PÓLO DO I.S.T

# On minimizing a quadratic function on Stiefel manifolds [1]

Marija Dodig Marko Stošić[2] João Xavier

April 2009

ISR — Torre Norte
Av. Rovisco Pais, 1
1049-001 Lisboa
PORTUGAL

[1]This work was supported by Fundação para a Ciência e a Tecnologia (ISR/IST pluriannual funding) through the POS_Conhecimento Program (include FEDER funds). João Xavier and Marko Stošić were partially supported by grant SIPM PTDC/EEA-ACR/73749/2006, Marija Dodig was partially supported by the Fundação para a Ciência e Tecnologia (Portugal) through the project ISFL-1-1431 and scholarship SFRH/BPD/26607/2006.

[2]Corresponding author: mstosic@isr.ist.utl.pt

**Abstract**

We give an efficient, quick algorithm for the minimization of a quadratic function over Stiefel manifold. We reduce the original (nonconvex) problem, to an SDP, by computing a convex hull of the certain set of matrices.

**Keywords:** Convex, Semidefinite programming, Stiefel matrix.

# 1 Introduction

Many important optimization problems can be written as the minimization over the set of Stiefel matrices, i.e. the matrices whose columns form an orthonormal set. The important particular case is when the Stiefel matrices in question are of the size $3 \times 2$, i.e. when we are observing the set of all matrices $Q \in \mathbb{R}^{3\times 2}$, such that $Q^T Q = I_2$.

In this paper, we deal with the following Quadratic Programming problem:

$$(\text{arg}) \min_{q=\text{vec}(Q)} q^T C q, \tag{1}$$

where $Q \in \mathbb{R}^{3\times 2}$ runs through Stiefel matrices, and $C \in \mathbb{R}^{6\times 6}$ is a given matrix. This problem is indeed a quadratic programming problem since the vector $q \in \mathbb{R}^{6\times 1}$ is of the form $\text{vec}(Q)$ for some Stiefel matrix $Q$ iff

$$\begin{aligned} q^T \begin{bmatrix} I_3 & 0 \\ 0 & 0 \end{bmatrix} q &= 1, \\ q^T \begin{bmatrix} 0 & 0 \\ 0 & I_3 \end{bmatrix} q &= 1, \\ q^T \begin{bmatrix} 0 & I_3 \\ 0 & 0 \end{bmatrix} q &= 0. \end{aligned}$$

The set over which we are minimizing is non-convex, and we managed to compute its convex hull (tight convex relaxation), so that the problem becomes an SDP, and thus easily resolvable via SeDuMi MATLAB toolbox.

# 2 Solution

Since we are minimizing a quadratic function over the set given by three quadratic restrictions, it is beyond the scope of the known general techniques (see Polyak [1]). Thus, we needed to apply some new techniques.

Problem (1) can be re-written in the following way:

$$\min_{q=\text{vec}(Q)} q^T C q = \min_{q=\text{vec}(Q)} \text{Tr}(Cqq^T) = \min_{X\in S} \text{Tr}(CX),$$

where $S$ is the set of matrices $X$ of the form $qq^T$ with $q = \text{vec}(Q)$, i.e. $S$ is the set of all real symmetric 6 by 6 matrices $X = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}$, with

$X_{11} \in \mathbb{R}^{3\times 3}$, satisfying the following

$$X \succcurlyeq 0, \tag{2}$$

$$\mathrm{Tr}(X_{11}) = \mathrm{Tr}(X_{22}) = 1, \quad \mathrm{Tr}(X_{12}) = 0, \tag{3}$$

$$\mathrm{rank}\, X = 1. \tag{4}$$

Because of the rank constraint the set $S$ is non-convex. Since the cost function is linear we have

$$\min_{X\in S} \mathrm{Tr}(CX) = \min_{X\in \mathrm{co}(S)} \mathrm{Tr}(CX),$$

where $\mathrm{co}(S)$ is the convex hull of the set $S$, i.e. the set of all convex combinations $c_1X_1 + c_2X_2 + \ldots + c_kX_k$, where $X_i \in S$, $i = 1, \ldots, k$, and $c_i$'s are nonnegative such that $c_1 + c_2 + \cdots + c_k = 1$. In other words, the convex hull of the set $S$ is the smallest convex set (with respect to inclusion) that contains the set $S$.

