Approximation of sequences of symmetric matrices with the symmetric rank-one algorithm and applications

Sylvain Arguillere

Approximation of sequences of symmetric matrices with the symmetric rank-one algorithm and applications

Sylvain Arguillère, Université Pierre et Marie Curie (Univ. Paris 6),
CNRS UMR 7598, Laboratoire Jacques-Louis Lions, F-75005,
Paris, France (`sylvain.arguillere@upmc.fr`)

Abstract

The symmetric rank-one update method is well-known in optimization for its applications in the quasi-Newton algorithm. In particular, Conn, Gould, and Toint proved in 1991 that the matrix sequence resulting from this method approximates the Hessian of the minimized function. Extending their idea, we prove that the symmetric rank-one update algorithm can be used to approximate any sequence of symmetric invertible matrices, thereby adding a variety of applications to more general problems, such as the computation of constrained geodesics in shape analysis imaging problems. We also provide numerical simulations for the method and some of these applications.

1 Introduction

Let d be an integer, $f : \mathbb{R}^d \to \mathbb{R}$ be a $\mathcal{C}^2$ function, and consider the problem of minimizing f over $\mathbb{R}^d$. A well-known efficient algorithm to numerically solve this minimization problem is Newton's method: starting at some point x_0, it considers the sequence

$$x_{k+1} = x_k - h_k H(f)_{x_k}^{-1} \nabla f(x_k),$$

with ∇f the gradient of f, $H(f)$ its Hessian, and $h_k > 0$ some appropriate step.

However, very often the Hessian of f is too difficult to compute, leading to the introduction of the so-called quasi-Newton methods. The method defines a sequence

$$x_{k+1} = x_k - h_k B_k^{-1} \nabla f(x_k),$$

where (B_k) is a sequence of symmetric matrices such that

$$B_{k+1}(x_{k+1} - x_k) = \nabla f(x_{k+1}) - \nabla f(x_k). \tag{1}$$

Indeed, since

$$\nabla f(x_{k+1}) - \nabla f(x_k) = \left(\int_0^1 H(f)_{x_k+t(x_{k+1}-x_k)} dt \right)(x_{k+1} - x_k) \simeq H(f)_{x_k}(x_{k+1} - x_k),$$

we get

$$B_{k+1}(x_{k+1} - x_k) \simeq H(f)_{x_k}(x_{k+1} - x_k).$$

It is then expected that B_k is close to $H(f)_{x_k}$ in the direction $s_k = x_{k+1} - x_k$. See [3, 4, 5] for more.

There are many ways to build a matrix sequence (B_k) satisfying (1). However, it was proved in [2] and [8] that some of these methods let B_k approximate $H(f)_{x_k}$ in all directions instead of just one, i.e.

$$\|B_k - H(f)_{x_k}\| \underset{k\to\infty}{\to} 0 \quad \Longrightarrow \quad \|B_k - H(f)_{x_*}\|\|B_k - H(f)_{x_k}\| \underset{k\to\infty}{\to} 0.$$

with the additional assumption of the uniform linear independence of the sequence

$$s_k = x_{k+1} - x_k,$$

a notion that will be recalled later. In [2] for example, this is proved for the update of B_k by

$$y_k = \nabla f(x_{k+1}) - \nabla f(x_k) = A_k s_k, \quad r_k = B_k s_k - y_k, \quad B_{k+1} = B_k + \frac{r_k r_k^T}{r_k^T s_k}, \tag{2}$$

with

$$A_k = \int_0^1 H(f)_{x_k + t(x_{k+1} - x_k)} dt.$$

In this paper, our aim is to generalize the approach in [2] by defining the above symmetric rank-one algorithm for any sequences of symmetric matrices (A_k) and vectors (s_k), and derive a convergence result, opening a wider range of applications.

For instance, if a sequence A_k converges to an invertible matrix A_*, we can use the above algorithm to approximate the inverse A_*^{-1} of the limit A_*. Indeed, let $(e_0, \ldots, e_{d-1})$ be the canonical vector basis of $\mathbb{R}^d$, and define the sequence (s_k) in $\mathbb{R}^d$ by

$$s_k := A_k e_{k[d]}, \quad y_k = A_k^{-1} s_k = e_{k[d]}, \tag{3}$$

where $k[d]$ is the remainder of the Euclidean division of k by d. This sequence is uniformly linearly independent, hence the sequence B_k defined by (2) will converge to A_*^{-1}. This convergence might be a little slow depending on the dimension d and the rate of convergence of A_k, but B_k is much easier to compute than A_k^{-1}.

