Optimization over the Stiefel manifold

C. Fraikin[1], **K. Hüper**[2], and **P. Van Dooren**[*1]

[1] Université catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium

[2] NICTA, Canberra Research Laboratory, and Department of Information Engineering, The Australian National University, Canberra, Australia

In this note we parameterize the Stiefel manifold $St_{k,n}$ in a manner that allows to perform a constrained Newton step in a relatively simple way.

Copyright line will be provided by the publisher

We develop an approach to perform Newton steps for the following constrained optimization problem (see [1])

$$\min f\colon St_{k,n} \to \mathbf{R}, \quad U \mapsto \frac{1}{2}\mathrm{tr}(U^T AUB^T) \text{ with given } A \in \mathbf{R}^{n\times n}, B \in \mathbf{R}^{k\times k},$$

where typically $k \ll n$. Here $St_{k,n}$ is the Stiefel manifold of $n \times k$ orthonormal matrices, i.e. matrices $U \in \mathbf{R}^{n\times k}$ s.t. $U^T U = I_k$. The local parameterization of $St_{k,n}$ we use is based on the Cayley transform $C(\Omega)$ of the vector space $\mathcal{W}$ of skew symmetric matrices Ω with block structure as

$$C(\Omega) := (I+\Omega)(I-\Omega)^{-1}, \quad \Omega = \begin{bmatrix} \Omega_{11} & -\Omega_{21}^T \\ \Omega_{21} & 0 \end{bmatrix}, \quad \Omega_{11} = -\Omega_{11}^T \in \mathbf{R}^{k\times k}, \quad \Omega_{21} \in \mathbf{R}^{(n-k)\times k}. \tag{1}$$

The Stiefel manifold $St_{k,n}$ is a $d = k(n-k) + \frac{k(k-1)}{2}$ dimensional manifold and clearly, $\dim \mathcal{W} = d$. The space $\mathcal{W}$ can be used to parameterize the Stiefel manifold $St_{k,n}$ around Q close to $\begin{bmatrix} I_k & 0 \end{bmatrix}^T$ via the function $\varphi\colon \mathcal{W} \to St_{k,n}$, defined by $\Omega \mapsto C(\Omega)Q$. Partition $Q \in St_{k,n}$ and the transformation $U = C(\Omega)Q$ as follows :

$$Q = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}, \quad U = C(\Omega)\begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix}, \quad U_1, Q_1 \in \mathbf{R}^{k\times k}, \quad U_2, Q_2 \in \mathbf{R}^{(n-k)\times k}. \tag{2}$$

Therefore

$$(I+\Omega)\begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} = (I-\Omega)\begin{bmatrix} U_1 \\ U_2 \end{bmatrix} \quad \Longleftrightarrow \quad \Omega \begin{bmatrix} U_1+Q_1 \\ U_2+Q_2 \end{bmatrix} = \begin{bmatrix} U_1-Q_1 \\ U_2-Q_2 \end{bmatrix}$$

with

$$\begin{bmatrix} U_1 \\ U_2 \end{bmatrix} = \begin{bmatrix} -Q_1 \\ Q_2 \end{bmatrix} + 2\begin{bmatrix} I \\ \Omega_{21} \end{bmatrix} S_c^{-1}(Q_1 - \Omega_{21}^T Q_2), \quad S_c := I_k - \Omega_{11} + \Omega_{21}^T\Omega_{21}. \tag{3}$$

The inverse map φ^{-1} can be defined for all U for which $\det(U_1+Q_1) \neq 0$:

$$\begin{bmatrix} \Omega_{11} \\ \Omega_{21} \end{bmatrix} = \begin{bmatrix} (U_1^T+Q_1^T)^{-1}(Q_1^T U_1 + U_2^T Q_2 - U_1^T Q_1 - Q_2^T U_2) \\ U_2 - Q_2 \end{bmatrix}(U_1+Q_1)^{-1}. \tag{4}$$

Assume $n \geq 2k$. Then it can easily be shown that for a given $Q \in St_{k,n}$ the subset of all those $U \in St_{k,n}$ for which $\det(U_1+Q_1) = 0$ is a subset of measure zero. In particular, this means that if Q is sufficiently close to $\begin{bmatrix} I_k & 0 \end{bmatrix}^T$, φ is not just a local parameterization around Q, but almost all of $St_{k,n}$ can be parameterized via φ. Notice, however, that if Q is *not* close to $\begin{bmatrix} I_k & 0 \end{bmatrix}^T$, then those points on $St_{k,n}$ which are not in the image of φ might get arbitrarily close to Q. Clearly, the image of φ is always connected. Notice that the complexity of applying the transformation $C(\Omega)$ to Q requires only $8nk^3 + O(k^3)$ floating point operations because of the use of the Schur complement S_c.