The "standard" convex relaxation - simply loosing the rank constraint is not the correct choice, as can be shown by the following example:

**Example 1** *The matrix*

$$M = \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right],$$

*satisfies the conditions (2) and (3). However, it doesn't belong to the convex hull of $S$. Indeed, if there exists nonnegative numbers $c_1, \ldots, c_k$ such that $c_1 + c_2 + \cdots + c_k = 1$, and the matrices $M_1, \ldots, M_k \in S$, such that*

$$M = c_1M_1 + \cdots + c_kM_k,$$

*then we would have that in all matrices $M_i$, entries at the positions (2,2), (3,3), (5,5) and (6,6) are zero, and consequently all entries in the second, third, fifth and sixth rows and columns are zero (all matrices are positive semi-definite). However, matrix $M_1$ being from $S$ is of the form $qq^T$ for some $q = \mathrm{vec}(Q)$, and thus the corresponding matrix $Q$ would be of the form*

$$Q = \begin{bmatrix} * & * \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

*The last is impossible, since $Q$ is the Stiefel matrix.*

So, we want to compute co($S$), and if possible, to describe it by linear matrix inequalities. The set $S$ is not convex because of the rank condition (4), and hence we want to find its relaxation, by introducing some further constraints.

To that end, let $Q \in \mathbb{R}^{3\times 2}$ be a Stiefel matrix, and denote its columns by $q_1$ and $q_2$. Then the vector $q = \mathrm{vec}(Q) \in \mathbb{R}^{6\times 1}$ is given by $q^T = [q_1^T \;\; q_2^T]$, and the matrix $X = qq^T$ belongs to $S$.

The vectors $q_1$, $q_2$ and their cross-product $q_1 \times q_2$ form an orthonormal basis, and consequently the sum of projectors to these three vectors is equal to the identity matrix $I_3$. Moreover, we have an access to the entries of $q_1 \times q_2$ as linear functions of the entries of the off-diagonal block $X_{12}$. So, we have that the matrices $X \in S$ satisfy

$$vv^T + X_{11} + X_{22} = I_3, \tag{5}$$

where

$$v = v(X) := \begin{bmatrix} b_{23} - b_{32} \\ b_{31} - b_{13} \\ b_{12} - b_{21} \end{bmatrix} \tag{6}$$

with $X_{12} = [b_{ij}]$.

Thus, we have that every matrix $X \in \mathrm{co}(S)$ satisfy the following additional condition

$$vv^T + X_{11} + X_{22} \preccurlyeq I_3, \tag{7}$$

where $v$ is defined as above. Indeed, as we saw, all matrices from $S$ satisfy (7). Moreover, if matrices $X'$ and $X''$ satisfy (7) (the corresponding vectors $v(X')$ and $v(X'')$ are denoted by $v_1$ and $v_2$, respectively), and if $c_1$ and $c_2$ are nonnegative and such that $c_1 + c_2 = 1$, then the matrix $Y := c_1X' + c_2X''$ also satisfy (7):

$$\begin{aligned} v(Y)v(Y)^T + Y_{11} + Y_{22} &= (c_1v_1 + c_2v_2)(c_1v_1^T + c_2v_2^T) + c_1X'_{11} + c_1X'_{22} + \\ &\quad + c_2X''_{11} + c_2X''_{22} = \\ &= c_1(v_1v_1^T + X'_{11} + X'_{22}) + c_2(v_2v_2^T + X''_{11} + X''_{22}) - \\ &\quad - c_1c_2(v_1 - v_2)(v_1 - v_2)^T \preccurlyeq \\ &\preccurlyeq c_1I_3 + c_2I_3 = I_3. \end{aligned}$$

We can write the formula (7) as Linear Matrix Inequality

$$\begin{bmatrix} I_3 - X_{11} - X_{22} & v \\ v^T & 1 \end{bmatrix} \succcurlyeq 0. \tag{8}$$

It is straightforward to see that this new condition easily discards the matrix from Example 1.

Thus we have proved the following:

**Proposition 1** *The convex hull* co$(S)$ *satisfies the following:*

$$co(S) \subset \left\{ X \;\middle|\; \begin{array}{c} X \succcurlyeq 0, \\ \mathrm{Tr}(X_{11}) = \mathrm{Tr}(X_{22}) = 1, \\ \mathrm{Tr}(X_{12}) = 0, \\ \begin{bmatrix} I_3 - X_{11} - X_{22} & v \\ v^T & 1 \end{bmatrix} \succcurlyeq 0 \end{array} \right\}.$$

Moreover, we conjecture that the converse is also valid:

**Conjecture 1** *Let* $\Sigma$ *be the following set of symmetric* $6 \times 6$ *matrices:*

$$\Sigma = \left\{ X \;\middle|\; \begin{array}{c} X \succcurlyeq 0, \\ \mathrm{Tr}(X_{11}) = \mathrm{Tr}(X_{22}) = 1, \\ \mathrm{Tr}(X_{12}) = 0, \\ \begin{bmatrix} I_3 - X_{11} - X_{22} & v \\ v^T & 1 \end{bmatrix} \succcurlyeq 0 \end{array} \right\}.$$

*Then we have*

$$\mathrm{co}(S) = \Sigma.$$

Although we don't have the complete rigorous proof of this conjecture, we have some quite strong evidence of its validity. First of all, we run the tests on very large number of randomly genereated matrices (10000), and the results were always correct, i.e. randomly generated matrix from $\Sigma$ was always in the convex hull of the set $S$.

In order to prove $\Sigma \subset \mathrm{co}(S)$, we need to prove that every matrix $X \in \Sigma$ can be written as a convex combination of the matrices from $S$, i.e. that there exist positive real numbers $c_1, \dots, c_k \geq 0$, with $\sum_{i=1}^k c_i = 1$, and matrices $X_1, \dots, X_k \in S$, such that

$$X = \sum_{i=1}^{k} c_i X_i.$$

So, let $X \in \Sigma$ be arbitrary. First of all, note that if $P \in \mathbb{R}^{3\times3}$ is the orthogonal matrix, then the matrix $\begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}$ is from $\Sigma$ (or from $S$) if and only if the matrix

$$\begin{bmatrix} P & 0 \\ 0 & P \end{bmatrix} \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} \begin{bmatrix} P^T & 0 \\ 0 & P^T \end{bmatrix}, \tag{9}$$

is from $\Sigma$ (or respectively from $S$). Hence, it is enough to show that for some orthogonal matrix $P$, the matrix (9) is from $\mathrm{co}(S)$.

Denote by $v = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$ the vector that corresponds to $X$ by (6) (for a matrix $Y \in \Sigma$, we shall denote the corresponding vector by $v(Y)$). Then, since the condition (8) can be written as

$$X_{11} + X_{22} + vv^T \preccurlyeq I_3,$$

and since $X_{11}$ and $X_{22}$ are positive semi-definite, we have that $||v|| \leq 1$.

We have managed to give the complete proofs of the conjecture in some particular cases. Below, we include the proofs for two of them: the first one when $||v|| = 1$ and the second one when $\mathrm{rank}(X_{11}+X_{22}) = 2$. As can be seen from these proofs, they are quite involved and messy, and we expect that the general proof will be along the same lines.

**Proof of the two particular cases:**

**Case 1:** $||v|| = 1$

Then $x^2+y^2+z^2 = 1$, and there exists the orthogonal matrix $P$, such that $Pv = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$. Moreover, denote $D = B - B^T$. Then, $D$ is skew-symmetric, and it has the form:

$$D = \begin{bmatrix} 0 & z & -y \\ -z & 0 & x \\ y & -x & 0 \end{bmatrix}.$$

Then we have

$$D^2 = vv^T - ||v||^2 I_3 = vv^T - I_3.$$

Moreover, if $D' = PDP^T$, and $w = v(D')$, then we have that $||D'||^2 = ||D||^2 = 2$, and so $||w|| = 1$. Also, as above we have

$$D'^2 = ww^T - ||w||^2 I_3 = ww^T - I_3,$$

and so

$$(Pv)(Pv)^T = ww^T.$$

Hence we have

$$w = \pm Pv = \begin{bmatrix} 0 \\ 0 \\ \pm 1 \end{bmatrix}.$$

If the sign is positive, we define $P' = P$, while if it is negative, we define

$$P' = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} P.$$

Then, by applying the similarity operation (9) on the matrix $X$ with the orthogonal matrix $P'$ instead of $P$, we obtain the matrix $X'$ of the form:

$$X' = \begin{bmatrix} A' & B' \\ B'^T & C' \end{bmatrix}, \tag{10}$$

with

$$B' - B'^T = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

So, we are left with proving that such matrices $X' \in \Sigma$ belong to $\mathrm{co}(S)$. From the conditions (8) and the trace condition in the Conjecture 1 we have that

$$A' + C' = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Moreover, since they are positive semi-definite, we have that $A'_{33} = C'_{33} = 0$, and hence the whole third and sixth row and column of $X'$ are equal to zero. Also, the only matrices from $S$ which can make the convex combination of $X'$ must have the same property, and from now on we can restrict only to the submatrix of $X'$ formed by the first, second, fourth and fifth rows and columns. Then we have that the obtained matrix (still denoted by $X'$) is of the following form:

$$X' = \left[\begin{array}{cc|cc} a_1 & a_2 & b_1 & b_2+1 \\ a_2 & 1-a_1 & b_2 & -b_1 \\ \hline b_1 & b_2 & 1-a_1 & -a_2 \\ b_2+1 & -b_1 & -a_2 & a_1 \end{array}\right],$$

for some $a_1, a_2, b_1, b_2 \in \mathbb{R}$. From the positive semi-definiteness (non-negativity of the principal minors), we obtain the following inequalities for the minors of the dimension 1 and 2:

$$a_1(1-a_1) \geq a_2^2, \tag{11}$$

$$a_1^2 \geq (b_2+1)^2, \tag{12}$$

$$(1-a_1)^2 \geq b_2^2, \tag{13}$$

$$a_1(1-a_1) \geq b_1^2. \tag{14}$$

From the (11)-(13) we have $b_2 = a_1 - 1$. Moreover, the principal 3 by 3 minor, gives:

$$(a_1 - 1)(a_2 + b_1)^2 \geq 0.$$

The case $a_1 = 1$, automatically gives the matrix from $S$. If $a_1 < 1$, then we must have $b_1 = -a_2$, and so our matrix has the form

$$X' = \left[\begin{array}{cc|cc} a_1 & a_2 & -a_2 & a_1 \\ a_2 & 1-a_1 & a_1-1 & a_2 \\ \hline -a_2 & a_1-1 & 1-a_1 & -a_2 \\ a_1 & a_2 & -a_2 & a_1 \end{array}\right]. \tag{15}$$

The matrices from $S$ have the similar form:

$$\left[\begin{array}{cc|cc} cos^2\varphi & sin\varphi cos\varphi & -sin\varphi cos\varphi & cos^2\varphi \\ sin\varphi cos\varphi & sin^2\varphi & -sin^2\varphi & sin\varphi cos\varphi \\ \hline -sin\varphi cos\varphi & -sin^2\varphi & sin^2\varphi & -sin\varphi cos\varphi \\ cos^2\varphi & sin\varphi cos\varphi & -sin\varphi cos\varphi & cos^2\varphi \end{array}\right],$$

with $\varphi \in \mathbb{R}$. Hence we are left with proving that the point $(a_1, a_2) \in \mathbb{R}^2$ is in the convex hull of the set $K = \{(cos^2\varphi, sin\varphi cos\varphi) | \varphi \in [0, 2\pi]\}$. However, $K$ is in fact the circle given by the equation $(x - \frac{1}{2})^2 + y^2 = (\frac{1}{2})^2$, i.e. $x^2 + y^2 = x$, while from (11) we have that the point $(a_1, a_2)$ is inside this circle, which finishes our proof in this case.

Moreover, we have obtained that matrices of the form (15) with $a_1^2 + a_2^2 \leq a_1$ belong to co($S$) (remember that the third and the sixth rows and columns are zero). Analogously, we can obtain the matrices of the form

$$\left[\begin{array}{cc|cc} a_1 & a_2 & a_2 & -a_1 \\ a_2 & 1-a_1 & 1-a_1 & -a_2 \\ \hline a_2 & 1-a_1 & 1-a_1 & -a_2 \\ -a_1 & -a_2 & -a_2 & a_1 \end{array}\right], \tag{16}$$

with $a_1^2 + a_2^2 \leq a_1$ also belong to co($S$).

**Case 2:** rank$(X_{11} + X_{22}) = 2$

In this case, from the condition (8) we have that there exists orthogonal $P \in \mathbb{R}^{3\times 3}$ such that

$$P(X_{11} + X_{22})P^T = \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0, \end{array}\right],$$

and such that $PX_{11}P^T$ is diagonal. Again, as in the previous case, this implies that the third and sixth rows and colums are zero, and hence we are left with proving that the positive semi-definite matrix of the form

$$X = \left[\begin{array}{cc|cc} a_1 & 0 & b_1 & b_2+x \\ 0 & 1-a_1 & b_2 & -b_1 \\ \hline b_1 & b_2 & 1-a_1 & 0 \\ b_2+x & -b_1 & 0 & a_1 \end{array}\right], \tag{17}$$

with $-1 < x < 1$ is from $\text{co}(S)$. We shall show that it can be written as the convex combination of the matrices of the form (15) and (16). Namely, we shall show that there exists $m_1, m_2, n_1, n_2 \in \mathbb{R}$, with $m_1^2 + m_2^2 \leq m_1$ and $n_1^2 + n_2^2 \leq n_1$ such that

