Notes on Optimization on Stiefel Manifolds

Hemant D. Tagare

Department of Diagnostic Radiology
Department of Biomedical Engineering
Yale University, New Haven, CT 06520.

Version: *01/24/2011*

1 Introduction

1.1 The optimization problem

Suppose we wish to find p orthonormal vectors in $\mathcal{R}^n$ that are optimal with respect to an objective function F. We pose this problem as follows: Let X be any $n \times p$ matrix satisfying $X^T X = I$. We take the columns of X to be p orthonormal vectors in $\mathcal{R}^n$ and we assume that F is a function from the space of $n \times p$ matrices to the real line. We form the optimization problem:

$$\min_{X \in \mathcal{R}^{n \times p}} F(X) \text{ s.t. } X^T X = I. \tag{1}$$

1.2 Overview of the algorithm

This note reviews a recent algorithm [1] for solving the above problem. The key ideas behind the algorithm are as follows:

1. The constraint set $X^T X = I$ is a submanifold of $\mathcal{R}^{n \times p}$ called the *Stiefel manifold*.

2. The algorithm works by finding the gradient of F in the tangent plane of the manifold at the point $X^{[k]}$ of the current iterate (see figure **??**). A curve is found on the manifold that proceeds along the projected negative gradient, and a curvilinear search is made along the curve for the next iterate $X^{[k+1]}$.

3. The search curve is not a geodesic, but is instead constructed using a Cayley transformation (explained in section 4). This has the advantage that matrix exponentials are not required. The algorithm only requires inversion of a $2p \times 2p$ matrix. The algorithm is especially suitable for use in high dimensional spaces when $p << n$.

This note explains details of the algorithm. The note should be accessible to any graduate student who is familiar with vector space theory and has some familiarity with differentiable manifolds (it is sufficient to know what a differentiable sub-manifold and its tangent space are).

We will begin with some preliminaries, and then move on to an introduction to Stiefel manifolds and its tangent spaces. Next, we will consider how the search curve is created via the Cayley transform. Finally, we present the numerical algorithm.

2 Preliminaries

2.0.1 Vector Spaces

We only consider finite-dimensional real vector spaces. I am assuming that you are familiar with the theory of these spaces. Key facts and terminology that we need are mentioned here for reference.

Orthonormal coordinates: If $v_1, v_2, \ldots, v_n$ form a basis of a vector space $\mathcal{V}$, then any vector $u \in \mathcal{V}$ can be expressed as $u = a_1 v_1 + a_2 v_2 + \ldots + a_n v_n$ for unique real numbers $a_1, a_2, \ldots, a_n$. We will call these real numbers the *coordinates* of u (in the basis $\{v_1, \ldots, v_n\}$). The basis is *orthonormal* with respect to an inner product $\langle ., . \rangle$ if $\langle v_i, v_j \rangle = \delta_{i,j}$, the Kronecker delta. Equivalently, the basis is orthonormal if for any $u = \sum_i a_i v_i$ and $v = \sum_i b_i v_i$, we have $\langle u, v \rangle = \sum_i a_i b_i$. If the basis is orthonormal, then the coordinates with respect to the basis are *orthonormal coordinates*.

Orthogonal complements: Two subspaces $\mathcal{A}$ and $\mathcal{B}$ of a vector space $\mathcal{V}$ are *orthogonal complements* in $\mathcal{V}$ if

1. $\mathcal{A}$ and $\mathcal{B}$ are orthogonal to each other, i.e for any $a \in \mathcal{A}$ and any $b \in \mathcal{B}$, $\langle a, b \rangle = 0$.

2. $\mathcal{V}$ is the direct sum of $\mathcal{A}$ and $\mathcal{B}$, i.e. any $v \in \mathcal{V}$ can be written as $v = a + b$ for some $a \in \mathcal{A}$ and $b \in \mathcal{B}$. *Check*

Dimension Counting: Because we are dealing with finite-dimensional subspaces, many arguments can be simplified by just counting dimensions. For example (we use these arguments later):

1. To show that two subspaces A and B are equal, it is sufficient to show that A is a subspace of B and that dim $(A) =$ dim (B).

