On Riemannian and non-Riemannian Optimisation, and Optimisation Geometry

Jeanne Lefevre, Florent Bouchard, Salem Said, Nicolas Le Bihan, Jonathan H Manton

Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 54-9 (2021) 578–583

On Riemannian and non-Riemannian Optimisation, and Optimisation Geometry

Jeanne Lefevre * **Florent Bouchard** ** **Salem Said** ***
Nicolas Le Bihan * **Jonathan H. Manton** ****

* *GIPSA-lab, Univ. Grenoble Alpes, CNRS, Grenoble INP, France*
** *LISTIC, Univ. Savoie Mont-Blanc, Annecy, France*
*** *IMS, Univ. Bordeaux, CNRS, France*
**** *University of Melbourne, Australia*

Abstract Optimisation algorithms such as the Newton method were first generalised to manifolds by generalising the components of the algorithm directly: gradients were replaced by Riemannian gradients, straight lines were replaced by geodesics, and so forth. This meant having to endow the manifold with a Riemannian metric. Traditionally then, attention focused on the geometry of the underlying manifold. However, we argue the geometry of the manifold is not the right geometry to focus on because it does not take the cost function into consideration. For online optimisation problems requiring the minimisation of many different cost functions, of most relevance is the geometry of the family of cost functions as a whole: if the cost functions fit together in a "nice" way, fast optimisation algorithms can be developed even if individual cost functions are difficult to optimise. In particular, non-convex problems are not necessarily difficult problems. This paper presents a Riemannian-based homotopy algorithm for solving such Optimisation Geometry problems and briefly explains how it can be generalised to a non-Riemannian (e.g., coordinate-adapted) algorithm.

Keywords: Global optimisation, differential geometry, homotopy methods.

1. INTRODUCTION

Numerical optimisation is a key tool in control, signal processing and machine learning. Engineering applications include communications, remote sensing, audio and biomedical signal analysis and computer vision. Online optimisation algorithms in these fields must solve an optimisation problem each time new data is observed. Denoting a (vector-valued or manifold-valued) observation by $\theta \in \Theta$, the algorithm must solve

$$x^* = \operatorname*{argmin}_{x \in X} \quad f(x, \theta) \tag{1}$$

each time a new observation θ arrives, where $f : X \times \Theta \to \mathbb{R}$ is some cost function, and X and Θ are manifolds. For example, if θ represents a block of received data, the transmitted data x can be found by minimising the negative log-likelihood function $f(x, \theta)$ characterising the channel model Kay (1993). Finite-horizon control problems are also of this form (see *e.g.* Garg et al. (2011)), as are various machine-learning algorithms such as Harandi et al. (2017).

Invariably, there are two approaches to solving (1). If the observations vary slowly in time, it is treated as a tracking problem: the minimum x^* for the previous observation θ is used as a starting point for finding the minimum for the current observation. Alternatively, if there is little relationship between successive observations, (1) is treated as a standard optimisation problem for the individual cost function $f(\cdot, \theta)$ where θ is fixed. In particular, if each individual cost function is not convex, the optimisation problem is usually considered difficult. *But this ignores we have prior knowledge of the family of cost functions.*

As pointed out in Manton (2013), if the cost functions fit together nicely then (1) can be readily solved even though every individual cost function is difficult to optimise. To exemplify this, consider the trivial case $f(x, \theta) = g(x - \theta)$. If g is difficult to optimise then so is each individual cost function $f(\cdot, \theta)$. But given we are allowed to make a finite number of precomputations at the algorithm design stage, we can (somehow) optimise g. If $x^{0\,*}$ is the minimum of g then $x^* = x^{0\,*} + \theta$ is the minimum of f; a single addition is all that is needed to solve (1) in this case.

In more realistic applications, greater effort will be required to go from the minimum of one cost function to the minimum of a neighbouring cost function. One aim of this paper is to give a rigorous algorithm for doing just this. It is based on a simple homotopy idea, but importantly, this can be augmented to give guaranteed performance, as outlined in Manton (2013). Guaranteed performance will be the subject of a future paper: ultimately, an upper bound can be given on the number of computational steps required to find the *global* minimum to within a prescribed level of accuracy.

Finding the global minimum is achieved by tracking all the local minima. This is well-suited to modern computing platforms that are highly parallel.

⋆ This material is based on a collaboration that was supported by the IDEX University of Grenoble Alpes.

10.1016/j.ifacol.2021.06.119

The natural setting for studying the "geometry" of the cost functions is to generalise (1) to a certain type of constrained optimisation problem on a fibre bundle Manton (2013). For simplicity of presentation, we refrain from doing this here, working instead with the trivial fibre bundle $M = X \times \Theta$. This suffices for studying the crux of the problem.

