Non-parametric online change point detection on Riemannian manifolds

Xiuheng Wang, Ricardo Augusto Borsoi, Cédric Richard

Non-parametric Online Change Point Detection on Riemannian Manifolds

Xiuheng Wang [1] **Ricardo Augusto Borsoi** [2] **Cédric Richard** [1]

Abstract

Non-parametric detection of change points in streaming time series data that belong to Euclidean spaces has been extensively studied in the literature. Nevertheless, when the data belongs to a Riemannian manifold, existing approaches are no longer applicable as they fail to account for the structure and geometry of the manifold. In this paper, we introduce a non-parametric algorithm for online change point detection in manifold-valued data streams. This algorithm monitors the generalized Karcher mean of the data, computed using stochastic Riemannian optimization. We provide theoretical bounds on the detection and false alarm rate performances of the algorithm, using a new result on the non-asymptotic convergence of the stochastic Riemannian gradient descent. We apply our algorithm to two different Riemannian manifolds. Experimental results with both synthetic and real data illustrate the performance of the proposed method.

1. Introduction

Change point detection (CPD) is the problem of finding abrupt variations in a property of time series data, which may be indicative of transitions between different states (Aminikhanghahi & Cook, 2017). This question often plays a central role in the modeling, analysis and prediction of time series data, and it has been addressed in many applications ranging from remote sensing (Zeng et al., 2020) and climatology (Reeves et al., 2007) to financial data analysis (Bai & Perron, 1998). Research on this topic can be categorized into two primary branches: offline and online. Offline CPD necessitates access to all received samples, as extensively covered in the literature (Truong et al., 2020). In contrast, online CPD methods process data in real-time and aim to detect change points with minimal delay after their occurrence. In numerous real-world scenarios, the pursuit of non-parametric CPD is also highly relevant since it can be challenging to possess prior knowledge of the data distribution. However, even with the longstanding history and continued interest in CPD techniques, it is noteworthy that the overwhelming majority of existing algorithms assume that the data resides in Euclidean spaces.

Recent developments in statistical learning and signal processing have increasingly confronted the analysis of data residing in non-Euclidean spaces. Among these spaces, Riemannian manifolds have garnered attention due to their wide-ranging applications, such as in diffusion tensor imaging (Pennec et al., 2006), pedestrian detection (Tuzel et al., 2008), and human behavior understanding (Kacem et al., 2018). Consequently, Riemannian optimization (Absil et al., 2009; Boumal, 2023) has emerged as an area of significant interest, offering essential and potent tools for handling data on manifolds, especially with the recent advancements in Riemannian stochastic gradient descent (R-SGD) algorithms (Bonnabel, 2013; Zhang & Sra, 2016). While the detection of change points in Euclidean spaces has been notably successful, it is noteworthy that only a limited number of CPD methods have been specifically crafted for Riemannian manifolds (Bouchard et al., 2020; Dubey & Müller, 2020; Wang et al., 2023a), and these still lack theoretical analyses or online operation. The main hurdles stem from the need to account for the intrinsic non-linear geometry of these spaces and the absence of a vector space structure in the data, making the adaptation of algorithms originally conceived for Euclidean spaces a complex undertaking.

In response to the aforementioned challenges, the objective of this paper is to introduce a unified framework for non-parametric and online CPD on Riemannian manifolds. Our contributions are as follows:

1. **General non-parametric framework:** We propose a comprehensive non-parametric framework for CPD by monitoring central values within Riemannian manifolds. Our framework places particular emphasis on the generalized Karcher mean. We update two estimates of the generalized Karcher mean using the R-SGD algorithm, each with distinct constant step sizes. These two estimates, one with longer memory and the other more adaptive, are compared to construct an online CPD statistic.

[1]Université Côte d'Azur, CNRS, OCA, France [2]Université de Lorraine, CNRS, CRAN, Vandoeuvre-lès-Nancy, France. Correspondence to: Xiuheng Wang <xiuheng.wang@unice.fr>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

2. **Theoretical analyses:** We provide theoretical analyses related to the proposed CPD statistic. We establish non-asymptotic convergence results for R-SGD with a curvature-dependent linear rate under the condition of constant step size (Theorem 4.1). Additionally, in the absence of any change, we derive an upper bound for the false alarm rate (Theorem 4.2). Furthermore, in the presence of a change, we establish a lower bound for the detection rate (Theorem 4.3).

3. **Application to specific manifolds:** We tailor our algorithm to suit two common instances of Riemannian manifolds, specifically, the manifold of symmetric positive definite (SPD) matrices and the Grassmann manifold. We then provide empirical illustrations of the performance of our CPD algorithm on these manifolds through numerical experiments on synthetic and real-world datasets.

By introducing this framework and offering theoretical insights into its performance, we aim to contribute to the advancement of non-parametric and online CPD methods for data residing on Riemannian manifolds, which can impact a range of applications such as, e.g., voice activity detection, pedestrian detection and subspace change detection.

2. Related work

In this section, we review related works on online change point detection and Riemannian optimization which are connected to the proposed approach.

Online CPD: Online CPD methods can be broadly categorized into two main groups: parametric and non-parametric, depending on whether prior knowledge about the data distribution is available. Parametric CPD techniques, illustrated by methods such as the cumulative sum (CUSUM) (Page, 1954; Tartakovsky et al., 2014) and the generalized likelihood ratio test (GLRT) (Gustafsson, 1996), assume that the data distribution conforms to a known parametric family.

In many applications, prior knowledge of the data distribution cannot be guaranteed, leading to the development of non-parametric methods. These approaches encompass various techniques, including monitoring changes in the mean or variance of a data stream, as seen in methods like the Exponentially Weighted Moving Average (EWMA) (Costa & Rahim, 2006), and the use of kernel maximum mean discrepancy (MMD) derived from the data stream (Gretton et al., 2006). Recent advancements in this field have introduced innovative non-parametric methods. For instance, the NEWMA algorithm (Keriven et al., 2020) was introduced to detect change points without the necessity of retaining historical data samples. This is achieved by comparing two EWMAs of data stream statistics, each computed with distinct forgetting factors. The non-parametric kernel MMD statistic initially introduced for hypothesis testing in (Gretton et al., 2006) has recently been widely employed in the context of kernel CPD with both offline (Harchaoui et al., 2008; Sinn et al., 2012) as well as online algorithms (Gong et al., 2012; Li et al., 2019). Kernel extensions of the CUSUM statistic have also been considered in (Madrid Padilla et al., 2023; Arlot et al., 2019; Wei & Xie, 2022). A computationally efficient approximation of the kernel MMD based on the neural tangent kernel has also been proposed (Cheng & Xie, 2021). Another non-parametric online algorithm was developed in (Ferrari et al., 2022), making use of adaptive kernel density ratio estimation. The capabilities of neural networks were explored in (Wang et al., 2023b) to enhance the effectiveness of non-parametric online CPD.

These algorithms, however, assume that the data belongs to an Euclidean space. While some non-parametric online CPD algorithms have been extended to specific non-Euclidean domains, such as graphs (Ferrari & Richard, 2020) and categorical data (Ienco et al., 2014), very few works have investigated scenarios where the data belongs to a Riemannian manifold. In (Bouchard et al., 2020), an online CPD algorithm was specifically designed for the compound Gaussian distribution, which, however, is parametric and not broadly applicable. For data lying on manifolds, a non-parametric offline algorithm (Duan et al., 2019) was developed to detect change points of rigid body motions in the special Euclidean group. Another non-parametric technique, monitoring changes in the Fréchet means and variances, was proposed in (Dubey & Müller, 2020). However, it can only detect a single change point and operates offline. A work extending NEWMA to manifolds was introduced in (Wang et al., 2023a), but the algorithm is not general and does not have any theoretical analyses.

This paper presents a general formulation for CPD on manifolds based on the generalized Karcher mean, a discussion of its related existence and uniqueness questions, theoretical results related to the convergence, false alarm, and detection performance of the algorithm, and exemplifies its application to different manifolds with challenging examples.

Riemannian optimization: Riemannian optimization has recently garnered significant interest as it takes into account the geometry of data manifolds, which is prevalent in many practical applications. Both the books (Absil et al., 2009) and (Boumal, 2023) provide detailed presentations on Riemannian optimization. Substantial work has also been undertaken in order to extend optimization algorithms that were originally developed in Euclidean spaces, such as steepest descent (Smith, 1994) and quasi-Newton (Huang et al., 2015) algorithms, to Riemannian manifolds, as well as to study their convergence behavior.

The R-SGD algorithm, as presented in (Bonnabel, 2013),

has gained significant attention for its capability to handle noisy gradient estimates. Sophisticated variance reduction techniques have been recently introduced to provide algorithms with accelerated convergence rate (Zhang et al., 2016; 2018; Zhou et al., 2019). While the asymptotic convergence of the R-SGD was studied in (Bonnabel, 2013) for diminishing step sizes, explicit convergence rates were not provided. Results on the sublinear convergence rates of first-order Riemannian optimization on geodesically convex problems were recently obtained in (Zhang & Sra, 2016). However, these rates were derived under the assumption of diminishing step sizes or deterministic gradients.

3. Background

This section introduces some basic concepts of Riemannian geometry, focusing on the essential tools for optimization on manifolds. Detailed presentations can be found in (Absil et al., 2009) and (Boumal, 2023).

A *Riemannian manifold* $(\mathcal{M}, g)$ is a constrained set $\mathcal{M}$ endowed with a *Riemannian metric* $g_{\boldsymbol{x}}(\cdot,\cdot) : T_{\boldsymbol{x}}\mathcal{M}\times T_{\boldsymbol{x}}\mathcal{M} \to \mathbb{R}$, defined for every point $\boldsymbol{x} \in \mathcal{M}$, with $T_{\boldsymbol{x}}\mathcal{M}$ the so-called *tangent space* of $\mathcal{M}$ at $\boldsymbol{x}$. A *geodesic* $\gamma : [0,1] \to \mathcal{M}$ is the curve of minimal length linking two points $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{M}$ such that $\boldsymbol{x} = \gamma(0)$ and $\boldsymbol{y} = \gamma(1)$, with $\boldsymbol{v} \in T_{\boldsymbol{x}}\mathcal{M}$ the velocity of γ at 0 denoted by $\dot{\gamma}(0)$. The *geodesic distance* $d_{\mathcal{M}}(\cdot,\cdot) : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ is defined as the length of the geodesic linking two points $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{M}$. It satisfies all the conditions to be a metric.

The *exponential map* $\boldsymbol{w} = \exp_{\boldsymbol{x}}(\boldsymbol{v})$ is defined as the point $\boldsymbol{w} \in \mathcal{M}$ located on the unique geodesic $\gamma(t)$ with endpoints $\boldsymbol{x} = \gamma(0)$, $\boldsymbol{w} = \gamma(1)$ and velocity $\boldsymbol{v} = \dot{\gamma}(0)$. Since calculating the exponential map can be computationally demanding, in practice it is common to employ a *retraction* $R_{\boldsymbol{x}} : T_{\boldsymbol{x}}\mathcal{M} \to \mathcal{M}$ instead, defined at every $\boldsymbol{x} \in \mathcal{M}$, which consists of a second-order approximation to the exponential map, satisfying $d_{\mathcal{M}}(R_{\boldsymbol{x}}(t\boldsymbol{v}), \exp_{\boldsymbol{x}}(t\boldsymbol{v})) = O(t^3)$. Consider a smooth function $f : \mathcal{M} \to \mathbb{R}$. The *Riemannian gradient* of f at $\boldsymbol{x} \in \mathcal{M}$ is defined as the unique tangent vector $\nabla f(\boldsymbol{x}) \in T_{\boldsymbol{x}}\mathcal{M}$ satisfying $\frac{d}{dt}\big|_{t=0} f(\exp_{\boldsymbol{x}}(t\boldsymbol{v})) = \langle \nabla f(\boldsymbol{x}), \boldsymbol{v}\rangle_{\boldsymbol{x}}$, for all $\boldsymbol{v} \in T_{\boldsymbol{x}}\mathcal{M}$.

4. Proposed method

4.1. Problem Background

Consider a sequence of statistically independent samples $\boldsymbol{x}_t$ belonging to a Riemannian manifold $\mathcal{M}$. The Riemannian CPD problem consists of estimating time index $t_r \in \mathbb{N}$, referred to as the *change point*, at which the probability measure of $\boldsymbol{x}_t$ undergoes a change (Pennec, 2004):

$$\begin{aligned} t < t_r &: \ \boldsymbol{x}_t \sim P_1(\boldsymbol{x}) \,, \\ t \geq t_r &: \ \boldsymbol{x}_t \sim P_2(\boldsymbol{x}) \,. \end{aligned} \tag{1}$$

Here, $P_1(\boldsymbol{x})$ and $P_2(\boldsymbol{x})$ are probability measures on $\mathcal{M}$, such that $P_1(\boldsymbol{x}) \neq P_2(\boldsymbol{x})$, representing how $\boldsymbol{x}_t$ is distributed before and after the change point, respectively. Throughout this paper, it is assumed that the difference between the generalized Karcher means of $P_1(\boldsymbol{x})$ and $P_2(\boldsymbol{x})$ (see Section 4.2 for a definition) is sufficiently large, to make this problem tractable. While the CPD problem as defined in (1) presents only a single change point for simplicity, CPD algorithms are typically designed to handle multiple change points.

CPD algorithms aim to compute an estimate $\hat{t}_r$ of the change point. These algorithms have two primary objectives: first, to minimize the delay between the occurrence of a change point and its detection by the algorithm, and second, to minimize the probability of generating false alarms on time steps when no actual change has occurred. While various CPD algorithms have been proposed for Euclidean spaces, the constraint that the data $\boldsymbol{x}_t$ belongs to a Riemannian manifold $\mathcal{M}$, which typically lacks a vector space structure, presents challenges for algorithm design. Furthermore, many applications involve streaming data and require the online resolution of the CPD problem. In other words, an algorithm must determine whether a recent time index $t' \leq t$ is a change point based solely on past data $\{\boldsymbol{x}_s\}_{s=1}^{t}$ for every $t \in \mathbb{N}_+$ while minimizing the detection delay.