2. To show that two subspaces $\mathcal{A}$ and $\mathcal{B}$ are orthogonal complements in $\mathcal{V}$, it is sufficient to show that $\mathcal{A}$ is orthogonal to $\mathcal{B}$, and that dim $\mathcal{V} =$ dim $\mathcal{A} +$ dim $\mathcal{B}$.

Representation: If $\langle ., . \rangle$ is an inner product defined on $\mathcal{V}$, then corresponding to any linear functional $L : \mathcal{V} \to \mathcal{R}$ there is a $v \in \mathcal{V}$ with the property that

$$\langle v, u \rangle = L(u) \text{ for all } u \in \mathcal{V}.$$

The vector v is the *representation* of L. The representation of L depends on the inner product.

Automorphisms: An invertible linear map $L : V \to W$ between two vector spaces V, W is an *isomorphism*. An isomorphism between two spaces shows that the spaces are essentially identical as far as their vector space structure is concerned. An isomorphism between a space and itself is an *automorphism*.

2.0.2 Vector Spaces of Matrices

The vector space of $n \times p$ matrices is $\mathcal{R}^{n \times p}$. The Euclidean inner product for this vector space is

$$\langle A, B \rangle = \text{tr } (A^T B) = \sum_{i,j} a(i,j) b(i,j).$$

The matrix $I_{n,p} \in \mathcal{R}^{n\times p}$ is the "truncated" identity matrix

$$I_{n,p}(i,j) = \delta_{i,j}.$$

The matrix $I_{n,n}$ is the standard identity matrix. We will usually write the standard identity matrix without subscripts. Its dimensions will be clear from context.

The set of symmetric $n \times n$ matrices forms a subspace of $\mathcal{R}^{n\times n}$. It is denoted Sym_n. The dimension of Sym_n is $\frac{1}{2}n(n+1)$. The projection of a matrix $A \in \mathcal{R}^{n\times n}$ on Sym_n is $\text{sym}(A) = \frac{1}{2}(A + A^T)$.

The set of all skew-symmetric $n \times n$ matrices also forms a subspace of $\mathcal{R}^{n\times n}$. It is denoted Skew_n. The dimension of Skew_n is $\frac{1}{2}n(n-1)$. The projection of a matrix $A \in \mathcal{R}^{n\times n}$ on Skew_n is $\text{skew}(A) = \frac{1}{2}(A - A^T)$.

The subspaces Sym_n and Skew_n are orthogonal complements in $\mathcal{R}^{n\times n}$. This is easily verified by showing that any two elements of Sym_n and Skew_n are orthogonal, and that $\mathcal{R}^{n\times n}$ is the direct sum of Sym_n and Skew_n (since $A = \text{sym}(A) + \text{skew}(A)$).

2.1 Manifold Structure of $X^T X = I$

The most basic fact we need is, of course, that the set $\{X \in \mathcal{R}^{n\times p} \mid X^T X = I\}$ is a manifold. A proof can be found in [2]. This manifold is the *Stiefel manifold*. The standard notation for a Stiefel manifold of p orthonormal vectors in $\mathcal{R}^n$ is $\mathcal{V}_p(\mathcal{R}^n)$. It has dimension equal to $np - \frac{1}{2}p(p+1)$. We will view this manifold as an embedded sub-manifold of $\mathcal{R}^{n\times p}$. This means that we identify tangent vectors to the manifold with $n \times p$ matrices.

2.2 The Tangent Space

Our next concern is to understand the tangent space to $\mathcal{V}_p(\mathcal{R}^n)$ at X. The tangent space at X is denoted $\mathcal{T}_X\mathcal{V}_p(\mathcal{R}^n)$. Vectors in the tangent space are characterized by

Lemma 1. *Any $Z \in \mathcal{T}_X\mathcal{V}_p(\mathcal{R}^n)$, then Z (as an element of $\mathcal{R}^{n\times p}$) satisfies*

$$Z^T X + X^T Z = 0.$$

That is, $Z^T X$ is a skew-symmetric $p \times p$ matrix.

