mathematics

MDPI

Article

When Will a Sequence of Points in a Riemannian Submanifold Converge?

Tuyen Trung Truong

Department of Mathematics, University of Oslo, Blindern, 0851 Oslo, Norway; tuyentt@math.uio.no

Received: 26 August 2020; Accepted: 19 October 2020; Published: 3 November 2020

Abstract: Let X be a Riemannian manifold and x_n a sequence of points in X. Assume that we know a priori some properties of the set A of cluster points of x_n. The question is under what conditions that x_n will converge. An answer to this question serves to understand the convergence behaviour for iterative algorithms for (constrained) optimisation problems, with many applications such as in Deep Learning. We will explore this question, and show by some examples that having X a submanifold (more generally, a metric subspace) of a good Riemannian manifold (even in infinite dimensions) can greatly help.

Keywords: compact metric space; deep neural networks; random dynamical systems; global convergence of Gradient Descent; iterative optimisation; Nash conjecture; Nash's embedding theorem

In this paper we will explain how the geometry of submanifolds of $\mathbb{R}^k$ is useful to optimisation problems in Deep Learning, and we explore similar properties for other manifolds.

1. Motivation

Optimisation is important in many aspects of engineering and computer sciences. For a modern example, one can mention deep neural networks, which can solve effectively several tasks (image/video classification, natural language processing and so on) that posed enormous challenges for the old paradigm of based-rule learning. For deep neural networks to be able to work, one has to solve large scale and non-convex optimisation. For example, modern state-of-the-art deep neural network architectures give rise to optimisation problems (finding minima) in hundred of million variables.

For an explicit example, we consider the case of recognising of hand written digits, useful for example when scanning postal packages. A well-known dataset is MNIST [1]. A sample is in Figure 1, several of them can be challenging even for human beings. This task is extremely difficult for the old paradigm of based-rule learning, but is considered a simple task for deep neural networks. For this task, one can use a "simple" deep neural network, which gives rise to an optimisation problem in about 12,000 variables.

Because of this large scale and non-convex feature of associated optimisation problems, one must confine with iterative numerical methods. One would like the method to guarantee convergence to local minima. This can be divided into two steps. First, show that the method converge, and then show that the limit point is a local minimum. Since saddle points are dominant for functions in higher dimensions [2,3], for the second step it is important to guarantee that the limit point is not a saddle point.

It could be considered as a lucky fact that from beginning Gradient Descent methods (GD) have been used in deep neural networks. At first, it could be the fact that GD is easy to implement and is not costly to run in large scale optimisation. Then, even though it dates back more than 170 years [4], only gradually (with some results announced only very recently) it has been shown that GD has good properties: it can avoid saddle points [5,6]. While its standard version does not guarantee convergence to critical points, its Backtracking version [7] does [8–10] (the latter paper consists of the more experimental part of arXiv:1808.05160, in combination with arXiv:2001.02005 and arXiv:2007.03618) and can be implemented in deep neural networks with very good performance on CIFAR10 and CIFAR100 image datasets [10,11]. Some further modifications of Backtracking GD can avoid saddle points as well [12,13].

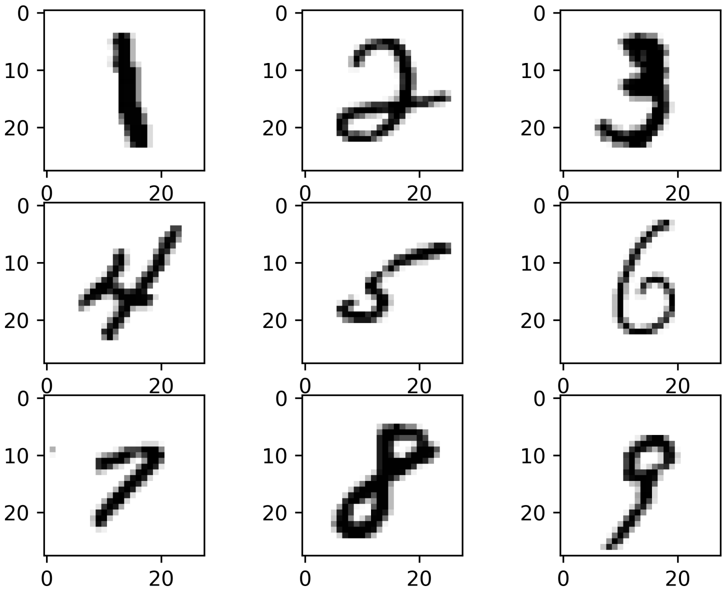

Figure 1. Some samples from the MNIST hand-written digit dataset. (Source: [1]).