4.2. The algorithm

In this study, we introduce a non-parametric CPD strategy designed for situations where there is no prior knowledge about the probability measures of the data. In Euclidean spaces, this has been accomplished in particular by monitoring changes in either the mean or the variance (Costa & Rahim, 2006), or in a generalized statistics (Gretton et al., 2006) of the data stream. We propose to extend such strategies to Riemannian manifolds by monitoring changes in a generalized moment of $\boldsymbol{x}_t \in \mathcal{M}$. This generalized moment can include the Fréchet mean (Fréchet, 1948), which extends the concept of the Euclidean mean to metric spaces. In a broader sense, we consider a *generalized Fréchet mean* of $\mathcal{M}$, as defined in (Schötz, 2019):

$$\boldsymbol{m}^* \in \arg\min_{\boldsymbol{m}\in\mathcal{M}} f(\boldsymbol{m}) \,, \tag{2}$$

where $f(\boldsymbol{m})$ is given by:

$$f(\boldsymbol{m}) = \mathbb{E}_{\boldsymbol{x}\sim P(\boldsymbol{x})}\big\{c(\boldsymbol{x},\boldsymbol{m})\big\} = \int c(\boldsymbol{x},\boldsymbol{m}) dP(\boldsymbol{x}) \,,$$

with $c : \mathcal{M} \times \mathcal{M} \to [0, +\infty)$ the cost. This framework generalizes several important central values on Riemannian manifolds, including the Fréchet mean by considering $c(\boldsymbol{x},\boldsymbol{m}) = d^2_{\mathcal{M}}(\boldsymbol{x},\boldsymbol{m})$ where $d_{\mathcal{M}}(\boldsymbol{x},\boldsymbol{m})$ denotes the geodesic distance between $\boldsymbol{x}$ and $\boldsymbol{m}$, and the median by setting $c(\boldsymbol{x},\boldsymbol{m}) = d_{\mathcal{M}}(\boldsymbol{x},\boldsymbol{m})$.

The existence and uniqueness of minimizers for (2) is not guaranteed in general, even in the case of the Fréchet mean. When $c = d^2_{\mathcal{M}}$, the *Karcher mean* relaxes this definition by considering the local optima of $f(\boldsymbol{m})$ rather than only the global one. This allows us to establish existence and uniqueness conditions (Kendall, 1990) and compute $\boldsymbol{m}$ by solving (2) locally using Riemannian optimization methods (Pennec, 2004). In particular, if the support of $P(\boldsymbol{x})$ is included in a regular geodesic ball (Pennec, 2006, definition 5), then the Karcher mean exists and is unique (Kendall, 1990). This condition is satisfied for connected manifolds with non-positive curvature (Afsari, 2011), referred to as *Hadamard manifolds* (Shiga, 1984). In this work, we extend this concept by defining the *generalized Karcher mean* as the set of local minimizers of (2) with various central values. Although our framework is considered in a broader sense, we will focus on the case of Karcher mean in Section 4.3, as discussed in (Wang et al., 2023a), for the sake of convenience and to facilitate the theoretical analysis.

The proposed CPD strategy on manifolds will be designed to monitor abrupt changes in a generalized Karcher mean. An important requirement is that change points must be detected in an online manner, meaning that they are based only on past data. Consequently, we will adopt stochastic Riemannian optimization to estimate the generalized Karcher mean of the streaming data $\boldsymbol{x}_t$ in an online manner. This will constitute a fundamental component of our approach.

4.2.1. Online estimation

In a non-parametric setting, it is not possible to compute the solution to the optimization problem in (2) explicitly because $P(\boldsymbol{x})$ is unknown. Instead, we assume that we have access to observations $\boldsymbol{x}_t$ and can evaluate both the cost function $c(\boldsymbol{m}, \boldsymbol{x}_t)$ and its Riemannian gradient for any parameter $\boldsymbol{m}$ and sample $\boldsymbol{x}_t$. This enables us to construct a stochastic approximation of the gradient $\nabla f(\boldsymbol{m})$ using the input $\boldsymbol{x}_t$. Consequently, we can utilize the R-SGD algorithm (Bonnabel, 2013) to compute an online solution to (2). An update of $\boldsymbol{m}$ can be computed on $\mathcal{M}$ as:

$$\boldsymbol{m}_{t+1} = \exp_{\boldsymbol{m}_t}\big(-\alpha H(\boldsymbol{m}_t, \boldsymbol{x}_t)\big)\,, \tag{3}$$

with $\alpha > 0$ a constant step size. In this expression, $\exp_{\boldsymbol{m}}$ denotes the exponential map at $\boldsymbol{m}$, and $H(\boldsymbol{m}, \boldsymbol{x})$ is the stochastic Riemannian gradient, assumed to be an unbiased estimate of the full gradient $\nabla f(\boldsymbol{m})$,

$$\mathbb{E}_{\boldsymbol{x}\sim P(\boldsymbol{x})}\big\{H(\boldsymbol{m}, \boldsymbol{x})\big\} = \int H(\boldsymbol{m}, \boldsymbol{x}) dP(\boldsymbol{x}) = \nabla f(\boldsymbol{m})\,.$$

The exponential map in (3) can also be replaced by a computationally simpler retraction $R_{\boldsymbol{m}_t}$. It is important to note that we are considering R-SGD in a non-standard setting. The estimates provided by this algorithm should be able to adapt to changes in the data distribution and, consequently, to the underlying cost function $f(\boldsymbol{m})$. This necessitates the use of a constant (instead of diminishing) step size α, which will impact the theoretical analysis in Section 4.3 since non-asymptotic convergence results will be required.

4.2.2. An adaptive CPD

Our goal is to detect change points by monitoring abrupt changes in the value of $\boldsymbol{m}$ over time. In simpler terms, we label a time index t' as a change point if there is a sudden shift in the value of $\boldsymbol{m}$ at that time. This requires knowledge of two quantities of interest, $\boldsymbol{m}_{\text{bef}}$ and $\boldsymbol{m}_{\text{aft}}$, which respectively represent the generalized Karcher mean before and after a candidate change point t'. First, we propose an approach to compute estimates of these quantities, denoted as $\widehat{\boldsymbol{m}}_{\text{bef}}$ and $\widehat{\boldsymbol{m}}_{\text{aft}}$. Then, a test statistic is designed to compare these two quantities using the Riemannian distance, specifically $d_{\mathcal{M}}(\widehat{\boldsymbol{m}}_{\text{bef}}, \widehat{\boldsymbol{m}}_{\text{aft}})$. The larger the Riemannian distance between the generalized Karcher mean estimates before and after time instant t', the higher the likelihood of identifying t' as a change point.

The challenge is to find a computationally efficient online method for calculating $\widehat{\boldsymbol{m}}_{\text{bef}}$ and $\widehat{\boldsymbol{m}}_{\text{aft}}$. Previous work (Dubey & Müller, 2020) proposed dividing a data stream $\{\boldsymbol{x}_t\}_{t=1}^N$ with N samples into two segments, $\{1, \ldots, t'-1\}$ and $\{t', \ldots, N\}$ for every t', and testing for differences between their Karcher mean and variance. However, this approach is not suitable for processing data streams on the fly or detecting multiple change points. In (Keriven et al., 2020), estimates of $\widehat{\boldsymbol{m}}_{\text{bef}}$ and $\widehat{\boldsymbol{m}}_{\text{aft}}$ were computed considering the data $\boldsymbol{x}_t$ to belong to a Euclidean space. This was achieved using two EWMAs with different forgetting factors: one adapting quickly to track $\widehat{\boldsymbol{m}}_{\text{aft}}$ after a change point, and another adapting slowly to keep track of $\widehat{\boldsymbol{m}}_{\text{bef}}$. However, this approach cannot be directly applied to Riemannian manifolds due to its lack of accounting for manifold geometry. Instead, we propose using two iterative estimates computed using R-SGD algorithms, described in Section 4.2.1, with two different fixed step sizes $\lambda < \Lambda$. The generalized Karcher mean estimates are updated according to (3) as:

$$\boldsymbol{m}_{\lambda,t+1} = \exp_{\boldsymbol{m}_{\lambda,t}}\big(-\lambda H(\boldsymbol{m}_{\lambda,t}, \boldsymbol{x}_t)\big)\,, \tag{4}$$

$$\boldsymbol{m}_{\Lambda,t+1} = \exp_{\boldsymbol{m}_{\Lambda,t}}\big(-\Lambda H(\boldsymbol{m}_{\Lambda,t}, \boldsymbol{x}_t)\big)\,, \tag{5}$$

with initialization $\boldsymbol{m}_{\lambda,0} = \boldsymbol{m}_{\Lambda,0} = \boldsymbol{x}_0$. The convergence of the updates (4) and (5) is directly influenced by λ and Λ, with a larger step size typically resulting in faster convergence, as we will demonstrate in Theorem 4.1 in the next section. Therefore, having $0 < \lambda < \Lambda$ means that $\boldsymbol{m}_{\Lambda,t}$ is more likely to adapt to new data and quickly approximate $\widehat{\boldsymbol{m}}_{\text{aft}}$, while $\boldsymbol{m}_{\lambda,t}$ has a longer memory and is better suited for estimating $\widehat{\boldsymbol{m}}_{\text{bef}}$. Using constant step sizes is

Algorithm 1 Online CPD on Riemannian manifolds

Input: $\{\boldsymbol{x}_t\}$, step sizes λ, Λ, threshold ξ.
Initialization: $\boldsymbol{m}_{\lambda,0} = \boldsymbol{m}_{\Lambda,0} = \boldsymbol{x}_0$.
for $t = 1, 2, 3, \ldots$ **do**
  Update the generalized Karcher mean estimates $\boldsymbol{m}_{\lambda,t}$ and $\boldsymbol{m}_{\Lambda,t}$ using (4) and (5);
  Compute the test statistic $g_t = d_{\mathcal{M}}(\boldsymbol{m}_{\lambda,t}, \boldsymbol{m}_{\Lambda,t})$;
  if $g_t > \xi$ **then**
    Flag t as a change point;
  end if
end for

crucial to allow the algorithm to adapt to changes in the data distribution and detect multiple change points.

Based on the estimates provided in (4) and (5), we can formulate an adaptive CPD statistic by calculating the difference between $\boldsymbol{m}_{\lambda,t}$ and $\boldsymbol{m}_{\Lambda,t}$ using the Riemannian distance on $\mathcal{M}$ as follows:

$$g_t = d_{\mathcal{M}}(\boldsymbol{m}_{\lambda,t}, \boldsymbol{m}_{\Lambda,t})\,. \tag{6}$$

The CPD procedure involves comparing the statistic g_t to a given threshold ξ. The complete CPD procedure is outlined in Algorithm 1. It is important to note that the selection of ξ directly impacts its average run length and detection delay, as will be shown in Theorems 4.2 and 4.3, which give bounds on the probability of a false alarm and of detecting a true change point. Moreover, as in (N)EWMA methods, the time interval between change points must be sufficiently large so that the algorithms converge to obtain adequate detection and false alarm performance.

4.3. Theoretical analysis

In this section, we will assess the performance of the proposed CPD algorithm in two main aspects: i) the likelihood of a false alarm, which refers to the probability of incorrectly identifying a time step as a change point, and ii) the probability of correctly identifying a change point when there is a shift in the generalized Karcher mean of the data stream. To achieve this, we will also need a supplementary outcome, iii) the non-asymptotic convergence analysis of the R-SGD algorithm with a constant step size.

For the sake of feasibility in our theoretical analysis, we will concentrate on the Karcher mean with $c = d^2_{\mathcal{M}}$, and the R-SGD cost function $f(\boldsymbol{m}) = \mathbb{E}\big\{d^2_{\mathcal{M}}(\boldsymbol{m}, \boldsymbol{x})\big\}$, which corresponds to the Karcher variance, which is minimized using the iteration (3). However, it is important to note that our convergence analysis of R-SGD will not be limited to this particular cost function. We will also focus on Hadamard manifolds as in (Zhang & Sra, 2016). The proofs of all the results are provided in Appendix A. Before presenting the theoretical results, let us introduce some definitions related to the cost function f and its properties as follows.

Definition 1 (Geodesically strong convexity) A function $f : \mathcal{M} \to \mathbb{R}$ is geodesically μ-strongly convex if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{M}$, we have:

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \exp_{\boldsymbol{x}}^{-1}(\boldsymbol{y}) \rangle + \frac{\mu}{2} \| \exp_{\boldsymbol{x}}^{-1}(\boldsymbol{y}) \|^2\,. \tag{7}$$

Definition 2 (Lipschitz gradients) The gradient of a function $f : \mathcal{M} \to \mathbb{R}$ is said to be L-Lipschitz if, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{M}$ in the domain of f, it satisfies:

$$\big\| \nabla f(\boldsymbol{x}) - \Gamma_{\boldsymbol{y}}^{\boldsymbol{x}} \nabla f(\boldsymbol{y}) \big\| \leq L\, d_{\mathcal{M}}(\boldsymbol{x}, \boldsymbol{y})\,, \tag{8}$$

where $\Gamma_{\boldsymbol{y}}^{\boldsymbol{x}}$ denotes the parallel transport from $\boldsymbol{y}$ to $\boldsymbol{x}$.