Proof: Let $Y(t)$ be a curve in $\mathcal{V}_p(\mathcal{R}^n)$. Then, $Y^T(t)Y(t) = I$. Differentiating this w.r.t. t gives $Y'^T(t)Y(t) + Y^T(t)Y'(t) = 0$. At $t = 0$, setting $Y(0) = X$ and $Y'(0) = Z$ gives the result. □

Next, we will find a useful represent for tangent vectors. We will proceed as follows: Noting that the tangent space $\mathcal{T}_X\mathcal{V}_p(\mathcal{R}^n)$ is a subspace of $\mathcal{R}^{n\times p}$, we will first find a representation for elements of $\mathcal{R}^{n\times p}$, and then specialize the representation to $\mathcal{T}_X\mathcal{V}_p(\mathcal{R}^n)$.

Recall that if $X \in \mathcal{V}_p(\mathcal{R}^n)$, then its columns are orthonormal vectors in $\mathcal{R}^n$. For any such X, we can always find additional $n-p$ vectors in $\mathcal{R}^n$ (by the Gram-Schmidt procedure, for example) which when combined with the p columns of X form an orthonormal basis of $\mathcal{R}^n$. Let $X_\perp$ denote the $n \times n-p$ matrix whose columns are these

new $n-p$ vectors, and consider the matrix $[XX_\perp]$ (the concatenation of columns of X and $X_\perp$. Not to be confused with the Lie bracket). This is an $n \times n$ orthonormal matrix.

Lemma 2. *The matrix $[XX_\perp]$ (viewed as a linear operator) is an automorphism $\mathcal{R}^{n\times p}$.*

Proof: Let $W = [XX_\perp]$, then W is an $n \times n$ orthonormal matrix, and hence is invertible. Further, since multiplying an $n \times n$ matrix by an $n \times p$ matrix gives an $n \times p$ matrix, $W\mathcal{R}^{n\times p} \subset \mathcal{R}^{n\times p}$. Using the same argument for W^{-1} shows $W^{-1}\mathcal{R}^{n\times p} \subset \mathcal{R}^{n\times p}$. Hence W is an automorphism of $\mathcal{R}^{n\times p}$. □

Hence any element $U \in \mathcal{R}^{n\times p}$ can be written as $U = [XX_\perp]C$, where C is an $n \times p$ matrix. Splitting C as

$$C = \begin{bmatrix} A \\ B \end{bmatrix},$$

where A is a $p\times p$ matrix and the matrix B is a $(n-p)\times p$ matrix, then $U = XA+X_\perp B$. This is the representation we want for elements of $\mathcal{R}^{n\times p}$. Since

$$\begin{aligned} \text{tr}\,(U^TU) &= \text{tr}\,((XA+X_\perp B)^T(XA+X_\perp B)) = \text{tr}\,(A^TA+B^TB) \\ &= \sum_{i,j} a^2(i,j)+b^2(i,j), \end{aligned}$$

the elements of A and B are orthonormal coordinates for $\mathcal{R}^{n\times p}$ (for a fixed X).

The representation $U = XA+X_\perp B$ is specialized to the tangent space $\mathcal{T}_X\mathcal{V}_p(\mathcal{R}^n)$ by

Lemma 3. *A matrix $Z = XA+X_\perp B$ belongs to the tangent space $\mathcal{T}_X\mathcal{V}_p(\mathcal{R}^n)$ if and only if A is skew symmetric.*

Proof: Let S be the subspace of $\mathcal{R}^{n\times p}$ consisting of all matrices Z which can be expressed as $Z = XA+X_\perp B$ with A skew-symmetric. Since A is $p \times p$ and B is a $(n-p)\times p$ matrix, dim $(S) = \frac{1}{2}p(p-1)+(n-p)p = np-\frac{1}{2}p(p+1)$.

