Constrained Gradient Descent and Line Search for Solving Optimization Problem with Elliptic Constraints

Ali A. Hasan[†] and Mohammed A. Hasan[‡]
[†]College of Electronic Engineering, Bani Waleed, Libya
[‡] Department of Electrical & Computer Engineering, University of Minnesota Duluth
E.mail:mhasan@d.umn.edu

Abstract

Finding global minima and maxima of constrained optimization problems is an important task in engineering applications and scientific computation. In this paper, the necessary conditions of optimality will be solved sequentially using a combination of gradient descent and exact or approximate line search. The optimality conditions are enforced at each step while optimizing along the direction of the gradient of the Lagrangian of the problem. Among many applications, this paper proposes learning algorithms which extract adaptively reduced rank canonical variates and correlations, reduced rank Wiener filter, and principal and minor components within similar framework.

1. Introduction

The problem of minimizing functionals over a sphere or an ellipse arises in many applications in engineering and applied sciences. Some of these applications include solving linear and nonlinear eigenvalue problems, major and minor component estimation [1], and canonical correlation analysis [2]. Even the SVD and EVD in matrix computation can be formulated as optimization problems over a sphere [3].

In the signal processing field, there are numerous problems that can be formulated as optimization problems over orthogonal constraints. These problems include

1. Minimum Subspace Computation
2. Minor and principal Subspace Tracking
3. Adaptive Subspace Computation
4. Computing the first r dominant eigenpairs
5. Canonical correlation analysis
6. Reduced Rank Wiener Filtering

We present in this paper new methods of computing and solving optimization problems using constrained gradient descent of the Lagrangian in conjunction with exact and approximate line search. Thus these approaches may be considered as constrained iterative gradient descent methods.

2. Problem Formulation

The idea of using an approximate or exact line search for solving constrained optimization problem with orthogonal constraints will be applied to several practical problems. Consider the following minimization problem

$$\text{Minimize} \quad F(x) \;\text{ subject to } x^Tx = I_r, \tag{1}$$

where F is at least twice continuously differentiable real valued function, $x \in \mathbb{R}^{m\times r}$, and I_r stands for the identity matrix of size r. Define the Lagrangian as

$$\mathcal{L}(x,\lambda) = F(x) - trace\{(x^Tx - I_r)\frac{\lambda}{2}\}, \tag{2}$$

where λ is a matrix of Lagrange multipliers. The necessary condition for optimality is that $\nabla\mathcal{L} = 0$, where

$$\nabla\mathcal{L} = \left\{ \begin{array}{c} \nabla_x F(x) - x\lambda \\ x^Tx - I_r \end{array} \right\}. \tag{3a}$$

If (x,λ) is an optimal solution, then λ may be expressed as

$$\lambda = x^T\nabla_x F(x). \tag{3b}$$

Substituting this expression in (3a) yields

$$\begin{aligned} \nabla_x\mathcal{L} &= \nabla_x F(x) - xx^T\nabla_x F(x) \\ &= (I_r - xx^T)\nabla_x F(x). \end{aligned} \tag{4}$$

Conversely, if we assume that $\nabla\mathcal{L} = 0$, then $x^T\nabla\mathcal{L} = (x^Tx - I_r)x^T\nabla F(x)$. Thus if λ is non-singular (which is the case in many applications listed in Section 3), then $(I_r - x^Tx)\lambda = 0$, and hence $x^Tx = I_r$.

Now assume that an approximate solution matrix x is given and assume that λ has been computed as in (3b). For a given nonzero direction matrix h, we are interested in computing $\alpha \in \mathbb{R}^{r\times r}$ so that $\mathcal{L}(x + h\alpha)$ is minimum. Clearly, the Taylor expansion of $\mathcal{L}(x + h\alpha)$ around x is given by

$$\begin{aligned} \mathcal{L}(x + h\alpha) = F(x) + D_xF(x)h\alpha + \frac{1}{2}\alpha^Th^T\nabla_x^2F(x)h\alpha + h.o.t. \\ - trace\{x^Tx\frac{\lambda}{2} + \alpha^Th^Tx\frac{\lambda}{2} + x^Th\alpha\frac{\lambda}{2} + \alpha^Th^Th\alpha\frac{\lambda}{2}\}, \end{aligned} \tag{5a}$$

where h.o.t. stands for cubic and higher order terms. By ignoring h.o.t., it follows that

$$(\frac{\partial\mathcal{L}(x + h\alpha)}{\partial\alpha})^T = D_xF(x)h + h^T\nabla_x^2F(x)h\alpha - h^Tx\alpha - h^Th\alpha\lambda. \tag{5b}$$