Definition 3 (Smoothness) Any differentiable function $f : \mathcal{M} \to \mathbb{R}$ is geodesically L-smooth if its gradient is L-Lipschitz, i.e., for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{M}$, it satisfyies:

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \exp_{\boldsymbol{x}}^{-1}(\boldsymbol{y}) \rangle + \frac{L}{2} \| \exp_{\boldsymbol{x}}^{-1}(\boldsymbol{y}) \|^2\,. \tag{9}$$

4.3.1. Non-asymptotic convergence of R-SGD

The following theorem shows that the R-SGD algorithm (3) with a fixed step size $\alpha > 0$ has a curvature-dependent linear rate of convergence for geodesically strongly convex and smooth functions on Riemannian manifolds.

Theorem 4.1. *Assuming that $f : \mathcal{M} \to \mathbb{R}$ is geodesically μ-strongly convex with geodesically L-Lipschitz gradient, the diameter of the domain is bounded by D, the sectional curvature of the manifold is bounded below by κ, and the stochastic gradient is an unbiased estimator of the gradient, namely, $\mathbb{E}\big\{H(\boldsymbol{m}_t, \boldsymbol{x}_t)\big\} = \nabla f(\boldsymbol{m}_t)$ with variance $\mathbb{E}\{\|\nabla f(\boldsymbol{m}_t) - H(\boldsymbol{m}_t, \boldsymbol{x}_t)\|^2\} \leq \sigma^2$ and magnitude bounded by $\|H(\boldsymbol{m}_t, \boldsymbol{x}_t)\| < \rho$. We assume that the step size satisfies $0 < \alpha \leq \min\{\frac{1}{2L}, \frac{I}{\rho}\}$, where I is the injectivity radius of $\mathcal{M}$. Then, for any $s \in \mathbb{N}_*$, the stochastic Riemannian gradient descent algorithm satisfies:*

$$\mathbb{E}\{f(\boldsymbol{m}_s) - f(\boldsymbol{m}^*)\} \leq \frac{(1-\epsilon)^{(s-1)} D^2}{2\alpha} + \frac{\alpha \sigma^2}{2\epsilon}\,, \tag{10}$$

with $\epsilon = \min\{\frac{1}{\zeta(\kappa, D)}, \alpha\mu\}$ and $\zeta(\kappa, D) = \frac{\sqrt{|\kappa|}D}{\tanh(\sqrt{|\kappa|}D)}$.

The proof is provided in Appendix A.1, based on certain results in (Boumal, 2023) and the trigonometric distance bound, specifically, Corollary 8 in (Zhang & Sra, 2016). However, it is important to note that Theorem 4.1 differs from Theorem 14 (diminishing step sizes) and 15 (deterministic optimization) in (Zhang & Sra, 2016). In our case, we consider a stochastic optimization method with a constant step size to compute the CPD statistics g_t. If f is geodesically strongly convex and smooth and the manifold satisfies the conditions in Theorem 4.1, convergence can be guaranteed for sufficiently small step sizes α.

4.3.2. PERFORMANCE GUARANTEE

We now provide two performance guarantees of our CPD statistics g_t as defined in (6). These guarantees consist of an upper bound on the false alarm rate under the null hypothesis (i.e., when no change point has occurred) and a lower bound on the detection rate under the alternative hypothesis.

Theorem 4.2. *We assume that, under the null hypothesis* H_0, $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1}$ *are drawn i.i.d. from* $P(\boldsymbol{x})$ *with the Karcher mean* $\boldsymbol{m}^*$. *We also assume that the conditions in Theorem 4.1 on* $f(\boldsymbol{m})$, $H(\boldsymbol{m}, \boldsymbol{x}_t)$, $\mathcal{M}$ *and the step sizes* λ *and* Λ *hold. At a steady state (i.e., when* $t \to \infty$*), the false alarm rate can be upper bounded by:*

$$\mathbb{P}(g_\infty \geq \xi | \mathsf{H}_0) \leq \frac{2}{\xi}\left(f(\boldsymbol{m}^*) + \frac{(\lambda+\Lambda)\sigma^2}{4\epsilon}\right)^{\frac{1}{2}}, \quad (11)$$

with $\epsilon = \min\left\{\frac{1}{\zeta(\kappa, D)}, \lambda\mu\right\}$ *and* ξ *the detection threshold.*

The proof of this theorem is provided in Appendix A.2. Theorem 4.2 shows that when no change occurs, a higher detection threshold ξ leads to a lower upper bound on the false alarm rate. It is worth noting that the bound on the false alarm rate is influenced by the Karcher variance term, which implies that the bound will be tighter when the data distribution has lower dispersion. Smaller values of λ and Λ are also recommended for a tighter bound because they reduce the impact of gradient noise captured by σ^2. However, choosing larger detection thresholds and smaller step sizes also reduces the probability of detecting an actual change point, as indicated by the following theorem.

Theorem 4.3. *We assume that, under the alternative hypothesis* H_1, $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-B-1}$ *are drawn i.i.d. from* $P_1(\boldsymbol{x})$ *with Karcher mean* $\boldsymbol{m}_1^*$, *and* $\boldsymbol{x}_{t-B}, \boldsymbol{x}_{t-B+1}, \ldots, \boldsymbol{x}_{t-1}$ *are drawn i.i.d. from* $P_2(\boldsymbol{x})$ *with Karcher mean* $\boldsymbol{m}_2^*$. *We also assume that the conditions in Theorem 4.1 on* $f(\boldsymbol{m})$, $H(\boldsymbol{m}, \boldsymbol{x}_t)$, *the manifold* $\mathcal{M}$ *and the step sizes* λ *and* Λ *hold, and that* t *is sufficiently large such that the algorithms converged before the change point. Then, the detection rate can be lower bounded as:*

$$\mathbb{P}(g_t > \xi | \mathsf{H}_1) \geq \frac{d_{\mathcal{M}}(\boldsymbol{m}_1^*, \boldsymbol{m}_2^*) - \psi(\lambda) - \phi(\Lambda) - \xi}{D - \xi}, \quad (12)$$

$$\textit{where} \quad \psi(\lambda) = \left(2f_{\text{bef}}(\boldsymbol{m}_1^*) + \frac{\lambda\sigma^2}{\epsilon}\right)^{\frac{1}{2}} + \lambda\rho B,$$

$$\phi(\Lambda) = \left(2f_{\text{aft}}(\boldsymbol{m}_2^*) + \frac{(1-\epsilon)^B D^2}{\Lambda} + \frac{\Lambda\sigma^2}{\epsilon}\right)^{\frac{1}{2}},$$

with $f_{\text{bef}}(\boldsymbol{m}_1^*) = \min_{\boldsymbol{m}\in\mathcal{M}} \mathbb{E}_{\boldsymbol{x}\sim P_1(\boldsymbol{x})}\{d_{\mathcal{M}}^2(\boldsymbol{m}, \boldsymbol{x})\}$ *and* $f_{\text{aft}}(\boldsymbol{m}_2^*) = \min_{\boldsymbol{m}\in\mathcal{M}} \mathbb{E}_{\boldsymbol{x}\sim P_2(\boldsymbol{x})}\{d_{\mathcal{M}}^2(\boldsymbol{m}, \boldsymbol{x})\}$ *the Karcher variances of the data before and after the change point.*

The proof of this theorem is provided in Appendix A.3. Theorem 4.3 shows that smaller values of ξ and larger values of $d_{\mathcal{M}}(\boldsymbol{m}_1^*, \boldsymbol{m}_2^*)$ make the lower bound on the detection rate tighter when a change point occurs. Moreover, the bound also gets tighter as λ gets smaller and Λ gets bigger, which is intuitive since, when B is not too large, a small λ assures $\boldsymbol{m}_{\lambda,t}$ will still be close to the Karcher mean of the data before the change point, whereas a large Λ means that $\boldsymbol{m}_{\Lambda,t}$ will converge faster to the Karcher mean of the data after the change point, their distance being thus more effective for change detection. However, one should note that λ being too small can hurt the adaptability of the method and its capability to detect multiple change points. Thus, the step sizes should be selected to ensure a sufficiently fast speed of convergence for the desired application.

The increase in the number of samples B after a change point has a twofold effect on the lower bound to the detection rate in (12). On the one hand, the estimate of the Karcher means before the change point from the "slow" algorithm gets polluted by samples following the post-change distribution, causing the term $\psi(\lambda)$ to *increase* with B. On the other hand, the "fast" algorithm will converge to the Karcher means of the post-change data, causing the term $\phi(\Lambda)$ to *decrease* with B. The bound also gets larger as the Karcher variances of the data, the gradient noise, and the bound on the diameter of the domain decrease. Note that these quantities are the main sources of stochasticity in the proposed algorithm, and as the uncertainty decreases the theoretical detection performance of the algorithm improves. A similar behavior is also observed for the upper bound to the false alarm rate in (11).

One challenge in applying CPD algorithms is the selection of the detection threshold ξ without prior knowledge of the data distribution. In real use cases, a simple yet effective procedure is to adjust ξ so as to achieve some desired performance in the absence of change points. We provide an heuristic procedure for the adaptive threshold selection based on this idea in Appendix B.

5. Application to specific manifolds

In this section, we tailor Algorithm 1 to two common instances of Riemannian manifolds for the case of the Karcher mean cost function $c = d_{\mathcal{M}}^2$, which will later be illustrated through numerical experiments in Section 6. The first one is the manifold of $p \times p$ SPD matrices, denoted by $\mathcal{S}_p^{++}$. The second is the Grassmann manifold, a set of k-dimensional linear subspaces of $\mathbb{R}^p$, denoted by $\mathcal{G}_p^k$. We refer the interested reader to (Boumal, 2023; Collas, 2022) for more details. Note that although $\mathcal{G}_p^k$ is not a Hadamard manifold, Algorithm 1 still performs empirically well as will be presented in Section 6. In practice, these manifolds can appear as natural representations of the data (e.g., in diffusion tensor imaging) or as feature embeddings thereof. For computational simplicity, we will replace the exponential

maps in the R-SGD updates (4) and (5) with approximate retractions $R_{\boldsymbol{m}_{\lambda,t}}$ and $R_{\boldsymbol{m}_{\Lambda,t}}$ as in (Bonnabel, 2013).

5.1. The manifold of SPD matrices

The manifold $\mathcal{S}_p^{++}$ consists of the set of SPD matrices endowed with an appropriate metric. When considering the affine invariant metric, the geodesic distance between two SPD matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_t \in \mathcal{S}_p^{++}$ can be computed as (Pennec et al., 2006):

$$d_{\mathcal{S}_p^{++}}(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_t) = \left\| \log(\boldsymbol{\Sigma}_t^{-\frac{1}{2}} \boldsymbol{\Sigma} \boldsymbol{\Sigma}_t^{-\frac{1}{2}}) \right\|_F , \tag{13}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. In this case, the Riemannian gradient $H(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_t)$ of the loss function $d^2_{\mathcal{S}_p^{++}}(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_t)$ at $\boldsymbol{\Sigma} \in \mathcal{S}_p^{++}$ is obtained by applying the transformation $\frac{1}{2}\boldsymbol{\Sigma}\left(\boldsymbol{G}^T + \boldsymbol{G}\right)\boldsymbol{\Sigma}$ to its Euclidean gradient $\boldsymbol{G}$ (Bhatia, 2009), which gives us:

$$H(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}_t) = 2\log(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_t^{-1})\boldsymbol{\Sigma} . \tag{14}$$

Finally, a second-order retraction on $\mathcal{S}_p^{++}$ is given by:

$$R_{\boldsymbol{\Sigma}, \mathcal{S}_p^{++}}(\boldsymbol{\xi}) = \boldsymbol{\Sigma} + \boldsymbol{\xi} + \frac{1}{2}\boldsymbol{\xi}\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi} . \tag{15}$$

With $\{\boldsymbol{\Sigma}_t\}_{t\in\mathbb{N}}$ lying in $\mathcal{S}_p^{++}$ and the metric defined in (13), the Karcher means were estimated by minimizing the following objective function $f(\boldsymbol{\Sigma}) = \mathbb{E}_{\boldsymbol{\Sigma}_t \sim P(\boldsymbol{\Sigma})}\left\{ \left\| \log(\boldsymbol{\Sigma}_t^{-\frac{1}{2}} \boldsymbol{\Sigma} \boldsymbol{\Sigma}_t^{-\frac{1}{2}}) \right\|_F^2 \right\}$. Note this cost function is known to be geodecially strong convex and smooth as discussed in (Zhang & Sra, 2016). The R-SGD algorithms in (4) and (5) with the stochastic gradient (14) and the retraction (15) were used to compute the online CPD statistic in (6).