Next, let Z be an element of $\mathcal{T}_X\mathcal{V}_p(\mathcal{R}^n)$. Since Z also belongs to $\mathcal{R}^{n\times p}$ it can be expressed as $Z = XA+X_\perp B$. Then, the condition $Z^TX+X^TZ = 0$ gives $A^T+A = 0$, showing that A is skew-symmetric. Thus Z belongs to S, showing that $\mathcal{T}_X\mathcal{V}_p(\mathcal{R}^n)$ is a subspace of S. However,

$$\text{dim}\,(\mathcal{T}_X\mathcal{V}_p(\mathcal{R}^n)) = \text{dim}\,\mathcal{V}_p(\mathcal{R}^n) = np-\frac{1}{2}p(p+1) = \text{dim}\,(S)$$

showing that $S = \mathcal{T}_X\mathcal{V}_p(\mathcal{R}^n)$. □

The elements of A above the principle diagonal and all elements of B are coordinates of the tangent vector Z. The representation of tangent vectors given in lemma 3 is used below to gain insight into different inner products defined on the tangent space $\mathcal{T}_X\mathcal{V}_p(\mathcal{R}^n)$.

2.3 Euclidean and Canonical Inner Products for the Tangent Space

The Stiefel manifold becomes a Riemannian manifold by introducing an inner product in its tangent spaces. There are two natural inner products for tangent spaces of Stiefel manifolds: the Euclidean inner product and the canonical inner product.

The Euclidean inner product: Let $Z_1, Z_2 \in \mathcal{T}_X \mathcal{V}_p(\mathcal{R}^n)$. Then define the Euclidean inner product

$$\langle Z_1, Z_2 \rangle_e = \text{tr}\,(Z_1^T Z_2).$$

This inner product is the inner product of the ambient space $\mathcal{R}^{n\times p}$. The metric induced by this inner product is the *Euclidean metric*.

Setting $Z = XA + X_\perp B$ where A is a $p \times p$ skew symmetric matrix and B is a $(n-p) \times p$ arbitrary matrix gives

$$\begin{aligned}
\langle Z, Z \rangle_e &= \text{tr}\,(B^T X_\perp^T + A^T X^T)(XA + X_\perp B) \\
&= \text{tr}\,(B^T X_\perp^T XA + B^T X_\perp^T X_\perp B + A^T X^T XA + A^T X^T X_\perp B) \\
&= \text{tr}\,(A^T A + B^T B) = \text{tr}\,(A^T A) + \text{tr}\,(B^T B),
\end{aligned}$$

where we have used $X^T X = I$, $X_\perp^T X_\perp = I$, and $X^T X_\perp = 0$. Notice that tr $(A^T A) = \sum_{i>j} 2a^2(i,j)$ and tr $(B^T B) = \sum_{i,j} b^2(i,j)$, so that

$$\langle Z, Z \rangle_e = \sum_{i>j} 2a^2(i,j) + \sum_{i,j} b^2(i,j).$$

Recall that the elements of A above the principle diagonal and all elements of B are coordinates of Z. The Euclidean metric weighs these coordinates unequally; it weighs the "A" coordinates twice as much as the "B" coordinates.

The canonical inner product: The canonical inner product weighs the coordinates equally. Loosely speaking, the idea is to find the "A" matrix of the tangent vector Z and weigh it by $\frac{1}{2}$ in the inner product. This is done by the following argument: Since $Z = XA + X_\perp B$, the matrix A is given by $A = X^T Z$, and XA is given by and $XA = XX^T Z$. Thus $(I - \frac{1}{2}XX^T)Z = Z - \frac{1}{2}XX^T Z = XA + X_\perp B - \frac{1}{2}XA = \frac{1}{2}XA + X_\perp B$, and

$$\begin{aligned}
\text{tr}\,(Z^T(I - \frac{1}{2}XX^T)Z) &= \text{tr}\,(XA + X_\perp B)^T(\frac{1}{2}XA + X_\perp B) \\
&= \frac{1}{2}\text{tr}\,A^T A + \text{tr}\,B^T B \\
&= \sum_{i>j} a^2(i,j) + \sum_{i,j} b^2(i,j),
\end{aligned}$$

which gives equal weight to elements of A and B.

