))}
\]
so that by the Gronwall inequality we have
\[
\|\dot{\hat{w}}_{t_n}(t)\| \leq c_T e^{L_T}, \quad t \in [t_n, t_{n+T}).
\]
Let $C_T = c_T e^{LT} + \|\nabla f(0)\|$. Integrating (32) between $t_{n+k}$ to $t \in [t_{n+k}, t_{n+k+1})$ with $0 \leq k \leq m - 1$,
\[
\|\hat{w}_{t_n}(t) - \hat{w}_{t_n}(t_{n+k})\| \leq \|\int_{t_{n+k}}^{t} \nabla f(\hat{w}_{t_n}(s))ds\|,
\]
\[
\leq C_T(t - t_{n+k}).
\]
This gives the required bound:
\[
\|\int_{t_n}^{t_{n+m}} \{\nabla f(\hat{w}_{t_n}(t)) - \nabla f(\hat{w}_{t_n}(t))\}dt\| \leq L \int_{t_n}^{t_{n+m}} \|\dot{\hat{w}}_{t_n}(t) - \hat{w}_{t_n}(t)\|dt
\]
\[
\leq L \sum_{k=0}^{m-1} \int_{t_{n+k}}^{t_{n+k+1}} \|\dot{\hat{w}}_{t_n}(t) - \hat{w}_{t_n}(t_{n+k})\|dt \leq L \sum_{k=0}^{m-1} a_{n+k}^2,
\]
20
so that
\[
\| \int_{t_n}^{t_{n+m}} \{ \nabla f(\hat{w}_{t_n}(t)) - \nabla f(\hat{w}_{t_n}(t)) \} dt \| \leq LCT \sum_{k=0}^{\infty} a_{n+k}^2, \tag{45}
\]

Subtracting (38) and (44) and using the fact that \( \hat{w}_n^i = \hat{w}^n(t_n) \), we get:
\[
\hat{w}_n^{i+m} - \hat{w}_{t_n}(t_{n+m}) \leq \beta_{n,n+m} + \gamma_{n,n+m} + \kappa_{n,n+m} - \int_{t_n}^{t_{n+m}} \{ \nabla f(\hat{w}_{t_n}(t)) - \nabla f(\hat{w}_{t_n}(t)) \} dt,
\]
\[
- \sum_{k=0}^{m-1} \{ a_{n+k} \nabla f(\hat{w}_{n+k}^i) - \nabla f(\hat{w}_{t_n}(t_{n+k})) \}.
\]

Using the triangle equality of expectation norm,
\[
\mathbb{E}[\| \hat{w}_n^{i+m} - \hat{w}_{t_n}(t_{n+m}) \|^2]^{\frac{1}{2}} \leq \mathbb{E}[\| \beta_{n,n+m} \|^2]^{\frac{1}{2}} + \mathbb{E}[\| \gamma_{n,n+m} \|^2]^{\frac{1}{2}} + \mathbb{E}[\| \kappa_{n,n+m} \|^2]^{\frac{1}{2}} + LCT \sum_{k=0}^{\infty} a_{n+k}^2
\]
\[
+ L \sum_{k=0}^{m-1} a_{n+k} \mathbb{E}[\| \hat{w}_{n+k}^i - \hat{w}_{t_n}(t_{n+k}) \|^2]^{\frac{1}{2}}.
\]

We have used (41) and (45) in the last term of the above inequality. Using the expression for \( K_{T,n} \) defined earlier in (43), we have,
\[
\mathbb{E}[\| \hat{w}_n^{i+m} - \hat{w}_{t_n}(t_{n+m}) \|^2]^{\frac{1}{2}} \leq K_{T,n} + LCT \sum_{k=0}^{\infty} a_{n+k}^2 + L \sum_{k=0}^{m-1} a_{n+k} \mathbb{E}[\| \hat{w}_{n+k}^i - \hat{w}_{t_n}(t_{n+k}) \|^2]^{\frac{1}{2}}.
\]