5.2. The Grassmann manifold

We consider the Grassmann manifold $\mathcal{G}_p^k$ endowed with the canonical metric. The Grassmann manifold is typically characterized as a smooth quotient of the Stiefel manifold $\mathcal{S}_p^k = \{\boldsymbol{U} \in \mathbb{R}^{p\times k} : \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_k\}$. This way, by defining the surjective map $\pi : \mathcal{S}_p^k \to \mathcal{G}_p^k$ as follows: $\pi(\boldsymbol{U}) = \{\boldsymbol{U}\boldsymbol{O} : \boldsymbol{O} \in \mathbb{R}^{k\times k}, \boldsymbol{O}^T\boldsymbol{O} = \boldsymbol{I}_k\}$, every point $\pi(\boldsymbol{U}) \in \mathcal{G}_p^k$ can be equivalently represented by the orthonormal matrix $\boldsymbol{U}$ whose columns form its basis. We spare the reader of the technical details, which can be found in (Absil et al., 2009; Boumal, 2023). To proceed, let us first denote by $\boldsymbol{V}_1 \operatorname{diag}(\boldsymbol{\theta}_t)\boldsymbol{V}_2^T$ the singular value decomposition (SVD) of $\boldsymbol{U}^T\boldsymbol{U}_t$. The geodesic distance between $\pi(\boldsymbol{U}) \in \mathcal{G}_p^k$ and $\pi(\boldsymbol{U}_t) \in \mathcal{G}_p^k$ can be defined as (Edelman et al., 1998):

$$d_{\mathcal{G}_p^k}(\boldsymbol{U}, \boldsymbol{U}_t) = \|\cos^{-1}(\boldsymbol{\theta}_t)\|_2 . \tag{16}$$

The Riemmanian gradient $H(\boldsymbol{U}, \boldsymbol{U}_t)$ of the loss function $d^2_{\mathcal{G}_p^k}(\boldsymbol{U}, \boldsymbol{U}_t)$ at $\pi(\boldsymbol{U}) \in \mathcal{G}_p^k$ can be computed by applying the transformation $(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^T)\boldsymbol{G}$ to its Euclidean gradient $\boldsymbol{G}$. Using results from matrix calculus, this results in:

$$H(\boldsymbol{U}, \boldsymbol{U}_t) = \\ -(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^T)\boldsymbol{U}_t\boldsymbol{V}_2 \operatorname{diag}\left(2\left(1-\boldsymbol{\theta}_t^2\right)^{-\frac{1}{2}}\right)\boldsymbol{V}_1^T . \tag{17}$$

Let $\boldsymbol{\xi} \in T_{\pi(\boldsymbol{U})}\mathcal{G}_p^k$, and let $\boldsymbol{X}\boldsymbol{\Upsilon}\boldsymbol{Y}^T = \boldsymbol{U} + \boldsymbol{\xi}$ be the thin SVD of $\boldsymbol{U} + \boldsymbol{\xi} \in \mathbb{R}^{p\times k}$. A second-order retraction on the Grassmann manifold is given by (Boumal, 2023)

$$R_{\pi(\boldsymbol{U})}(\boldsymbol{\xi}) = \pi\left(\boldsymbol{X}\boldsymbol{Y}^T\right) . \tag{18}$$

With $\{\pi(\boldsymbol{U}_t)\}_{t\in\mathbb{N}}$ lying in $\mathcal{G}_p^k$ and the metric defined in (16), the Karcher means were estimated by minimizing the objective function $f(\pi(\boldsymbol{U})) = \mathbb{E}_{\pi(\boldsymbol{U}_t)\sim P(\pi(\boldsymbol{U}))}\left\{ \|\cos^{-1}(\boldsymbol{\theta}_t)\|_2^2 \right\}$. Accordingly, the R-SGD algorithms in (4) and (5) with the stochastic gradient (17) and the retraction (18) were used to compute the online CPD statistic in (6).

6. Experiments

In this section, we present numerical experiments using the manifolds $\mathcal{S}_p^{++}$ and $\mathcal{G}_p^k$ discussed in Section 5. Our method was implemented in Python using Pymanopt (Townsend et al., 2016). The step sizes of our method were set as $\lambda = 0.01$ and $\Lambda = 0.02$. Open-source code to reproduce the results is publicly available at `https://github.com/xiuheng-wang/CPD_manifold_release`. Here we briefly describe the baselines and evaluation metrics.

Baselines: We selected four CPD methods Scan-B (Li et al., 2019), NEWMA (Keriven et al., 2020), the Fréchet CPD (F-CPD) (Dubey & Müller, 2020) and NODE (Wang et al., 2023b) as baselines for comparison with our method. Scan-B, NEWMA, and NODE are online algorithms originally designed for Euclidean spaces but were adapted to the manifold setting in this study. We applied Scan-B, NEWMA, and NODE to the vectorization of the lower triangular portion of each SPD matrix $\boldsymbol{\Sigma}_t$ and to each entire matrix $\boldsymbol{U}_t$ for the SPD and Grassmann manifolds, respectively. In Scan-B, the number of reference blocks was set to 3. NEWMA was implemented with Random Fourier features using the Gaussian kernel. The window size of Scan-B and NEWMA were both set to 50. The reference and test window lengths of NODE were both set to 64. F-CPD was designed to operate on manifolds but can only detect a single change point and operates offline. To address these limitations, we computed statistics in F-CPD to compare data distributions in two consecutive sliding windows, each with 64 samples. We provide a discussion on the computational complexity of our method compared to baselines in Appendix C.

Metrics: To evaluate the performance of the methods, we considered three metrics: the Average Run Length (ARL),

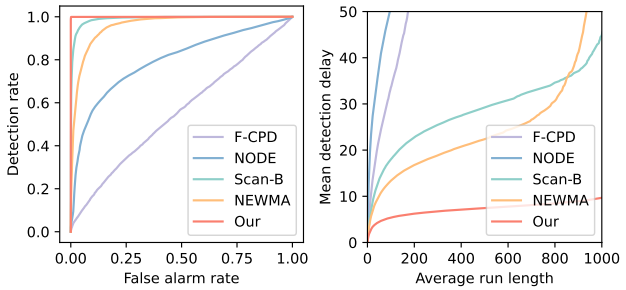

Figure 1. ROC curves, ARL versus MDD for the compared algorithms on synthetic data on $\mathcal{S}_p^{++}$.

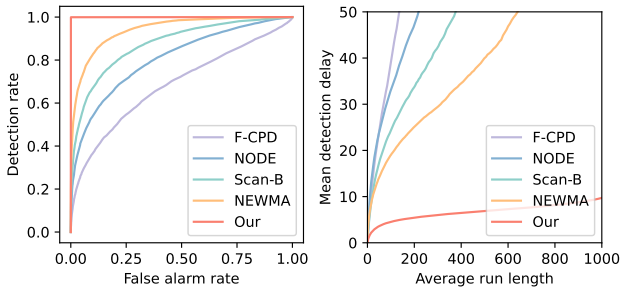

Figure 2. ROC curves, ARL versus MDD for the compared algorithms on synthetic data on $\mathcal{G}_p^k$.

Mean Detection Delay (MDD), and Receiver Operating Characteristic (ROC) curves. ARL represents the expected time before incorrectly announcing a change point when none has occurred, and is related to the false alarm rate. MDD signifies the expected time the algorithm needs to flag a detection after a change point occurs, reflecting its sensitivity. The ROC curve is a graphical representation of the detection rate versus the false alarm rate.

6.1. Validations on synthetic data

We first present results over sequences of i.i.d. synthetically generated data in $\mathcal{S}_p^{++}$ and $\mathcal{G}_p^k$.

Manifold $\mathcal{S}_p^{++}$: We sampled matrices $\boldsymbol{\Sigma}_t \in \mathcal{S}_p^{++}$ with $p = 8$ from a Wishart distribution with a randomly generated scaling matrix $\boldsymbol{V}$ and $p + 2$ degrees of freedom. We generated 2000 samples and set a change point at $t_r = 1500$ where we reset $\boldsymbol{V}$.

Manifold $\mathcal{G}_p^k$: The data $\pi(\boldsymbol{U}_t) \in \mathcal{G}_p^k$ with $p = 15$, $k = 5$ was generated in two steps. First, we generated matrices $\boldsymbol{Z}_t$ following a matrix Gaussian distribution (Gupta & Nagar, 1999) with random mean and row/column covariance matrices. Then, the orthonormal matrices $\boldsymbol{U}_t$ were generated as the left singular vectors corresponding to the k largest singular values of $\boldsymbol{Z}_t$. We generated 2000 samples and set a change point at $t_r = 1500$ where we reset the mean of the matrix Gaussian distribution of $\boldsymbol{Z}_t$.

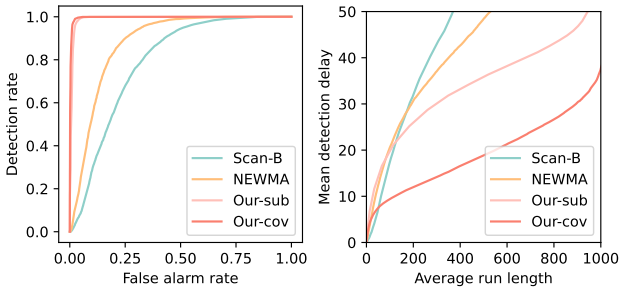

Figure 3. ROC curves, ARL versus MDD for the compared algorithms on real data for voice activity detection.

Results: The ROCs, MDD as a function of ARL for all methods, averaged over 10^4 Monte Carlo runs, are depicted in Figure 1 and 2 for both manifolds. It is evident that the proposed method results in a significantly lower detection delay for a fixed ARL when compared to Euclidean methods Scan-B, NEWMA, and NODE, which does not consider manifold geometry, and F-CPD, which does not benefit from long time series through a recursive operation. The compared methods exhibited similar behavior in both manifolds, although the proposed method resulted in slightly lower MDDs for $\mathcal{G}_p^k$. This underscores the importance of accounting for manifold geometry and utilizing an efficient online estimation framework. Illustrations of the mean and standard deviation and further comparisons between histograms of the test statistics for all compared methods are provided in Appendix D.1 and D.2, respectively.

6.2. Voice activity detection

We now present results on real data on both $\mathcal{S}_p^{++}$ and $\mathcal{G}_p^k$ by considering the task of voice activity detection on audio signals. We first added 4 seconds of real speech extracted from the TIMIT database (Garofolo, 1993) to 15 seconds of background noises in real street environments from the QUT-NOISE database (Dean et al., 2010), with -3 dB Signal-to-Noise Ratio. The goal is to detect the speech segments in the noise background. Then, we used the Short Time Fourier Transform (STFT) (Cohen, 1995) on a one-dimensional audio signal to extract on-the-fly frequency information and form a $d = 128$ dimensional time series $\boldsymbol{s}_t \in \mathbb{R}^d$. The two methods with the best performance in the experiments with synthetic data, Scan-B and NEWMA, were used as baselines in this experiment. They were directly applied on $\boldsymbol{s}_t$ as they are designed to operate on Euclidean spaces.

Manifold $\mathcal{S}_p^{++}$: We averaged the neighboring channels of $\boldsymbol{s}_t$ in the frequency domain to obtain its down-sampled version with 16 channels. We then generated data points $\boldsymbol{\Sigma}_t \in \mathcal{S}_p^{++}$ with $p = 16$ by computing the covariance matrices in sliding windows, each with 32 samples. The proposed method on such covariance descriptors is denoted as "Our-cov".

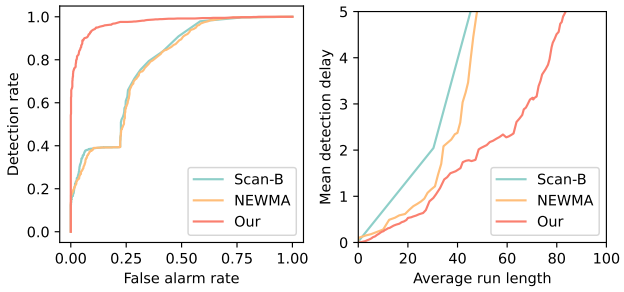

Figure 4. ROC curves, ARL versus MDD for the compared algorithms on real data for skeleton-based action recognition.

Manifold $\mathcal{G}_p^k$: We also applied the truncated SVD with $k = 1$ singular values to the samples in the same sliding windows to obtain orthonormal matrices $\boldsymbol{U}_t$ defining the subspaces $\pi(\boldsymbol{U}_t) \in \mathcal{G}_p^k$. We denote our method on these subspaces as "Our-sub".

Results: The ROCs and MDD as a function of ARL for all methods, averaged over 10^4 Monte Carlo runs, are depicted in Figure 3. It is important to note that the problem setting is challenging due to the complexity of real acoustic signals and the non-i.i.d. nature of the extracted features. Nevertheless, one can observe that the proposed strategy exhibits a higher detection rate for a given false alarm rate and better performance on MDD versus ARL when compared to both Scan-B and NEWMA, except for very small ARLs where Scan-B has a lower MDD. This behavior occurs since both the covariance and subspace descriptors are computed over a sliding window, which introduces a small detection delay in our method when the ARL is small[1]. However, its performance is significantly better for larger ARLs. This illustrates the superior performance of our method. Furthermore, the performance was slightly superior in the covariance descriptors on $\mathcal{S}_p^{++}$ compared to the subspace representations on $\mathcal{G}_p^k$.

6.3. Skeleton-based action recognition

We also present results on real data on the $\mathcal{S}_p^{++}$ manifold by considering the problem of detecting change points in skeleton-based action recognition using the HDM05 motion capture database (Müller et al., 2007). In this database, we identified action categories and preprocessed the data as described in (Huang & Van Gool, 2017) to generate data points $\boldsymbol{\Sigma}_t \in \mathcal{S}_p^{++}$ with $p = 93$ by computing the joint covariance descriptor (Hussein et al., 2013) of 3D coordinates of the 31 joints. The aim is to flag a change point at the border of two different action categories. We randomly selected the sequences corresponding to action categories containing more than 200 samples and then concatenated them. The parameters of the compared algorithms were appropriately re-adjusted for this example since there were fewer samples between change points, requiring a faster convergence.

Results: The ROC and MDD versus ARL curves of the compared methods (Scan-B, NEWMA, and our algorithm), averaged over 10^3 Monte Carlo runs, can be seen in Figure 4. Note that the problem setting is challenging due to the high data dimension. It can be seen that our method achieves a significantly higher detection rate compared to the Scan-B and NEWMA, which had very similar ROCs[2]. Moreover, for ARLs smaller than 40 samples, the proposed method and NEWMA obtained similar MDDs. However, when the ARL was higher, the proposed method performed significantly better. This further illustrates the effectiveness of our method.

