Steepest descent method on a Riemannian manifold: the convex case

Julien Munier

Steepest descent method on a Riemannian manifold: the convex case

Julien Munier

Abstract: In this paper we are interested in the asymptotic behavior of the trajectories of the famous steepest descent evolution equation on Riemannian manifolds. It writes

$$\dot{x}(t) + \mathrm{grad}\phi(x(t)) = 0.$$

It is shown how the convexity of the objective function ϕ helps in establishing the convergence as time goes to infinity of the trajectories towards points that minimize ϕ. Some numerical illustrations are given for the Rosenbrock's function.

Keywords: Riemannian manifold, gradient flow, steepest descent, convexity, minimization

AMS Subject Classification: 34C40,37N40, 37M05,65J15

1 Introduction

Let $(M, \langle ., . \rangle)$ be a manifold endowed with a Riemannian metric. Reference on this field are Do Carmo [8] or [10]. The associated length of a curve is denoted $\ell(.)$, and the distance $\rho(.,.)$. Let $\phi : M \to \mathbb{R}$ be a differentiable function, we denote $\mathrm{grad}\phi(x)$ the gradient with respect to the Riemannian inner product $\langle ., . \rangle_x$, which means that for any $u \in T_x M$, the tangent space to the manifold at point x, following holds th

$$d\phi(x)\, u = \langle \mathrm{grad}\phi(x), u \rangle_x.$$

We will consider in this paper the following evolution equation

$$\dot{x}(t) + \mathrm{grad}\phi(x(t)) = 0 \tag{1}$$

under the assumptions

$$(\mathbf{H_1}) \quad \begin{cases} (M, \rho) \text{ is complete,} \\ \phi \in C^1 \text{ with a locally Lipschitz continuous gradient} \\ \phi \text{ bounded from below} \end{cases}$$

with the following definition, which appears in [2]

Definition 1 *A vector field X is L-*Lipschitz continuous *on a subset $U \subset M$ if for all geodesic curve $\gamma : [0,1] \to M$ with endpoints in U*

$$|P_{\gamma,1,0} X(\gamma(1)) - X(\gamma(0))|_{\gamma(0)} \leq L\ell(\gamma).$$

Here $P_{\gamma,1,0}$ is the parallel transport from time 1 to time 0 along γ (see [8]).

Equation (1) generalizes to a Riemannian setting the continuous version of the famous *steepest descent* equation.This dynamical system is intemely linked with the problem of minimizing ϕ, see [4, 5]. The framework of manifolds is interesting, for instance in order to deal with non-linear equality constraints. Indeed, the following problems

$$\min_{x \in \mathbb{R}^n, f(x)=0} \phi(x)$$

and

$$\min_{x \in M} \phi(x)$$

are equivalent, if we denote $M = f^{-1}(0) \subset \mathbb{R}^n$. If the function f is regular, the set M can be endowed with the Riemannian structure of submanifold of $\mathbb{R}^n$.

This evolution equation has already been studied by Udrişte in [15] chap.7 §1. under the name "minus gradient line". Theorems 1 and 2 are proved under slightly stronger assumptions, but our other results are new (up to our knowledge). The case of a Morse-Bott objective function ϕ is studied in [11]. In the case of an open subset of $\mathbb{R}^n$, links between trajectories of (1) and central paths are made in [12] (see also [3, 11] and references therein). For a submanifold of $\mathbb{R}^n$ which is the intersection of an open convex set with an affine subspace can be found in [1]. The Riemannian structure there comes from the Hessian of a Legendre barrier function. Some discrete versions of (1) appear in [15, 6, 7, 9, 14].

2 Existence and first results

All the results that appear in this section are true under the assumptions $(\mathbf{H_1})$. They are similar to results that exist in Hilbert spaces.