This can be used to compute geodesics constrained to embedded submanifolds of Riemannian spaces. Indeed, to obtain a geodesic between two fixed points of a submanifold, we need to find a converging sequence of maps $t \mapsto \lambda_k(t)$ given implicitly by an equation of the form ([1])

$$A_k(t)\lambda_k(t) = c_k(t),$$

where $A_k(t)$ is a convergent sequence of symmetric, positive definite matrices of high dimension. The λ_k are the Lagrange multipliers induced by the equations of the submanifold. It is very consuming to solve such a linear system for every time t and every step k. Instead, we can take

$$\lambda_k(t) = B_k(t)c_k(t),$$

with $B_k(t)$ obtained by applying the symmetric rank-one algorithm described in the previous paragraph. This is particularly useful in Shape Spaces, where the studied manifolds have a very high dimension and a very complex metric. The present article was actually motivated by such a problem appearing in shape analysis, investigated in [1].

This paper is structured as follows. We give the general framework in Section 2, then state the main result after recalling two equivalent definitions of the uniform linear independence of a sequence of vectors in Section 3. Then, Section 4 is dedicated to intermediary results that will, along with notions developed in Section 5, derive the proof of our theorem.

2 Notations and symmetric rank-one algorithm

Consider a sequence $(A_k)_{k \in \mathbb{N}}$ of real square symmetric matrices of size d. Assume that this sequence converges to some matrix A_*, i.e.

$$\|A_k - A_*\| \underset{k \to \infty}{\to} 0,$$

where $\|\cdot\|$ is the operator norm on $M_d(\mathbb{R})$ induced by the canonical Euclidean norm $|\cdot|$ on $\mathbb{R}^d$. Then define

$$\eta_{k,l} = \sup_{k \leq i \leq l} \|A_i - A_k\|, \quad \text{and} \quad \eta_{k,*} = \sup_{i \geq k} \|A_i - A_k\|$$

for all $k \leq l \in \mathbb{N}$. Note that

$$\forall k \leq l \in \mathbb{N}, \quad \eta_{k,l} \leq \eta_{k,*} \quad \text{and} \quad \eta_{k,*} \to 0 \quad \text{as} \quad k \to \infty.$$

Now let $(s_k)_{k \in \mathbb{N}}$ be a sequence of vectors of $\mathbb{R}^d$.

Recall that we want to find a somewhat simple sequence $(B_k)_{k \in \mathbb{N}}$ of symmetric matrices such that $B_k \to A_*$, using only s_k and $y_k = A_k s_k$.

We use the symmetric rank-one update method from [2]. Start with $B_0 = I_d$. Then, define for $k \in \mathbb{N}$

$$y_k = A_k s_k, \quad r_k = (A_k - B_k)s_k = y_k - B_k s_k,$$

then take

$$B_{k+1} = B_k + \frac{r_k r_k^T}{r_k^T s_k}.$$

It is of course required that $r_k^T s_k \neq 0$ for every k.

3   Main Result

For every k, we have

$$B_{k+1}s_k = B_k s_k + r_k = B_k s_k + y_k - B_k s_k = y_k,$$

so

$$A_k s_k = B_{k+1} s_k.$$

The main idea is that if A_k, A_{k+1}, ..., A_{k+m} are not too far apart (i.e., for k large enough), we expect $B_{k+m}s_{k+i}$ to be relatively close to $A_{k+m}s_{k+i}$ for $i \leq m$. Then, if we can extract from every finite subsequence $(s_k, \ldots, s_{k+m})$ a vector basis of $\mathbb{R}^d$, we will obtain the desired convergence.

For a more precise statement, we next define the notion of uniform linear independence. The most intuitive and geometric definition is the following.

Definition 1 *Take a sequence $s = (s_k)_{k \in \mathbb{N}}$ of vectors in $\mathbb{R}^d$, $d \in \mathbb{N}^*$, and let $m \geq d$ be an integer. Then s is said to be m-uniformly linearly independent if for some constant $\alpha > 0$, and for all $k \in \mathbb{N}$, there are d integers $k \leq k_1 < \cdots < k_d \leq k + m$ such that*

$$|\det(s_{k_1}, \ldots, s_{k_d})| \geq \alpha |s_{k_1}| \ldots |s_{k_d}|.$$

In other words, from every finite segment of (s_k) of length m, we can extract a linear basis $s_{k_1}, \ldots, s_{k_d}$ that will, once normalized, form a parallelepiped that does not become flat as k goes to infinity.