$$X = \frac{1+x}{2}\left[\begin{array}{cc|cc} m_1 & m_2 & -m_2 & m_1 \\ m_2 & 1-m_1 & m_1-1 & m_2 \\ \hline -m_2 & m_1-1 & 1-m_1 & -m_2 \\ m_1 & m_2 & -m_2 & m_1 \end{array}\right] + \frac{1-x}{2}\left[\begin{array}{cc|cc} n_1 & n_2 & n_2 & -n_1 \\ n_2 & 1-n_1 & 1-n_1 & -n_2 \\ \hline n_2 & 1-n_1 & 1-n_1 & -n_2 \\ -n_1 & -n_2 & -n_2 & n_1 \end{array}\right]. \tag{18}$$

Straightforward computation gives the unique solution of (18):

$$\begin{aligned} m_1 &= \frac{a_1+b_2+x}{1+x}, \\ m_2 &= -\frac{b_1}{1+x}, \\ n_1 &= \frac{a_1-b_2-x}{1-x}, \\ n_2 &= \frac{b_1}{1-x}. \end{aligned}$$

Hence, we are left with proving that $m_1^2 + m_2^2 \leq m_1$ and $n_1^2 + n_2^2 \leq n_1$, i.e.

$$b_1^2 \leq \min\{(a_1+b_2+x)(1-a_1-b_2), (a_1-b_2-x)(1-a_1+b_2)\}. \tag{19}$$

In order, to prove this, we will use that $X \succcurlyeq 0$, and in particular that its determinant is nonnegative, which gives:

$$b_1^4 - 2(a_1(1-a_1) - b_2(b_2+x))b_1^2 + (1-a_1-b_2)(1-a_1+b_2)(a_1-b_2-x)(a_1+b_2+x) \geq 0. \tag{20}$$

Denote $A = a_1(1-a_1) - b_2(b_2+x)$ and $B = (1-a_1-b_2)(1-a_1+b_2)(a_1-b_2-x)(a_1+b_2+x)$. Then we have

$$A^2 - B = (a_1 b_2 - (1-a_1)(b_2+x))^2,$$

and so (20) is equivalent to

$$b_1^2 \le A - \sqrt{A^2 - B} \quad \text{or} \quad b_1^2 \ge A + \sqrt{A^2 - B}. \tag{21}$$

However, from the non-negativity of the principal 3 by 3 minors we obtain that

$$b_1^2 \le a_1(1-a_1) - \max\left\{ \frac{a_1 b_2^2}{1-a_1}, \frac{(1-a_1)(b_2+x)^2}{a_1} \right\}.$$

Since the geometric mean of the two expression in which maximum we are interested in is $|b_2(b_2+x)|$, we obtain

$$b_1^2 \le a_1(1-a_1) - |b_2(b_2+x)| \le A.$$

Thus (21), we have

$$b_1^2 \le a_1(1-a_1) - b_2(b_2+x) - |a_1 b_2 - (1-a_1)(b_2+x)|.$$

Finally, this gives

$$\begin{aligned} b_1^2 &\le \min\{a_1(1-a_1) - b_2(b_2+x) - a_1 b_2 + (1-a_1)(b_2+x), \\ &\quad a_1(1-a_1) - b_2(b_2+x) + a_1 b_2 - (1-a_1)(b_2+x)\} = \\ &= \min\{(a_1+b_2+x)(1-a_1-b_2), (a_1-b_2-x)(1-a_1+b_2)\}, \end{aligned}$$

which gives (19), as wanted.

■

# 3 Algorithm and numerical experiments

By using the results from the prevous section, we can replace our problem (1) with the equivalent problem of finding the minimum of a linear function ($\mathrm{Tr}(CX)$) on a convex set ($\Sigma = \mathrm{co}(S)$), which is given only by LMI's. Hence, this problem can be easily solved by semi-definite programming (SDP).

We implemented this algorithm in SeDuMi toolbox of MATLAB, and we quickly obtain the solution matrix $X$ of rank 1.

We run more than 10000 experiments with randomly generated matrix $C$, and in 100% of cases our algorithm always returned a minimization matrix $X$ of rank 1, and thus belonging to the set $S$, as wanted.

By factorizing $X = qq^T$, we obtain the wanted Stiefel matrix $Q \in \mathbb{R}^{3\times 2}$ as $Q = \mathrm{vec}^{-1}(q)$.

# References

[1] B.T. Polyak: Convexity of Quadratic Transformation and Its Use in Control and Optimization, *Journal of Optimization Theory and Application*, Vol 99, No. 3 (1998), 553-583.