Applying the discrete Gronwall inequality (21, p. 146) to the above we get,
\[
\mathbb{E}[\| \hat{w}_n^{i+m} - \hat{w}_{t_n}(t_{n+m}) \|^2]^{\frac{1}{2}} \leq (K_{T,n} + LCT \sum_{k=0}^{\infty} a_{n+k}^2) \times \exp\{LT\},
\]

Since \( K_{T,n} = O(\epsilon) \) from (43) and \( \lim_{n} \sum_{k} a_{n+k}^2 \to 0 \), the theorem is proved. Note that this proves the theorem only for \( t = t_n \uparrow \infty \). However, we can extend it for general \( t \uparrow \infty \) by exploiting the fact that \( a_n \to 0 \), so that \( t_{n+1} - t_n \to 0 \).

Let \( \mathcal{A} \) denote the equilibrium set,
\[
\mathcal{A} \doteq \{ w \in \mathcal{M} : \nabla f(w) = 0 \}.
\]

and \( \rho(w, \mathcal{A}) \doteq \min_{\bar{w} \in \mathcal{A}} d(w, \bar{w}) \) denote the distance of \( w \in \mathcal{M} \) from \( \mathcal{A} \). By (A4) we have \( \mathcal{A} \subset \mathcal{X} \) where \( \mathcal{X} \) is a compact set. This implies that,
\[
\sup_k \mathbb{E}[\rho(w_k^i, \mathcal{A})^2 | \{ \rho(w_k^i, \mathcal{A}) \geq R \}] = 0.
\]

for a large enough \( R > 0 \) for all \( i \). Also, let
\[
\mathcal{A}^c \doteq \{ w \in \mathcal{M} : \rho(w, \mathcal{A}) \leq \epsilon \}.
\]

21
\textbf{Theorem 8.} Let (A1)-(A6) and (N1)-(N2) hold. Then, we have
\[
\limsup_{n \to \infty} \mathbb{E}[\rho(w_n^i, A \cap D)^2]^{\frac{1}{2}} = O(\epsilon)
\]
for all i.

\textit{Proof.} Convergence to D follows from Lemma \ref{lem:conv_to_D}. Take a large enough \( t = \sum_{m=0}^{k} a_m \) in Lemma \ref{lem:conv_to_D} so that,
\[
\mathbb{E}[d^2(w(t+T), w(t + T))]^{\frac{1}{2}} < K\epsilon.
\]
for some constant \( K \). Note that we drop the superscript in \( w^i(t) \). Also, since \( A \) is the globally asymptotic stable compact attractor of \( (32) \) from Proposition \ref{prop:attractor} we have for a large enough \( T \) and any \( w'(t) = w(t) \in \mathcal{X} - A \),
\[
\rho(w(t + T), A) \leq \frac{1}{2} \rho(w(t), A) \quad (46)
\]
for any \( t > 0 \). Let \( T_k = \inf \{ \ell : \sum_{j=k}^{\ell} a_j = T \} \) by adjusting \( T \) if necessary. Then we have
\[
\mathbb{E}[\rho(w_{k+T_k}, A)^2]^{\frac{1}{2}} = \mathbb{E}[\rho(w(t+T), A)^2]^{\frac{1}{2}} \leq \mathbb{E}[\rho(w(t+T), A)^2 I\{w(t) \in A^c\}]^{\frac{1}{2}} + \mathbb{E}[\rho(w(t+T), A)^2 I\{w(t) \in \mathcal{X} - A^c\}]^{\frac{1}{2}}. \quad (47)
\]
We consider both terms in the RHS separately. Consider the fist term, by Lemma \ref{lem:conv_to_D} we have
\[
\mathbb{E}[\rho(w(t+T), A)^2 I\{w(t) \in A^c\}]^{\frac{1}{2}} \leq \mathbb{E}[\rho(w(t+T), A)^2]^{\frac{1}{2}} + \epsilon \leq K\epsilon
\]
for some appropriate \( K \). To get the last inequality we have used \( (46) \) and the fact that \( w(t) = w^i(t) \). Now consider the second term in \( (47) \),
\[
\mathbb{E}[\rho(w(t+T), A)^2]^{\frac{1}{2}} \leq \mathbb{E}[\rho(w(t+T), A)^2 I\{w(t) \in \mathcal{X} - A^c\}]^{\frac{1}{2}} + \epsilon \
\leq \frac{1}{2} \mathbb{E}[\rho(w(t), A)^2]^{\frac{1}{2}} + K\epsilon = \frac{1}{2} \mathbb{E}[\rho(w_k, A)^2]^{\frac{1}{2}} + K\epsilon,
\]
where we have used Lemma \ref{lem:conv_to_D} in the first inequality, \( (46) \) in the second inequality and \( w'(t) = w_k \) in the last equality. Using the above two bounds in \( (47) \) we get,
\[
\mathbb{E}[\rho(w_{k+T_k}, A)^2]^{\frac{1}{2}} \leq (K + K)\epsilon + \frac{1}{2} \mathbb{E}[\rho(w_k, A)^2]^{\frac{1}{2}}. \quad (48)
\]
Iterating the above we have,
\[
\limsup_{n \to \infty} \mathbb{E}[\rho(w_{k+nT_k}, A)^2]^{\frac{1}{2}} \leq 2(K + K)\epsilon.
\]
which establishes the result. \hfill \Box