Another definition was given in [2] after [7] as follows.

Definition 2 *A sequence $s = (s_k)_{k \in \mathbb{N}}$ of vectors in $\mathbb{R}^d$, $d \in \mathbb{N}^*$, is said to be (m, β)-uniformly linearly independent, where $d \leq m \in \mathbb{N}$ and $\beta \in \mathbb{R}$, if for all $k \in \mathbb{N}$, there are d integers $k \leq k_1 < \cdots < k_d \leq k + m$ such that*

$$\left| \lambda \left(\frac{s_{k_1}}{|s_{k_1}|}, \ldots, \frac{s_{k_d}}{|s_{k_d}|} \right) \right| \geq \beta,$$

where $\lambda(M)$ is the complex eigenvalue of the square matrix M with smallest module.

Remark: A sequence $s = (s_k)$ in $\mathbb{R}^d$ is (m, β)-uniformly linearly independent for some $m \geq d$ and $\beta > 0$ if and only it is m-uniformly linearly independent in the sense of Definition 1. Indeed, let $v_1, \ldots, v_d \in \mathbb{R}^d$, and denote $V = \left(\dfrac{v_1}{|v_1|}, \ldots, \dfrac{v_d}{|v_d|} \right)$. If $|\lambda(V)| \geq \beta > 0$, then $\det(V) \geq \beta^d$, which proves the first part of the equivalence. On the other hand, we know that the eigenvalue of V with largest modulus has modulus less than $\sqrt{d} \max\limits_{i=1,\ldots,d} \dfrac{|s_{k_i}|}{|s_{k_i}|} = \sqrt{d}$. Now, assume that $\det(V) \geq \alpha > 0$. Then $|\lambda(V)| \geq \frac{\alpha}{d^{\frac{d-1}{2}}}$, ensuring the second part of the equivalence.

This definition is sufficient to state our main result.

Theorem 1 *Let (A_k), (s_k), (y_k), (r_k) and (B_k) be defined as in Section 2, with (A_k) having a limit A_*. Assume that for some fixed constant $c > 0$,*

$$|r_k^T s_k| \geq c |r_k| |s_k|.$$

Then, for every $\beta > 0$ such that (s_k) is (m, β)-uniformly linearly independent in the sense of Definition 2, we have for all $k \in \mathbb{N}$ the quantitative estimates

$$\|B_{k+m} - A_*\| \leq \left(1 + \left(\frac{2+c}{c} \right)^{m+1} \right) \frac{\sqrt{d}}{\beta} \eta_{k,*}. \tag{4}$$

The next sections are dedicated to the proof of this theorem.

4   First estimates

In this section, we give upper bounds on

$$\left| (B_{k+m} - A_k) \frac{s_k}{|s_k|} \right|,$$

and deduce estimates on

$$\left| \frac{(B_{k+m} - A_*)x}{|x|} \right|$$

for a particular set of $x \in \mathbb{R}^d$.

Proposition 1 *Let $(A_k)_{k\in\mathbb{N}}$ be a sequence of real symmetric matrices in $M_d(\mathbb{R})$, $d \in \mathbb{N}$, and (s_k) be any sequence in $\mathbb{R}^d$. Define y_k, B_k and r_k as above. Assume that for some fixed constant $0 < c \leq 1$ and for all $k \in \mathbb{N}$,*

$$r_k^T s_k \geq c |r_k| |s_k|.$$

Then, for all $l \geq k+1$,

$$|(A_k - B_l)s_k| \leq \left(\frac{2+c}{c} \right)^{l-k-1} \eta_{k,l-1} |s_k|.$$

Proof: We prove this inequality by induction on l, with $k \in \mathbb{N}$ fixed. For $l = k+1$, we know that $B_{k+1}s_k = A_k s_k = y_k$, hence

$$|(A_k - B_{k+1})s_k| = 0.$$

We will use the notation

$$IH(l) := \left(\frac{2+c}{c} \right)^{l-k-1} \eta_{k,l-1} |s_k|,$$

where IH stands for Induction Hypothesis. Now, assume the result to be true for some $l \geq k+1$, i.e.

