Optimization over Grassmann manifolds

Kerstin Johnsson

July 4, 2012

The purpose of this paper is to explain the theory behind the R package `grassopt`, which provides functions for minimizing a function over a Grassmann manifold. For details of the functions we refer to the manual; this is a more general introduction to the theory behind them. The theory mainly comes from [1] by Edelman et al., where Newton's method and the conjugate gradient method is adapted from the usual Euclidean space to the Stiefel and Grassmann manifolds. These manifolds are viewed as Riemannian manifolds with canonical metrics, which means that we get a way to define the gradient and the Hessian of a function on the manifold as well as geodesics on the manifold. The background needed is some Riemannian geometry, including knowledge of tangent spaces and geodesics on Riemannian manifolds. An accessible introduction to Riemannian geometry is [2].

Newton's method

Newton's method minimizes a convex function f over $\mathbb{R}^n$ through the following steps:

1. Start with an initial guess x_0 and a tolerance ϵ.
2. Repeat:
   (a) Compute the Newton step $\Delta x = -(\mathrm{Hess}(f)^{-1} \cdot \nabla f)\big|_{x_n}$.
   (b) Use line search to find the step size t.
   (c) Let $x_n = x_{n-1} + t \cdot \Delta x$.

   until $\nabla f(x_n)^T \Delta x < \epsilon$.

When f is not convex, i.e. when $\text{Hess}(f)$ is not positive semidefinite everywhere, Δx is not always a descent direction. In order to work also for non-convex functions the Hessian can be replaced by $B = \text{Hess}(f) + \epsilon I$, where ϵ is chosen so that B is positive definite. This case is not treated in [1], but it is implemented in `grassopt` and we will return to it in the end.

In their extension of Newton's method to functions on Stiefel and Grassmann manifolds Edelman et al. used the gradient and the Hessian from Riemannian geometry to define the Newton step.

Gradient and Hessian on a Riemannian manifold

Let (M, g) be a Riemannian manifold with metric g. Then the gradient of a function $f\colon M \to \mathbb{R}$ at $x \in M$ is defined as the unique tangent ∇f such that

$$g(\nabla f, v) = \partial_v f := \frac{d}{dt}(f(\gamma(t)))\Big|_{t=0} \tag{1}$$

for any tangent vector v. $\gamma(t)$ is any curve in M such that $\gamma(0) = x$ and $\dot{\gamma}(0) = v$. The Hessian of f is a tensor field of type $(2, 0)$, i.e. $\text{Hess}(f)\colon T^*M \otimes T^*M \to \mathbb{R}$. It can for example be defined by

$$\text{Hess}(f)(X, Y) = X(Y(f)) - df(\nabla_X Y)\,,$$

where ∇ is the Levi-Civita connection on (M, g). If $X = Y = \dot{\gamma}(0)$ where γ is a geodesic on (M, g), then $\nabla_X Y = 0$, so

$$\text{Hess}(f)(X, Y) = X(Y(f)) = \frac{d^2}{dt^2} f(\gamma(t))\Big|_{t=0}\,. \tag{2}$$

The Newton step is determined as the unique tangent vector V_{nt} such that

$$\text{Hess}(f)(V_{nt}, V) = g(-\nabla f, V) \tag{3}$$

for all tangent vectors V. Since the Hessian is symmetric and bilinear, (2) can be used to compute the Hessian for any X and Y. By polarization we get

$$\text{Hess}(f)(X, Y) = \frac{1}{4}\left\{\text{Hess}(f)(X + Y, X + Y) - \text{Hess}(f)(X - Y, X - Y)\right\}.$$

Lines in $\mathbb{R}^n$ corresponds to geodesics on a Riemannian manifold, so when updating x, instead of moving along a line we move along the geodesic with tangent V_{nt}.