\section{Gradient tracking}

In this section we use the concept of gradient tracking (see \cite{24}) to replace the step (S2) of the algorithm. The proposed alternative requires only one round of communication per iteration as opposed to \( n_k \). To accomplish this, we replace (S2) with the following iteration:
\[
g^i_k = \sum_{j=1}^{N} q_{ij} T^{v^i_k}_{v^j_{k-1}} g^j_{k-1} + \text{grad} f^i(v^i_k) - T^{v^i_k}_{v^j_{k-1}} \text{grad} f^j(v^j_{k-1}), \quad (49)
\]
where $\nabla f^i(v^i_k) = \nabla f^i(v^i_k) + \bar{M}^i_{k+1}$ and $g^i_0 = \nabla f^i(v^i_0)$. We assume $\bar{M}^i_{k+1} = 0$ to simplify the analysis. Update (14) then becomes

$$w^i_{k+1} = R_{v^i_k}(-a_k g^i_k).$$

We show that the auxiliary sequence $\{g^i_k\}$ tracks the average gradient closely enough at each node. More precisely, we prove that

$$\|g^i_k - \nabla f(v^i_k, v^i_k)\| \to 0. \quad (50)$$

We remark that (50) is all that is needed for the results of the previous section to hold. For this section we assume the following mild condition on the time step $a_k$:

$$\frac{a_k}{a_{k+1}} \leq c'$$

for some constant $c' > 0$.

Let $\bar{g}^i_k = \frac{1}{N} \sum_{j=1}^N T_{v^i_k}^i g^i_j$. It is easy to verify from (49) that

$$\bar{g}^i_k = T_{v^i_k}^{i-1} g^i_{k-1} + \nabla f(v^i_k, v^i_k) - T_{v^i_k}^{i-1} \nabla f(v^i_{k-1}, v^i_{k-1}). \quad (52)$$

Since $g^i_0 = \nabla f^i(v^i_0)$, we have $\bar{g}^i_1 = \nabla f(v^i_1, v^i_1)$. Using (52), one can deduce by induction that

$$\bar{g}^i_k = \nabla f(v^i_k, v^i_k) \quad (53)$$

Before proving (50), we prove that

$$\|\nabla f^i(v^i_k) - T_{v^i_k}^{i-1} \nabla f^i(v^i_{k-1})\| \leq d(v^i_k, v^i_{k-1}) = O(a_k). \quad (54)$$