Based on the above argument, define the *canonical inner product*

$$\langle Z_1, Z_2 \rangle_c = \text{tr}\,(Z_1^T(I - \frac{1}{2}XX^T)Z_2),$$

and the *canonical metric* $\langle Z, Z \rangle_c$.

Which inner product? The canonical inner product and metric are exclusively used in the discussion below. Of course, the Eucldiean inner product can be used as well. The canonical inner product seems natural because it gives equal weight to the elements of A and B (there may be deeper reason for choosing the canonical inner product, but I am not aware of it at the moment).

3 Differentials and Gradients

Now that we have some understanding of the tangent spaces of Stiefel manifolds, we can return to the optimization problem of equation(1).

If F is a function from $\mathcal{R}^{n\times p}$ to $\mathcal{R}$, and $X, Z \in \mathcal{R}^{n\times p}$, then $DF_X : \mathcal{R}^{n\times p} \to \mathcal{R}$, called the *differential* of F, gives the derivative of F in the Z direction at X by

$$\begin{aligned} DF_X(Z) &= \sum_{i,j} \frac{\partial F}{\partial X_{i,j}} Z_{i,j} && (2) \\ &= \text{tr}\,(G^T Z), && (3) \end{aligned}$$

where $G = \left[\frac{\partial F}{\partial X_{i,j}}\right] \in \mathcal{R}^{n\times p}$. From now on, we will reserve the symbol G to represent this matrix.

Because DF_X is a linear functional on $\mathcal{R}^{n\times p}$ a representation of DF_X in $\mathcal{R}^{n\times p}$ of any of its subspaces is the *gradient* of F in $\mathcal{R}^{n\times p}$ or in the subspace. Equation (2) shows that G is the representation of DF_X under the Euclidean inner product for $\mathcal{R}^{n\times p}$ (recall the discussion of representations in section 2.0.1).

Now suppose that X is a point in the Stiefel manifold $\mathcal{V}_p(\mathcal{R}^n)$. Then DF_X is also linear functional on the tangent space $\mathcal{T}_X\mathcal{V}_p(\mathcal{R}^n)$. Hence DF_X restricted to $\mathcal{T}_X\mathcal{V}_p(\mathcal{R}^n)$ has a representation. The representation is:

Lemma 4. *Under the canonical inner product, the vector AX with $A = (GX^T - XG^T)$ represents the action of DF_X on the tangent space $\mathcal{T}_X\mathcal{V}_p(\mathcal{R}^n)$.*

Proof: Because G is in $\mathcal{R}^{n\times p}$ it can be expressed as $G = XG_A + X_\perp G_B$. Suppose Z is a tangent vector to the Steifel manifold at X, then $Z = XZ_A + X_\perp Z_B$, where Z_A is skew symmetric. Therefore,

$$\begin{aligned} DF_X(Z) &= \text{tr}\,(G^T Z) \\ &= \text{tr}\,((XG_A + X_\perp G_B)^T (XZ_A + X_\perp Z_B)) \\ &= \text{tr}\,(G_A^T Z_A) + \text{tr}\,(G_B^T Z_B). \end{aligned}$$

Writing G_A as $G_A = \text{sym}\, G_A + \text{skew}\, G_A$, we get $\text{tr}\,(G_A^T Z_A) = \text{tr}\,((\text{skew}\, G_A)^T Z_A)$, so that

$$DF_X(Z) = \text{tr}\,((\text{skew}\, G_A)^T Z_A) + \text{tr}\,(G_B^T Z_B). \quad (4)$$