Theorem 1 *Assume* $(\mathbf{H_1})$ *hold. For any starting point* $x_0 \in M$*, the Cauchy problem*

$$\begin{cases} \dot{x}(t) + \text{grad}\phi(x(t)) = 0 \\ x(0) = x_0 \end{cases}$$

has a unique solution $x(.)$ *defined on* $[0, +\infty)$*. This solution is continuously differentiable.*

Proof: Local existence and uniqueness comes when writing (1) in a local chart. It becomes a system of n ODEs and we can apply Cauchy results. This provides a maximal solution defined on an interval of type $[0, T)$. Assume by contradiction that $T < +\infty$. As

$$\frac{d}{dt}\phi(x(t)) = \langle \text{grad}\phi(x(t)), \dot{x}(t) \rangle_{x(t)} = -|\dot{x}(t)|^2_{x(t)} \tag{2}$$

we see that

$$\int_0^T |\dot{x}(t)|_{x(t)} dt \leq \sqrt{T\left(\phi(x_0) - \inf_M \phi\right)}.$$

Thus $\dot{x}(t)$ is integrable, and $x(.)$ has a limit when t tends to T, because of completeness of M. We can then extend this solution, which contradicts the maximality. □

Theorem 2 *Let* $x(.)$ *be a solution of (1). Under assumptions* $(\mathbf{H_1})$*, it satisfies*

(i) $\phi(x(.))$ *decreases and converges,*

(ii) $|\dot{x}(.)|_{x(.)} \in L^2(0,+\infty)$.

Proof: It comes from (2). As ϕ is bounded from below, $\phi(x(t))$ has a limit, say ϕ_∞, when t goes to $+\infty$. □

Theorem 3 *Assume* $(\mathbf{H_1})$ *hold. Let* $x(.)$ *be a solution of (1) assumed bounded. It satisfies*

$$\lim_{t\to+\infty} |\text{grad}\phi(x(t))|_{x(t)} = 0.$$

Proof: Let L be the Lipschitz constant for gradϕ on the bounded set $\{x(t), t \in [0,+\infty)\}$. Denote, for $s \leq t$, by $\gamma : [0,1] \to M$ a minimizing geodesic such that $\gamma(0) = x(s)$ and $\gamma(1) = x(t)$. Since $P_{\gamma,1,0}$ is an isometry, we have the following

$$\begin{aligned}
\left||\dot{x}(t)|_{x(t)} - |\dot{x}(s)|_{x(s)}\right| &\leq |P_{\gamma,1,0}\dot{x}(t) - \dot{x}(s)|_{x(s)} \\
&\leq L\ell(\gamma) \\
&\leq L\int_s^t |\dot{x}(\tau)|_{x(\tau)} d\tau \\
&\leq \left(L\int_0^{+\infty} |\dot{x}(\tau)|^2_{x(\tau)} d\tau\right)\sqrt{t-s}
\end{aligned}$$

Thus $|\dot{x}(.)|_{x(.)}$ is uniformly continuous and square integrable, and necessarily $\lim_{t\to+\infty} |\dot{x}(t)|_{x(t)} = 0$. This achieves the proof. □

3 Convex case

In this section, assumptions $(\mathbf{H_1})$ still hold, and we assume moreover that

$$(\mathbf{H_2}) \quad \begin{cases} \phi \text{ convex on } M, \\ \text{argmin}\phi \neq \emptyset. \end{cases}$$

We recall here the definition of a convex function. It comes from [15]. A subset $A \subset M$ is called totally convex if for all geodesic curve γ, we have $(\gamma(0), \gamma(1)) \in A \times A \Rightarrow \forall t \in [0,1], \gamma(t) \in A$.

Definition 2 $f : A \to \mathbb{R}$ *is said* convex *on* A*, a totally convex set, if for all geodesic curve* γ *with* $(\gamma(0), \gamma(1)) \in A \times A$ *and for all* $t \in [0,1]$ *we have*

$$f(\gamma(t)) \leq (1-t) f(\gamma(0)) + t f(\gamma(1)).$$

We will remark at the end of the section that this definition can be weakened, in what concerns this work.

Denote for some y in M and for every t, $u(t)$ a vector of $T_{x(t)}M$ such that

$$\begin{array}{l} \exp_{x(t)}(u(t)) = y \\ \rho(x(t), y) = |u(t)|_{x(t)} \end{array} \tag{3}$$

Such a vector always exists if M is complete. It may not be unique, but this is not a problem. The curve $[0,1] \ni s \mapsto \exp_{x(t)}(su(t))$ is thus a minimizing geodesic that joins $x(t)$ and y.