2. Convergence Results

We will, for the remainder of this section, discuss convergence to critical points for GD methods, which uses special geometrical properties. First, we recall in detail the update rule in GD. We consider a function $f : \mathbb{R}^k \to \mathbb{R}$, which is assumed to be C^1. We want to find minima of f. One starts from a random initial point x_0, and construct a sequence $x_{n+1} = x_n - \delta_n \nabla f(x_n)$, where $\delta_n > 0$ is an appropriate number. There are many ways to choose δ_n. In the Standard GD scheme, one chooses δ_n to be a constant δ_0. A disadvantage of Standard GD is that it does not guarantee convergence, to have good behaviour one must assume that f is in $C^{1,1}_L$, that is ∇f is globally Lipschitz continuous with the Lipschitz constant L, and further assume that δ_0 is in the order of $1/L$. There are many popular modifications trying to overcome this, such as Adam, Adadelta, Nesterov Accelerated Gradient, Momentum and so on (see [14] for a review), none of these are guaranteed to converge in general either. To date, only Backtracking GD is guaranteed to converge: see Chapter 12 in [9], in particular Proposition 12.6.1 there, for the case $f \in C^{1,1}_L$ and has compact sublevels and has at most countably many critical points, see [8] when f is real analytic (or more generally satisfies the so-called Losjasiewicz gradient inequality), and see [10] for the general case of f being C^1 only and has at most countably many critical points. Note that the assumption in the last paper is not too restrictive: indeed, it is known from transversality results that such an assumption is satisfied by a generic C^1 function (for example, by Morse's functions, which are a well-known class of functions in geometry and analysis). Since the real analyticity assumption in [8] is quite special, we will not discuss about it in the below, trying to provide only the most general ideas.

Both [9,10] start from the following property: If $\{x_n\}$ is constructed as above, and $\{x_{n_j}\}$ is a convergent subsequence, then $\lim_{j\to\infty} \nabla f(x_{n_j}) = 0$. This is classically known (see [15]), the main idea is as follows: if $\lim_{j\to\infty} x_{n_j} = x_\infty$ and $\nabla f(x_\infty) \neq 0$, then $\liminf_{j\to\infty} \delta_{n_j} > 0$. The latter is a contradiction, when Armijo's condition is taken into account.

Now, one needs also the following property:

Property 1. *Either* $\lim_{n\rightarrow\infty} f(x_n) = -\infty$, *or* $\lim_{n\rightarrow\infty} ||x_{n+1} - x_n|| = 0$.

In case f has compact sublevels, then this is easily proven [9]. For the general case, see [10] for a proof.

Now, one needs a special property of compact metric spaces [16]. We recall that given a sequence $\{x_n\}$, its set of cluster points consists of points x so that there is a subsequence $\{x_{n_j}\}$ for which $\lim_{j\rightarrow\infty} x_{n_j} = x$.

Theorem 1. *Let* (X, d) *be a compact metric space. If* $\{x_n\} \subset X$ *is a sequence so that* $\lim_{n\rightarrow\infty} d(x_{n+1}, x_n) = 0$, *then the set of cluster points of* $\{x_n\}$ *is connected.*

In the setting of [9], one can finish the proof of convergence as follows: Let $X = \{x \in \mathbb{R}^k : f(x) \leq f(x_0)\}$. Since f is assumed to have compact sublevels, it follows that X is compact. Let d be the restriction to X of the usual metric on $\mathbb{R}^k$, then (X, d) is a compact metric space. Since we know that $\lim_{n\rightarrow\infty} d(x_n, x_{n+1}) = 0$ for the constructed sequence, we can then apply Theorem 1 for the sequence $\{x_n\}$, and have that the set of cluster points $\mathcal{D}$ of $\{x_n\}$ is connected. Since we also know that $\mathcal{D}$ must be contained in the set of critical points $\mathcal{C} = \{x : \nabla f(x) = 0\}$ and by assumption $\mathcal{C}$ is countable, it follows that $\{D\}$ is also countable. Since a countable and connected set must be either empty or one point, it follows that either $\lim_{n\rightarrow\infty} ||x_n|| = \infty$ (the first case) or $\{x_n\}$ converges. Note that in this case, since $\{x_n\}$ is bounded, it follows that only the second case happens, that is $\{x_n\}$ converges.