The Grassmann Manifold

A point on the Grassmann manifold $\mathrm{Gr}(p, n)$ is represented by an orthonormal basis for the subspace, i.e. an $n \times p$-matrix Y such that $Y^T Y = I$. Clearly this representation is not unique, each Y corresponds to one point on the Stiefel manifold $V_p(\mathbb{R}^n)$ of which $\mathrm{Gr}(n, p)$ is a quotient space. If $p = n$, then $V_p(\mathbb{R}^n) = O(n)$. The equivalence relation yielding $\mathrm{Gr}(p, n)$ as a quotient space of $V_p(\mathbb{R}^n)$ is

$$Y \sim Y' \iff \exists Q \in O(p) \colon Y = Y'Q \ .$$

This means that two matrices are equivalent if their columns span the same subspace of $\mathbb{R}^n$, hence a point in the quotient space represents a p-plane in $\mathbb{R}^n$.

The Tangent Space

The quotient space representation $\mathrm{Gr}(p, n) = V_p(\mathbb{R}^n)/{\sim}$ can be used to find the tangent space of $\mathrm{Gr}(p, n)$. By differentiating $Y^T Y = I$ we see that the tangent space of the Stiefel manifold at Y consists of all matrices V such that $Y^T V$ is skew symmetric. If $Y_\perp$ is any n-by-$(n-p)$ matrix such that $[Y \;\; Y_\perp] \in O(n)$, a tangent can also be written as

$$V = YA + Y_\perp B \ ,$$

where A is skew-symmetric (p-by-p) and B is arbitrary ($(n-p)$-by-p).

The p-by-p skew-symmetric matrices constitute the tangent space to $O(p)$, hence the tangents $V = YA$ are also tangents to the submanifold

$$M = \{Y' : Y' \sim Y\} \ .$$

The tangents $V = Y_\perp B$ belong to the normal space of this submanifold under the Euclidean inner product metric $g(A, B) = \mathrm{tr}(A^T B)$, which is the canonical metric for the Grassmann manifold and the one used here. The tangent space of M at Y is called the *vertical space* and the normal space of M at Y is called the *horizontal space*. Vectors fields the horizontal space represent tangents of the quotient manifold [3], hence $V = Y_\perp B$ are the tangents of the Grassmann manifold. An equivalent characterization is that V is a tangent to $\mathrm{Gr}(p, n)$ at Y if

$$Y^T V = 0 \ .$$

It is easy to see that the orthogonal projection onto the tangent space of $\mathrm{Gr}(p, n)$ at Y is $\pi(Z) = (I - YY^T)Z$.

The Gradient

Since we use the Euclidean inner product metric on the Grassmann manifold, from (1) we get that the gradient of the function F on the Grassmann manifold is

$$\nabla F = \pi(F_Y) = (I - YY^T)F_Y \ ,$$

where

$$(F_Y)_{ij} = \frac{\partial F}{\partial Y_{ij}} \ . \tag{4}$$

The Geodesics

Geodesics on the Grassman manifold have a very nice form when we represent points by $n \times p$ matrices. If H is a tangent in the horizontal space at Y_0 and $H = U\Sigma V^T$ is the compact singular value decomposition of H, then the geodesic which has direction H at Y_0 is

$$Y(t) = (Y_0 V \ \ U) \begin{pmatrix} \cos \Sigma t \\ \sin \Sigma t \end{pmatrix} V^T. \tag{5}$$

To arrive at this expression Edelman et al. first find that geodesics in the Stiefel manifold satisfy the differential equation

$$\ddot{Y} + Y(\dot{Y}^T \dot{Y}) = 0 \ , \tag{6}$$

then they show that $Y^T \dot{Y}$ is constant along a geodesic on the Stiefel manifold. This means that if $\dot{Y}(t)$ is horizontal for some t, then $Y^T \dot{Y} = 0$ along the geodesic, i.e. $\dot{Y}(t)$ is horizontal for all t. Hence a geodesic on the Stiefel manifold with $\dot{Y}(0) = H$ is a geodesic along the Grassmann manifold. Now, in order to see that (5) is a geodesic it is enough to verify that it satisfies (6).