The first inequality is obvious from the smoothness of $f^i(\cdot)$. By triangle inequality we have,

$$d(v^i_k, v^i_{k-1}) \leq d(v^i_k, v^i_k) + d(w^i_k, v^i_{k-1}) \leq \epsilon \|\nabla w^i_k \varphi(w^i_k)\| + a_k \|g^i_{k-1}\|,$$

where we have used (2) in the second inequality. The above implies from (40) and (41) that

$$d(v^i_k, v^i_{k-1}) = a_k (\epsilon C + c' \|g^i_{k-1}\|) = O(a_k).$$

We now prove (50). We do this by proving $\|g^i_k - \bar{g}^i_k\| \to 0$ for all $i$ which implies (50) from (53). Setting $\mu^i_k = \nabla f^i(v^i_k) - \nabla f^i(v^i_{k-1})$, we have by iterating (49) the following equation

$$g^i_{k+m} = \sum_{j=1}^N q_{ij} T_{v^i_k}^{j+m} g^i_j + \sum_{j=1}^N \sum_{p=1}^{k+m} q_{ij} T_{v^i_k}^{j+m-p} T_{v^i_k}^j \mu^j_p.$$  

We then have

$$\|g^i_{k+m} - T_{v^i_k}^{i-1} \bar{g}^i_k\| \leq \|\sum_{j=1}^N q_{ij} T_{v^i_k}^{i+m} g^i_j - \frac{1}{N} \sum_{j=1}^N T_{v^i_k}^{i+m} g^i_j\| + \|\sum_{p=k+1}^{k+m} \sum_{j=1}^N q_{ij} T_{v^i_k}^{j+m-p} T_{v^i_k}^j \mu^j_p\|$$

$$\leq \kappa \beta^{-m} \|g^i_k\| + O(\sum_{p=k+1}^{k+m} a_p).$$

We have used (8) and (54) in bounding the first and second terms of the above inequality. Letting $k \to \infty$ followed by $m \to \infty$, we note that the second and the first term go to zero respectively. So, for any limit point $g^i_k$ of the sequence $\{g^i_k\}$ that

$$\|g^i_k - \bar{g}^i_k\| \to 0.$$
6 Numerical Experiments

There are a lot of problems that fit the framework presented and hence could be solved using Algorithm 1. These include regression on manifolds, principal component analysis (PCA), matrix completion, computing the leading eigenvector of distributed data, among many others. In this section we demonstrate the proposed algorithm on the spherical and the Grassman manifolds. The algorithm is simulated in a MATLAB environment using the Manopt toolbox [25].

6.1 Computing the leading eigenvector

The problem entails computing the leading eigenvector of data matrix whose entries are stored at different nodes of a communication network. The problem can be precisely stated as

\[
\begin{align*}
\text{minimize} & \quad \left\{ -w^T A w \right\} \\
\text{subject to} & \quad w \in S^n,
\end{align*}
\]

where \( z_i \in \mathbb{R}^{n+1} \) is the data entry located at the \( i \)’th node and \( S^n \) is the \( n \)-dimensional sphere defined as,

\[ S^n = \{ w \in \mathbb{R}^{n+1} : w^T w = 1 \}. \]

The tangent plane \( T_w S^n \) at a point \( w \) is given by,

\[ T_w S^n = \{ x \in \mathbb{R}^{n+1} : x^T w = 0 \}. \]

