1 Lagrange multipliers

For a matrix A, recall the definition of matrix two norm:

$$\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2$$

This is indeed a constraint optimization problem. Consider A = $\begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix}$

We want to maximize a function f on $\mathbb{R}^2$ subject to a constraint $g(x_1, x_2) = 0$ In our matrix norm example,

$$\begin{aligned} f(x_1, x_2) &= \|Ax\|_2^2 \\ &= (x_1 + 2x_2)^2 + x_1^2 \\ &= 2x_1^2 + 4x_1x_2 + 4x_2^2 \end{aligned}$$

and

$$g(x_1, x_2) = \|x\|_2^2 - 1 = x_1^2 + x_2^2 - 1$$

The ellipses are the level curves of f, i.e. the curves $f(x_1, x_2) =$ constant. The gradient vector field of f,

$$\nabla f(x_1, x_2) = \frac{\delta f(x_1, x_2)}{\delta x_1}\hat{i} + \frac{\delta f(x_1, x_2)}{\delta x_2}\hat{j}$$

indicates the direction and rate of the fastest increase of f at the point (x_1, x_2). It is always perpendicular to the level curves of f. The thick circle in the figure is the curve $T : g(x_1, x_2) = 0$, the constraint. We are able to find the maximum of f on that curve. Look at the point $P \in T$ and at the gradient vector ∇f at point P. This vector is not perpendicular to T and therefore has a non-zero component tangent to T. It follows that f increases along T, in the direction of $\nabla f(B)_{tangent}$, and that therefore f(P) is not a maximum of f on T. For the same reason , f(P) is not a minimum.
A maximum (or minimum) of f on T can only occur at points where ∇f is perpendicular to T and therefore has zero tangential component. Since ∇_g is also always perpendicular to the curve T, we are looking for points (x_1, x_2) at which ∇_f and $\nabla_g$are parallel, meaning that

$$\nabla_f(x_1, x_2) = \lambda \nabla_g(x_1, x_2) \tag{1}$$

Those points of course must lie on T

$$g(x_1, x_2) = 0 \tag{2}$$

Equation (1) and (2) form a system of 3 (usually nonlinear) equations for the three unknowns x_1, x_2, λ. In the figure the four points M, M', m, m' satisfy all 3 conditions. The maximum of f is assumed at M and M', and the minimum of m and m'. Function of more variables, say $f(x_1, \ldots, x_n)$ subject to multiple constrains (say $g_1 = g_2 = \ldots = g_k = 0$) are treated in a similar manner.

$$\nabla f(x_1, ..., x_n) = \sum_{j=1}^{k} \lambda_j \nabla_{g_j}(x_1, ..., x_n) \tag{3}$$

$$\begin{aligned} g_1(x_1, ..., x_n) &= 0 \\ &.... \\ g_k(x_1, ..., x_n) &= 0 \end{aligned}$$