$$|(A_k - B_l)s_k| \leq \left(\frac{2+c}{c} \right)^{l-k-1} \eta_{k,l-1} |s_k| = IH(l). \tag{5}$$

4

Let us prove that

$$|(A_k - B_{l+1})s_k| \le \left(\frac{2+c}{c}\right)^{l-k} \eta_{k,l}|s_k| = IH(l+1).$$

Note that

$$\begin{aligned}
|(A_k - B_{l+1})s_k| &= |A_k s_k - (B_l + \frac{r_l r_l^T}{r_l^T s_l})s_k| \\
&= |A_k s_k - B_l s_k - \frac{r_l r_l^T s_k}{r_l^T s_l}| \\
&\le |(A_k - B_l)s_k| + \frac{|r_l||r_l^T s_k|}{c|r_l||s_l|} \\
&\le IH(l) + \frac{|r_l^T s_k|}{c|s_l|}.
\end{aligned}$$

(6)

Let us find a bound for $\frac{|r_l^T s_k|}{c|s_l|}$, the second term of the right-hand side. First we have

$$\begin{aligned}
|r_l^T s_k| &= |y_l^T s_k - s_l^T B_l s_k| \\
&\le |y_l^T s_k - s_l^T y_k| + |s_l^T(y_k - B_l s_k)| \\
&= |y_l^T s_k - s_l^T y_k| + |s_l^T(A_k - B_l)s_k| \\
&\le |y_l^T s_k - s_l^T y_k| + |s_l|IH(l).
\end{aligned}$$

However, since A_l is symmetric and $y_l = A_l s_l$,

$$|y_l^T s_k - s_l^T y_k| = |s_l^T(A_l - A_k)s_k| \le \eta_{k,l}|s_l||s_k|,$$

from which we deduce

$$|r_l^T s_k| \le \eta_{k,l}|s_l||s_k| + IH(l)|s_l|.$$

Going back to Inequality (6), we get

$$\begin{aligned}
|(A_k - B_{l+1})s_k| &\le IH(l) + \frac{|r_l^T s_k|}{c|s_l|} \\
&\le IH(l) + \frac{1}{c}\eta_{k,l}|s_k| + \frac{1}{c}IH(l) \\
&= (1 + \frac{1}{c})IH(l) + \frac{1}{c}\eta_{k,l}|s_k| \\
&= \frac{1+c}{c}\left(\frac{2+c}{c}\right)^{l-k-1}\eta_{k,l-1}|s_k| + \frac{1}{c}\eta_{k,l}|s_k| \\
&\le \left(\frac{2+c}{c}\right)^{l-k}\eta_{k,l}|s_k| = HI(l+1),
\end{aligned}$$

where the last inequality comes from the simple fact that $\eta_{k,l-1} \le \eta_{k,l}$. $\square$

This proposition shows that if A_k, $A_{k+1}, \ldots,$ A_l are not too far away from each other (i.e. if $\eta_{k,l}$ is small), then $B_l s_k$ stays quantifiably close to $A_k s_k$.

Now, note that $\|A_* - A_k\| \le \eta_{k,*}$, and $\eta_{k,*}$ decreases to 0 as k goes to infinity. Keeping the same assumptions, we obtain the following result.

Corollary 1 *Take $m, k \in \mathbb{N}$, and let $x \in \mathbb{R}^d$ be in the span of $s_k, \ldots, s_{k+m}$. If*

$$\frac{x}{|x|} = \sum_{i=0}^{m} \lambda_i \frac{s_{k+i}}{|s_{k+i}|}, \quad \lambda_0, \ldots, \lambda_m \in \mathbb{R},$$

then

$$\frac{|B_{k+m}x - A_*x|}{|x|} \le \eta_{k,*}\left(1 + \left(\frac{2+c}{c}\right)^{m+1}\right)\sum_0^m |\lambda_i|.$$

Proof: First, it follows from Lemma 1 that

$$\begin{aligned}
\frac{|B_{k+m}x - A_*x|}{|x|} &\le \sum_{i=0}^m \frac{|\lambda_{k+i}|}{|s_{k+i}|}|B_{k+m}s_{k+i} - A_*s_{k+i}| \\
&\le \sum_{i=0}^m \frac{|\lambda_{k+i}|}{|s_{k+i}|}\left(|B_{k+m}s_{k+i} - A_{k+i}s_{k+i}| + |A_{k+m}s_* - A_{k+i}s_{k+i}|\right) \\
&\le \sum_{i=0}^m |\lambda_i|\left(\left(\frac{2+c}{c}\right)^{i+1}\eta_{k,k+m-1} + \eta_{k,*}\right).
\end{aligned}$$