2. RIEMANNIAN VS NON-RIEMANNIAN

The classical Newton iterate for finding a critical point is

$$x_{k+1} = N_f(x_k), \quad N_f(x) = x - [\mathrm{Hess}\, f(x)]^{-1} \nabla f(x). \quad (2)$$

Although the gradient ∇f and the Hessian $\mathrm{Hess}\, f$ depend on the chosen inner product, the Newton iterate does not. It was generalised to manifolds by Gabay (1982), who replaced the gradient and Hessian by their Riemannian counterparts: for $f\colon X \to \mathbb{R}$, its gradient $\nabla f(x) \in T_xX$ is defined implicitly by $Df(x) \cdot z = \langle \nabla f(x), z \rangle_x$ for all $z \in T_xX$. Note T_xX is the tangent space at $x \in X$ of the manifold X. The Riemannian Hessian $\mathrm{Hess}\, f_\theta(x)$ at $x \in X$ is the linear map defined as

$$\begin{aligned} \mathrm{Hess}\, f_\theta(x) :\ T_xX &\to T_xX \\ \xi_x &\mapsto \nabla_{\xi_x} \nabla f_\theta(x) \end{aligned}$$

where $\nabla : TX \times TX \to TX$ is the Levi-Civita connection. This means $v = -[\mathrm{Hess}\, f(x)]^{-1} \nabla f(x)$ is an element of T_xX. Since $N_f(x) = x + v$ can be interpreted as starting at x and moving in a straight line for one unit of time with constant velocity v, it generalises to moving along a geodesic, expressed via the Riemannian exponential map:

$$N_f(x) = \exp_x \left(-[\mathrm{Hess}\, f(x)]^{-1} \nabla f(x) \right). \quad (3)$$

Except in special cases when the cost function directly relates to the geometry of the manifold (e.g., Manton (2004)), there is no performance benefit in using a Riemannian approach: following geodesics can be computationally intensive, and the behaviour of the cost function itself is normally ignored when choosing the Riemannian structure to endow X with. Instead, a coordinate-adapted approach is generally preferred Manton (2002, 2015). Its derivation starts with the observation that applying the Newton method in a different coordinate system takes the form

$$x_{k+1} = (\phi \circ N_{f \circ \phi} \circ \phi^{-1})(x_k) \quad (4)$$

where ϕ is a change of coordinates. The key idea is to allow different coordinate changes at each step:

$$x_{k+1} = (\phi_k \circ N_{f \circ \phi_k} \circ \phi_k^{-1})(x_k). \quad (5)$$

Importantly, this can be applied to functions on a manifold if the ϕ_k are chosen to be local parametrizations: functions from $\mathbb{R}^n$ to a neighbourhood of x_k on the manifold. If the ϕ_k are chosen to be Riemannian normal coordinates then the Riemannian Newton method is recovered.

Going further, it is argued in Manton (2015) that any Newton method on a manifold must be of the form

$$x_{k+1} = (\psi_k \circ N_{f \circ \phi_k} \circ \phi_k^{-1})(x_k) \quad (6)$$

where ψ_k is an approximation of ϕ_k, again allowing for computationally more efficient choices.

The remainder of the paper takes a Riemannian approach for simplicity of presentation: it is "cleaner" to work with. Importantly though, the ideas immediately generalise to a coordinate-adapted framework.

3. OPTIMISATION GEOMETRY

The goal of optimisation geometry is to use precomputations on a family of functions to simplify the quest for the local or global minima of one element in this family. A first step in doing so is to show that for nice enough functions, the critical point fit together in a nice way meaning it's possible to define a smooth path of critical points from one critical point to another.

3.1 Smooth path between critical points

First, we define a nice class of function on which the method we propose has guaranteed performance. This is done in Definition 1.

Definition 1. Let X, Θ be two smooths compact manifolds of dimensions k and n respectively, and $M = X \times \Theta$, be their Cartesian product of dimension $k + n$. Let $f\colon X \times \Theta \to \mathbb{R}$ be a smooth cost function on M. f is called *fibre-wise Morse* on M if the restriction

$$\begin{aligned} f_\theta :\ X &\to \mathbb{R} \\ x &\mapsto f(x; \theta) \end{aligned}$$

is a Morse function for every $\theta \in \Theta$. We recall that a Morse function is a smooth real-valued function that has no degenerate critical points Milnor et al. (1969).

A direct consequence of the definition of a Morse function is that its critical points are isolated. A less direct but still close consequence is that Morse functions on compact sets can only have a finite number of critical points. These two elements are of importance in the definition of an homotopy method in section 3.2. The next theorem shows how critical points of a smooth family of Morse functions fit together in a nice way, and is the first step in showing why the family of function rather than just the individual cost function should be considered.

Theorem 2. A point (x^*, θ) is called *fibre-wise critical* if x^* is a critical point of f_θ i.e $Df_\theta(x^*) = 0$. The set $\tilde{N}$ of fibre-wise critical points of the fibre-wise Morse function $f\colon M = X \times \Theta \to \mathbb{R}$ is a n-dimensional smooth submanifold of M in the sense of Guillemin and Pollack (2010). Furthermore, it is topologically closed and has no boundaries.