When F is quadratic, the Kronecker product may be used to obtain exact solution to the system; $\frac{\partial\mathcal{L}(x+h\alpha)}{\partial\alpha} = 0$. Specifically,

$$((I_r\otimes h^TD_x^2F(x)h) + (\lambda^T\otimes h^Th))\text{vec}(\alpha) = \text{vec}(h^Tx\lambda - h^T\nabla_xF(x)). \tag{6}$$

Here vec stands for the operation of stacking the columns of a matrix into one column, and $\otimes$ denotes the Kronecker product.

If h is chosen as $h = \nabla_x \mathcal{L}(x, \lambda)$, then $x^T h = 0$. Therefore, if $r = 1$, i.e., x is a vector, then the scalar α can be obtained as

$$\alpha = -(h^T \nabla_x^2 \mathcal{L} h)^{-1} h^T h, \qquad (7a)$$

where $\nabla_x^2 \mathcal{L} = \nabla_x^2 F(x) - \lambda I_r$, and x can be updated as

$$x' = x + h\alpha. \qquad (7b)$$

If α is chosen to be fixed at each stage, then the above procedure reduces to the constrained gradient descent.

3. Examples

In this section we present a few signal processing applications where the proposed methods can be utilized.

Example 1: Major Subspace Computation

The above ideas of line search minimization can be used for principal component analysis. The first principal component optimizes the reconstruction mean-square-error by choosing the best one dimensional subspace to project the signal into. The criterion for the first component w is that it minimizes

$$\begin{aligned} J_1(w) &= E(||x - ww^T x||) \\ &= E\{||x||^2\} - 2w^T R_{xx} w + (w^T R_{xx} w) w^T w, \end{aligned} \qquad (8a)$$

where $||.||$ is the Euclidean norm. The solution to this minimization problem is $w = q_1$, where q_1 is the eigenvector corresponding to the largest eigenvalue of the auto-covariance matrix R_{xx}. For the rth component criterion, one may solve the minimization problem

$$\text{Minimize } J_r(w) = E(||x - ww^T(I - W_r^T W_r)x||^2), \qquad (8b)$$

where $W_r = [\, w_1 \quad \cdots \quad w_r \,]^T$. Thus

$$\begin{aligned} J_r &= E(||x||^2 - 2w^T(I - W_r^T W_r) R_{xx} w + \\ &\quad (w^T(I - W_r W_r^T) R_{xx} (I - W_r W_r^T)) w) \\ &= E(||x||^2) - 2w^T P_r R_{xx} w + (w^T P_r R_{xx} P_r w)||w||^2, \end{aligned} \qquad (9)$$

where $P_r = I - W_r^T W_r$. To minimize J_1 using the gradient method, we have

$$\nabla J_1 = -4R_{xx} w + 2w^T w R_{xx} w + 2(w^T R_{xx} w) w. \qquad (10)$$

If $\nabla J_1 = 0$, it can be verified that $w^T w = 1$. Now let us set $h = \nabla J_1$ and determine α so that $J_1(w + \alpha h)$ is minimum. It can be shown that α is a solution for the cubic equation

$$0 = f(\alpha) = a_0 \alpha^3 + a_1 \alpha^2 + a_2 \alpha + a_3, \qquad (11)$$

where

$$\begin{aligned} a_0 &= 4(h^T h)(h^T R_{xx} h), \\ a_1 &= 2(h^T h)(h^T R_{xx} w + w^T R_{xx} h) + 3(h^T w + w^T h)(h^T R_{xx} h), \\ a_2 &= 2(h^T w + w^T h)(w^T R_{xx} h + h^T R_{xx} w) + 2(h^T h)(w^T R_{xx} w) \\ &\quad + 2(w^T w - 2)(h^T R_{xx} h), \\ a_3 &= (h^T w + w^T h)(w^T R_{xx} w) + (w^T w - 2) \\ &\quad \times (h^T R_{xx} w + w^T R_{xx} h). \end{aligned} \qquad (12a)$$