In the general case, the above proof does not go through, since the set X in the above may not be bounded. In [10], a way to go around is as follows. We let $(\mathbb{P}^k, d)$ be the real projective space with its canonical metric (the spherical metric). We let $(\mathbb{R}^k, ||.||)$ be the usual Euclidean metric on $\mathbb{R}^k$, and let $(\mathbb{R}^k, d)$ be the restricted metric of d. Then, **topologically** , the two spaces $(\mathbb{R}^k, ||.||)$ are homeomorphic. In particular, convergence properties of a sequence $\{x_n\}$ in $(\mathbb{R}^k, ||.||)$ can be translated to that of the same sequence but considered in $(\mathbb{R}^k, d)$. Even though they are not isometric, one can check that if $\lim_{n\rightarrow\infty} ||x_{n+1} - x_n|| = 0$, then also $\lim_{n\rightarrow\infty} d(x_n, x_{n+1}) = 0$. In addition, $(\mathbb{P}^k, d)$ is a compact metric space. Hence, one can apply Theorem 1, and have that for the constructed sequence $\{x_n\}$, considered in $(\mathbb{P}^k, d)$ the set of cluster points $\overline{\mathcal{D}}$ is connected. Since $\overline{\mathcal{D}} \cap \mathbb{R}^k = \mathcal{D} \subset \mathcal{C}$ and $\mathcal{C}$ is countable, it follows that if $\mathcal{D}$ is non-empty then it is a point. Hence, we obtain the same conclusions as before. Note that here the case $\lim_{n\rightarrow\infty} ||x_n|| = \infty$ can happen, for example for the function $f(x) = x^3$.

3. Riemannian Manifold Optimisation

Now we discuss the general setting of optimisation on manifolds. Let X be a real manifold and $f : X \rightarrow \mathbb{R}$. We want to find (local) minima of f. Here one could, as before, try to use Standard GD. To prove good properties for Standard GD, one could, as before, restrict the discussion to $C_L^{1,1}$ functions. However, here it is very cumbersome to define a global $C_L^{1,1}$ notion in the manifold setting. It is better to switch to Backtracking GD, which is local in nature, and hence has a natural extension to the manifold setting. We let it to the readers to state a specific version of Backtracking GD in the manifold setting, by working on small coordinate charts.

Taking the ideas from the proof in [10], as presented above, we obtain the following general result. We say that a sequence $\{x_n\}$ in a metric space X diverges to infinity if $\{x_n\}$ eventually leaves every bounded subset of X.

Theorem 2. *Let X be a Riemmanian manifold, with the induced metric d. Assume that there is a compact metric space (Z, d_Z), together with a homeomorphism $h : X \rightarrow h(X) \subset Z$ such that $d_Z(h(x_1), h(x_2)) \leq d(x_1, x_2)$ for all $x_1, x_2 \in X$. Let $f : X \rightarrow \mathbb{R}$ be a C^1 function, and $\{x_n\}$ a sequence constructed by Backtracking GD. Assume that f has at most countably many critical points. Then, either $\{x_n\}$ converges to a critical point of f, or $\{x_n\}$ diverges to infinity.*

For clarity of the next discussion, we define a class of Riemannian manifolds.

Definition 1. *A Riemannian manifold* (X, d) *is called non-expandingly homeomorphically compactible if there is a compact metric space* (Z, d_Z)*, together with a homeomorphism* $h : X \to h(X) \subset Z$ *such that* $d_Z(h(x_1), h(x_2)) \leq d(x_1, x_2)$ *for all* $x_1, x_2 \in X$*.*

To be able to apply Theorem 2, it remains to find what manifolds are non-expandingly homeomorphically compactible. We give first some examples.

Example 1. *X is a bounded subset of $\mathbb{R}^l$. This is the case in [9].*

Example 2. *X is a subset of $\mathbb{R}^l$. This is the case in [10] and is more general than Example 1, when X is unbounded. On the other hand, it is interesting to note that an X from Example 2 is* ***homeomorphic*** *to a one in Example 1. Indeed, by Theorem 3.4.4 in [17], the real projective space $\mathbb{P}^l$ is* ***affine****, that is isometric to a compact subset of $\mathbb{R}^{(l+1)^2}$.*

Remark 1. *In a very recent paper [18], the author has been able to extend previously mentioned results, including Theorem 2 and a modification of Newton's method, to general Riemannian manifolds. There, the readers can consult more general algorithms on Riemannian manifolds. In particular, by using Nash's embedding theorem [19–21], every Riemannian manifold is non-expandingly homeomorphically compactible and hence Theorem 2 holds in general.*

From the above discussion, we are motivated to state the following question. Even partial answers to it will be greatly helpful for optimisation on manifolds.

Question 1. In the statement of Theorem 2, if one does not assume that f has at most countably many critical points, will the conclusions of Theorem 2 still hold?