We refer the reader to Manton (2013) for a proof of this theorem in a more general case. In the following property, the shape of $\tilde{N}$ is further specified. It sits in M over Θ, and can be locally parametrized by $\theta \in \Theta$.

Proposition 3. Each connected component of $\tilde{N}$ locally defines a smooth section over Θ. Numbering the connected components of $\tilde{N}$ as $\tilde{N}_j$, there is an open set U such that the map

$$\begin{aligned} \pi_j : U \cap \tilde{N}_j &\to \Theta \\ (x, \theta) &\mapsto \theta \end{aligned}$$

is a smooth diffeomorphism.

This theorem and its corollary show that the shape of the submanifold of fibre-wise critical points is deeply connected and constrained by the topology of M itself. Indeed, $\tilde{N}$ is closed and has no boundaries, and because it is locally parametrizable by θ, it can have no turning point with

regard to Θ. An illustrative way to see it is that the submanifold $\tilde{N}$ sits over the base space Θ in M, like a topological copy of Θ. Hence, there can be no boundaries or limit points in θ. A direct consequence is that all Morse function of the family $\{f_\theta\}$ have the same number of critical points. As an example, let $M = S^1 \times S^1$ the two-dimensional torus parametrized by (x, θ), where x describes a rotation along horizontal circles (see figure 1) and θ is a rotation along vertical axis. Let f a fibre-wise Morse function over the torus. The set of fibre-wise critical points of f is a one-dimensional closed differentiable submanifold of M with no boundaries. Hence, it is made of one or several loops (closeness condition) winding their way around the torus and never intersecting each other (submanifold condition). Furthermore, It is locally parametrizable by the vertical circles θ (vertical circle in Figure 1) which means in particular it can have no turning point with respect to this coordinate. This condition excludes for instance a horizontal circle around the Torus. If all the functions of the family f_θ are the same however, then $\tilde{N}$ is a vertical circle. The only way to cross the same angle θ twice is by turning around the torus, with the angle θ only increasing or only decreasing modulo 2π. Then, $\tilde{N}$ can only be made of one or several circles winding their ways around the torus the same number of times.

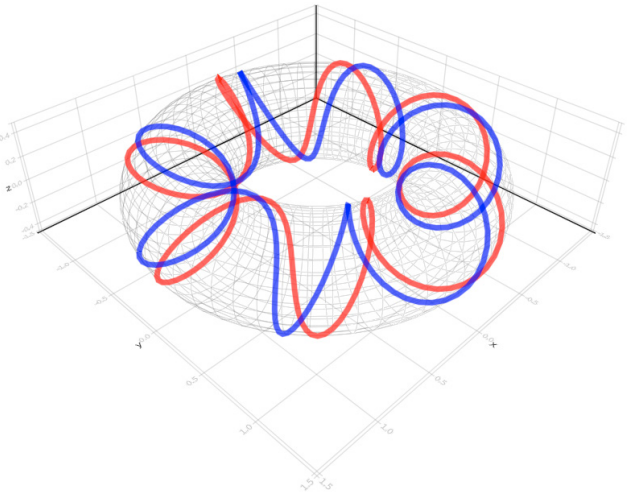

Figure 1. A possible shape for the fibre-wise critical points of a f on the torus. The submanifold $\tilde{N}$ is represented by the red and blue lines and is made of two connected components winding their way seven times around the Torus. The number K of critical points in this example is 14, it is the number of intersection of $\tilde{N}$ with any horizontal circle at θ constant.

The following corollary shows that if $\tilde{N}$ is made of finitely many connected components, then just some of them contain local minima. It ensures in particular that the index of a critical point along a connected component is constant.

Corollary 4. There is a submanifold $N \subset \tilde{N}$ made of finitely many connected components and containing only fibre-wise local minima.

Assume that in real time a data θ^1 is observed so that we have to optimise on f_{θ^1} while having precomputed information on f_{θ^0}. If we build a path between θ^0 and θ^1, can we deduce a path between a local minima of f_{θ^0} and f_{θ^1} so that every element on the path is a local minima for some f_θ ? The next corollary shows that the answer is yes, and that this path can be smooth.

Corollary 5. Let θ^0 and θ^1 be two sets of data in Θ. Let $\theta\colon [0,1] \to \Theta$ a smooth diffeomorphism joining θ^0 and θ^1, *i.e.* $\theta(0) = \theta^0$ and $\theta(1) = \theta^1$. Let x^{0*} be a local minimum of f_{θ^0}. and let $N_0 \subset N$ be the connected component containing (θ^0, x^{0*}). Then there is a unique smooth path $x^*\colon [0,1] \to X$ satisfying $(\theta(t), x^*(t)) \in N_0, \forall t$. The endpoint $x^*(1) = x^{1*}$ is a local minimum of f_{θ^1}.

Figure 2 shows an example of how a path in θ induces a lift in N. Note that the lift is not unique as long as x^{0*} is not defined, and this is true even if N has only one connected component. Indeed, in the Torus example where N winds its ways several times around the Torus, a connected component can intersect several points of the shape (θ, x_i) where θ is fixed.