If h is chosen to be orthogonal to w, then $h^T w = w^T h = 0$ and hence (12a) simplifies to

$$\begin{aligned} a_1 &= 2(h^T h)(w^T R_{xx} h), \\ a_2 &= 2(h^T h)(w^T R_{xx} w) + 2(w^T w - 2)(h^T R_{xx} h), \\ a_3 &= (w^T w - 2)(h^T R_{xx} w + w^T R_{xx} h). \end{aligned} \qquad (12b)$$

The polynomial $f(\alpha)$ has at least one real zero. If all zeros are real, one may choose the step size α to be $\max\{\alpha_i\}_{i=1}^3$. Theoretically w converges to a unit vector, however to speed up convergence, one should normalize w to a unit vector in each step. Similarly, J_r may be minimized by replacing R_{xx} with $\bar{R}_{xx} = P_r R_{xx} P_r$.

Algorithm 1 (Constrained Gradient Descent-Line-Search) (CGD-LS)

The inputs for this algorithm are $R_{xx} \in \mathbb{R}^{m \times m}$, and $W_r \in \mathbb{R}^{m \times r}$, where W_r is a matrix whose columns are the first r principal components. The output is the $(r+1)$th principal component w_{r+1}.

1. Choose a nonzero random initial vector w. Normalize w so that $w^T w = 1$.
2. Let $h = \nabla J_r$ as in (10) with R_{xx} being replaced with $\bar{R}_{xx} = P_r R_{xx} P_r$, where $P_r = I - W_r W_r^T$.
3. Compute a_0, a_1, a_2 and a_3 as in (12a) and solve (11) for α_i. Set $\alpha = \max\{\alpha_i\}_{i=1}^3$.
4. Update w using $w' = w + h\alpha$.
5. Orthogonalize w' with respect to W_r using the formula $w'' = (I - W_r W_r^T) w'$.
6. Normalize w'', i.e., set $w = \frac{w''}{\sqrt{w''^T w''}}$.
7. Repeat Steps 2-5 until convergence.

This algorithm can be slightly modified for adaptive computation of principal subspace. This can be established by updating the auto-covariance R_{xx} so that $\hat{R}_{xx} = (1 - \beta) R_{xx} + \beta x x^T$ for some number $0 < \beta \leq 1$. Here x is the new observation vector.

Example 2: Canonical Correlation Analysis

The main idea in two-set canonical correlation analysis (CCA) is to investigate the relationship between two sets of variables. It finds corresponding sets of linear combinations of the original two groups of variables. CCA is first developed in [4]. As indicated in [5], canonical correlations and variates can be found by solving the maximization problem:

$$\begin{aligned} &\text{Maximize } w^T R_{xy} v \\ &\text{subject to } w^T R_{xx} w = 1, \ v^T R_{yy} v = 1, \end{aligned} \qquad (13)$$

where $R_{xx} = E(xx^T)$, $R_{xy} = E(xy^T) = R_{yx}^T$, $R_{yy} = E(yy^T)$, $\{.\}^T$ denotes matrix transpose, and $E\{.\}$ denotes expectation. The proposed method of Section 2 can be applied to serially compute the canonical correlations and variates. Specifically, let W_r and U_r be the first r canonical coordinates so that

$$U_r^T R_{yx} W_r = K_r, \ W_r^T R_{xx} W_r = I_r, \ U_r^T R_{yy} U_r = I_r, \qquad (14)$$

where K_r is a diagonal matrix; $K_r = diag(\kappa_1, \kappa_2, \cdots, \kappa_r)$, where $\kappa_1 \geq \kappa_2 \geq \cdots \geq \kappa_r$. Here κ_i is the ith canonical correlation. To determine κ_{r+1}, w_{r+1}, and v_{r+1} one can solve the optimization problem

$$\begin{aligned}&\text{maximize} \quad w^T R_{xy} u\\ &\text{subject to}\\ &W_r^T R_{xy} u = 0,\ w^T R_{xy} U_r = 0,\\ &w^T R_{xx} W_r = 0,\ u^T R_{yy} U_r = 0,\\ &w^T R_{xx} w = 1,\ u^T R_{yy} u = 1.\end{aligned} \tag{15}$$