Now, if we take

$$C(m) = \left(1 + \left(\frac{2+c}{c}\right)^{m+1}\right)$$

and use the fact that $\eta_{k,k+m} \le \eta_{k,*}$, then we get

$$\frac{|B_{k+m}x - A_*x|}{|x|} \le \eta_{k,*}C(m)\sum_0^m |\lambda_i|.$$

The result follows. $\square$

In particular, if we can let k go to infinity while keeping $\sum_{i=0}^m |\lambda_i|$ bounded, then we obtain $B_{k+m}x \to A_*x$. Thus, if we can do it for all $x \in \mathbb{R}^d$, we will have proved that $B_k \to A_*$.

In other words, we need every normalized vector $x \in \mathbb{R}^d$ to be a uniformly bounded linear combination of $s_k, \ldots, s_{k+m}$ as k goes to infinity. In the next section of this paper, we will define a third notion of uniform linear independence of a sequence directly related to to this property and prove that it is equivalent to the previous definitions.

5   Uniform m-span of a sequence and applications

In order to investigate the subspace on which $B_k \to A_*$, we need a notion that is more precise than uniform linear independence.

Definition 3 *Let $s = (s_k)_{k\ge 0}$ be a sequence in $\mathbb{R}^d$, and let $m \in \mathbb{N}$. We say that a vector x in $\mathbb{R}^d$ is uniformly in the $m-$span of s if for some fixed $\gamma_x > 0$,*

$$\forall k \in \mathbb{N}, \quad \exists \lambda_0, \ldots, \lambda_m \in \mathbb{R} \quad \frac{x}{|x|} = \sum_{i=0}^m \lambda_i \frac{s_{k+i}}{|s_{k+i}|} \quad and \quad \sum_{i=0}^m |\lambda_i| \le \gamma_x. \tag{7}$$

We denote by $US_m(s)$ the set of all such vectors.

$US_m(s)$ is a vector sub-space of $\mathbb{R}^n$. Moreover, there exists a constant $\gamma > 0$ such that Property (7) holds for all $x \in US_m(s)$ with $\gamma_x = \gamma$, i.e.

$$\exists \gamma > 0, \quad \forall k \in \mathbb{N}, \ x \in US_m(s), \quad \exists \lambda_0, \ldots, \lambda_m \in \mathbb{R}, \quad \frac{x}{|x|} = \sum_{i=0}^m \lambda_i \frac{s_{k+i}}{|s_{k+i}|} \quad and \quad \sum_{i=0}^m |\lambda_i| \le \gamma. \tag{8}$$

To prove the existence of γ in (8), it suffices to consider an orthonormal basis $(x_i)_i$ of $US_m(s)$, associated with some constants $(\gamma_{x_i})_{1\le i\le d}$, in Property (7). Then we can just take $\gamma = \gamma_{x_1} + \cdots + \gamma_{x_d}$.

Remark: There holds $US_m(s) \subset \bigcap\limits_{k=0}^{\infty} \mathrm{span}(s_k, \ldots, s_{k+m})$.

Example: Define the sequence $s = (s_k)$ by

$$
s_k = \begin{cases} e_{k[d]} & \text{when} \quad k[d] \neq n-1, \\ e_0 + \dfrac{1}{k}e_{d-1} & \text{when} \quad k[d] = n-1, \end{cases}
$$

where $k[d]$ is the remainder of the Euclidean division of k by d. Then

$$
US_m(s) = \begin{cases} \{0\} & \text{if} \quad 0 \le m \le n-2 \\ \mathrm{span}(e_0, \ldots, e_{d-2}) & \text{otherwise.} \end{cases}
$$

Using this definition, a simple application of Corollary 1 gives the following result.