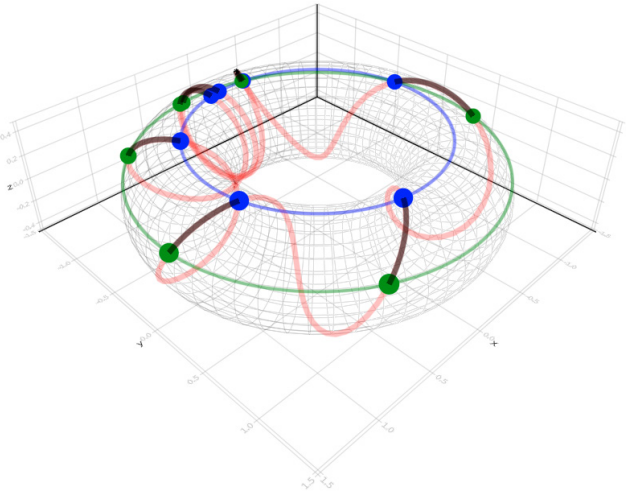

Figure 2. Illustration of Corollary 5 on the torus where an admissible submanifold of fibre-wise critical points is represented in red. The torus is parameterized by (x, θ) where x is the angle along the horizontal circles and θ the angle along the vertical circles. The green circle is the set of points with a fixed $\theta = \theta^0$ and the blue circle has a fixed $\theta = \theta^1$. Green dots are fibre-wise critical points of f_{θ^0} while blue dots are fibre-wise critical points of f_{θ^1}. The dark lines are the lifts in N of a curve $\theta(t)$ that goes from θ^0 to θ^1. Note that the lift generates as many possible $x^*(t)$ as the number of local minima of f_{θ^0}. However, fixing a starting point (one of the green dots) uniquely determines a section of dark line, i.e a function $x^*(t)$.

3.2 A homotopy method for optimisation geometry

Previous section has shown the existence of a smooth path made of local minima, joining the two point of interest in our problem. A homotopy method can take advantage of this path as long as it is a natural attractor for some recursive algorithm. We will show how this is the case for x^*. Our work differs from previous work on homotopy methods, a summary of which can be found in Allgower and Georg (2012), which usually focus on

ordinary differential equations or level sets of 0 by the homotopy map. Being on the tangent bundle TX, we consider the level set of the submanifold made of the zeros of the different tangent spaces in TX. This whole section uses Riemannian tools and framework. Let $f : X \times \Theta \to \mathbb{R}$ be a fibre-wise Morse function as in Definition 1. Let θ^0, $\theta^1 \in \Theta$ and let $x^{0\,*} \in X$ be a local minimum of f_{θ^0}. We aim at finding a local minimum $x^{1\,*}$ of f_{θ^1} by exploiting the known local minimum $x^{0\,*}$ of f_{θ^0}. To do so, we first need to define a diffeomorphic curve $\theta : [0,1] \to \Theta$ such that $\theta(0) = \theta^0$ and $\theta(1) = \theta^1$. For example, if Θ is a Riemannian manifold, one can choose the curve $\theta : [0,1] \to \Theta$ as the geodesic joining θ^0 and θ^1. From Corollary 5, we know that there exists a curve $x^* : [0,1] \to X$ of local minima of the functions $f_{\theta(t)}$. Thus, starting from $x^{0\,*}$, we propose to follow the curve $x^* : [0,1] \to X$ corresponding to the chosen curve $\theta : [0,1] \to \Theta$ in order to find a local minimum $x^*(1) = x^{1\,*}$ of f_{θ^1}.

The curve $x^* : [0,1] \to X$ is implicitly defined, hence one cannot expect to obtain it in closed form in general and an iterative method is needed to estimate it. To construct $x^* : [0,1] \to X$, we exploit its graph, which is defined as

$$G(x^*) = \{(t, x^*(t)) : t \in [0,1]\} \subset [0,1] \times X.$$

To do so, we rely on the characterisation of $G(x^*)$ provided in Proposition 6. This property shows that locally, the curve x^* exactly contains all and only the antecedent of points of the shape $0_x \in TX$ for H. The function H thus introduced is the homotopy map we will use in this section.

Proposition 6. Let the mapping

$$\begin{aligned} H : [0,1] \times X &\to TX \\ (t, x) &\mapsto \nabla f_{\theta(t)}(x) \in T_x X. \end{aligned}$$

Then, every point $(t, x^*(t))$ satisfies the relation

$$H(t, x) = 0_x$$

Furthermore, for every point $(t, x^*(t))$, there is a neighbourhood V around this point such that $\{(t,x) \in [0,1] \times X : H(t,x) = 0_x\} \cap V$ contains only points in the graph of x^*, i.e, only points of the shape $(t, x^*(t))$ for some $t \in [0,1]$.