The Hessian

Let $Y(t)$ be the geodesic (5). Define F_{Y_0} from (4) and $F_{Y_0Y_0}$ from

$$(F_{YY})_{ij,kl} = \frac{\partial F}{\partial Y_{ij} \partial Y_{kl}}, \quad F_{YY}(H_1, H_2) = \sum_{ij,kl} (F_{YY})_{ij,kl} (H_1)_{ij} (H_2)_{kl} \ .$$

Then by (2)

$$\begin{aligned} \mathrm{Hess}(F)(H,H) &= \frac{d^2}{dt^2} F(Y(t))\Big|_{t=0} = \frac{d}{dt} \mathrm{tr}(F_{Y(t)}^T \dot{Y}(t))\Big|_{t=0} \\ &= F_{Y_0Y_0}(H,H) + \mathrm{tr}(F_{Y_0}^T \ddot{Y}(0)) \\ &= F_{Y_0Y_0}(H,H) - \mathrm{tr}(F_{Y_0}^T Y_0 H^T H) \ , \end{aligned}$$

and from polarization

$$\mathrm{Hess}(F)(H_1, H_2) = F_{Y_0Y_0}(H_1, H_2) - \mathrm{tr}(F_{Y_0}^T Y_0 H_1^T H_2) \ .$$

Newton step on the Grassmannian

Now we have the ingredients in equation (3) for the Grassmann manifold. Let $\Pi_T = I - YY^T$ denote the projection matrix for the projection onto the tangent space. The Newton step for the function F at the point Y on the Grassmann manifold is the unique tangent V_{nt} such that

$$F_{YY}(V_{nt}, V) - \mathrm{tr}(F_Y^T Y V_{nt}^T V) = -g(\Pi_T F_Y, V) \tag{7}$$

for all tangents V at Y. If we let $\widehat{F_{YY}}(V)$ be the matrix defined by

$$(\widehat{F_{YY}}(V))_{kl} = \sum_{ij} (F_{YY})_{ij,kl} V_{ij} \ ,$$

the left-hand side of this equation can be written as

$$\begin{aligned} F_{YY}(V_{nt}, V) - \mathrm{tr}(F_Y^T Y V_{nt}^T V) &= \mathrm{tr}(\widehat{F_{YY}}(V_{nt})^T V) - \mathrm{tr}((V_{nt} Y^T F_Y)^T V) \\ &= g(\Pi_T \widehat{F_{YY}}(V_{nt}) - V_{nt} Y^T F_Y, V) \ . \end{aligned}$$

This means that (7) holds for all tangents V if and only if

$$\Pi_T \widehat{F_{YY}}(V_{nt}) - V_{nt} Y^T F_Y = -\Pi_T F_Y \ . \tag{8}$$

This is a system of np linear equations and np unknowns, and we can write it as

$$(-\Pi_T F_Y)_{ml} = \sum_{ij} \left(\sum_k (\Pi_T)_{mk} (F_{YY})_{ij,kl} - I_{\{i=m\}} (Y^T F_Y)_{jl} \right) (V_{nt})_{ij} \, , \tag{9}$$

where I_{cond} is the indicator function, i.e. $I_a = 1$ if a is true and $I_a = 0$ otherwise. The system (8) can be underdetermined, but using the additional constraint $Y^T V_{nt} = 0$, V_{nt} is determined uniquely.

Newton's method for non-convex functions

In Euclidean space, if the Hessian is not positive definite we replace it with $B = \text{Hess}(f) + \epsilon I$ when we compute the Newton step, i.e. $\Delta x = (\text{Hess}(f) + \epsilon I)^{-1} \nabla f$. Then Δx is always a descent direction. When we get close to a local minimum, the Hessian is positive definite, so we get the same Newton step as in the convex case.

To find the analogous method for optimization over a Grassmann manifold we first consider the matrix A that maps tangent vectors to tangent vectors through

$$\text{vec}(V) \mapsto \text{vec}(\Pi_T \widehat{F_{YY}}(V) - V Y^T F_Y) \, .$$

The operator vec: $\mathbb{R}^{n \times p} \to \mathbb{R}^{np}$ takes the columns in a matrix and put them after each other into one column vector. From (9) we see that

$$(A)_{(l-1)n+m,(j-1)n+m} = \sum_k (\Pi_T)_{mk} (F_{YY})_{ij,kl} - I_{\{i=m\}} (Y^T F_Y)_{jl} \, .$$

This matrix represents the Hessian:

$$\text{Hess}(F)(V_1)(V_2) = \text{vec}(V_1)^T A \, \text{vec}(V_2)$$

and when we find the Newton step we solve

$$A \cdot \text{vec}(V_{nt}) = -\text{vec}(\nabla F) \, , \quad Y^T V_{nt} = 0 \, .$$

In order to get rid of the constraint $Y^T V_{nt} = 0$ we consider the matrix representing the Hessian of the projection of two arbitrary matrices.