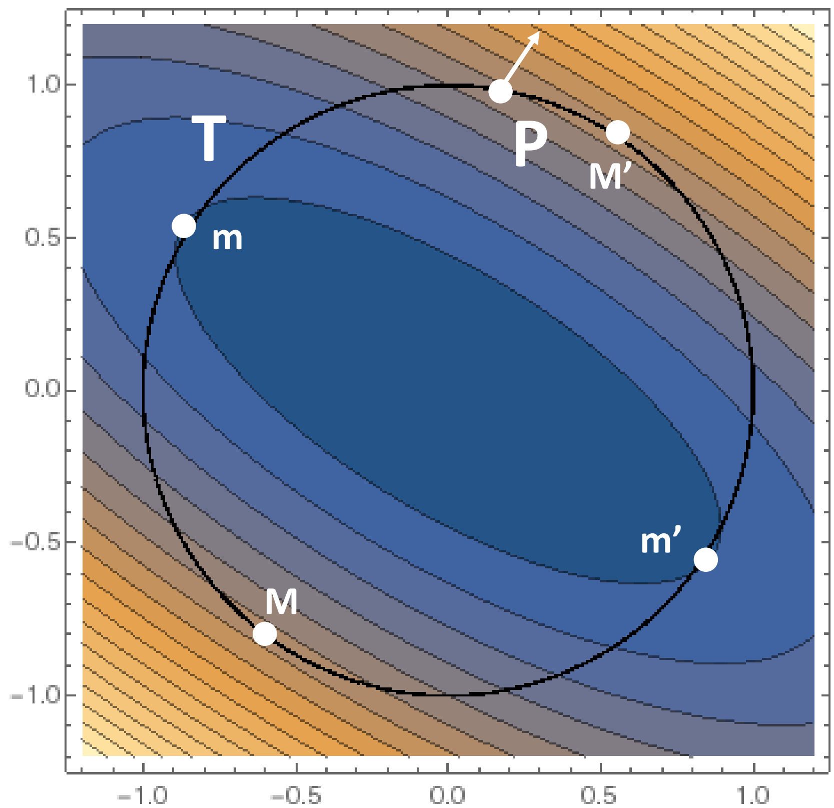

Figure 1: Level set for $f(x_1, x_2)$

and test the points found to be the maxima and minima. As in single-variable calculus, there are "inflection points" where $\nabla f = 0$ and $\lambda = 0$ or exceptional points where $\nabla g = 0$ and λ might have to be ∞ but those will play no role in our example. The proportionality constants $\lambda_1, ..., \lambda_n$ are called Lagrange multipliers. The solutions of (3) are called critical points of the constrained function f. In general optimization problem, this will be reformulated as (first order) **Karush-Kuhn-Tucker condition** or **KKT** condition in short.

Karush-Kuhn-Tucker(KKT) condition

Recall that the general convex optimization problem can be formulated as:

$$\begin{aligned} \min \quad & f_0(x) \\ s.t. \quad & f_i(x) \leq 0, \ \forall i = 1, \ldots, m \\ & h_j(x) = 0, \ \forall j = 1, \ldots, p \end{aligned} \tag{4}$$

The *Lagrangian* of the constrained optimization problem (4) is denoted as:

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{l} \mu_i h_i(x) \tag{5}$$

Now assume f_0, f_i, h_i are differentiable. We say that x satisfies the (first order) Karush-Kuhn-Tucker(KKT) condition if:

$$\begin{aligned} \nabla_x L(x, \lambda, \mu) &= 0 \\ g_i(x) &\leq 0, i = 1, \ldots, m \\ h_i(x) &= 0, i = 1, \ldots, l \\ \lambda_i g_i(x) &= 0, i = 1, \ldots, m \\ \lambda_i &\geq 0, i = 1, \ldots, m \end{aligned}$$

More detailed discussion will show this is a necessary condition for a convex function to reach the optimum, i.e. if $x^\star$ is an optimal solution of the minimization problem, then $x^\star$ must satisfy the KKT conditions above. Consider the following system

$$2x_1 + 2x_2 = \lambda x_1 \tag{6}$$
$$2x_1 + 4x_2 = \lambda x_2 \tag{7}$$

together with the constraint

$$x_1^2 + x_2^2 = 1 \tag{8}$$

Now the equations (6) and (7) are an eigenvalue problem with solutions

$$\lambda_+ = 3 + \sqrt{5} \tag{9}$$
$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = c \begin{bmatrix} 1 \\ (\lambda_+ - 2)/2 \end{bmatrix} \tag{10}$$

and

$$\lambda_+ = 3 - \sqrt{5} \tag{11}$$
$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = c \begin{bmatrix} 1 \\ (\lambda_- - 2)/2 \end{bmatrix} \tag{12}$$

Here notice that the equation (8) forces us to choose c in such a way that the eigenvector has length 1. For each λ there are 2 unit eigenvectors which are negative of each other.

We now have found four solutions x_1, x_2, λ of the Lagrange multiplier problem. Two give the minimum of $2x_1^2 + 4x_1x_2 + 4x_2^2$ on $x_1 + x_2 = 1$, the other 2 the maximum. The desired norm of A will be the square root of that maximum, since the norm is

$$\max \sqrt{2x_1^2 + 4x_1x_2 + 4x_2^2}$$

The key question is: what is the matrix $\begin{bmatrix} 2 & 2 \\ 2 & 4 \end{bmatrix}$ that appeared in the eigenvalue problem (6) and (7).

$$\|Ax\|_2^2 = (Ax.Ax) = (A^tAx.x) \tag{13}$$

Set B $= A^tA$. This is a symmetric matrix. If B $= \begin{bmatrix} a & b \\ b & c \end{bmatrix}$ then

$$Bx.x = ax_1^2 + 2bx_1x_2 + cx_2^2 \tag{14}$$

Lagrange's method applied to (14) with constraint $x_1^2 + x_2^2 = 1$, yields the two equations

$$ax_1 + bx_2 = \lambda x_1 \tag{15}$$

$$bx_1 + cx_2 = \lambda x_2 \tag{16}$$

This is precisely the eigenvalue problem for the symmetric matrix B.

$$B = A^tA = \begin{bmatrix} 1 & 1 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 4 \end{bmatrix}$$

The Lagrange mulitpliers, alias the eigenvalues of (6) and (7) are precisely the maximum and minimum of $\|Ax\|_2^2$ on $\|x\|_2^2 = 1$, obviously the larger one $3 + \sqrt{5}$ is the maximum. The maximum of $\|Ax\|_2$ is then the square root $\|A\|_2 = \sqrt{3 + \sqrt{5}}$