An equation for the curve $x^* : [0,1] \to X$ is not known in closed form, so that in order to be able to follow this curve we need its tangent. Given $(t, x^*(t)) \in G(x^*)$, we are able to define the tangent space $T_{(t,x^*(t))}G(x^*)$, which is a one-dimensional subspace of $\mathbb{R} \times T_{x^*(t)}X$. To do so, we need to differentiate the function H defined in Proposition 6. Differentiating H with respect to $t \in [0,1]$ in direction $dt \in \mathbb{R}$ yields $\frac{\partial}{\partial t} \nabla f_{\theta(t)}(x) dt$, while differentiating it with respect to $x \in X$ in direction $dx \in T_x X$ yields the Riemannian Hessian $\operatorname{Hess} f_{\theta(t)}(x) dx$ of $f_{\theta(t)}$, *i.e.*,

$$DH(t,x)(dt, dx) = \frac{\partial}{\partial t} \nabla f_{\theta(t)}(x) dt + \operatorname{Hess} f_{\theta(t)}(x) dx.$$

It is readily checked that, as expected, $DH(t,x)$ is a mapping from $\mathbb{R} \times T_x X$ onto $T_x X$. Proposition 7 provides a closed form expression of the tangent space $T_{(t,x^*(t))}G(x^*)$ of $G(x^*)$ at $(t, x^*(t))$ and an illustration is proposed in Figure 3.2.

Lemma 7. The tangent space $T_{(t,x^*(t))}G(x^*)$ of the graph $G(x^*)$ at $(t, x^*(t))$ is

$$T_{(t,x^*(t))}G(x^*) = \ker(DH(t, x^*(t))).$$

Furthermore, the kernel of the mapping $DH(t, x^*(t))$ is

$$\ker(DH(t, x^*(t))) = \{\lambda(1, \xi) : \lambda \in \mathbb{R}\},$$

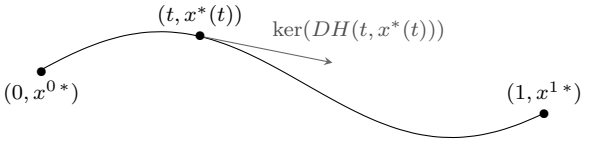

Figure 3. Schematic illustration of the graph $G(x^*)$ of the curve $x^* : [0,1] \to X$ of local minima corresponding to the curve $\theta : [0,1] \to \Theta$. At $(t, x^*(t)) \in G(x^*)$, the tangent space of the graph, which is a one-dimensional subspace of $\mathbb{R} \times T_{x^*(t)}X$, is given by $\ker(DH(t, x^*(t)))$.

where $\xi \in T_{x^*(t)}X$ is the unique solution to

$$\operatorname{Hess} f_{\theta(t)}(x^*(t))\xi = -\frac{\partial}{\partial t} \nabla f_{\theta(t)}(x^*(t)). \tag{7}$$

These results are sufficient to be able to develop an iterative algorithm that estimates the curve $x^* : [0,1] \to X$ of local minima. Given a predetermined sequence $\{t_k\}_{0 \le k \le K}$ arranged in increasing order such that $t_0 = 0$ and $t_K = 1$, it returns the sequence $\{x_k^*\}$ of local minima of the functions f_{θ_k}, where $\theta_k = \theta(t_k)$. The proposed method is described in Algorithm 1 and a schematic illustration is provided in Figure 3.2. Every iteration can be decomposed into two steps. The first one is the prediction step (lines 2-4 in Algorithm 1), which consists in following the direction dx_k^* in $T_{x_k^*}X$ provided by the tangent space of $G(x^*)$ at (t_k, x_k^*). It yields $y_k \in X$, which is obtained by taking the Riemannian exponential mapping of dx_k^* at x_k^*. The second one is the correction step (line 5 in Algorithm 1), where x_{k+1}^* is obtained by projecting y_k on the curve $x^* : [0,1] \to X$. This is achieved by minimising $f_{\theta_{k+1}}$ with the Newton method initialised at y_k.

Algorithm 1 Optimisation geometry algorithm

Require: $\{t_k\}_{0 \le k \le K}$ arranged in increasing order, with $t_0 = 0$ and $t_K = 1$; $\{\theta_k\}_{0 \le k \le K}$ such that $\theta_k = \theta(t_k)$; local minimum $x^{0\,*}$ corresponding to θ^0.

Ensure: Sequence $\{x_k^*\}_{0 \le k \le K}$ of local minima corresponding to each θ_k.