7. Discussion

This paper presented a general approach for non-parametric online CPD on Riemannian manifolds. An adaptive test statistic was computed using stochastic Riemannian optimization to monitor the generalized Karcher mean of data streams. Performance guarantees for detection and false alarm rates were established based on a theoretical analysis of the non-asymptotic convergence of the R-SGD algorithm. Experimental results on the manifold of SPD matrices and the Grassmann manifold demonstrated the superiority of the proposed algorithm on synthetic and real-world datasets. We also identify the main limitations of our work:

- The number of samples needs to be large enough for the "slow" algorithm to converge to the Karcher mean of the data before a new change point occurs to perform well. This is a limitation of our method and also of other recursive algorithms.
- Although $\mathcal{S}_p^{++}$ and $\mathcal{G}_p^k$ are selected to illustrate our approach, our framework is more general and can indeed be applied to other manifolds. The main possible hurdle is related to the convergence rate of the R-SGD algorithm affected by the manifold curvature, being slower for higher curvature values. This in turn can negatively impact the detection delay.
- For the theoretical analysis, we make additional assumptions on the manifold (e.g., Hadamard). Although this does not limit the practical applicability to other manifolds, manifolds with complex geometries can introduce additional challenges such as non-convexity of the cost function.

[1] Although a shorter sliding window is preferred to introduce a smaller delay, the window length has to be long enough to provide enough samples for an accurate estimation of these statistical descriptors.

[2] The mean and standard deviation of the test statistic of all methods for this example can also be seen in Appendix D.1.

Impact statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

The authors would like to thank the reviewers for their constructive feedback. The work of Cédric Richard was supported in part by the French Government through the 3IA Côte d'Azur Investments in the Future Project under grant ANR-19-P3IA-0002, and in part by grant ANR-19-CE48-0002. The work of Ricardo Borsoi was supported in part by the French National Research Agency, under grants ANR-23-CE23-0024, ANR-23-CE94-0001, and by the National Science Foundation, under grant NSF 2316420.

References

Absil, P.-A., Mahony, R., and Sepulchre, R. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

Afsari, B. Riemannian ℓ^p center of mass: existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society*, 139(2):655–673, 2011.

Aminikhanghahi, S. and Cook, D. J. A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367, 2017.

Arlot, S., Celisse, A., and Harchaoui, Z. A kernel multiple change-point algorithm via model selection. *Journal of machine learning research*, 20(162):1–56, 2019.

Bai, J. and Perron, P. Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1): 47–78, 1998.

Bhatia, R. *Positive definite matrices*. Princeton university press, 2009.

Bonnabel, S. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58 (9):2217–2229, 2013.

Bouchard, F., Mian, A., Zhou, J., Said, S., Ginolhac, G., and Berthoumieu, Y. Riemannian geometry for compound gaussian distributions: Application to recursive change detection. *Signal Processing*, 176:107716, 2020.

Boumal, N. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

Chen, L., Keilbar, G., and Wu, W. B. Recursive quantile estimation: Non-asymptotic confidence bounds. *Journal of Machine Learning Research*, 24(91):1–25, 2023.

Cheng, X. and Xie, Y. Neural tangent kernel maximum mean discrepancy. *Advances in Neural Information Processing Systems*, 34:6658–6670, 2021.

Cohen, L. *Time-frequency analysis*, volume 778. Prentice hall New Jersey, 1995.

Collas, A. *Riemannian geometry for statistical estimation and learning: application to remote sensing*. PhD thesis, université Paris-Saclay, 2022.

Costa, A. and Rahim, M. A single EWMA chart for monitoring process mean and process variance. *Quality Technology & Quantitative Management*, 3(3):295–305, 2006.

Dean, D., Sridharan, S., Vogt, R., and Mason, M. The qut-noise-timit corpus for evaluation of voice activity detection algorithms. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association*, pp. 3110–3113. International Speech Communication Association, 2010.

Duan, X., Sun, H., and Zhao, X. A matrix information-geometric method for change-point detection of rigid body motion. *Entropy*, 21(5):531, 2019.

Dubey, P. and Müller, H.-G. Fréchet change-point detection. *The Annals of Statistics*, 48(6):3312–3335, 2020.

Edelman, A., Arias, T. A., and Smith, S. T. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Ferrari, A. and Richard, C. Non-parametric community change-points detection in streaming graph signals. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5545–5549. IEEE, 2020.

Ferrari, A., Richard, C., Bourrier, A., and Bouchikhi, I. Online change-point detection with kernels. *Pattern Recognition*, pp. 109022, 2022.

Fréchet, M. Les éléments aléatoires de nature quelconque dans un espace distancié. In *Annales de l'institut Henri Poincaré*, volume 10, pp. 215–310, 1948.

Garofolo, J. S. Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium, 1993*, 1993.

Gong, D., Medioni, G., Zhu, S., and Zhao, X. Kernelized temporal cut for online temporal segmentation and recognition. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III 12*, pp. 229–243. Springer, 2012.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006.

Gupta, A. K. and Nagar, D. K. *Matrix variate distributions*, volume 104. CRC Press, 1999.

Gustafsson, F. The marginalized likelihood ratio test for detecting abrupt changes. *IEEE Transactions on automatic control*, 41(1):66–78, 1996.

Harchaoui, Z., Moulines, E., and Bach, F. Kernel change-point analysis. *Advances in neural information processing systems*, 21, 2008.

Huang, W., Gallivan, K. A., and Absil, P.-A. A Broyden class of quasi-Newton methods for Riemannian optimization. *SIAM Journal on Optimization*, 25(3):1660–1685, 2015.

Huang, Z. and Van Gool, L. A Riemannian network for SPD matrix learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Hussein, M. E., Torki, M., Gowayyed, M. A., and El-Saban, M. Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. In *Twenty-third international joint conference on artificial intelligence*, 2013.

Ienco, D., Bifet, A., Pfahringer, B., and Poncelet, P. Change detection in categorical evolving data streams. In *29th annual ACM symposium on applied computing*, pp. 792–797, 2014.

Kacem, A., Daoudi, M., Amor, B. B., Berretti, S., and Alvarez-Paiva, J. C. A novel geometric framework on gram matrix trajectories for human behavior understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(1):1–14, 2018.

Kendall, W. S. Probability, convexity, and harmonic maps with small image I: uniqueness and fine existence. *Proceedings of the London Mathematical Society*, 3(2):371–406, 1990.

Keriven, N., Garreau, D., and Poli, I. NEWMA: a new method for scalable model-free online change-point detection. *IEEE Transactions on Signal Processing*, 68: 3515–3528, 2020.

Li, S., Xie, Y., Dai, H., and Song, L. Scan b-statistic for kernel change-point detection. *Sequential Analysis*, 38 (4):503–544, 2019.

Madrid Padilla, C. M., Xu, H., Wang, D., Madrid Padilla, O. H., and Yu, Y. Change point detection and inference in multivariate non-parametric models under mixing conditions. *Advances in Neural Information Processing Systems*, 36, 2023.

Müller, M., Röder, T., Clausen, M., Eberhardt, B., Krüger, B., and Weber, A. Documentation mocap database HDM05. Technical Report CG-2007-2, Universität Bonn, June 2007.

Page, E. S. Continuous inspection schemes. *Biometrika*, 41 (1/2):100–115, 1954.

Pennec, X. Probabilities and statistics on Riemannian manifolds: A geometric approach. Technical Report 5093, INRIA, 2004.

Pennec, X. Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25:127–154, 2006.

Pennec, X., Fillard, P., and Ayache, N. A Riemannian framework for tensor computing. *International Journal of computer vision*, 66(1):41–66, 2006.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Reeves, J., Chen, J., Wang, X., Lund, R., and Lu, Q. Q. A review and comparison of changepoint detection techniques for climate data. *Journal of applied meteorology and climatology*, 46(6):900 – 915, 2007. ISSN 1558-8424.

Schötz, C. Convergence rates for the generalized Fréchet mean via the quadruple inequality. *Electronic Journal of Statistics*, 13:4280–4345, 2019.

Shiga, K. Hadamard manifolds. *Geometry of Geodesics and Related Topics*, 3:239–282, 1984.

Sinn, M., Ghodsi, A., and Keller, K. Detecting change-points in time series by maximum mean discrepancy of ordinal pattern distributions. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pp. 786–794, 2012.

Smith, S. Optimization techniques on Riemannian manifolds. *Fields Institute Communications*, 3, 1994.

Tartakovsky, A., Nikiforov, I., and Basseville, M. *Sequential analysis: Hypothesis testing and changepoint detection*. CRC press, 2014.

Townsend, J., Koep, N., and Weichwald, S. Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation. *Journal of Machine Learning Research*, 17(137):1–5, 2016. URL http://jmlr.org/papers/v17/16-177.html.

Truong, C., Oudre, L., and Vayatis, N. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.

Tuzel, O., Porikli, F., and Meer, P. Pedestrian detection via classification on Riemannian manifolds. *IEEE transactions on pattern analysis and machine intelligence*, 30 (10):1713–1727, 2008.

Wang, X., Borsoi, R., and Richard, C. Online change point detection on Riemannian manifolds with Karcher mean estimates. In *IEEE European Signal Processing Conference (EUSIPCO)*, 2023a.

Wang, X., Borsoi, R. A., Richard, C., and Chen, J. Change point detection with neural online density-ratio estimator. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2023b.

Wei, S. and Xie, Y. Online kernel CUSUM for change-point detection. *arXiv preprint arXiv:2211.15070*, 2022.

Zeng, L., Wardlow, B. D., Xiang, D., Hu, S., and Li, D. A review of vegetation phenological metrics extraction using time-series, multispectral satellite data. *Remote Sensing of Environment*, 237:111511, 2020.

Zhang, H. and Sra, S. First-order methods for geodesically convex optimization. In *Conference on Learning Theory*, pp. 1617–1638, 2016.

Zhang, H., Reddi, S. J., and Sra, S. Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. In *Advances in Neural Information Processing Systems*, pp. 4592–4600, 2016.

Zhang, J., Zhang, H., and Sra, S. R-spider: A fast Riemannian stochastic optimization algorithm with curvature independent rate. *arXiv preprint arXiv:1811.04194*, 2018.

Zhou, P., Yuan, X., Yan, S., and Feng, J. Faster first-order methods for stochastic non-convex optimization on Riemannian manifolds. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):459–472, 2019.

A. Proofs of the main results

A.1. Theorem 4.1

Assume f is a geodesically L-smooth function, that is, its gradient is geodesically L-Lipschitz. As this property is related to deterministic gradient $\nabla f(x)$, we shall first reformulate it with respect to the stochastic gradient. Replacing $y = \boldsymbol{m}_{t+1}$, $x = \boldsymbol{m}_t$ in (9), denote $\Delta_t = f(\boldsymbol{m}_t) - f(\boldsymbol{m}^*)$, and considering the fact $\langle a, b\rangle \leq \frac{\alpha}{2}a^2 + \frac{1}{2\alpha}b^2$, we have:

$$\begin{aligned}
\Delta_{t+1} - \Delta_t &= f(\boldsymbol{m}_{t+1}) - f(\boldsymbol{m}_t) \\
&\leq \langle \nabla f(\boldsymbol{m}_t), \exp_{\boldsymbol{m}_t}^{-1}(\boldsymbol{m}_{t+1})\rangle + \frac{L}{2}\|\exp_{\boldsymbol{m}_t}^{-1}(\boldsymbol{m}_{t+1})\|^2 \\
&= \langle H(\boldsymbol{m}_t, \boldsymbol{x}_t), \exp_{\boldsymbol{m}_t}^{-1}(\boldsymbol{m}_{t+1})\rangle + \langle \nabla f(\boldsymbol{m}_t) - H(\boldsymbol{m}_t, \boldsymbol{x}_t), \exp_{\boldsymbol{m}_t}^{-1}(\boldsymbol{m}_{t+1})\rangle + \frac{L}{2}\|\exp_{\boldsymbol{m}_t}^{-1}(\boldsymbol{m}_{t+1})\|^2 \\
&\leq \langle H(\boldsymbol{m}_t, \boldsymbol{x}_t), \exp_{\boldsymbol{m}_t}^{-1}(\boldsymbol{m}_{t+1})\rangle + \frac{\alpha}{2}\|\nabla f(\boldsymbol{m}_t) - H(\boldsymbol{m}_t, \boldsymbol{x}_t)\|^2 + \left(\frac{L}{2} + \frac{1}{2\alpha}\right)\|\exp_{\boldsymbol{m}_t}^{-1}(\boldsymbol{m}_{t+1})\|^2. \qquad (19)
\end{aligned}$$

Assuming $\|H(\boldsymbol{m}_t, \boldsymbol{x}_t)\| < \rho$ and $0 < \alpha \leq \frac{I}{\rho}$ where I is the injectivity radius of $\mathcal{M}$, we have $\|\alpha H(\boldsymbol{m}_t, \boldsymbol{x}_t)\| < I$. By Proposition 10.22 of (Boumal, 2023), $\exp_{\boldsymbol{m}_t}^{-1}(\boldsymbol{m}_{t+1}) = \exp_{\boldsymbol{m}_t}^{-1}(\exp_{\boldsymbol{m}_t}(-\alpha H(\boldsymbol{m}_t, \boldsymbol{x}_t)) = -\alpha H(\boldsymbol{m}_t, \boldsymbol{x}_t)$, taking the expectation w.r.t. $\{\boldsymbol{x}_s\}_{s=0}^t$, one obtains:

$$\begin{aligned}
\mathbb{E}\Delta_{t+1} - \mathbb{E}\Delta_t &\leq -\alpha\mathbb{E}\|H(\boldsymbol{m}_t, \boldsymbol{x}_t)\|^2 + \frac{\alpha\sigma^2}{2} + \left(\frac{L}{2} + \frac{1}{2\alpha}\right)\alpha^2\mathbb{E}\|H(\boldsymbol{m}_t, \boldsymbol{x}_t)\|^2 \\
&= \frac{\alpha\sigma^2}{2} + \Big(\frac{\alpha L + 1}{2} - 1\Big)\alpha\mathbb{E}\|H(\boldsymbol{m}_t, \boldsymbol{x}_t)\|^2. \qquad (20)
\end{aligned}$$

Assuming $0 \leq \alpha \leq \frac{1}{2L}$, we have:

$$\mathbb{E}\Delta_{t+1} - \mathbb{E}\Delta_t \leq \frac{\alpha\sigma^2}{2} - \frac{\alpha}{4}\mathbb{E}\|H(\boldsymbol{m}_t, \boldsymbol{x}_t)\|^2. \qquad (21)$$