To establish a Newton-type method on $St_{k,n}$ exploiting the parameterization φ we proceed as follows. For any $\Delta \in \mathcal{W}$ we compute the directional derivative

$$\mathrm{D}(f\circ\varphi)(\Omega)\Delta = \frac{\mathrm{d}}{\mathrm{d}\,\varepsilon}(f\circ\varphi)(\Omega+\varepsilon\Delta)\Big|_{\varepsilon=0} = \mathrm{tr}\,(G_\Omega)^T\,\Delta, \tag{5}$$

* Corresponding author: e-mail: paul.vandooren@uclouvain.be, Phone: +00 32 10 478040, Fax: +00 32 10 472180

Copyright line will be provided by the publisher

with

$$G_\Omega := (I+\Omega)^{-1}(AUB^T + A^TUB)Q^T(I+\Omega)^{-1} \tag{6}$$

and where we used

$$\mathrm{D}\, C(\Omega)\Delta = \frac{\mathrm{d}}{\mathrm{d}\,\varepsilon} C(\Omega+\varepsilon\Delta)\Big|_{\varepsilon=0} = 2(I-\Omega)^{-1}\Delta(I-\Omega)^{-1}. \tag{7}$$

As $f\circ\varphi$ is a function on $\mathcal{W}$ we establish an explicit expression for the gradient of $f\circ\varphi$ using the metric induced by the inner product $\langle X,Y\rangle_{\mathcal{W}} := \mathrm{tr}(X^TY)$ for all $X,Y\in\mathcal{W}$. That is

$$\mathrm{D}(f\circ\varphi)(\Omega)\Delta = \mathrm{tr}\left(\Delta^T\, \mathrm{grad}(f\circ\varphi)(\Omega)\right) \tag{8}$$

with

$$\mathrm{grad}(f\circ\varphi)(\Omega) = \frac{1}{2}\left((G_\Omega - G_\Omega^T) - \begin{bmatrix}0 & 0\\ 0 & I_{n-k}\end{bmatrix}(G_\Omega - G_\Omega^T)\begin{bmatrix}0 & 0\\ 0 & I_{n-k}\end{bmatrix}\right), \tag{9}$$

being the image of G_Ω under the orthogonal projection onto $\mathcal{W}$. Accordingly, an explicit expression for the Hessian operator

$$\mathrm{Hess}_{(f\circ\varphi)(\Omega)} \colon \mathcal{W}\to\mathcal{W} \tag{10}$$

can be achieved via computing the directional derivative of the gradient

$$\mathrm{Hess}_{(f\circ\varphi)(\Omega)}\,\Delta = \mathrm{D}(\mathrm{grad}(f\circ\varphi)(\Omega))\Delta = \frac{\mathrm{d}}{\mathrm{d}\,\varepsilon}\,\mathrm{grad}(f\circ\varphi)(\Omega+\varepsilon\Delta)\Big|_{\varepsilon=0}. \tag{11}$$

Now using

$$\mathrm{D}\, U(\Omega)\Delta = (\mathrm{D}\, C(\Omega)\Delta)Q = 2(I-\Omega)^{-1}\Delta(I-\Omega)^{-1}Q \tag{12}$$

and the abbreviations

$$\hat{A} := (I+\Omega)^{-1}A(I-\Omega)^{-1}, \quad \hat{B} := (I-\Omega)^{-1}QBQ^T(I+\Omega)^{-1} \tag{13}$$

we get

$$\mathrm{D}\, G_\Omega\Delta = 2\hat{A}\Delta\hat{B}^T + 2\hat{A}^T\Delta\hat{B} - (I+\Omega)^{-1}\Delta G_\Omega - G_\Omega\Delta(I+\Omega)^{-1}. \tag{14}$$

For each Newton step we need to solve for a skew symmetric Δ the linear equation

$$\mathrm{Hess}_{(f\circ\varphi)(\Omega)}\,\Delta = -\,\mathrm{grad}(f\circ\varphi)(\Omega). \tag{15}$$

This is a linear equation on the space of skew symmetric matrices. If the Hessian $\mathrm{Hess}_{(f\circ\varphi)(\Omega)}$, now considered as the quadratic form $\mathcal{W}\times\mathcal{W}\to\mathbb{R}$, is invertible, the linear system has a unique solution in terms of a $\Delta\in\mathcal{W}$.

The corresponding algorithm was implemented and numerical experiments showed quadratic convergence for a starting point in the basin of attraction as expected.

Acknowledgements This paper presents research supported by the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office and by the National Science Foundation under contract OCI-03-24944. The scientific responsibility rests with its authors. NICTA is funded by the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council.

References

[1] C. Fraikin, Yu. Nesterov and P. Van Dooren, Optimizing the coupling between two isometric projections of matrices. Submitted to SIAM Journal on Matrix Analysis and Applications, 2005.

Copyright line will be provided by the publisher