1: **for** $k = 0$ **to** K **do**
2: Solve $\operatorname{Hess} f_{\theta_k}(x_k^*)\xi_k = -\frac{\partial}{\partial t} \nabla f_{\theta_k}(x_k^*)$ for ξ_k.
3: Compute $dx_k^* = dt_k \xi_k$, where $dt_k = t_{k+1} - t_k$.
4: Compute $y_k = \exp_{x_k^*}^X(dx_k^*)$.
5: Compute x_{k+1}^* by solving $\operatorname{argmin}_{x \in X} f_{\theta_{k+1}}(x)$ with the Newton method initialised at y_k.
6: **end for**

3.3 Convergence analysis

In this part we give the sketch of a proof for the convergence of the homotopy method provided in Algorithm 1. A more rigorous and complete proof is left for future publications, as well as a discussion about the rate of convergence. Given a fibre-wise Morse function $f : X \times \Theta \to \mathbb{R}$, a fibre-wise local minimum $(\theta^0, x^{0\,*})$ of f and a path $\theta : [0,1] \to \Theta$ satisfying conditions of Corollary 5 such that $\theta(0) = \theta^0$ and $\theta(1) = \theta^1$, the aim is to show that there exist an integer K and a sequence $\{t_k\}_{0 \le k \le K}$ such that Algorithm 1 to the local minimum $x^{1\,*}$.

The idea is to show that there exists $\delta_0 > 0$ such that, for all k, we can choose $dt_k = t_{k+1} - t_k \ge \delta_0$ allowing to

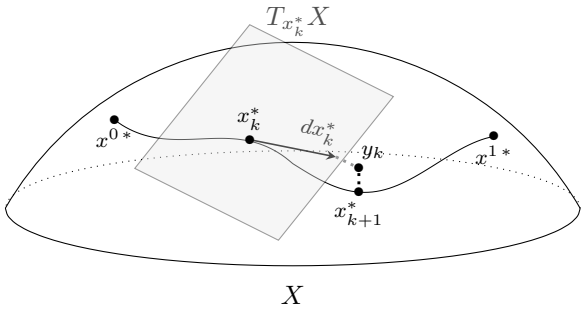

Figure 4. Schematic illustration of the procedure of Algorithm 1. Given iterate $x_k^* \in X$, a prediction step is achieved by computing $y_k \in X$ through the Riemannian exponential of $dx_k^* \in T_{x_k^*}X$, which is provided by the tangent space of the graph $G(x^*)$ at (t_k, x_k^*). A correction step is then performed in order to obtain the next iterate x_{k+1}^* by projecting the predicted point y_k onto $x^* : [0,1] \to X$.

predict a point $y_k = \exp^X_{x_k^*}(dt_k \xi_k) \in X$ sufficiently close to the local minimum x_{k+1}^* of $f_{\theta_{k+1}}$. By sufficiently close, we mean that there exists $\epsilon_0 > 0$ such that $d(y_k, x_{k+1}^*) < \epsilon_0$, where d is the Riemannian distance on X, and that the Newton method minimising $f_{\theta_{k+1}}$ converges to x_{k+1}^* for any starting point y satisfying $d(y, x_{k+1}^*) < \epsilon_0$ [1]. It follows that for Algorithm 1 to return $x^{1\,*}$, a sequence $\{t_k\}$ of at most $\left\lfloor \frac{1}{\delta_0} \right\rfloor + 1$ elements is required, where $\lfloor \cdot \rfloor$ is the floor function. To show this convergence result, the following points are to be proven:

(1) The convergence to a point on the curve $x^* : [0,1] \to X$ of the Newton method on line 5 of Algorithm 1 needs to be guaranteed when initialised with the predicted point y_k. We can show that there exists $\epsilon_0 > 0$ such that the Newton method converges to x_{k+1}^* if the initial point y satisfies $d(y, x_{k+1}^*) < \epsilon_0$. As ϵ_0 must not depend on k, this step basically requires f to admit basins of attractions of constant size across X and Θ.
(2) As $y_k = \exp^X_{x_k}(dt_k \xi_k)$, the only parameter we can pilot to ensure $d(y_k, x_{k+1}^*) < \epsilon_0$ is the step-size dt_k. We need to show that $d(y_k, x_{k+1}^*)$ can be bounded by an expression involving only constants and dt_k.
(3) The latter expression is then used to find a step-size dt_k that puts y_k in the convergence radius ϵ_0 of x_{k+1}^*.
(4) Finally, we need to check that the sequence $\{t_k\}$ does not converges to $l < 1$, *i.e.* that $x^{1\,*}$ can be reached in a finite number of steps. Hence, it is needed to prove that there exists $\delta_0 > 0$ such that, for all k, we have $d(y_k, x_{k+1}^*) < \epsilon_0$ for $dt_k \geq \delta_0$.

The remaining of this section details how the four previous points can be proven.

Point 1: By definition of fibre-wise Morse functions, f_θ is locally convex around each critical point x^*. Thus, the Newton method has a basin of convergence around each critical point x^*. Let ϵ_k be the radius of the basin around x_k^*. From the Morse condition, $\epsilon_k > 0$. It is then needed to show that the infimum ϵ_0 of ϵ_k over all possible x_k^* is strictly greater than zero. This can be achieved by (i) proving that ϵ_k is greater than a strictly positive continuous function in $X \times \Theta$ and (ii) using the fact that a continuous function on a compact set always reaches its bounds. Thus, we have $0 < \epsilon_0 < \epsilon_k$ for all possible ϵ_k. The minoring function in (i) can be built by exploiting implicit functions sending θ over a zero of the eigenvalue $\lambda_i(\mathrm{Hess}\, f_\theta(x))$.