2 Solving System of Linear Algebraic Equations

2.1 Under-constrained System

The SVD of a matrix A also yields valuable geometric information about solution of a system of linear (algebraic) equations.
Consider :

$$A\mathbf{x} = \mathbf{y}$$
$$U\Sigma V^T\mathbf{x} = \mathbf{y}$$
$$\Sigma V^T\mathbf{x} = U^T\mathbf{y}$$
$$\widetilde{\mathbf{x}} = \Sigma^{-1}\widetilde{\mathbf{y}}$$
$$\widetilde{\mathbf{x}}_i = \widetilde{\mathbf{y}}_i/\sigma_i$$

Example.

$$\begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}\mathbf{x} = \mathbf{y}$$

Express the 2×2 *matrix in terms of its SVD,* $U\Sigma V^T$

$$\begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\mathbf{x} = \mathbf{y}$$

Invert U and combine Σ and V^T

$$\begin{bmatrix}1 & 1\\0 & 0\end{bmatrix}\mathbf{x} = \begin{bmatrix}1\\5\end{bmatrix}\begin{bmatrix}1 & 2\\-2 & 1\end{bmatrix}\mathbf{y}$$
$$\begin{bmatrix}x_1 + x_2\\0\end{bmatrix} = \frac{1}{5}\begin{bmatrix}y_1 + 2y_2\\-2y_1 + y_2\end{bmatrix}\mathbf{y}$$

Clearly, a solution only exists when $y_2 = 2y_1$, i.e. when y lives in the range of A.

The problem here is that A is rank deficient, i.e. not full-rank.
Consider the general case where A, is an $n \times n$, and has rank k. This means that A has $n-k$ singular values which are zero:

$$\Sigma = \begin{bmatrix}\sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_k & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0\end{bmatrix} = \begin{bmatrix}\Sigma_k & 0\\0 & 0\end{bmatrix}$$

If we try to solve the equations as in Equation 1, we hit a snag:

$$\begin{aligned}Ax &= y\\ U\Sigma V^T\mathbf{x} &= \mathbf{y}\\ \begin{bmatrix}\Sigma_k & 0\\0 & 0\end{bmatrix}\widetilde{\mathbf{x}} &= \widetilde{\mathbf{y}}\\ \widetilde{x}_i &= \widetilde{y}_i/\sigma_i \text{ , for } i \le M \text{ only}\end{aligned} \tag{17}$$

In the SVD-rotated space, we can only access the first k elements of the solution.
If there are more unknowns M than equations N, the problem is under-constrained or ill-posed :

$$Ax = y$$

where A is order $N \times M$, x is order $M \times 1$ and y is order $N \times 1$, where $N < M$
Using our previous results on SVD, we can rewrite the linear system as

$$\sum_{i=1}^{N} \sigma_i u_i ({v_i}^T x) = y \tag{18}$$

In other words, the only part of x that matters is the component that lies in the N-dimensional subspace of $\mathbb{R}^M$ spanned by the first N columns of V. Thus, the addition of any component that lies in the null-space of A will make no difference: if x^* is any solution to Equation 18, so is $x^* + \sum_{i=N+1}^{M} \alpha_i v_i$, for any α_i.

Example. *Consider*

$$\begin{bmatrix}1 & 1\end{bmatrix}\mathbf{x} = 4$$

The SVD of this 1×2 matrix shows:

$$\begin{bmatrix}1\end{bmatrix}\begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}1 & 1\\1 & -1\end{bmatrix}\mathbf{x} = 4$$

The null-space (i.e. the set of vector $\mathbf{x}$ such that $Ax = 0$) of A is $\alpha\begin{bmatrix}1\\-1\end{bmatrix}$. And we know $\mathbf{x} = \begin{bmatrix}2\\2\end{bmatrix}$ is one solution. Therefore , in this under-constrained scenario, the solution will be :

$$x = \begin{bmatrix}2\\2\end{bmatrix} + \alpha\begin{bmatrix}1\\-1\end{bmatrix}$$

2.1.1 Regularization

In general, under-constrained problems can be made well-posed by the addition of a regularization, i.e. a cost-function that we'd like the solution to minimize. In the case of under-constrained linear equations, we know that the solution space lies in an N dimensional subspace of $\mathbb{R}^M$. One obvious regularization would be to pick the solution that has the minimum square norm, i.e. the solution that is closest to the origin. The new, well-posed version of the problem can now be stated as follows:

$$\begin{aligned} &\min_{x \in \mathbb{R}^M} x^T x \\ &s.t. \quad Ax = y \end{aligned} \tag{19}$$