Proposition 2 *Let (A_k), (s_k), (y_k), (r_k) and (B_k) be defined as in Section 2, assuming that (A_k) has a limit A_* and that $|r_k^T s_k| \ge c|r_k||s_k|$ for some fixed constant $c > 0$.*
Then, for every $m \in \mathbb{N}$

$$
\sup_{x \in US_m(\gamma)} \frac{|B_{k+m}x - A_*x|}{|x|} \le C(m)\gamma\eta_{k,*}, \tag{9}
$$

where γ is taken from (8) and

$$
C(m) = \left(1 + \left(\frac{2+c}{c}\right)^{m+1}\right).
$$

Finally the main result follows by combining this proposition with the following lemma.

Lemma 1 *Let $s = (s_k)_{k\ge 0}$ be a sequence in $\mathbb{R}^d$, and let $m \in \mathbb{N}$. Then s is (m, β)-uniformly linearly independent if and only if $US_m(s) = \mathbb{R}^d$. Moreover, we can take $\gamma = \frac{\sqrt{d}}{\beta}$ in (8).*

Proof: Let $v_1, \ldots, v_d$ be linearly independent elements of $\mathbb{R}^d$ and define the invertible matrix

$$
V = \left(\frac{v_1}{|v_1|}, \ldots, \frac{v_d}{|v_d|}\right).
$$

Let $\Lambda = (\lambda_1, \ldots, \lambda_d) \in \mathbb{R}^d$, and define $x \in \mathbb{R}^d$ a normalized vector such that

$$
x = \sum_{i=1}^{d} \lambda_i v_i = V\Lambda.
$$

Then

$$
\sum_{i=1}^{d} |\lambda_i| \le \sqrt{d}|\Lambda| = \sqrt{d}|V^{-1}x| \le \frac{\sqrt{d}}{|\lambda(V)|}.
$$

This proves that if a sequence $s = (s_k)$ in $\mathbb{R}^d$ is (m, β)-uniformly linearly independent, then $US_m(s) = \mathbb{R}^d$ and we can take $\gamma_m(s) = \frac{\sqrt{d}}{\beta}$.

On the other hand, take $x \in \mathbb{R}^d$ a normalized such that

$$
V^{-1T}V^{-1}x = \frac{1}{\lambda(V)^2}x.
$$

Then, if denoting $(\lambda_1, \ldots, \lambda_d) = \Lambda = V^{-1}x$,

$$\frac{1}{|\lambda(V)|} = |\lambda(V)|\frac{1}{|\lambda(V)|^2} = |\lambda(V)||V^{-1T}V^{-1}x| = |\lambda(V)||V^{-1T}\Lambda| \leq |\Lambda| \leq \sum_{i=1}^{d}|\lambda_i|,$$

which proves the converse. $\square$

Our main result is proved.

6 Examples of applications and numerical simulations

In this section, after running numerical simulations of the algorithm on random symmetric matrices, we check that the inverse of a sequence of matrices can indeed be approximated. Then we give an application for computing constrained geodesics between two fixed points in Riemannian manifolds.

All simulations were done using Matlab.

6.1 Approximation of a sequence of matrices

Here we test the algorithm on random symmetric matrices with coefficients generated by a normalized Gaussian law. Let $d \in \mathbb{N}^*$, which will denote the size of the matrices.

First we define a square symmetric matrix $A_* = \frac{1}{2}(M + M^T)$, where the entries of the $d \times d$ matrix M were chosen at random using the normalized Gaussian law. Then, fix $0 < \lambda < 1$, and define the sequence (A_k) of symmetric matrices by perturbating A_* as follows

$$A_k = A_* + \frac{\lambda^k}{2}(M_k + M_k^T),$$

where M_k is a matrix with random coefficients taken uniformly in $[0, 1]$.

Obviously, $A_k \to A_*$ linearly as $k \to \infty$.

Now, we define the sequence (B_k) thanks to the symmetric rank-one algorithm, starting with $B_0 = I_d$, and the sequence (s_k) by the formula

$$s_k = e_{k \bmod d}, \quad k \in n,$$

where $(e_0, \ldots, e_{d-1})$ is the canonical basis of $\mathbb{R}^d$.

Using the classical norm given by the Euclidean product $\langle X, Y \rangle = tr(X^TY)$ on the space of square matrices of size n, we give in the following table the distance between B_k and A_*. We took $d = 10$, and several values of λ for a various number of steps.