Proposition 1 *Under assumptions* $(\mathbf{H_1})$*, if moreover* ϕ *is convex on* M *and* y *belong to the set* $L = \{y \in M, \phi(y) \leq \phi_\infty\}$*, the function* $t \mapsto \rho(x(t), y)$ *decreases, and thus converges.*

Proof: For fixed t and h, in order to simplify the notations, x, g, u and x_h will respectively denote $x(t)$, $\mathrm{grad}\phi(x(t))$, $u(t)$ and $x(t+h)$. Consider the geodesic $\gamma(s) = \exp_x(su)$. We have $\gamma(0) = x$, $\dot{\gamma}(0) = u$ and $\gamma(1) = y$. Since $\phi \circ \gamma$ is convex, we get

$$\langle g, u \rangle_x \leq \phi(y) - \phi(x).$$

The latter is nonpositive, and is null iff $\phi(x) = \phi(y) \leq \phi_\infty \leq \phi(x)$. Thus $\phi(x(.))$ would be constant on $[t, +\infty)$, which means with (1) and (2) that $g = 0$. In this case, the trajectory is stationary, and the proposition is easy. We can restrict to the case when $\langle g, u \rangle_x < 0$.
Since the opposite of u belongs to the normal cone at the point x of the set $\{z \in M, \rho(y,z) \leq \rho(y,x)\}$, this suffices to show the desired decreasingness. We give in the following a geometrical idea of this fact. As both paths $h \mapsto x_h$ and $h \mapsto \exp_x(-hg)$ are C^1, and have the same initial condition of order 0 and 1, we have

$$\rho(x_h, \exp_x(-hg)) = o(h). \tag{4}$$

An argument of same type gives

$$\rho(\exp_x(-hg), \exp_x(h\lambda u)) = |-hg - h\lambda u|_x + o(h). \tag{5}$$

Choose λ such that $\langle -g - \lambda u, -g \rangle_x = 0$ (that is $\lambda = \frac{|g|_x^2}{-\langle u, g \rangle_x} > 0$). It gives $|-hg - h\lambda u|_x = h\left((\lambda|u|_x)^2 - |g|^2\right)^{\frac{1}{2}}$. Finally

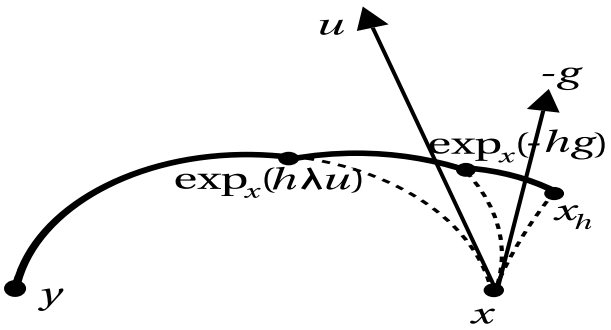
$$\rho(\exp_x(h\lambda u), y) = (1 - h\lambda)|u|_x. \tag{6}$$

Combining (4), (5) and (6), we are now able to construct a continuous and piecewise differentiable curve α (in bold in the figure below) that joins x_h to y of length

$$\ell(\alpha) = |u|_x - h[\lambda|u|_x - \left((\lambda|u|_x)^2 - |g|^2\right)^{\frac{1}{2}}] + o(h).$$

The bracket just above is positive. Indeed $(\lambda|u|_x)^2 > (\lambda|u|_x)^2 - |g|^2 \geq 0$. Thus, for h small enough, we have

$$\rho(x_h, y) \leq \ell(\alpha) \leq |u|_x = \rho(x, y).$$



□

According to this result, we can state the following theorem. Such a result was proved in [5] in the case of a Hilbert space.