Currently, we do not know about the answer to Question 1. We have just a couple of comments. First, by transversality theorem, functions satisfying Theorem 2 is dense, and hence for practical purposes the Theorem is applicable in realistic setttings. Second, we have done many examples on various types of benchmark functions, and found that the sequence constructed by Backtracking GD always satisfies the conclusions of Theorem 2. Hence, we conjecture that the answer to Question 1 is affirmative, at least when the initial point x_0 is randomly chosen.

Funding: This work is supported by Young Research Talents grant number 300814 from Research Council of Norway.

Acknowledgments: The author thanks the editors of the special issue "Riemannian geometry of submanifolds" in the journal Mathematics (MDPI), for invitation and help. The author would like to thank the referees for helpful feedbacks.

Conflicts of Interest: The author declares no conflict of interest.

References

1. LeCun, Y.; Cortes, C.; Burges, C. MNIST Handwritten Digit Database. 2010. Available online: http://yann.lecun.com/exdb/mnist/ (accessed on 1 April 2020).
2. Bray, A.J.; Dean, D.S. Statistics of critical points of gaussian fields on large-dimensional spaces. *Phys. Rev. Lett.* **2006**, *98*, 150201. [CrossRef] [PubMed]
3. Dauphin, Y.N.; Pascanu, R.; Gulcehre, C.; Cho, K.; Ganguli, S.; Bengjo, Y. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In Proceedings of the 27th International Conference on Neural Information Processing Systems NIPS' 14, Montreal, QC, Canada, 8–13 December 2014; Volume 2, pp. 2933–2941.
4. Cauchy, A.Method général pour la résolution des systemes d'équations simulanées. *C. R. Math.* **1847**, *25*, 536.

5. Lee, J.D.; Simchowitz, M.; Jordan, M.I.; Recht, B. Gradient descent only converges to minimizers. *JMRL* **2016**, *49*, 1–12.
6. Panageas, I.; Piliouras, G. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS 2017), Berkeley, CA, USA, 9–11 January 2017; Volume 2, pp. 1–12.
7. Armijo, L. Minimization of functions having Lipschitz continuous first partial derivatives. *Pac. J. Math.* **1966**, *16*, 1–3. [CrossRef]
8. Absil, P.-A.; Mahony, R.; Andrews, B. Convergence of the iterates of descent methods for analytic cost functions. *SIAM J. Optim.* **2005**, *16*, 531–547. [CrossRef]
9. Lange, K. *Optimization*, 2nd ed.; Springer Texts in Statistics; Springer: New York, NY, USA, 2013.
10. Truong, T.T.; Nguyen, T.H. Backtracking Gradient Descent method and some applications to Large scale optimisation. Part 2: Algorithms and experiments. *Appl. Math. Optim.* **2020**. [CrossRef]
11. Vaswani, S.; Mishkin, A.; Laradji, I.; Schmidt, M.; Gidel, G.; Lacoste-Julien, S. Painless Stochastic Gradient: interpolation, line-search and convergence rates. paper #8630. NeurIPS 2019. *arXiv* **2019**, arXiv:1905.09997.
12. Truong, T.T. Some convergent results for Backtracking Gradient Descent method on Banach spaces. *arXiv* **2020**, arXiv: 2001.056768.
13. Truong, T.T. Convergence to minima for the continuous version of Backtracking Gradient Descent. *arXiv* **2019**, arXiv: 1911.04221.
14. Ruder, S. An overview of gradient descent optimisation algorithms. *arXiv* **2017**, arXiv:1609.04747.
15. Bertsekas, D.P. *Nonlinear Programming*, 2nd ed.; Athena Scientific: Belmont, MA, USA, 1999.
16. Asic, M.D.; Adamovic, D.D. Limit points of sequences in metric spaces. *Am. Math. Month.* **1970**, *77*, 613–616. [CrossRef]
17. Bochnak, J.; Coste, M.; Roy, M.-F. *Real Algebraic Geometry*; A Series of Modern Surveys in Mathematics; Springer: Berlin/Heidelberg, Germany, 1998; Volume 36.
18. Truong, T.T. Unconstrained optimisation on Riemannian manifolds. *arXiv* **2006**, arXiv:2008.11091.
19. Kuiper, N.H. On C^1-isometric imbeddings, I and II. *Indag. Math. Proc.* **1955**, *58*, 545–556, 683–689. [CrossRef]
20. Nash, J. The Imbedding problem for Riemannian manifolds. *Ann. Math.* **1956**, *63*, 20–63. [CrossRef]
21. Nash, J. C^1-isometric imbeddings. *Ann. Math.* **1955**, *60*, 383–396. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.