Number of steps	10	20	50	100
λ =0.9	4	2	0.1	0.005
λ =0.5	1	1e-3	1e-12	0
λ =0.1	0.02	2e-12	0	0

In the case of the usual quasi-Newton method ([7], the goal is to approximate the inverse of the Hessian $H(f)$ of some function f on $\mathbb{R}^d$. For this, we get a sequence of points x_k converging to the minimum x_*, and we define

$$s_k := x_{k+1} - x_k,$$

$$A_k := \int_0^1 H(f)_{x_k + ts_k}dt,$$

$$y_j := A_k s_k = \nabla g(x_{k+1}) - \nabla g(x_k).$$

Then $\lim_{k \to \infty} A_k = H(x_*)$.

This is a perfect example of a situation where it is easy to compute $A_k s_k$ for some particular s_k, but where it might be much harder to compute the actual A_k (or just $H(x_k)$ for that matter). A large number of numerical simulations showing the efficiency of the symmetric rank-one algorithm can be found in [2].

8

6.2 Approximation of the inverse of a sequence

As mentioned in the introduction, another application is the computation of the inverse A_*^{-1} of the limit A_*, provided A_* is invertible.

Indeed, consider the following sequences for the symmetric rank-one algorithm

$$s_k := A_k e_{k \bmod d}, \ y_k = A_k^{-1} s_k = e_{k \bmod d}. \tag{10}$$

Then the sequence (s_j) is (d, β)-linearly independent for some $\beta > 0$ (at least starting at some k_0 large enough). Therefore, the sequence B_k will converge to A_*^{-1}. This convergence might be a bit slow depending on the dimension d and the rate of convergence of A_k, but B_k is much easier to compute than A_k^{-1}.

This can be useful when solving approximately converging sequences of linear equations, as we will show in the next section.

In the following numerical simulation, we used the same sequence (A_k) with random coefficients as in the previous section, with (A_k) converging linearly to a random matrix A_* with rate $\lambda = 0.5$. We then computed the distance between B_k and A_*^{-1}. We also added an extra test. Indeed, the form of the sequence s_k in (10) has no reason to be particularly good (i.e. uniformly linearly independent with a nice constant). Therefore, we applied the algorithm by taking a random vector y_k with coefficients taken along a normal Gaussian law at each step and $s_k = A_k y_k$.

Number of steps	10	20	50	100
$y_k = e_{k \bmod d}$	0.5 (300)	0.001 (0.2)	1e-7 (1e-4)	1e-13 (1e-11)
y_k random	1 (500)	0.01 (0.5)	1e-6 (1e-3)	1e-13 (1e-10)

For each case, we performed twenty experiments. Each entry in the previous table gives the mean value of the distance between B_k and A_*^{-1}, with the highest value obtained in parentheses. This number can be significantly larger than the mean because of the randomness of A_*, which can cause it to be almost singular, leading the algorithm to behave badly as the s_k are less uniformly linearly independent with this method.

This experiment shows that taking y_k random is not as efficient as taking y_k to periodically be equal to the canonical basis of $\mathbb{R}^d$.

6.3 An application: constrained optimisation

Consider the following control system on $\mathbb{R}^d$

$$\dot{x} = K_{x(t)} u(t), \quad u \in \mathbb{R}^d,$$

with K_x a semi-positive symmetric matrix with coefficients of class $\mathcal{C}^2$. This corresponds to a sub-Riemannian control system where u is the momentum of the trajectory and K_x the co-metric at x ([6]). This is a very natural formulation for problems of shape analysis, see [1].

Take $C \in M_{l,d}(\mathbb{R})$ such that the $l \times l$ matrix

$$A_x = C K_x C^T$$

is invertible for every $x \in \mathbb{R}^d$, and take an initial point $x_0 \in \mathbb{R}^d$ such that $C x_0 = 0$.

We consider the optimal control problem of minimizing

$$L(u) = \frac{1}{2} \int_0^1 u(t)^T K_{x(t)} u(t) dt + g(x(1)), \quad \text{where} \quad \dot{x} = K_{x(t)} u(t),$$

over all possible $u \in L^2([0,1], \mathbb{R}^d)$ such that

$$C K_{x(t)} u(t) = 0 \quad a.e. \ t \in [0,1].$$

9

This is the same as minimizing $L(u)$ over trajectories that stay in the sub-space $\ker(C)$. According to the Pontryagin Maximum Principle ([1, 9]), if u is optimal for this constrained problem, then there exists $p \in W^{1,2}([0,1], \mathbb{R}^d)$ such that $p(1) + dg_{x(1)} = 0$, and