The Riemannian metric is the usual inner product inherited from \( \mathbb{R}^{n+1} \). The sectional curvature is constant with \( \kappa_{\text{min}} = \kappa_{\text{max}} = 1 \). Also, \( \text{inj} S^n = r_c = \pi/2 \). The Riemannian gradient \( \text{grad}_w f(\cdot) \) for any function \( f(\cdot) \) can be obtained by the orthogonal projection \( \mathcal{P}_{T_w} (\cdot) \) of the Euclidean gradient \( \nabla f(w) \) on the tangent space,

\[
\text{grad}_w f(w) = \mathcal{P}_{T_w} (\nabla f(w)) = (I_{n+1} - ww^T) \nabla f(w),
\]

where \( I_{n+1} \) is the \((n+1)\times(n+1)\) identity matrix. The geodesic \( t \rightarrow x(t) \) expressed as a function of \( x_0 \in S^{n-1} \) and \( \dot{x}_0 \in T_{x_0} S^{n-1} \) is given by (Example 5.4.1, [2]) :

\[
x(t) = x_0 \cos(\|\dot{x}_0 t\|) + \dot{x}_0 \frac{1}{\|\dot{x}_0\|} \sin(\|\dot{x}_0\| t).
\]

An alternative to this is provided by approximating the geodesic with a retraction which just involves normalizing (i.e., divide by the norm) \( w_n \) at each step. The Riemannian distance between any two points is given as

\[
d(x, y) = \cos^{-1}(x^T y).
\]

The vector transport associated along the geodesic which we employ here is given by,

\[
T_{x_0}^{x(t)} (\xi_x) = \left\{ u \cos(\|\dot{x}_0\| t) x^T (0) - \sin(\|\dot{x}_0\| t) xu^T + (I_{n+1} - uu^T) \right\} \xi_x,
\]

where \( u = \frac{x_0}{\|x_0\|} \).

\[ ^8 \text{available at www.manopt.org} \]
Figure 1: (a) The network Graph (b) This shows the plot of the distance between the various agents vs. the number of iterations. We fix an agent $N$ and plot the value $d^2(w^i, w^N)$. (c) This shows the function value, $f(w^i)$ for all $i$, against the iteration count. (d) This figure demonstrates a failure to achieve consensus due to agents straying out of $S_{conv}$ when $w_0^i$’s are initialized too far away from each other.
The matrix $Q$ is generated using Metropolis weights and for $N = 30$ is shown in Figure 1a. The adjacency matrix satisfies assumptions (N1)-(N2). We consider $N = 30$ agents on the $S^3$ manifold. To generate the initial conditions, a random point $w_0 \in S^3$ is selected and the initial conditions at each node are set to $w^i_0 = \text{Exp}_{w_0}(v^i)$, where $\{v^i\}_{i=1}^N$ are tangent vectors drawn from an isotropic Gaussian distribution with standard deviation $\sigma = 0.2$. The algorithm is run for $k = 10^3$ iterations (although convergence is achieved pretty quickly due to a small matrix size) with step sizes $\epsilon = 0.1$ and $a_k = 1/k$. The results are plotted in Figure 1:

- The agents achieve consensus relatively early and stay in that configuration for the rest of the run. As expected, the number of iterations required increases with the dimension of the problem.
- Figure 1c shows the function value, $f(w^i)$ for all $i$, against the iteration count. Note that as the agents achieve consensus, the values decrease in cohesion for all agents.
- Figure 1d demonstrates a failure to achieve consensus due to agents straying out of $S_{\text{conv}}$ when $w_0^i$’s are initialized too far away from each other. Note that this situation is specific to manifolds with bounded convexity radius and would never occur in the Euclidean case.

6.2 Principal Component Analysis

We next consider a generalization of the previous problem known as Principal Component Analysis (PCA). PCA entails the computation of the $r$ principal eigenvectors of a $n \times n$ covariance matrix $A$ given by

$$A = \mathbb{E}[z_k z_k^T]$$

with $z_1, ..., z_k, ...$ being a stream of uniformly bounded $n$-dimensional data vectors. The cost function for PCA is:

$$C(W) = -\frac{1}{2} \mathbb{E}[z^T W^T W z] = -\frac{1}{2} \text{Tr}(W^T A W),$$