Assume f is a geodesically μ-strongly convex function, replacing $y = \boldsymbol{m}^*$, $x = \boldsymbol{m}_t$ in (7), we have:

$$\begin{aligned}
f(\boldsymbol{m}_t) - f(\boldsymbol{m}^*) &\leq \langle -\nabla f(\boldsymbol{m}_t), \exp_{\boldsymbol{m}_t}^{-1}(\boldsymbol{m}^*)\rangle - \frac{\mu}{2}\|\exp_{\boldsymbol{m}_t}^{-1}(\boldsymbol{m}^*)\|^2 \\
&= \langle -\nabla f(\boldsymbol{m}_t), \exp_{\boldsymbol{m}_t}^{-1}(\boldsymbol{m}^*)\rangle - \frac{\mu}{2}d_{\mathcal{M}}^2(\boldsymbol{m}_t, \boldsymbol{m}^*)\,. \qquad (22)
\end{aligned}$$

Assume the diameter of the domain is bounded above by D, and the sectional curvature lower-bounded by $\kappa < 0$, use the trigonometric distance bound, i.e., Corollary 8 in (Zhang & Sra, 2016), we have:

$$\langle -\nabla f(\boldsymbol{m}_t), \exp_{\boldsymbol{m}_t}^{-1}(\boldsymbol{m}^*)\rangle \leq \frac{1}{2\alpha}\big(d_{\mathcal{M}}^2(\boldsymbol{m}_t, \boldsymbol{m}^*) - d_{\mathcal{M}}^2(\boldsymbol{m}_{t+1}, \boldsymbol{m}^*)\big) + \frac{\zeta(\kappa, D)\alpha}{2}\|\nabla f(\boldsymbol{m}_t)\|^2. \qquad (23)$$

Combining (22) and (23), we have:

$$\begin{aligned}
\mathbb{E}\Delta_t = \mathbb{E}\{f(\boldsymbol{m}_t) - f(\boldsymbol{m}^*)\} &\leq \Big(\frac{1 - \alpha\mu}{2\alpha}\Big)\mathbb{E}d_{\mathcal{M}}^2(\boldsymbol{m}_t, \boldsymbol{m}^*) - \frac{1}{2\alpha}\mathbb{E}d_{\mathcal{M}}^2(\boldsymbol{m}_{t+1}, \boldsymbol{m}^*) + \frac{\zeta(\kappa, D)\alpha}{2}\mathbb{E}\|\nabla f(\boldsymbol{m}_t)\|^2 \\
&\leq \Big(\frac{1 - \alpha\mu}{2\alpha}\Big)\mathbb{E}d_{\mathcal{M}}^2(\boldsymbol{m}_t, \boldsymbol{m}^*) - \frac{1}{2\alpha}\mathbb{E}d_{\mathcal{M}}^2(\boldsymbol{m}_{t+1}, \boldsymbol{m}^*) + \frac{\zeta(\kappa, D)\alpha}{2}\mathbb{E}\|H(\boldsymbol{m}_t, \boldsymbol{x}_t)\|^2\,. \qquad (24)
\end{aligned}$$

Multiplying (21) by $2\zeta(\kappa, D)$ and adding to (24), we have:

$$2\zeta(\kappa, D)\mathbb{E}\Delta_{t+1} - (2\zeta(\kappa, D) - 1)\mathbb{E}\Delta_t \leq \Big(\frac{1 - \alpha\mu}{2\alpha}\Big)\mathbb{E}d_{\mathcal{M}}^2(\boldsymbol{m}_t, \boldsymbol{m}^*) - \frac{1}{2\alpha}\mathbb{E}d_{\mathcal{M}}^2(\boldsymbol{m}_{t+1}, \boldsymbol{m}^*) + \alpha\sigma^2\zeta(\kappa, D)\,. \qquad (25)$$

Multiplying (25) by $(1 - \epsilon)^{-t}$, we have:

$$\begin{aligned}
2(1 - \epsilon)^{-t}\zeta(\kappa, D)\mathbb{E}\Delta_{t+1} - 2(1 - \epsilon)^{-t}\Big(1 - \frac{1}{2\zeta(\kappa, D)}\Big)\zeta(\kappa, D)\mathbb{E}\Delta_t &\leq (1 - \epsilon)^{-t}\big(1 - \alpha\mu\big)\frac{1}{2\alpha}\mathbb{E}d_{\mathcal{M}}^2(\boldsymbol{m}_t, \boldsymbol{m}^*) \\
&\quad -(1 - \epsilon)^{-t}\frac{1}{2\alpha}\mathbb{E}d_{\mathcal{M}}^2(\boldsymbol{m}_{t+1}, \boldsymbol{m}^*) + (1 - \epsilon)^{-t}\alpha\sigma^2\zeta(\kappa, D)\,. \qquad (26)
\end{aligned}$$

We want to sum (26) from $t = 0$ to $t = s - 1$. However, to simplify the summation, we consider the case $t = 0$ and $t \geq 1$ separately, because in the latter case, we can get a simpler upper bound. First, let us consider the case $t \geq 1$. Let $\epsilon = \min\{\frac{1}{2\zeta(\kappa,D)}, \alpha\mu\}$ (Zhang & Sra, 2016), this implies $\epsilon \leq \frac{1}{2\zeta(\kappa,D)}$ and $\epsilon \leq \alpha\mu$. For $t \geq 1$, from (26) we have:

$$\begin{aligned} 2(1-\epsilon)^{-t}\zeta(\kappa,D)\mathbb{E}\Delta_{t+1} - 2(1-\epsilon)^{-(t-1)}\zeta(\kappa,D)\mathbb{E}\Delta_t \leq (1-\epsilon)^{-(t-1)}\frac{1}{2\alpha}\mathbb{E}d^2_{\mathcal{M}}(\boldsymbol{m}_t,\boldsymbol{m}^*) \\ - (1-\epsilon)^{-t}\frac{1}{2\alpha}\mathbb{E}d^2_{\mathcal{M}}(\boldsymbol{m}_{t+1},\boldsymbol{m}^*) \\ + (1-\epsilon)^{-t}\alpha\sigma^2\zeta(\kappa,D)\,. \end{aligned} \tag{27}$$

Now, let us consider the case $t = 0$. This case is simple, directly from (26) we have:

$$2\zeta(\kappa,D)\mathbb{E}\Delta_1 - (2\zeta(\kappa,D)-1)\mathbb{E}\Delta_0 \leq \Big(\frac{1-\alpha\mu}{2\alpha}\Big)\mathbb{E}d^2_{\mathcal{M}}(\boldsymbol{m}_0,\boldsymbol{m}^*) - \frac{1}{2\alpha}\mathbb{E}d^2_{\mathcal{M}}(\boldsymbol{m}_1,\boldsymbol{m}^*) + \alpha\sigma^2\zeta(\kappa,D)\,. \tag{28}$$

Finally, summing (26) over t from $t = 0$ to $t = s - 1$, and using the previous results, we have:

$$\begin{aligned} 2(1-\epsilon)^{-(s-1)}\zeta(\kappa,D)\mathbb{E}\Delta_s - (2\zeta(\kappa,D)-1)\mathbb{E}\Delta_0 &\leq \Big(\frac{1-\alpha\mu}{2\alpha}\Big)\mathbb{E}d^2_{\mathcal{M}}(\boldsymbol{m}_0,\boldsymbol{m}^*) - \frac{(1-\epsilon)^{-(s-1)}}{2\alpha}\mathbb{E}d^2_{\mathcal{M}}(\boldsymbol{m}_s,\boldsymbol{m}^*) \\ &\quad + \sum_{t=0}^{s-1}(1-\epsilon)^{-t}\alpha\sigma^2\zeta(\kappa,D) \\ &\leq \Big(\frac{1-\alpha\mu}{2\alpha}\Big)\mathbb{E}d^2_{\mathcal{M}}(\boldsymbol{m}_0,\boldsymbol{m}^*) + \sum_{t=0}^{s-1}(1-\epsilon)^{-t}\alpha\sigma^2\zeta(\kappa,D)\,, \end{aligned} \tag{29}$$

and plugging in $d_{\mathcal{M}}(\boldsymbol{m}_0,\boldsymbol{m}^*) \leq D$ (the diameter of the domain is bounded above by D), we have:

$$\begin{aligned} 2(1-\epsilon)^{-(s-1)}\zeta(\kappa,D)\mathbb{E}\Delta_s - (2\zeta(\kappa,D)-1)\mathbb{E}\Delta_0 &\leq \Big(\frac{1}{2\alpha} - \frac{\mu}{2}\Big)D^2 + \sum_{t=0}^{s-1}(1-\epsilon)^{-t}\alpha\sigma^2\zeta(\kappa,D) \\ &\leq \frac{D^2}{2\alpha} + \sum_{t=0}^{s-1}(1-\epsilon)^{-t}\alpha\sigma^2\zeta(\kappa,D)\,. \end{aligned} \tag{30}$$

Replacing $y = \boldsymbol{m}_0$, $x = \boldsymbol{m}^*$ in (9), considering an alternative definition of geodesic L-smoothness (Proposition 4.5 and 4.6. of (Boumal, 2023)) and plugging in $d_{\mathcal{M}}(\boldsymbol{m}_0,\boldsymbol{m}^*) \leq D$ and $\nabla f(\boldsymbol{m}^*) = 0$, we have:

$$\begin{aligned} \Delta_0 = f(\boldsymbol{m}_0) - f(\boldsymbol{m}^*) &\leq \langle \nabla f(\boldsymbol{m}^*), \exp^{-1}_{\boldsymbol{m}^*}(\boldsymbol{m}_0)\rangle + \frac{L}{2}\|\exp^{-1}_{\boldsymbol{m}^*}(\boldsymbol{m}_0)\|^2 \\ &= \langle \nabla f(\boldsymbol{m}^*), \exp^{-1}_{\boldsymbol{m}^*}(\boldsymbol{m}_0)\rangle + \frac{L}{2}d^2_{\mathcal{M}}(\boldsymbol{m}_0,\boldsymbol{m}^*) \leq \frac{LD^2}{2}\,. \end{aligned} \tag{31}$$

This ensures $\mathbb{E}\Delta_0 \leq \frac{LD^2}{2} \leq LD^2$ so that we have $\mathbb{E}\Delta_0 \leq \frac{D^2}{2\alpha}$ since $0 \leq \alpha \leq \frac{1}{2L}$, one can obtain from (30) that

$$\begin{aligned} \mathbb{E}\Delta_s = \mathbb{E}\{f(\boldsymbol{m}_s) - f(\boldsymbol{m}^*)\} &\leq \frac{(1-\epsilon)^{(s-1)}D^2}{2\alpha} + \sum_{t=0}^{s-1}(1-\epsilon)^t\frac{\sigma^2}{2} \\ &\leq \frac{(1-\epsilon)^{(s-1)}D^2}{2\alpha} + \sum_{t=0}^{\infty}(1-\epsilon)^t\frac{\sigma^2}{2} \\ &\leq \frac{(1-\epsilon)^{(s-1)}D^2}{2\alpha} + \frac{\alpha\sigma^2}{2\epsilon}\,, \end{aligned} \tag{32}$$

as desired.

A.2. Theorem 4.2

Using Markov's inequality with $\xi > 0$,

$$\mathbb{P}(g_t \geq \xi | H_0) \leq \frac{1}{\xi}\mathbb{E}\{g_t | \mathsf{H}_0\}\,. \tag{33}$$

Now, it remains to find an upper bound to $\mathbb{E}\{g_t|\mathsf{H}_0\}$. Let us ignore the conditioning of the expectation on H_0 to simplify the notation. The rest of the analysis is built upon the triangle inequality and the definition of g_t, which is,

$$g_t = d_{\mathcal{M}}(\boldsymbol{m}_{\lambda,t}, \boldsymbol{m}_{\Lambda,t}) \leq d_{\mathcal{M}}(\boldsymbol{m}_{\lambda,t}, \boldsymbol{x}) + d_{\mathcal{M}}(\boldsymbol{m}_{\Lambda,t}, \boldsymbol{x}) \tag{34}$$

for any $\boldsymbol{x} \in \mathcal{M}$. Take the expectation w.r.t. $\{\boldsymbol{x}_s\}_{s=0}^{t-1}$, with Theorem 4.1, Jensen's inequality and the fact $\left(\frac{\sqrt{a}+\sqrt{b}}{2}\right)^2 \leq \frac{a+b}{2}$ for nonnegative a and b, we can upper bound $\mathbb{E}\{g_t\}$ as

$$\begin{aligned}
\mathbb{E}\{g_t\} &\leq \mathbb{E}\{d_{\mathcal{M}}(\boldsymbol{m}_{\lambda,t}, \boldsymbol{x})\} + \mathbb{E}\{d_{\mathcal{M}}(\boldsymbol{m}_{\Lambda,t}, \boldsymbol{x})\} && (35)\\
&\leq \mathbb{E}\{d^2_{\mathcal{M}}(\boldsymbol{m}_{\lambda,t}, \boldsymbol{x})\}^{\frac{1}{2}} + \mathbb{E}\{d^2_{\mathcal{M}}(\boldsymbol{m}_{\Lambda,t}, \boldsymbol{x})\}^{\frac{1}{2}} && (36)\\
&= (\mathbb{E}\{f(\boldsymbol{m}_{\lambda,t})\})^{\frac{1}{2}} + (\mathbb{E}\{f(\boldsymbol{m}_{\Lambda,t})\})^{\frac{1}{2}} && (37)\\
&\leq \left(f(\boldsymbol{m}^*) + \frac{(1-\epsilon)^{t-1}D^2}{2\lambda} + \frac{\lambda\sigma^2}{2\epsilon}\right)^{\frac{1}{2}} + \left(f(\boldsymbol{m}^*) + \frac{(1-\epsilon')^{t-1}D^2}{2\Lambda} + \frac{\Lambda\sigma^2}{2\epsilon'}\right)^{\frac{1}{2}} && (38)\\
&\leq 2\left(f(\boldsymbol{m}^*) + \frac{(1-\epsilon)^{t-1}(\lambda+\Lambda)D^2}{4\lambda\Lambda} + \frac{(\lambda+\Lambda)\sigma^2}{4\epsilon}\right)^{\frac{1}{2}}, && (39)
\end{aligned}$$

with $\epsilon' = \min\{\frac{1}{\zeta(\kappa,D)}, \Lambda\mu\}$ satisfying $\epsilon' \geq \epsilon$ due to the step size condition $\lambda < \Lambda$. Taking the limit as $t \to \infty$, we get the following bound for $\mathbb{E}\{g_t\}$ at steady state:

$$\lim_{t\to\infty} \mathbb{E}\{g_t\} \leq 2\left(f(\boldsymbol{m}^*) + \frac{(\lambda+\Lambda)\sigma^2}{4\epsilon}\right)^{\frac{1}{2}}. \tag{40}$$

Combining this bound with (33) we obtain the desired result.