$$
\begin{cases}
\dot{x} = K_x \left(p - C' A_x^{-1} C K_x p \right), \\
\dot{p} = -\dfrac{1}{2} \left(p - C' A_x^{-1} C K_x p \right)^T K_x \left(p - C' A_x^{-1} C K_x p \right).
\end{cases}
\tag{11}
$$

for almost every $t \in [0,1]$. Moreover,

$$
L(u) = \tilde{L}(p(0)) = \frac{1}{2} \left(p(0) - C' A_{x(0)}^{-1} C K_{x(0)} p(0) \right)^T K_{x(0)} \left(p(0) - C' A_{x(0)}^{-1} C K_{x(0)} p(0) \right) + g(x(1)).
\tag{12}
$$

Since (11) is an ordinary differential equation, and since the minimization of L reduces to the minimization of $\tilde{L}$ with respect to the initial momentum $p_0 = p(0)$. Then, the computation the gradient of $\tilde{L}$ requires solving an adjoint equation with coefficients depending on the derivatives of the right-hand side of (11). This is described in more details in [1].

One of the most time-consuming aspects of this method is the computation, at each time step, of the inverse of A_x. Therefore, we applied the Quasi-Newton Algorithm as follows.

For any $k \in \mathbb{N}$, define $y_k = e_{k \bmod d}$. We start with the initial momentum $p_0 = 0$, and let $B_0(t) = Id_l$ for all $t \in [0,1]$. Then, assuming we have constructed an initial momentum p_k and a family of matrices $B_k(t)$, $t \in [0,1]$, we use (11) to compute a trajectory $x_k(t)$, replacing A_x^{-1} by $B_k t$. Finally, at each time t, we define

$$
s_k(t) = A_{x_k(t)} y_k,
$$
$$
r_k(t) = B_k(t) s_k(t) - y_k,
$$
$$
B_{k+1}(t) = B_k(t) + \frac{r_k(t) r_k^T(t)}{r_k^T(t) s_k(t)}.
$$

We can then compute the gradient of $\tilde{L}$ with an adjoint equation, where any derivative

$$
\partial_x (A_{x_k(t)})^{-1} = -A_{x_k(t)}^{-1} \partial_x A_{x_k(t)} A_{x_k(t)}^{-1}
$$

is replaced by $-B_k(t) \partial_x A_{x_k(t)} B_k(t)$. This allows the minimization of $\tilde{L}$ using gradient descent or a regular quasi-Newton algorithm.

As long as the algorithm gives a converging sequence of initial momenta p_k, the trajectories $x_k(t)$ will also converge to a trajectory $x_*(t)$, making each $A_{x_k(t)}$, with $t \in [0,1]$ fixed, a converging sequence, with invertible limit $A_*(t)$. Therefore, each $B_k(t)$, $t \in [0,1]$ fixed, converges to $A_*(t)$ as $k \to \infty$. In other words, as $k \to \infty$, we are indeed computing the true gradient of $\tilde{L}$.

References

[1] S. Arguillère, E. Trélat, A. Trouvé, and L. Younès. Control theory and shape analysis. *Preprint*, 2013.

[2] A. R. Conn, N. I. Gould, and P. L. Toint. Convergence of quasi-newton matrices generated by the symmetric rank one update. *Mathematical Programming*, (2):177–195, 1991.

[3] J. E. Dennis, Jr. and Jorge J. Moré. Quasi-Newton methods, motivation and theory. *SIAM Rev.*, 19(1):46–89, 1977.

[4] J. E. Dennis, Jr. and Robert B. Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*, volume 16 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996. Corrected reprint of the 1983 original.

[5] Philip E. Gill, Walter Murray, and Margaret H. Wright. *Practical optimization*. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], London, 1981.

[6] R. Montgomery. *A tour of subriemannian geometries, their geodesics and applications*, volume 91 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2002.

[7] J. M. Ortega and W. C. Rheinboldt. *Iterative solution of nonlinear equations in several variables*, volume 30 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000. Reprint of the 1970 original.

[8] G. Schuller. On the order of convergence of certain quasi-Newton methods. *Numer. Math.*, 23:181–192, 1974.

[9] Emmanuel Trélat. *Contrôle optimal*. Mathématiques Concrètes. [Concrete Mathematics]. Vuibert, Paris, 2005. Théorie & applications. [Theory and applications].