If $V_1, V_2 \in \mathbb{R}^{n\times p}$ are two arbitrary $n \times p$ matrices, the Hessian of their projections onto the tangent space can be written as

$$
\begin{aligned}
&\mathrm{Hess}(F)(\Pi_T V_1, \Pi_T V_2) \\
&\quad = \sum_{ij,kl} (F_{YY})_{ij,kl} (\Pi_T V_1)_{ij} (\Pi_T V_1)_{kl} - \sum_{ij} (\sum_{l} (\Pi_T V_2)_{il} (Y^T F_Y)_{lj}) (\Pi_T V_1)_{ij} \\
&\quad = \sum_{ij,kl} (F_{YY})_{ij,kl} (\sum_{m} (\Pi_T)_{im} (V_1)_{mj}) (\sum_{n} (\Pi_T)_{kn} (V_2)_{nl}) \\
&\qquad\quad - \sum_{l} ((\sum_{n} (\Pi_T)_{in} (V_2)_{nl}) (Y^T F_Y)_{lj}) (\sum_{m} (\Pi_T)_{im} (V_2)_{mj}) \\
&\quad = \sum_{mj,nl} \left(\sum_{ij} (F_Y Y)_{ij,kl} (\Pi_T)_{im} (\Pi_T)_{kn} - \sum_{i} (\Pi_T)_{in} (Y^T F_Y)_{lj} (\Pi_T)_{im} \right) (V_1)_{mj} (V_2)_{nl} \\
&\quad = \mathrm{vec}(V_1)^T \tilde{A}\, \mathrm{vec}(V_2)
\end{aligned}
$$

where

$$
\begin{aligned}
(\tilde{A})_{(j-1)n+m,(l-1)n+n} &= \sum_{ik} (F_{YY})_{ij,kl} (\Pi_T)_{im} (\Pi_T)_{kn} - \sum_{i} (Y^T F_Y)_{lj} (\Pi_T)_{in} (\Pi_T)_{im} \\
&= \sum_{ik} (\Pi_T)_{im} (\Pi_T)_{kn} ((F_{YY})_{ij,kl} - (Y^T F_Y)_{lj} I_{\{i=k\}}) \\
&= (\Pi_T^T (F_{YY})_{\cdot j, \cdot l} \Pi_T)_{mn} - (Y^T F_Y)_{lj} (\Pi_T^T \Pi_T)_{mn} \ . \qquad (10)
\end{aligned}
$$

If $Y^T F_Y$ is not symmetric, $\tilde{A}$ might not be symmetric, however

$$
\begin{aligned}
&\mathrm{Hess}(F)(\Pi_T V_1, \Pi_T V_2) = \mathrm{Hess}(F)(\Pi_T V_2, \Pi_T V_1) \\
&\qquad \Rightarrow \mathrm{vec}(V_1)^T \tilde{A}\, \mathrm{vec}(V_2) = \mathrm{vec}(V_2)^T \tilde{A}\, \mathrm{vec}(V_1) = \mathrm{vec}(V_1)^T \tilde{A}^T\, \mathrm{vec}(V_2)
\end{aligned}
$$

so we can replace $\tilde{A}$ by $\bar{A} = (\tilde{A} + \tilde{A}^T)/2$. Now if $\tilde{V}_{nt}$ satisfies

$$\bar{A} \cdot \mathrm{vec}(\tilde{V}_{nt}) = -\mathrm{vec}(\nabla F) \ ,$$

$V_{nt} = \Pi_T \tilde{V}_{nt}$ satisfies

$$
\begin{aligned}
\mathrm{Hess}(F)(V_{nt}, V) = \mathrm{Hess}(F)(\Pi_T \tilde{V}_{nt}, \Pi_T V) &= \mathrm{vec}(V)^T \bar{A}\, \mathrm{vec}(\tilde{V}_{nt}) \\
&= -\mathrm{vec}(V)^T \mathrm{vec}(\nabla F) = g(-\nabla F, V)
\end{aligned}
$$