Suppose $U = XA + X_\perp B$ is the vector in the tangent space that represents the action of DF_X. Then,

$$\begin{aligned} \langle U, Z\rangle_c &= \text{tr}\,(U^T(I - \frac{1}{2}XX^T)Z) \\ &= \frac{1}{2}\text{tr}\,(A^T Z_A) + \text{tr}\,(B^T Z_B). \end{aligned} \tag{5}$$

Comparing equations (4) and (5), U can be made to represent DF_X by setting $A = 2\text{skew}\, G_A$ and $B = G_B$. Thus,

$$U = 2X\text{skew}\,(G_A) + X_\perp G_B.$$

But skew$(G_A) = \frac{1}{2}(G_A - G_A^T) = \frac{1}{2}(X^T G - G^T X)$, so that

$$\begin{aligned} U &= 2X\text{skew}\,(G_A) + X_\perp G_B \\ &= X(X^T G - G^T X) + X_\perp B \\ &= X(X^T G) + X_\perp G_B - XG^T X \\ &= XG_A + X_\perp G_B - XG^T X \\ &= G - XG^T X \\ &= GX^T X - XG^T X \\ &= (GX^T - XG^T)X. \quad \square \end{aligned}$$

We will denote the vector $AX = (GX^T - XG^T)X$ by $\nabla_c F$ to suggest that it is the gradient of F under the canonical metric. Note that A is a skew symmetric $n \times n$ matrix.

4 Cayley Transform and the Search Curve

Having found the gradient of F, we turn to generating the descent curve.

Let $X \in \mathcal{V}_p(\mathcal{R}^n)$, and W be any $n \times n$ skew-symmetric matrix. Consider the curve

$$Y(\tau) = \left(I + \frac{\tau}{2}W\right)^{-1}\left(I - \frac{\tau}{2}W\right)X. \tag{6}$$

This curve has the following properties:

1. It stays in the Stiefel manifold, i.e. $Y(\tau)^T Y(\tau) = I$.
2. Its tangent vector at $\tau = 0$ is $Y'(0) = -WX$.
3. If we set $W = A = GX^T - XG^T$, then the curve is a descent curve for F.

We can view $Y(\tau)$ as the point X transformed by $\left(I + \frac{\tau}{2}W\right)^{-1}\left(I - \frac{\tau}{2}W\right)$. The tranformation $\left(I + \frac{\tau}{2}W\right)^{-1}\left(I - \frac{\tau}{2}W\right)$ is called the *Cayley transformation*.

The rough sketch of the minimization algorithm using $Y(\tau)$ is as follows: Begin with some initial $X^{[1]}$. For $k = 1, \ldots$ generate $X^{[k+1]}$ from $X^{[k]}$ by a curvilinear

search along $Y(\tau) = \left(I + \frac{\tau}{2}W\right)^{-1}\left(I - \frac{\tau}{2}W\right)X^{[k]}$ by changing τ. The search is carried out using the Armijo-Wolfe rule. This is discussed below.

The formula for the curve $Y(\tau)$ requires an inversion of $I + \frac{\tau}{2}W$ which is an $n \times n$ matrix. This is computationally prohibitive. However, there is a technique for finding $Y(\tau)$ based on the Sherman-Morrison-Woodbury formula which only requires inverting a $2p \times 2p$ matrix.

5 The Sherman-Morrison-Woodbury Formula

The Sherman-Morrison-Woodbury (SMW) formula is a fast way for calculating matrix inverses of a certain form:

$$(B + \alpha UV^T)^{-1} = B^{-1} - \alpha B^{-1}U(I + \alpha V^T B^{-1}U)^{-1}V^T B^{-1}. \tag{7}$$

The SWM formula is important because it allows you to update the inverse of B to the inverse of $B + \alpha UV^T$ efficiently. If U, V are $n \times p,\ p < n$ matrices, then assuming that B^{-1} is available, calculating $(B + \alpha UV^T)^{-1}$ only requires inverting $(I + \alpha V^T B^{-1}U)$, which is a $p \times p$ matrix.