Theorem 4 *Under the complete assumptions* $(\mathbf{H_1})$-$(\mathbf{H_2})$, *every solution* $x(.)$ *of (1) has a limit* $\bar{x}$ *which belongs to* argminϕ

Proof: Under the assumptions, the set L is non-empty. Then $x(.)$ is bounded in M. According to Hopf-Rinow Theorem, there exists a sequence t_j such that $x(t_j)$ converges to some $\bar{x}$. But $\bar{x}$ belongs to L, and then $\rho(x(.), \bar{x})$ converges. The limit has to be 0 (consider the sequence t_j), that is to say $x(t)$ converges to $\bar{x}$.
Consider now some $y \in$ argminϕ. As previously, the convexity of ϕ implies that

$$\begin{aligned} 0 \leq \phi(x(t)) - \phi(y) &\leq \langle -\text{grad}\phi(x(t)), u(t)\rangle_{x(t)} \\ &\leq |\text{grad}\phi(x(t))|_{x(t)} \rho(x(t), y) \end{aligned}$$

One part tends to 0 (Theorem 3) and the other one is bounded in the latter, which shows that

$$\phi(\bar{x}) = \lim_{t\to+\infty} \phi(x(t)) = \phi(y) = \inf_M \phi.$$

□

The two following remarks allow to weaken the assumptions of this result.

Remark 1 *To ask ϕ to be convex on the whole manifold is strong, it suffices for ϕ to be convex on the set $\{z \in M, \phi(z) \leq \phi(x(0))\}$.*

Remark 2 *Another assumption here is too strong. The convexity definition is meaningless on compact manifolds. Indeed there is no non-constant convex function on such a manifold. So a better definition would have been*

$\phi : A \to \mathbb{R}$ is convex on A, a totally *minimizing* geodesic set, if the restriction to any *minimizing* geodesic is convex.

The proof of the preceding result remains true in that framework.

4 Example: Rosenbrock's function

This example comes from [15]. The function present a narrow valley with a "U" form (actually a parabola). It is plotted in the left hand side of figure 4 and is defined by

$$\begin{array}{rccl} \phi & : & \mathbb{R}^2 & \to & \mathbb{R} \\ & & (x_1, x_2) & \mapsto & 100\left(x_2 - x_1^2\right)^2 + (1 - x_1)^2. \end{array}$$

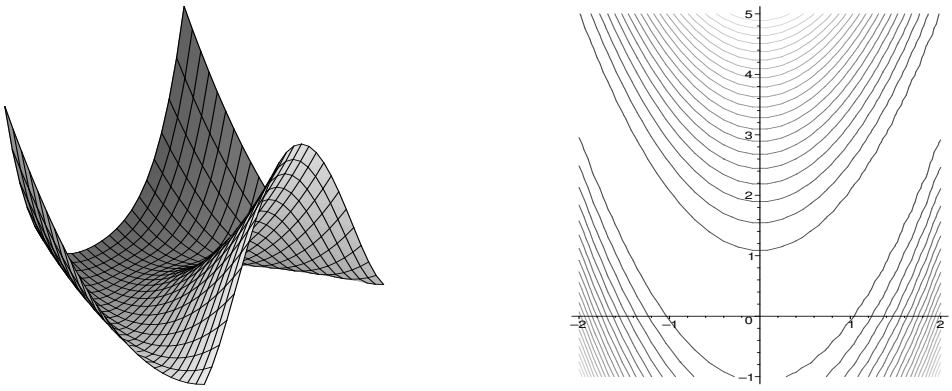


Figure 1: Rosenbrock's function and its the levelsets.

As we can see on the picture of the levelsets of ϕ, it is not convex when $\mathbb{R}^2$ is endowed with the Euclidean structure. But in the Riemannian manifold

$$\left(\mathbb{R}^2, \langle ., . \rangle\right) \text{ with } \langle u, v \rangle_x = u^{\mathrm{T}} A(x) v \text{ where } A(x) = \begin{pmatrix} 4x_1^2 + 1 & -2x_1 \\ -2x_1 & 1 \end{pmatrix}$$

it becomes convex.

Let us present some numerical computations made on this example. We construct the sequence of points obtained by an explicit discretization of equation (1)

$$x_{k+1} = x_k + \lambda_k v_k. \tag{7}$$

The starting point x_0 is given.
We discuss first about the choice of the stepsize λ_k. On one hand, it is chosen in the following manner

Algorithm 1 *Stepsize choice:*
1. $\lambda = 1$, *compute* $y = x_k + \lambda v_k$ *and* $z = x_k + \frac{\lambda}{2} v_k$.
2. *While* $\phi(y) > \phi(z)$ *do*
$\lambda = \frac{\lambda}{2}$,
$y = z$,
$z = x_k + \frac{\lambda}{2} v_k$.
3. *Finally* $\lambda_k = \lambda$ *and* $x_{k+1} = y$.