where $\text{Tr}(\cdot)$ denotes the trace of matrix and $W$ belongs to the Stiefel manifold $S_{n,r}$ defined as:

$$S_{n,r} = \{ X \in \mathbb{R}^{n \times r} : X^T X = I_r \}.$$  

Note that the cost function is invariant to the transformation $W \mapsto WO$ where $O \in O(r)$ which represents the orthogonal group. Thus the state space can identified with the Grassmann manifold:

$$\mathcal{G}(n,r) = \{ S_{n,r}/O(r) \},$$

so that any $[G] \in \mathcal{G}(n,r)$ is the equivalence class,

$$[G] = \{ W O, O \in O_r \}.$$  

For details on the geometry of this manifold we refer the reader to [27]. The Riemannian gradient of $C(W)$ under the sample $z$ is given by,

$$\text{grad}_W f(W) = (I_n - W W^T) z z^T W.$$  

From Theorem 8 the relevant ODE is

$$\dot{W} = (I_n - W W^T) A W.$$  

\[9\text{The code to generate it is borrowed from [26].}\]
This is the celebrated Oja’s algorithm [28]. To see how the above can be interpreted as an ODE on $G_{n,r}$, see ([29], Example 9.1). The ODE equilibrium points correspond to

$$AW^* = W^*(W^*)^TAW^*,$$

$$\Rightarrow AW^* = WM$$

for $M = (W^*)^TAW$, proving that any limit point $W^*$ of the algorithm becomes a basis of the dominant $r$-dimensional invariant subspace of $A$. A retraction which could be used here is given by (Example 4.1.3, [2]):

$$R_W(aH) = qf(W + aH),$$

where $qf(\cdot)$ gives the orthogonal factor in the QR-decomposition of its argument. The formula for the parallel transport on Grassmann manifolds is given in (Example 8.1.3 [2]) and for vector transport in (Example 8.1.10 [2]).

We consider a synthetic dataset of $d = 10^4$ measurements with dimension of each measurement being $n = 10^3$. The measurements are assumed to be distributed throughout a network of 10 nodes. The connectivity graph is generated in a similar manner as the previous example and the results are plotted in Figure 2. Note that this setup effectively amounts to running a batch gradient descent on the problem with each iteration representing a pass through the dataset. Another point to note here from Figure 2 is that consensus, as in the previous case, is achieved pretty early around $k = 100$ while the function value evaluated at the estimates keeps on decreasing till $k = 150$. This illustrates the two time scale effect mentioned earlier.

References

[1] A. Nedich et al., “Convergence rate of distributed averaging dynamics and optimization in networks,” Foundations and Trends® in Systems and Control, vol. 2, no. 1, pp. 1–100, 2015.

[2] P.-A. Absil, R. Mahony, and R. Sepulchre, Optimization algorithms on matrix manifolds. Princeton University Press, 2009.
[3] D. G. Luenberger, “The gradient projection method along geodesics,” Management Science, vol. 18, no. 11, pp. 620–631, 1972.

[4] U. Helmke and J. B. Moore, Optimization and dynamical systems. Springer Science & Business Media, 2012.

[5] C. Udriste, Convex functions and optimization methods on Riemannian manifolds. Springer Science & Business Media, 1994, vol. 297.

[6] J. N. Tsitsiklis, “Problems in decentralized decision making and computation.” Massachusetts inst. of tech. Cambridge Lab for information and decision systems, Tech. Rep., 1984.

[7] A. Nedic, A. Ozdaglar, and P. A. Parrilo, “Constrained consensus and optimization in multi-agent networks,” IEEE Transactions on Automatic Control, vol. 55, no. 4, pp. 922–938, 2010.

[8] J. Tsitsiklis, D. Bertsekas, and M. Athans, “Distributed asynchronous deterministic and stochastic gradient optimization algorithms,” IEEE transactions on automatic control, vol. 31, no. 9, pp. 803–812, 1986.