for all tangent vectors V. Furthermore, if ϵ is chosen such that $\bar{A} + \epsilon I$ is positive definite and

$$\mathrm{vec}(V_{nt}) = -(\bar{A} + \epsilon I)^{-1} \mathrm{vec}(\nabla F) \ ,$$

then V_{nt} is in the tangent space (since $\bar{A}\,\mathrm{vec}(V_{nt})$ and ∇F are) and it is a descent direction since

$$\begin{aligned} F_{V_{nt}} &= g(\nabla F, V_{nt}) = \mathrm{tr}(\nabla F^T, V_{nt}) = \mathrm{vec}(\nabla F)^T \mathrm{vec}(V_{nt}) \\ &= -\mathrm{vec}(\nabla F)^T (\bar{A} + \epsilon I)^{-1} \mathrm{vec}(\nabla F) < 0 \; . \end{aligned}$$

In the function `ntStep` in `grassopt`, the Newton step V_{nt} is determined the following way:

1. $\tilde{A}$ is computed from (10). $\bar{A} = (\tilde{A} + \tilde{A}^T)/2$.
2. Let λ_{min} be the smallest eigenvalue of $\bar{A}$ ($\lambda_{min} \leq 0$) and let $\epsilon = \lambda_{tol} - \lambda_{min}$.
3. $\mathrm{vec}(\tilde{V}_{nt}) = -(\bar{A} + \epsilon I)^{-1} \mathrm{vec}(\nabla F)$
4. $V_{nt} = \Pi_T \tilde{V}_{nt}$.

Example

Below is an example application, minimization of the function $F(Y) = \mathrm{tr}(Y^T A Y)/2$, where A is a symmetric matrix. For this function $F_Y = AY$ and $(F_{YY})_{ij,kl} = 0$ if $j \neq l$ and $(F_{YY})_{ij,kl} = A_{ik}$ otherwise. The minimum of this function is the sum of the p smallest eigenvalues. As can be seen in figure 1, after a certain threshold we get cubic convergence.

References

[1] A. Edelman, T. A. Arias, and S. T. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM Journal on Matrix Analysis and Applications*, vol. 20, no. 2, pp. 303–353, 1998.

[2] S. Gudmundsson, "An introduction to Riemannian geometry," 2012, http://www.matematik.lu.se/matematiklu/personal/sigma/.

[3] B. O'Neill, "The fundamental equations of a submersion." *The Michigan Mathematical Journal*, vol. 13, no. 4, pp. 459–469, 1966.

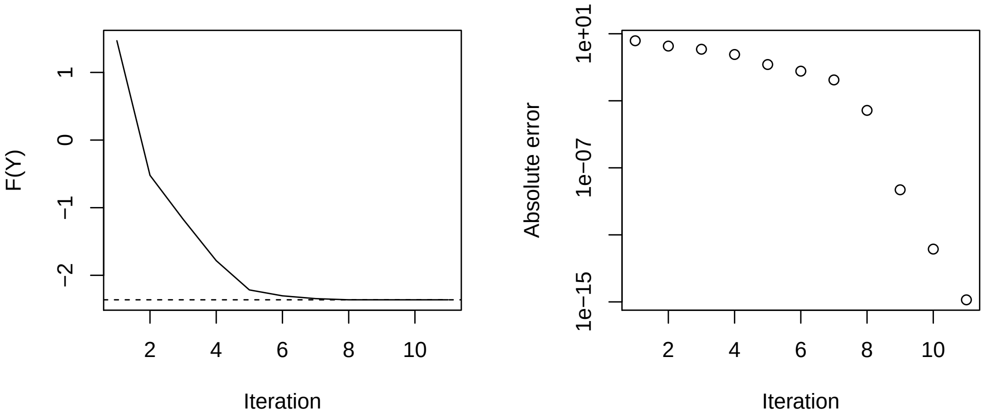

Figure 1: Example of optimization using `grassopt`. $F(Y) = \mathrm{tr}(Y^T AY)/2$, where A is a symmetric matrix. In this instance $n = 20$ and $p = 4$.