This is a type of optimal line search, reduced to the interval $[0, 1]$. Indeed, it approaches the theoretical choice

$$\lambda_k \in \operatorname{argmin}\{\phi(x_k + \lambda v_k), \lambda \in [0, 1]\}.$$

On the other hand, we compute the sequence obtained by (7) with a stepsize still given by a optimal line search, reduced to the interval $[0, 0.1]$. It suffices to change $\lambda = 1$ by $\lambda = 0.1$ in the previous algorithm.
The descent direction v_k is chosen first as the opposite of the Riemannian gradient

$$v_k = -\text{grad}\phi(x_k) \tag{8}$$

which provides the lines in figure 2, and secondly as the opposite of the Euclidean gradient

$$v_k = -\nabla\phi(x_k) \tag{9}$$

which provides the dotted trajectories.
As expected, the experiments we made, with starting points randomly chosen in $[-5, 5] \times [-5, 5]$, show that both sequences seem to converge to the minimum $(1, 1)$ (except cases were numerical instability occurs which will be discussed later). Indeed, (7) with (8) is a discretization of (1), whose trajectories converge to the minimum: ϕ is convex, apply then theorem 4. On the other hand, (7) with (9) is a discretization of

$$\dot{x}(t) + \nabla\phi(x(t)) = 0.$$

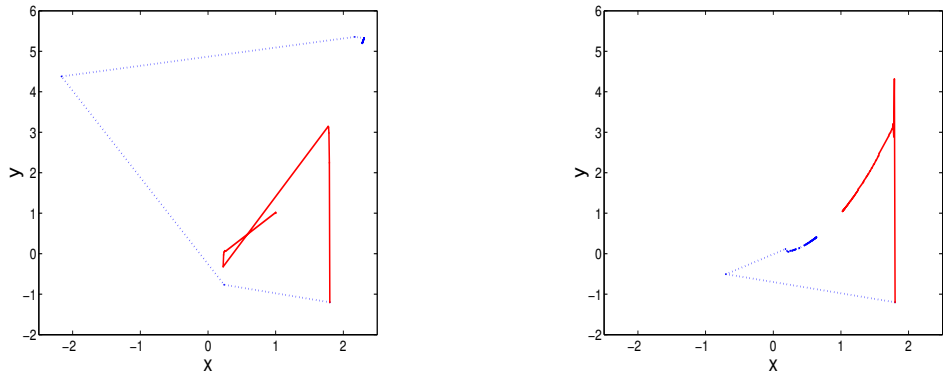


Figure 2: Trajectories obtained with a line search in $[0, 1]$ (left) and in $[0, 0.1]$ (right).

This (Euclidean) steepest descent is known to provide convergent trajectories, under analyticity assumptions on ϕ (see Łojasiewicz [13]). We wanted here to compare the two methods. Figure 3 presents the convergence curves obtained related to the experiments presented in figure 2: the value $\phi(x_i)$ is plotted versus the number i of the iterate.

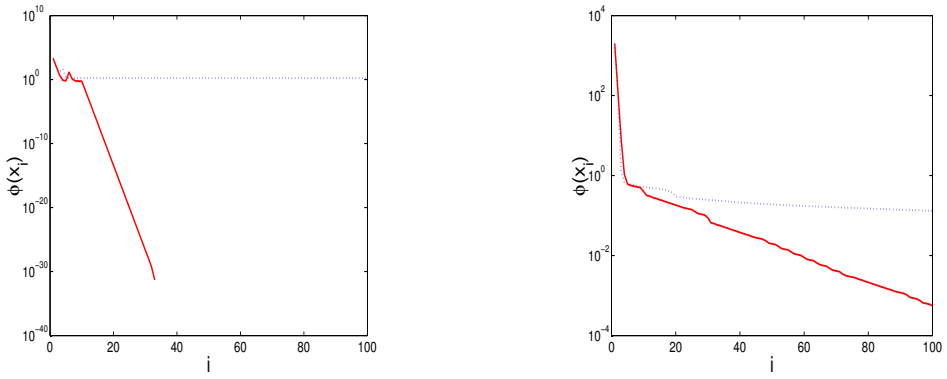


Figure 3: Convergence curves obtained with a line search in $[0, 1]$ (left) and in $[0, 0.1]$ (right).