A.3. Theorem 4.3

Let us ignore the conditioning of the expectation on H_1 to simplify the notation. Since the diameter of the domain is bounded above by D, $g_t \leq D$, thus, we can apply Markov's inequality to the nonnegative random variable $D - g_t$ to obtain

$$\mathbb{P}(D - g_t \geq D - \xi) \leq \frac{D - \mathbb{E}\{g_t\}}{D - \xi}\,, \tag{41}$$

which leads to

$$\mathbb{P}(g_t > \xi) \geq \frac{\mathbb{E}\{g_t\} - \xi}{D - \xi}\,. \tag{42}$$

We now have to lower bound $\mathbb{E}\{g_t\}$. Using the reverse triangle inequality:

$$\mathbb{E}\{g_t\} = \mathbb{E}\{d_{\mathcal{M}}(\boldsymbol{m}_{\lambda,t}, \boldsymbol{m}_{\Lambda,t})\} \geq \mathbb{E}\{d_{\mathcal{M}}(\boldsymbol{m}_{\lambda,t}, \boldsymbol{m}^*_2)\} - \mathbb{E}\{d_{\mathcal{M}}(\boldsymbol{m}_{\Lambda,t}, \boldsymbol{m}^*_2)\}\,, \tag{43}$$

with $\boldsymbol{m}^*_2$ being the Karcher mean after the change point.

Notice the procedure of optimizing the Karcher mean loss function $f_{\text{aft}}(\boldsymbol{m}) = \mathbb{E}_{\boldsymbol{x}\sim P_2(\boldsymbol{x})}\{d^2_{\mathcal{M}}(\boldsymbol{m}, \boldsymbol{x})\}$ after the change point (i.e., where the expectation is defined w.r.t. $P_2(\boldsymbol{x})$), with solution $\boldsymbol{m}^*_2$, by the SGD algorithms (4) and (5) can be recognized as started from $\boldsymbol{x}_{t-B}$. Let us take the expectation w.r.t. $\{\boldsymbol{x}_s\}_{s=t-B}^{t-1}$ in the following steps.

Now we can upper bound $\mathbb{E}\{d_{\mathcal{M}}(\boldsymbol{m}_{\Lambda,t}, \boldsymbol{m}_2^*)\}$ with Jensen's inequality, Theorem 4.1, and the fact $\left(\frac{\sqrt{a}+\sqrt{b}}{2}\right)^2 \leq \frac{a+b}{2}$ for nonnegative a and b, leading to

$$\begin{aligned}
\mathbb{E}\{d_{\mathcal{M}}(\boldsymbol{m}_{\Lambda,t}, \boldsymbol{m}_2^*)\} &\leq \mathbb{E}\{d_{\mathcal{M}}(\boldsymbol{m}_{\Lambda,t}, \boldsymbol{x})\} + \mathbb{E}\{d_{\mathcal{M}}(\boldsymbol{m}_2^*, \boldsymbol{x})\} && (44)\\
&\leq \mathbb{E}\{d^2_{\mathcal{M}}(\boldsymbol{m}_{\Lambda,t}, \boldsymbol{x})\}^{\frac{1}{2}} + \mathbb{E}\{d^2_{\mathcal{M}}(\boldsymbol{m}_2^*, \boldsymbol{x})\}^{\frac{1}{2}} && (45)\\
&\leq \left(f_{\text{aft}}(\boldsymbol{m}_2^*) + \frac{(1-\epsilon')^B D^2}{2\Lambda} + \frac{\Lambda\sigma^2}{2\epsilon'}\right)^{\frac{1}{2}} + (f_{\text{aft}}(\boldsymbol{m}_2^*))^{\frac{1}{2}} && (46)\\
&\leq \left(f_{\text{aft}}(\boldsymbol{m}_2^*) + \frac{(1-\epsilon)^B D^2}{2\Lambda} + \frac{\Lambda\sigma^2}{2\epsilon}\right)^{\frac{1}{2}} + (f_{\text{aft}}(\boldsymbol{m}_2^*))^{\frac{1}{2}} && (47)\\
&\leq \left(2f_{\text{aft}}(\boldsymbol{m}_2^*) + \frac{(1-\epsilon)^B D^2}{\Lambda} + \frac{\Lambda\sigma^2}{\epsilon}\right)^{\frac{1}{2}}, && (48)
\end{aligned}$$

where $\boldsymbol{x} \sim P_2(\boldsymbol{x})$, and $\epsilon' = \min\{\frac{1}{\zeta(\kappa,D)}, \Lambda\mu\}$ satisfying $\epsilon' \geq \epsilon$ due to the step size condition $\lambda < \Lambda$.

To lower bound $\mathbb{E}\{d_{\mathcal{M}}(\boldsymbol{m}_{\lambda,t}, \boldsymbol{m}_2^*)\}$, we can use the reverse triangle inequality, which gives us

$$\begin{aligned}
\mathbb{E}\{d_{\mathcal{M}}(\boldsymbol{m}_{\lambda,t}, \boldsymbol{m}_2^*)\} &\geq \mathbb{E}\{d_{\mathcal{M}}(\boldsymbol{m}_{\lambda,t-1}, \boldsymbol{m}_2^*)\} - \mathbb{E}\{d_{\mathcal{M}}(\boldsymbol{m}_{\lambda,t-1}, \boldsymbol{m}_{\lambda,t})\}\\
&\geq \mathbb{E}\{d_{\mathcal{M}}(\boldsymbol{m}_{\lambda,t-B-1}, \boldsymbol{m}_2^*)\} - \sum_{u=t-B}^{t} \mathbb{E}\{d_{\mathcal{M}}(\boldsymbol{m}_{\lambda,u-1}, \boldsymbol{m}_{\lambda,u})\}. && (49)
\end{aligned}$$

Using the stochastic gradient update equation, $\boldsymbol{m}_{\lambda,t} = \exp_{\boldsymbol{m}_{\lambda,t-1}}(-\lambda H(\boldsymbol{m}_{\lambda,t-1}, \boldsymbol{x}_{t-1}))$, we can express $d_{\mathcal{M}}(\boldsymbol{m}_{\lambda,u-1}, \boldsymbol{m}_{\lambda,u})$ as:

$$d_{\mathcal{M}}(\boldsymbol{m}_{\lambda,u-1}, \boldsymbol{m}_{\lambda,u}) = d_{\mathcal{M}}\big(\boldsymbol{m}_{\lambda,u-1}, \exp_{\boldsymbol{m}_{\lambda,u-1}}(-\lambda H(\boldsymbol{m}_{\lambda,u-1}, \boldsymbol{x}_{u-1}))\big). \tag{50}$$

Since the injectivity radius of the manifold is assumed to be globally bounded above by I, the condition $\lambda \leq \frac{I}{\rho}$ implies that $\|\lambda H(\boldsymbol{m}_{\lambda,u-1}, \boldsymbol{x}_{u-1})\| < I$. Thus, by proposition 10.22 of (Boumal, 2023),

$$\begin{aligned}
d_{\mathcal{M}}\big(\boldsymbol{m}_{\lambda,u-1}, \exp_{\boldsymbol{m}_{\lambda,u-1}}(-\lambda H(\boldsymbol{m}_{\lambda,u-1}, \boldsymbol{x}_{u-1}))\big) &= \lambda\|H(\boldsymbol{m}_{\lambda,u-1}, \boldsymbol{x}_{u-1})\| && (51)\\
&\leq \rho\lambda\,. && (52)
\end{aligned}$$

The term $\mathbb{E}\{d_{\mathcal{M}}(\boldsymbol{m}_{\lambda,t-B-1}, \boldsymbol{m}_2^*)\}$ can be lower bounded as

$$\mathbb{E}\{d_{\mathcal{M}}(\boldsymbol{m}_{\lambda,t-B-1}, \boldsymbol{m}_2^*)\} \geq d_{\mathcal{M}}(\boldsymbol{m}_1^*, \boldsymbol{m}_2^*) - \mathbb{E}\{d_{\mathcal{M}}(\boldsymbol{m}_{\lambda,t-B-1}, \boldsymbol{m}_1^*)\}\,, \tag{53}$$

with $\boldsymbol{m}_1^*$ being the Karcher mean of distribution $P_1(\boldsymbol{x})$ of the data before the change point.

Knowing that the change point occurred at time $t-B$, and since the algorithms are assumed to have asymptotically converged before the change point happened (i.e., $t-B-1$ is large), we can upper bound $\mathbb{E}\{d_{\mathcal{M}}(\boldsymbol{m}_{\lambda,t-B-1}, \boldsymbol{m}_1^*)\}$ in (53) using Jensen's inequality, Theorem 4.1, and the fact $\left(\frac{\sqrt{a}+\sqrt{b}}{2}\right)^2 \leq \frac{a+b}{2}$ for nonnegative a and b, which gives us

$$\begin{aligned}
\mathbb{E}\{d_{\mathcal{M}}(\boldsymbol{m}_{\lambda,t-B-1}, \boldsymbol{m}_1^*)\} &\leq \mathbb{E}\{d^2_{\mathcal{M}}(\boldsymbol{m}_{\lambda,t-B-1}, \boldsymbol{x}')\}^{\frac{1}{2}} + \mathbb{E}\{d^2_{\mathcal{M}}(\boldsymbol{m}_1^*, \boldsymbol{x}')\}^{\frac{1}{2}} && (54)\\
&\leq \left(f_{\text{bef}}(\boldsymbol{m}_1^*) + \frac{(1-\epsilon)^{t-B-1} D^2}{2\lambda} + \frac{\lambda\sigma^2}{2\epsilon}\right)^{\frac{1}{2}} + (f_{\text{bef}}(\boldsymbol{m}_1^*))^{\frac{1}{2}} && (55)\\
&\leq \left(2f_{\text{bef}}(\boldsymbol{m}_1^*) + \frac{\lambda\sigma^2}{\epsilon}\right)^{\frac{1}{2}}, && (56)
\end{aligned}$$

where $\boldsymbol{x}' \sim P_1(\boldsymbol{x})$ and the expectation above is now taken w.r.t. the distribution $P_1(\boldsymbol{x})$, before the change point; we used that fact $(1-\epsilon)^{t-B-1} \to 0$ due to the large $t-B-1$, and $f_{\text{bef}}(\boldsymbol{m}) = \mathbb{E}_{\boldsymbol{x}\sim P_1(\boldsymbol{x})}\{d^2_{\mathcal{M}}(\boldsymbol{m}, \boldsymbol{x})\}$ denotes the Karcher mean loss function before the change point (i.e., where the expectation is defined w.r.t. $P_1(\boldsymbol{x})$), with solution $\boldsymbol{m}_1^*$.

Algorithm 2 Adaptive threshold selection

Input: $\{g_t\}$, forgetting factor α, quantile q.
Initialization: $\beta_t^g = g_1$, $\gamma_t^g = g_1^2$.
for $t = 1, 2, 3, \ldots$ **do**
  $\beta_t^g = (1-\alpha)\beta_{t-1}^g + \alpha g_t$;
  $\gamma_t^g = (1-\alpha)\gamma_{t-1}^g + \alpha g_t^2$;
  $\hat{\xi}_t = \beta_t^g + \sqrt{\gamma_t^g - (\beta_t^g)^2}\sqrt{2}\mathrm{erf}^{-1}(2q-1)$;
end for

Combining the bounds in (49), (52), (53), (56) leads to the following lower bound:

$$\mathbb{E}\{d_{\mathcal{M}}(\boldsymbol{m}_{\lambda,t}, \boldsymbol{m}_2^*)\} \geq d_{\mathcal{M}}(\boldsymbol{m}_1^*, \boldsymbol{m}_2^*) - \left(2f_{\mathrm{bef}}(\boldsymbol{m}_1^*) + \frac{\lambda\sigma^2}{\epsilon}\right)^{\frac{1}{2}} - \rho\lambda B\,. \tag{57}$$

Finally, combining the bounds (42), (43), (48) and (57), we obtain

$$\mathbb{P}(g_t > \xi) \geq \frac{1}{D-\xi}\left[d_{\mathcal{M}}(\boldsymbol{m}_1^*, \boldsymbol{m}_2^*) - \left(2f_{\mathrm{bef}}(\boldsymbol{m}_1^*) + \frac{\lambda\sigma^2}{\epsilon}\right)^{\frac{1}{2}} - \rho\lambda B - \left(2f_{\mathrm{aft}}(\boldsymbol{m}_2^*) + \frac{(1-\epsilon)^B D^2}{\Lambda} + \frac{\Lambda\sigma^2}{\epsilon}\right)^{\frac{1}{2}} - \xi\right], \tag{58}$$

which is the desired result.