Point 2: Bounding the Riemannian distance $d(y_k, x_{k+1})$ can be done in two steps. First, we bound the distance $d(x_k^*, x_{k+1}^*)$ between points on $x^* : [0,1] \to X$ and the distance $d(x_k^*, y_k)$. Second, the triangular inequality is used to obtained the wished bound. The curve $x^* : [0,1] \to X$ admits a Lipschitz constant L with respect to the distance d on X. It comes from the fact that Dx^* is continuous and therefore bounded on $[0,1]$. We have:

$$d(x^*(t_{k+1}), x^*(t_k)) \leq dt_k L. \tag{8}$$

Moreover, as $y_k = \exp^X_{x_k^*}(dx_k^*)$, we have $d(x_k^*, y_k) = \|dx_k^*\|_x = dt_k \|\xi_k\|_x$. Notice that $\xi_k = \nabla x^*(x_k) \in T_x X$, thus

$$d(x_k^*, y_k) = dt_k \| \nabla x^*(x_k)\|_{x_k} \tag{9}$$

The operator norm on $\mathcal{L}(T_{x_k}X)$ and the vector norm on $T_{x_k}X$ both derive from the Riemannian metric and are compatible. It follows that $\| \nabla x^*(x_k)\|_{x_k} \leq L$. Combining (9) and (8) hence gives

$$d(y_k, x_{k+1}) \leq 2dt_k L. \tag{10}$$

Point 3: With Equation (10), we can set $dt_k = \frac{\varepsilon_0}{2L}$ in order to get $\mathrm{d}(y_k, x_{k+1}^*) \leq \varepsilon_0$

Point 4: In point 2, we could have bounded Dx^* only locally, which would have given a bigger (hence better) step-size in point 3. However, taking the supremum of the derivative on $[0,1]$ brings that the step-size $dt_k = \frac{\varepsilon_0}{2L}$ is independent of k. Hence, we can simply choose $\delta_0 = \frac{\varepsilon_0}{2L}$.

4. ILLUSTRATION

In this section, an illustration of the proposed optimisation geometry method is provided. Even though this example is trivial, it illustrates the interest of the method in practice. Let $X = [0,1]$, $\Theta = [0,1]$ and

$$\begin{aligned} f : X \times \Theta &\to \mathbb{R} \\ (x, \theta) &\mapsto \left(x - \theta^2\right)^2. \end{aligned} \tag{11}$$

Given $\theta \in \Theta$, the global minimum x^* of $f_\theta : x \mapsto f(x, \theta)$ is simply $x^* = \theta^2$. For the sake of the example, we will however apply the optimisation geometry method to obtain it.

First, we need to verify that f is fibre-wise Morse on $M = [0,1] \times [0,1]$. As it is polynomial in x and θ, it is smooth on M. Furthermore, for all $\theta \in \Theta$,

$$\frac{d^2 f_\theta}{dx^2} = 2.$$

Hence, for all $x \in X$, the second order derivative of f_θ is positive definite. It is in particular true at the critical point

[1] In Section 3.2, we consider $H(t,x) = 0_x$ for simplicity. We argue that having $\|H(t,x)\|_x \leq \epsilon$ is enough, where $\|\cdot\|_x$ is the norm on $T_x X$ and $\epsilon > 0$ is the tolerance. This is enough to ensure the convergence of the Newton method in a finite number of iteration.

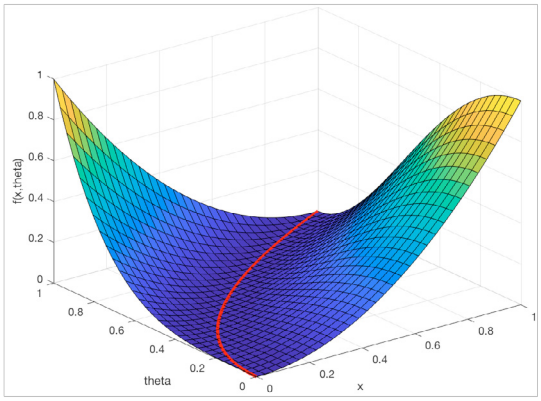

Figure 5. Graph of the function $f : X \times \Theta \to \mathbb{R}$ defined in (11) with $X = [0, 1]$ and $\Theta = [0, 1]$. The red curve corresponds to the submanifold of fibre-wise critical points of f embedded Guillemin and Pollack (2010) in the graph of f. Given θ^0 for which the minimum x^{0*} is known, the optimisation geometry method aims at following this curve in order to find the solution x^{1*} of the optimisation problem for θ^1.

of f_θ. It follows that f is indeed fibre-wise Morse on M and our optimisation geometry method can be employed.