Let $P_1 = I - W_r W_r^T R_{xx}$ and $P_2 = I - U_r U_r^T R_{yy}$ be projections so that for any two vectors (w, v) satisfying the constraints in (15), $P_1 w = w$, $P_2 u = u$, $P_1 W_r = 0$ and $P_2 U_r = 0$. Consequently, (15) may be expressed as

$$\begin{aligned}&\text{Maximize } w^T P_1^T R_{xy} P_2 v\\ &\text{subject to } w^T P_1^T R_{xx} P_1 w = 1,\ v^T P_2^T R_{yy} P_2 v = 1,\end{aligned} \tag{16}$$

Then the solutions must be orthogonalized with respect to the diagonal matrix K_r.

Example 3: Quadratics Over a Sphere

Minimization of a quadratic over a sphere arises in many algorithmic developments. For example, in the trust region algorithm [6]-[7] it is required to solve subproblems involving the minimization of the following quadratic over a sphere:

$$\begin{aligned}&\text{Minimize} \quad F(x) = trace\{x^T A x - 2b^T x\},\\ &\text{subject to } x^T x = I_r,\end{aligned} \tag{17}$$

where A is a symmetric $m \times m$ matrix, $b \in \mathbb{R}^m$, and T denotes matrix transpose. Problems of this form also arise in many other applications including regularization methods for ill-posed problems [8]. Several approaches have been developed to solve (17). In some of these approaches partial or complete diagonalization of A is used, however, this representation is practical only when m is small. The focus here is on large scale case. The Lagrangian of this problem is defined as

$$\mathcal{L}(x, \lambda) = \frac{1}{2} trace\{x^T A x - 2b^T x\} - trace\{(x^T x - I_r)\frac{\lambda}{2}\}, \tag{18}$$

where λ is the Lagrange multipliers matrix. Clearly,

$$\nabla_x \mathcal{L} = (Ax - b) - x\lambda. \tag{19a}$$

A necessary condition of optimality is that $\nabla_x \mathcal{L} = 0$ and therefore

$$\lambda = x^T A x - x^T b = F(x) + b^T x. \tag{19b}$$

Now let $h = Ax - b - x\lambda$ be a direction of descent matrix, then

$$\begin{aligned}\mathcal{L}(x + h\alpha) &= \frac{1}{2} trace\{(x + h\alpha)^T A (x + h\alpha) - 2b^T (x + h\alpha)\}\\ &- trace((x^T + \alpha^T h^T)(x + h\alpha) - I_r)\frac{\lambda}{2}.\end{aligned} \tag{20a}$$

Hence if α minimizes $\mathcal{L}(x + h\alpha)$, then it is a solution of the equation:

$$2h^T A x + 2h^T A h\alpha - 2(b^T h)^T - h^T x\lambda - h^T h\alpha\lambda = 0. \tag{20b}$$

Note that our choice of h and λ implies that $x^T h = 0$. Therefore, (20b) simplifies to

$$h^T h\alpha\lambda - h^T A h\alpha = b^T h - h^T A x. \tag{21}$$

This equation can be solved exactly as in (6) using the Kronecker product. If $r = 1$, an exact solution for α is

$$\alpha = \frac{b^T h - x^T A h}{h^T (A - \lambda I_m) h}. \tag{22}$$

An alternative approach for computing x for the case $r = 1$ is to consider $x = (A - \lambda I_m)^{-1} b$. Consequently,

$$1 = x^T x = b^T (A - \lambda I_m)^{-2} b. \tag{23}$$

The Newton-Gauss method can be applied so that

$$\lambda_{k+1} = \lambda_k - \frac{b^T (A - \lambda I_m)^{-2} b - 1}{2b^T (A - \lambda I_m)^{-3} b}.$$

The starting point λ_0 should be chosen appropriately so that λ_k converges to the smallest solution of (23). Then the optimal solution x is determined as $x = (A - \lambda I_m)^{-1} b$.