[9] S. S. Ram, A. Nedić, and V. V. Veeravalli, “Distributed stochastic subgradient projection algorithms for convex optimization,” Journal of optimization theory and applications, vol. 147, no. 3, pp. 516–545, 2010.

[10] K. Srivastava and A. Nedic, “Distributed asynchronous constrained stochastic optimization,” IEEE Journal of Selected Topics in Signal Processing, vol. 5, no. 4, pp. 772–790, 2011.

[11] P. Bianchi and J. Jakubowicz, “Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization,” arXiv preprint arXiv:1107.2526, 2011.

[12] A. S. Mathkar and V. S. Borkar, “Nonlinear gossip,” SIAM Journal on Control and Optimization, vol. 54, no. 3, pp. 1535–1557, 2016.

[13] R. Tron, B. Afsari, and R. Vidal, “Riemannian consensus for manifolds with bounded curvature,” IEEE Transactions on Automatic Control, vol. 58, no. 4, pp. 921–934, 2013.

[14] R. Olfati-Saber, “Swarms on sphere: A programmable swarm with synchronous behaviors like oscillator networks,” in Decision and Control, 2006 45th IEEE Conference on. IEEE, 2006, pp. 5060–5066.

[15] L. Scardovi, A. Sarlette, and R. Sepulchre, “Synchronization and balancing on the n-torus,” Systems & Control Letters, vol. 56, no. 5, pp. 335–341, 2007.

[16] A. Sarlette and R. Sepulchre, “Consensus optimization on manifolds,” SIAM Journal on Control and Optimization, vol. 48, no. 1, pp. 56–76, 2009.

[17] Y. Igarashi, T. Hatanaka, M. Fujita, and M. W. Spong, “Passivity-based attitude synchronization in se(3),” IEEE Transactions on Control Systems Technology, vol. 17, no. 5, pp. 1119–1134, 2009.

[18] A. Sarlette, S. Bonnabel, and R. Sepulchre, “Coordinated motion design on lie groups,” IEEE Transactions on Automatic Control, vol. 55, no. 5, pp. 1047–1058, 2010.
[19] M. P. Do Carmo, *Riemannian geometry*. Birkhauser, 1992.

[20] S. Gallot, D. Hulin, and J. Lafontaine, *Riemannian geometry*. Springer, 1990, vol. 3.

[21] V. S. Borkar, *Stochastic approximation: a dynamical systems viewpoint*. Cambridge University Press, 2008.

[22] B. Afsari, R. Tron, and R. Vidal, “On the convergence of gradient descent for finding the riemannian center of mass,” *SIAM Journal on Control and Optimization*, vol. 51, no. 3, pp. 2230–2260, 2013.

[23] S. M. Shah, “Stochastic approximation on riemannian manifolds,” *arXiv preprint arXiv:1711.10754*, 2017.

[24] A. Nedic, A. Olshevsky, and W. Shi, “Achieving geometric convergence for distributed optimization over time-varying graphs,” *SIAM Journal on Optimization*, vol. 27, no. 4, pp. 2597–2633, 2017.

[25] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre, “Manopt, a matlab toolbox for optimization on manifolds,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1455–1459, 2014.

[26] A. Simonetto and H. Jamali-Rad, “Primal recovery from consensus-based dual decomposition for distributed convex optimization,” *Journal of Optimization Theory and Applications*, vol. 168, no. 1, pp. 172–197, 2016.

[27] A. Edelman, T. A. Arias, and S. T. Smith, “The geometry of algorithms with orthogonality constraints,” *SIAM journal on Matrix Analysis and Applications*, vol. 20, no. 2, pp. 303–353, 1998.

[28] E. Oja, “Neural networks, principal components, and subspaces,” *International journal of neural systems*, vol. 1, no. 01, pp. 61–68, 1989.

[29] E. Hairer, C. Lubich, and G. Wanner, *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*. Springer Science & Business Media, 2006, vol. 31.