Proposition. *If $A \in \mathbb{R}^{N \times M}$, $N \leq M$, has full row-rank, then AA^T is invertible.*

Proof.

$$A = U \begin{bmatrix} \Sigma_N & 0 \end{bmatrix} V^T$$

where Σ_N is invertible. Therefore,

$$\begin{aligned} AA^T &= U \begin{bmatrix} \Sigma_N & 0 \end{bmatrix} V^T V \begin{bmatrix} \Sigma_N \\ 0 \end{bmatrix} U^T \\ &= U \begin{bmatrix} \Sigma_N & 0 \end{bmatrix} \begin{bmatrix} \Sigma_N \\ 0 \end{bmatrix} U^T \\ &= U {\Sigma_N}^2 U^T \end{aligned}$$

where ${\Sigma_N}^2$ is the $N \times N$ diagonal matrix with ${\sigma_i}^2$ on the ith diagonal. Since Σ_N has no zero elements on the diagonal, neither does Σ_N^2. Therefore AA^T is invertible (and symmetric, positive-definite). It is also worth noting here that since that each matrix in the last equation above is invertible, we can write down the SVD (and eigenvector decomposition) of $(AA^T)^{-1}$ by inspection: $(AA^T)^{-1} = U {\Sigma_N}^2 \; U^T$

□

We will now solve the problem stated in Equation 19 using Lagrange multipliers. We assume that A has full row-rank. Let

$$H = x^T x + \lambda^T (Ax - y) \tag{20}$$

The solution is found by solving the equation $\frac{\partial H}{\partial x} = 0$ and then ensuring that the constraint (Ax = y) holds. First solve for x:

$$\begin{aligned} \frac{\partial H}{\partial x} &= 0 \\ 2x^T + \lambda^T A &= 0 \\ x &= \frac{1}{2} A \lambda^T \end{aligned}$$

Now using the fact that AA^T is invertible, choose λ to ensure that the original equation holds:

$$\begin{aligned} Ax &= y \\ A(\frac{1}{2} A^T \lambda) &= y \\ \lambda &= 2(AA^T)^{-1} y \\ x &= A^T (AA^T)^{-1} y \end{aligned}$$

Denote the **right psuedo-inverse** as:

$$A_R{}^+ = A^T(AA^T)^{-1}$$

Proposition. *Let A be an $N \times M$ matrix, $N < M$, with full row-rank. Then the pseudo-inverse of A projects a vector from the range of A (A subset of $\mathbb{R}^N$) into the N-dimensional sub-space of $\mathbb{R}^M$ spanned by the columns of A: $A_R{}^+ = A^T(AA^T)^{-1}$*

Proof. In fact, we have:

$$\begin{aligned} A_R{}^+ &= A^T(AA^T)^{-1} \\ &= V \begin{bmatrix} \Sigma_N \\ 0 \end{bmatrix} U^T U {\Sigma_N}^{-2} U^T \\ &= V \begin{bmatrix} {\Sigma_N}^{-1} \\ 0 \end{bmatrix} U^T \\ &= \sum\nolimits_{i=1}^{N} {\sigma_i}^{-1} v_i {u_i}^T \end{aligned}$$

One should compare the sum of outer products in the deduction with those describing A in linear systems (Equation 18). This comparison drives home the geometric interpretation of the action of $A_R{}^+$. ☐

Proposition 2.1.1 means that the solution to the regularized problem of Equation 5, $x = A_R{}^+ y$, defines the unique solution x that is completely orthogonal to the null-space of A. This should make good sense: in Section 3.1 we found that any component of the solution that lies in the null-space is irrelevant, and the problem defined in Equation 5 was to find the smallest?? solution vector.
Finally then, we can write an explicit expression for the complete space of solutions to Ax = y, for A, $N \times M$, with full row-rank:

$$x = A_R^+ \, y + \sum_{i=N+1}^{M} \alpha_i v_i, \text{for any } \alpha_i$$

2.2 Over-Constrained System : Linear Regression

In the matrix form, it is

$$Ax = y$$

where A is order $N \times M$, x is order $M \times 1$, y is order $N \times 1$, where $N > M$. Find the vector x which minimizes $E = \sum_{i=1}^{N} \left({a_i}^T x - y\right)^2 = (Ax-y)^T(Ax-y)$ The column-rank of a matrix is equal to its row-rank and its rank, and all three equal the number of non-zero singular values.

$$A = U \begin{bmatrix} \Sigma_N & 0 \end{bmatrix} V^T$$

$$x = \left(A^T A\right)^{-1} A^T y$$

References

[Fl] Hermann Flaschka. *Principles of Analysis*, 1995

[SE] Stefan Evert. *Online Lectures by Stefan Evert, Osnabruck, Germany*