We use SMW to calculate $\left(I + \frac{\tau}{2}W\right)^{-1}$ for $W = A = GX^T - XG^T$. First we define $U = [G,\ X]$ and $V = [X,\ -G]$, so that $W = A = UV^T$. Then, using the SWM formula

$$(I + \tau UV^T)^{-1} = I - \frac{\tau}{2}U\left(I + \frac{\tau}{2}V^T U\right)^{-1}V^T. \tag{8}$$

Note that $V^T U$ and $I + \frac{\tau}{2}V^T U$ are $2p \times 2p$.

Using equation (8),

$$\begin{aligned}
&= \left(I + \frac{\tau}{2}W\right)^{-1}\left(I - \frac{\tau}{2}W\right) \\
&= \left(I + \frac{\tau}{2}UV^T\right)^{-1}\left(I - \frac{\tau}{2}UV^T\right) \\
&= \left(I - \frac{\tau}{2}U\left(I + \frac{\tau}{2}V^T U\right)^{-1}V^T\right)\left(I - \frac{\tau}{2}UV^T\right) \\
&= I - \frac{\tau}{2}U\left(I + \frac{\tau}{2}V^T U\right)^{-1}V^T - \frac{\tau}{2}UV^T + \frac{\tau^2}{4}U\left(I + \frac{\tau}{2}V^T U\right)^{-1}V^T UV^T \\
&= I - \frac{\tau}{2}U\left(I + \frac{\tau}{2}V^T U\right)^{-1}\left\{I + (I + \frac{\tau}{2}V^T U) - \frac{\tau}{2}V^T U\right\}V^T \\
&= I - \tau U\left(I + \frac{\tau}{2}V^T U\right)^{-1}V^T. \qquad (9)
\end{aligned}$$

From the above equation, we get

$$Y(\tau) = X - \tau U\left(I + \frac{\tau}{2}V^T U\right)^{-1}V^T X, \tag{10}$$

which is computationally simpler.

Next, we turn our attention to curvilinear search along $Y(\tau)$.

6 Curvilinear Search

Curvilinear search is just traditional linear search applied to the curve $Y(\tau)$. The search terminates when the Armijo-Wolfe conditions are satisfied. The Armijo-Wolfe conditions require two parameters $0 < \rho_1 < \rho_2 < 1$.

Curvilinear Search:

Initialize τ to a non-zero value.
Until $\{ F(Y(\tau)) \leq F(Y(0)) + \rho_1 \tau F'(Y(0))$ and
$F'(Y(\tau)) \geq \rho_2 F'(Y(0)) \}$
do $\tau \leftarrow \frac{\tau}{2}$.
Return $Y(\tau)$ as the curvilinear search "minimum".

To use curvilinear search, we need formulae for $F'(Y(\tau))$ and $F'(Y(0))$. (*Derive these*)

$$
\begin{aligned}
F'(Y(\tau)) &= \text{tr}\,(G^T Y'(\tau)), && (11)\\
\text{where} & \\
Y'(\tau) &= -\left(I + \frac{\tau}{2}A\right)^{-1} A \left(\frac{X + Y(\tau)}{2}\right), \\
Y'(0) &= -AX, && (12)
\end{aligned}
$$

where, as before, $A = GX^T - XG^T$.

7 The Algorithm

The algorithm is a straightforward application of curvilinear search.

Minimize on Steifel Manifold:

Initialize: Set $k = 1$ and $X^{[k]}$ to a point in the manifold.
Iterate till convergence: Calculate $G^{[k]} = DF(X^{[k]})$, the A, U and V matrices.
Use curvilinear search to obtain $X^{[k+1]}$.
Test for convergence.

References

[1] Z. Wen, W. Yin, "A Feasible Method For Optimization with Orthogonality Constraints", Rice University Technical Report, 2010.

[2] W. M. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Academic Press, 2002.