B. Adaptive threshold selection

One challenge with applying CPD algorithms is the selection of the detection threshold ξ for a given problem. A classical approach consists of adjusting ξ such that the algorithm achieves some desired performance under the null hypothesis (i.e., in the absence of change points), such as a given probability of false alarms (Keriven et al., 2020). For a false alarm rate of, e.g., 0.05, ξ can be set as the 95-th quantile of g_t. The performance of the algorithm under the null hypothesis can be computed using training data or based on a theoretical analysis, such as the result given in Theorem 4.2. However, threshold selection approaches based on theoretical analyses are hard to apply in practice as they require strong prior knowledge of the statistical distribution of the data, such as the Karcher variance $f(\boldsymbol{m})$ and gradient noise σ^2 in our case.

A more practical approach is to set ξ as an estimate of the q-th quantile of g_t obtained using a recursive algorithm. Although efficient algorithms have been proposed for recursive quantile estimation (Chen et al., 2023), we use a simpler alternative by approximating g_t by a Gaussian distribution (the validity of this hypothesis illustrated empirically in Figure 8), as also done in (Keriven et al., 2020). This way, computing only its first two moments is sufficient to compute the q-th quantile, which is given by the mean plus the standard deviation multiplied by $\sqrt{2}\mathrm{erf}^{-1}(2q-1)$, where erf is the Gauss error function. A simple recursive implementation of this strategy is shown in Algorithm 2, which is based on EWMAs of the first two moments of g_t. Experiments illustrating the validity of the Gaussian hypothesis over g_t and the performance of Algorithm 2 can be found in Appendix D.3.

C. Computational complexity

The computational complexity of our method consists mainly of the cost of implementing the two R-SGD algorithms used to estimate the generalized Karcher means. The R-SGD algorithm is a first-order method that is computationally efficient compared to other manifold optimization algorithms. It comprises two main steps: 1) computation of the Riemannian gradient of the loss function, and 2) computing the exponential or retraction to map the gradient back to the manifold.

The computational complexity involved with these steps depends on the choice of the manifold as it affects both the loss function (and therefore the gradient) and the retraction/exponential map. However, for many manifolds of great practical interest, including the SPD and the Grassmann, computing these operations is relatively efficient, and for these two manifolds, we can compute the complexity explicitly.

Complexity for the SPD manifold: The operations involved in implementing the R-SGD on the manifold of $p \times p$ SPD

matrices consist of five matrix multiplications, a matrix inverse, and a matrix logarithm. Thus, the computational cost is given by $O(p^3)$ operations.

Complexity for the Grassmann manifold: The operations involved in implementing the R-SGD on the Grassmann manifold of k-dimensional subspaces in $\mathbb{R}^p$ consists of two SVDs, five matrix products, and the evaluation of $O(k)$ arithmetic functions. Thus, the computational cost is given by $O(p^2k)$.

Comparison to baselines: We briefly compare the complexity with respect to the baselines F-CPD (Dubey & Müller, 2020) and NEWMA (Keriven et al., 2020). F-CPD is an offline method designed to operate on manifolds and detects a change point based on a two-sample test. For every candidate change point, the test statistic is computed as a function of the Karcher means and variances of the data before and after the candidate change point, which is computationally very intensive to implement. NEWMA, on the other hand, is an online method designed to operate on Euclidean spaces, by comparing exponentially weighted moving averages of generalized moments of the data computed based on the random features framework. Thus, the cost of NEWMA is dominated by the cost of computing the random features (Keriven et al., 2020). For random Fourier features (Rahimi & Recht, 2007), the computation complexity scales as $O(Sd)$ operations, where d is the dimension of the input data, and S is the number of random samples (the dimension of the feature space), which are sampled from a probability measure related to the kernel. Thus, depending on the choice of kernel and the feature dimension NEWMA can be efficient, although it does not take the manifold geometry into account.

D. Additional results

D.1. Mean and standard deviation of the test statistics

In Figure 5, we plot the mean and standard deviation of the test statistics of all the compared algorithms for the examples with synthetic data. It can be seen that for the synthetic example the test statistic of the proposed strategy required approximately 200 samples to converge after a change point occurs. The algorithm achieves good performance for detecting multiple change points as long as the interval between them is sufficiently large compared to the time it requires to converge. By comparison, we also plot in Figure 7 the test statistic for the compared methods for the skeleton-based action recognition example, in which the parameters of the algorithms had to be readjusted to achieve faster convergence since the number of samples between change points is smaller. It can be observed that the algorithms converge significantly faster (requiring only approximately 80 samples), however, the variances of the test statistic, particularly after the change point, are also much higher. This illustrates the trade-off between detection performance and adaptability of the proposed method.

D.2. Comparisons between the histograms of the test statistics on synthetic data

To get a deeper insight into the behavior of the ROC curves in the examples with synthetic data (Figures 1 and 2), where our method had an area under curve close to one, we compared the histograms of the test statistics of all methods under the null hypothesis and at their peak value after a change point. The result can be seen in Figure 6. One can observe that different from the competing methods, the histogram of the test statistic of our method under the null shows almost no overlap with its counterpart at peak value after a change point. This explains the behavior seen in the ROC curves.

D.3. Histogram of the test statistic, Gaussian fit, and illustration of the adaptive threshold procedure

To illustrate the validity of the Gaussian hypothesis of g_t, in Figure 8, we plot the histogram of g_t for 1000 Monte Carlo runs, computed based on samples of g_t under the null hypothesis when the algorithm is tested for the synthetic example on $\mathcal{S}_p^{++}$, after the algorithms converge (with step sizes $\lambda = 0.01$ and $\Lambda = 0.02$). It can be observed that the histogram and its Gaussian fit are very close, which justifies the approximations in Algorithm 2.

We illustrate the performance of Algorithm 2 with $\alpha = 0.005$ and $q = 0.95$ (5% of false alarms). We considered the same setup as in the synthetic example in $\mathcal{S}_p^{++}$, but here we added multiple change points, spaced by 200 samples to allow the algorithm to converge. The test statistic and the adaptive threshold are shown in Figure 9 (results are shown after the steady-state convergence of both the CPD and adaptive threshold selection algorithms), where it can be seen that the dynamic threshold can successfully adapt to detect multiple change points in a continuous run.

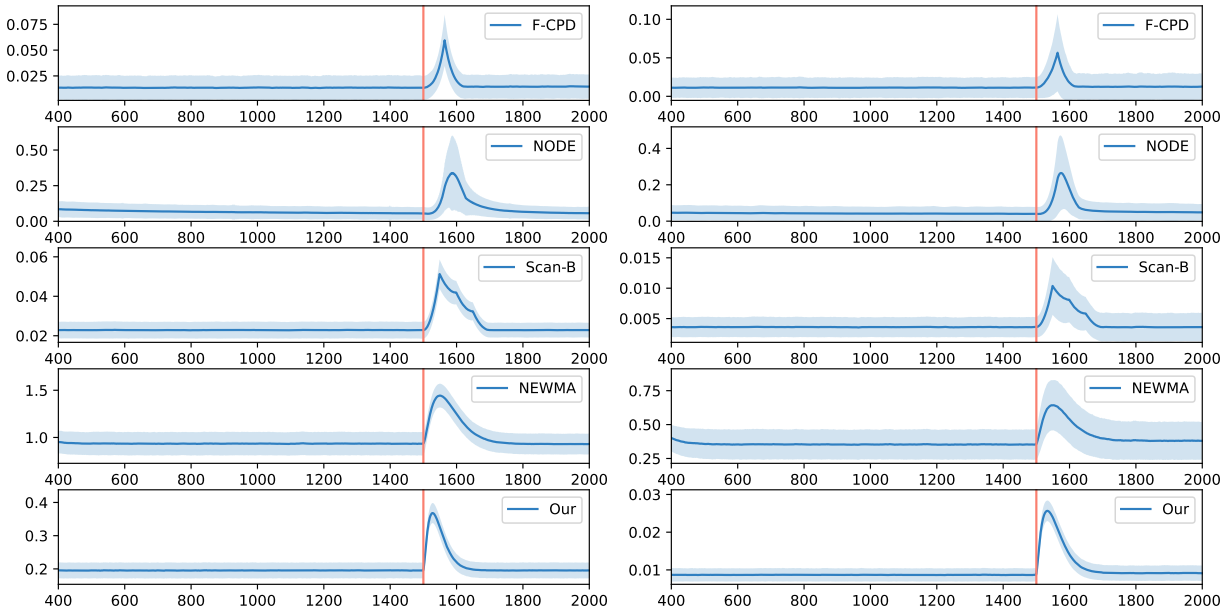

Figure 5. Illustration of the mean and standard deviation of all the compared detection statistics for the experiments on synthetic data on both $\mathcal{S}_p^{++}$ (left) and $\mathcal{G}_p^k$ (right). The red line indicates the change point.

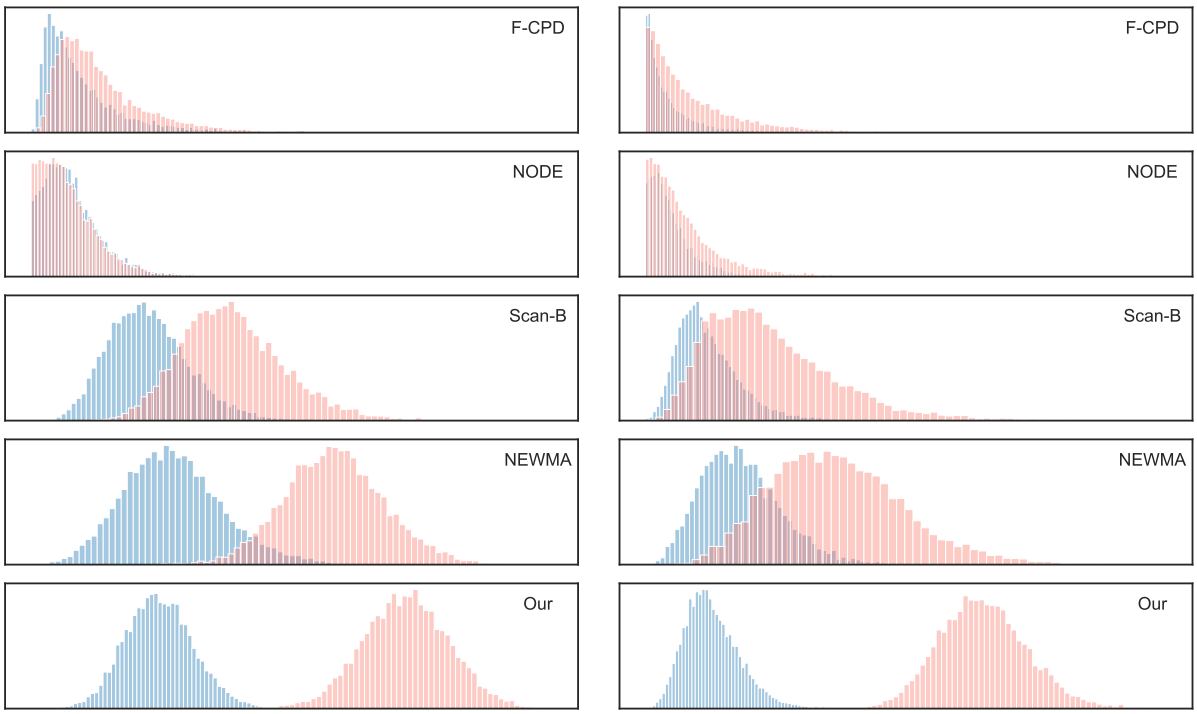

Figure 6. Histograms of all the compared detection statistics for the experiments on synthetic data on both $\mathcal{S}_p^{++}$ (left) and $\mathcal{G}_p^k$ (right). The blue histograms are under the null hypothesis and the pink histograms are at their peak values after the change point.

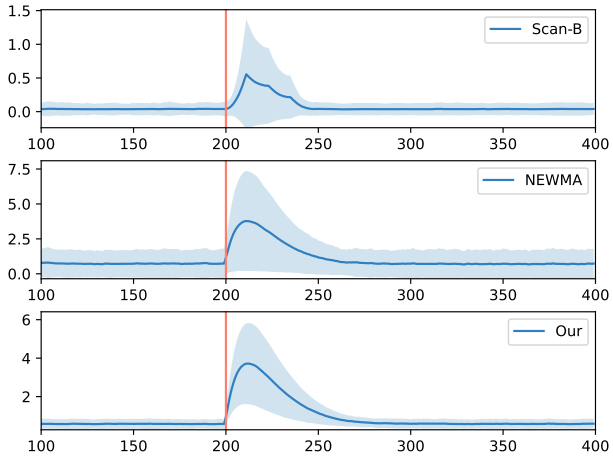

Figure 7. Illustration of the mean and standard deviation of the compared detection statistics for the experiments on real data for skeleton-based action recognition. The red line indicates the change point.

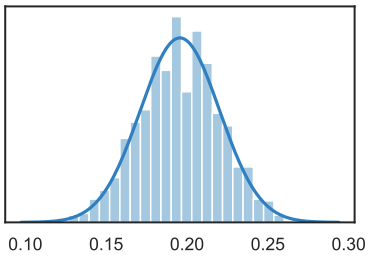

Figure 8. Histogram of g_t under the null hypothesis for synthetic data on $\mathcal{S}_p^{++}$ and its Gaussian fit.

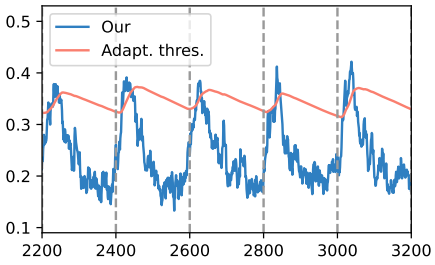

Figure 9. Illustration of the adaptive threshold procedure. The dotted gray lines indicate change points.