As a special case, when $r = 1$ and $b = 0$, (17) transcribes to the problem

$$\text{Minimize} \quad x^T A x \text{ subject to } ||x|| = 1. \tag{24}$$

It is well-known that a solution of this problem is any eigenvector associated with the smallest eigenvalue of A. Similarly, computing the r most sub-dominant eigenpairs involves the solution of (24) r times, where x in each case is restricted to the space orthogonal to the previous eigenvectors.

Example 4: Reduced Rank Wiener Filtering

The reduced rank Wiener filtering problem is to find the rank r minimizer W_r that minimizes

$$Q_{xx}(W_r) = E(||x(n) - W_r y(n)||^2) \quad W_r \text{ has rank r.} \tag{25a}$$

As in [9]and [10], this can be rewritten as

$$Q_{xx}(W_r) = Q_{xx}(W) + (R_{xy} - W_r R_{yy}) R_{yy}^{-1} (R_{xy} - W_r R_{yy})^T, \tag{25b}$$

where $Q_{xx} = R_{xx} - R_{xy} R_{yy}^{-1} R_{yx}$. Here $R_{xx} = E(xx^T) \in \mathbb{R}^{m \times m}$, $R_{xy} = E(xy^T) \in \mathbb{R}^{m \times n}$, and $R_{yy} = E(yy^T) \in \mathbb{R}^{n \times n}$.

To compute a reduced rank Wiener filter W_r, of rank r, assume that $W_r = U\Sigma V^T$ where $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{n \times r}$ are orthogonal and $\Sigma \in \mathbb{R}^{r \times r}$ is diagonal positive definite matrix. Therefore, the matrices U, Σ, and V are solutions of the minimization problem:

$$\begin{aligned}\text{Minimize } &\frac{1}{2} trace\{Q_{xx}(W) + (R_{xy} - U\Sigma V^H R_{yy}) R_{yy}^{-1}\\ &\times (R_{xy} - U\Sigma V^H R_{yy})^T\},\end{aligned} \tag{26}$$

subject to the constraints $U^T U = I$, $V^T V = I$ and Σ is positive definite diagonal matrix. To solve this problem, one can apply the techniques of Section 2 with some modification.

Example 5: Generalized Minimum Subspace Computation

There are many situations where it is required to obtain a generalized signal subspace or a small number of lowest or largest eigenvalues. One can repeatedly minimize a functional $F(x)$ restricting x to the space orthogonal to the previous subspace. In some cases, especially for multiple or clustered eigenvalues, it is advisable to compute the whole invariant subspace spanned by the corresponding eigenvectors. In this section we present an algorithm to compute the r smallest eigenvalues (with eigenvectors) simultaneously by considering the corresponding subspace as a whole. An r-dimensional subspace is spanned by r non-degenerate (column) vectors, which are combined into a rectangular matrix x. The general form of the generalized minimum subspace can be expressed as a minimization problem:

$$\text{Minimize}_x \quad Trace\{x^TAx\}, \quad x^TBx = I_r, \qquad (27a)$$

where A is symmetric and B is positive definite of size m. This problem can be shown to be equivalent to maximizing $trace\{(x^TAx)(x^TBx)^{-1}\}$ over all non zero vectors x.

Let $\mathcal{L}(x,\lambda) = \frac{1}{2}trace\{x^TAx\} - trace\{(x^TBx - I_r)\frac{\lambda}{2}\}$ be the Lagrangian, then a necessary condition for optimality is that

$$\nabla_x\mathcal{L}(x,\lambda) = Ax - Bx\lambda = 0.$$

A steepest descent method would search a new minimum along h, say $x' = x + h\alpha$ with a (small) coefficient matrix $\alpha \in \mathbb{R}^{r\times r}$. The matrix α is required to minimize $\mathcal{L}(x + h\alpha, \lambda)$ with respect to $r \times r$ matrix α, where $\lambda = (x^TAx)(x^TBx)^{-1}$. The exact solution can be computed from the equation