Let $\theta^0 \in \Theta$, for which the global minimum of f_{θ^0} is $x^{0*} = \theta^{0\,2}$, and $\theta^1 \in \Theta$, for which the corresponding minimum x^{1*} is assumed to be unknown. Let the curve

$$\begin{aligned} \theta : [0, 1] &\to \Theta \\ t &\mapsto t\theta^1 + (1 - t)\theta^0. \end{aligned}$$

The first order derivative of $f_{\theta(t)}$ is

$$df = 2\left(x - \theta(t)^2\right) dx.$$

It follows that the function $H : [0, 1] \times X \to T_x X \simeq \mathbb{R}$ defined in Proposition 6 is

$$H(t, x) = 2\left(x - \theta(t)^2\right).$$

Its first derivative is

$$DH(t, x)(dt, dx) = 2dx - 4\theta(t)\dot{\theta}(t)dt,$$

where $\dot{\theta}(t) = \theta^1 - \theta^0$. Thus, solving equation in Proposition 7, one can check that $\ker(DH(t, x)) = \{\lambda(1, \xi) : \lambda \in \mathbb{R}\}$, where

$$\xi = 2\theta(t)\dot{\theta}(t) = 2(\theta^1 - \theta^0)(\theta^0 + (\theta^1 - \theta^0)t).$$

Let $\theta^0 = 0$ ($x^{0*} = 0$), $\theta^1 = 1$ and $\{t_k\} = \{0, 0.2, \ldots, 0.8, 1\}$ (*i.e.*, $dt_k = 0.2$). Within these settings, $\theta_k = \theta(t_k) = t_k$ for all k. Algorithm 1 proceeds as follows at the k^{th} iteration:

- the direction ξ_k is given by $\xi_k = 2t_k$;
- the resulting predicted point is $y_k = x_k^* + 2t_k dt_k$;
- the corrected point $x_{k+1}^* = \theta_{k+1}^2$ is obtained with one iteration of the Newton method [2].

An illustration of the procedure can be found in Figure 4.

5. CONCLUSION

We presented in this paper a Riemannian-based homotopy algorithm that provides an original solution to optimisation problems over families of function. The main interest

[2] Note that the important point here is not the Newton method but the homotopy procedure.

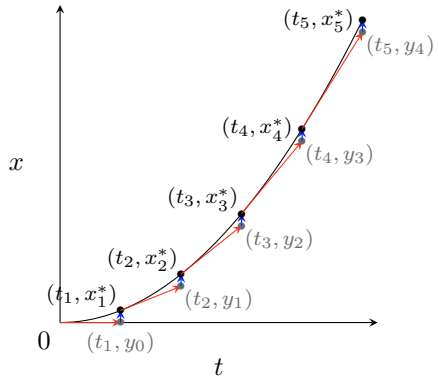

Figure 6. On this figure, all the steps of the homotopy algorithm are represented. The black curve is the graph of x^*, which usually has no closed form. The red arrow show the prediction step, in the direction of the tangent to the black curve. The blue arrow shows the projection back to the curve achieved by a descent method such as the Newton method. This is the correction step. Size of dt was taken constant equal to 0.2.

of this method is to focus on the complexity of the family of function rather than the complexity of an individual cost function.

REFERENCES

Allgower, E.L. and Georg, K. (2012). *Numerical continuation methods: an introduction*, volume 13. Springer Science & Business Media.

Gabay, D. (1982). Minimizing a differentiable function over a differential manifold. *Journal of Optimization Theory and Applications*, 37(2), 177–219. doi:10.1007/BF00934767.

Garg, D., Patterson, M.A., Francolin, C., Darby, C.L., Huntington, G.T., Hager, W.W., and Rao, A.V. (2011). Direct trajectory optimization and costate estimation of finite-horizon and infinite-horizon optimal control problems using a Radau pseudospectral method. *Computational Optimization and Applications*, 49(2), 335–358.

Guillemin, V. and Pollack, A. (2010). *Differential topology*, volume 370. American Mathematical Soc.

Harandi, M., Salzmann, M., and Hartley, R. (2017). Joint dimensionality reduction and metric learning: A geometric take. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1404–1413. JMLR. org.

Kay, S.M. (1993). *Fundamentals of statistical signal processing*. Prentice Hall PTR.

Manton, J.H. (2004). A globally convergent numerical algorithm for computing the centre of mass on compact Lie groups. In *ICARCV 2004 8th Control, Automation, Robotics and Vision Conference, 2004.*, volume 3, 2211–2216. IEEE.

Manton, J.H. (2015). A framework for generalising the Newton method and other iterative methods from Euclidean space to manifolds. *Numerische Mathematik*, 129(1), 91–125.

Manton, J. (2002). Optimization algorithms exploiting unitary constraints. *IEEE Transactions on Signal Processing*, 50(3), 635–650. doi:10.1109/78.984753.

Manton, J.H. (2013). Optimisation Geometry. In *Mathematical System Theory — Festschrift in Honor of Uwe Helmke on the Occasion of His Sixtieth Birthday*, 261–274. CreateSpace. Http://arxiv.org/abs/1212.1775.

Milnor, J.W., Spivak, M., and Wells, R. (1969). *Morse theory*, volume 1. Princeton university press Princeton.