For both stepsize choices, the Riemannian steepest descent (lines) is more efficient than the Euclidean one (dots).
The line search in $[0, 0.1]$ seems to stabilize the method while the points are not in the bottom of the valley, that means while the gradients are huge. Moreover, we can imagine that it reduces the number of computations done

in algorithm 1 providing the "optimal" stepsize. But it appears that once the point is in the bottom of the valley, steps of size 1 or at least between 0.1 and 1 speed up the convergence. Actually this is true for the Riemannian version, and after around thirty points we reach the limit of precision of the computer, whereas it can produce instability for the Euclidean version: in some cases, the sequence seem to skip from on side of the "U" of the valley to the other.

Remark 3 *A more accurate discretization of a continuous dynamical system on a Riemannian manifold should involve the geodesics or the exponential mapping, as in [6, 7, 2]. Here, we clearly make use of the vector space properties of* $\mathbb{R}^2$*, such structure provide a trivial retraction. It has the advantage of eluding the complicated computations of the exponential.*

References

[1] F. Alvarez, J. Bolte, and O. Brahic, *Hessian Riemannian gradient flows in convex programming*, SIAM J. Control Optim. **4** (2004), no. 2, 477–501.

[2] F. Alvarez, J. Bolte, and J. Munier, *A unifiying local convergence result for Newton's method in Riemannian manifolds*, Preprint INRIA RR-5381 (2004).

[3] J. Bolte and M. Teboulle, *Barrier operators and associated gradient-like dynamical systems for constrained minimization problems*, SIAM J. Control Optim. **42** (2003), 1266–1292.

[4] H. Brézis, *Opérateurs maximaux monotones et semi-groupes de contraction*, Lectures Notes No. 5, North Holland, 1973.

[5] R.E. Bruck, *Asymptotic convergence of nonlinear contraction semigroups in Hilbert space*, J. of Functional Analysis **18** (1975), 15–26.

[6] J.X. da Cruz Neto, L.L. de Lima, and P.R. Oliveira, *Geodesic algorithms in Riemannian geometry*, Balkan J. Geom. Appl. **3** (1998), 89–100.

[7] J.X. da Cruz Neto, O.P. Ferreira, and L.R. Lucambio Perez, *A proximal regularization of the steepest descent method in Riemannian manifold*, Balkan J. Geom. Appl. **4** (1999), 1–8.

[8] M.P. do Carmo, *Riemannian geometry*, Birkhäuser, Boston, 1992.

[9] O.P. Ferreira and P.R. Oliveira, *Proximal point algorithm on Riemannian manifolds*, Optimization **51** (2002), 257–270.

[10] S. Gallot, D. Hulin, and J. Lafontaine, *Riemannian geometry*, Springer, 2004.

[11] U. Helmke and J.B. Moore, *Optimization and dynamical systems*, Communication and control engineering series, Springer-Verlag, London, 1994.

[12] A.N. Iusem, B.F. Svaiter, and J.X. da Cruz Neto, *Central paths, generalized proximal point methods, and Cauchy trajectories in Riemannian manifolds*, SIAM J. Control Optim. **37** (1999), 566–588.

[13] S. Łojasiewicz, *Une proprit topologique des sous-ensembles analytiques rels*, Les Equations aux Drives Partielles (Centre National de la Recherche Scientifique, ed.), Paris, 1962, 1963, pp. 87–89.

[14] E.A.Papa Quiroz and P.R. Oliveira, *Proximal point method for a class of bregman distances on hadamard manifolds*, working paper.

[15] C. Udrişte, *Convex functions and optimization methods on Riemannian manifolds*, Mathematics and Its Applications, Vol. 297, Kluwer Academic Press, Dordrecht, 1994.

Julien Munier
Institut de Mathématiques et Modélisation de Montpellier UMR CNRS 5149
Université Montpellier II, cc 051, Pl. Eugène Bataillon, 34095 Montpellier Cedex 5, France
munier@math.univ-montp2.fr