$$h^TAx + h^TAh\alpha - h^TBx\lambda - h^TBh\alpha\lambda = 0, \qquad (27b)$$

by using similar approach given in (6). The matrix α can also be obtained approximately as

$$\alpha \approx -\frac{1}{2}(h^TAh)^{-1}(x^TAh + h^TAx), \qquad (27c)$$

where $h = \frac{1}{2}\nabla F(x)$ or $h = Ax - Bx(x^TBx)^{-1}x^TAx$. The old solution will be updated so that $x_1 = x + h\alpha$. This solution must then be normalized to a new matrix y so that $y^TBy = I$. One approach is to use

$$y = x_1(x_1^TBx_1)^{\frac{-1}{2}}.$$

Note that the matrices $(x_1^TBx_1)^{\frac{-1}{2}}$ and h^TAh are of order $r \times r$. Computing the inverse or the positive definite square root of such matrices is a small problem and can be solved by standard techniques.

Algorithm 2

Let $x(1)$ be any nonzero randomly generated vector. For $k = 1, 2, \cdots$ until convergence do

$$\begin{aligned}
\bar{x}(k) &= x(k)(x(k)^TBx(k))^{-\frac{1}{2}} \\
\lambda(k) &= \bar{x}(k)^TA\bar{x}(k) \\
h(k) &= A\bar{x}(k) - B\bar{x}(k)\lambda(k) \\
a(k) &= -(h(k)^TAh(k))(h(k)^TA\bar{x}(k)) \\
x(k+1) &= \bar{x}(k) + h(k)a(k)
\end{aligned}$$

4. Conclusion

In this paper we proposed a number of computational tools for solving optimization problems over spheres or ellipses. These include, among many other problems, the reduced rank canonical variates and correlations, reduced rank Wiener filters, reduced rank principal and minor component analysis. The main motivation of this work is the desire to solve linear systems of equations arising from the necessary conditions of optimality of these problems without inverting large scale matrices. Simulations have been conducted to examine the performance of each of the proposed methods, however, we did not include them here due to space limitation. Finally, the proposed approaches can be extended to complex functions after some modifications.

References

[1] S. Y. Kung, K. I. Diamantaras, J. S. Taur, "Adaptive Principal component EXtraction (APEX) and applications," Signal Processing, IEEE Transactions on, Volume: 42 Issue: 5, May 1994, Page(s): 1202-1217.

[2] T. W. Anderson, An Introduction to Multivariate Statistical Analysis, 2nd ed. New York: Wiley, 1984.

[3] G. H. Golub and C. F. Van Loan, Matrix Computations, 2nd ed. Baltimore, MD: Johns Hopkins Univ. Press, 1989.

[4] H. Hotelling, "Relations between two sets of variates," Biometrika, Vol. 28, pp. 321-377, 1936.

[5] S. V. Schell, Gardner, W.A., "Programmable canonical correlation analysis: a flexible framework for blind adaptive spatial filtering," Signal Processing, IEEE Transactions on, Volume: 43 Issue: 12, Dec. 1995, Page(s): 2898-2908.

[6] R. H. Byrd, R. B. Schnabel, and G. A. Schultz, "A trust region algorithm for nonlinearly constrained optimization," SIAM J. Numer. Anal., 24 (1987), pp. 1152-1170.

[7] J. J. More, Recent developments in algorithms and software for trust region methods, in Mathematical Programming: State of the Art, A. Bachem, M. Grotschel, and B. Korte, eds., Springer-Verlag, Berlin, 1983, pp. 258-287.

[8] A. Tarantola, Inverse Problem Theory, Elsevier, Amsterdam, The Netherlands, 1987.

[9] L. L. Scharf, Statistical Signal Processing, Detection, Estimation, and Time Series Analysis, Addison-Wesley, 1991.

[10] Mohammed A. Hasan, L. Scharf, and M.R. Azimi-Sadjadi, and Ali Pezeshki, "Fast Algorithms for Computing Full and Reduced Rank Wiener Filters," the Proceeding of ISCAS 2003, Bangkok, Thailand.