CME 108/MATH 114
Introduction to Scientific Computing
Summer 2020

Linear least squares problems

A special kind of optimization problem which comes up often in practice, especially in the context of modeling and fitting, is the least-squares (LS) problem. It is an optimization problem of the form $\min_x f(x)$ where $f(x) = ||g(x) - b||_2^2 = \sum_{i=1}^m (g_i(x) - b_i)^2$. Here $g : \mathbb{R}^n \to \mathbb{R}^m$ (typically with $m >> n$) and

$$g(x) = \begin{bmatrix} g_1(x) \\ \vdots \\ g_m(x) \end{bmatrix} \tag{1}$$

as usual

For a couple of applied contexts in which the LS problem arises, see HW6 and the additional practice problems.

There are two kinds of LS problems, classified by the structure of g. If $g(x) = Ax$ for some $m \times n$ matrix A (again typically $m >> n$ in model fitting), the problem is *linear* and its solution boils down to the solution of a *linear* system. If g is *nonlinear*, then solving the LS problem will require solving a *nonlinear* system equations.

In the modeling context, the LS problem arises as follows. Suppose we are given data points $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$. We propose a model (or classifier in some contexts) ϕ, which is a function of some weights $\vec{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}$.

We would like to find weight or parameter values such that $\phi_{\vec{w}}(x_i)$ is as close as possible to the target value y_i, for every $i = 1, 2, \ldots, m$. Here we use the square of the difference $(\phi_{\vec{w}}(x_i) - y_i)^2$ to measure how good is out fit.

Thus we set $g(\vec{w}) = \begin{bmatrix} \phi_{\vec{w}}(x_i) \\ \vdots \\ \phi_{\vec{w}}(x_m) \end{bmatrix}$ and obtain the LS problem $\min_{\vec{w}} ||g(\vec{w}) - \vec{y}||_2^2$.

Linear LS For now, let's suppose g is a linear function so $g(\vec{x}) = A\vec{x}$, where A is an $m \times n$ matrix (typically $m >> n$ in model fitting). In this special case, we'll see that the LS problem reduces to solving a *single* system of linear equations.

To find the solution to our LS problem, we seek to find the critical points of $f(x)$

We begin by re-writing f:

$$\begin{aligned} f(x) &= ||g(x) - b||^2 = ||Ax - b||^2 && (2)\\ &= (Ax - b)^T(Ax - b) && (3)\\ &= ((Ax)^T - b^T)(Ax - b) && (4)\\ &= x^T A^T Ax - x^T A^T b - b^T Ax + b^T b && (5)\\ &= x^T A^T Ax - 2b^T Ax + b^T b && (6) \end{aligned}$$

Therefore, (see "gradients involving matrices" notes)

$\nabla f(x)^T = ZA^T Ax - ZA^T b$. Hence we may solve the linear LS problem by solving the system $A^T Ax = A^T b$ (these are called the "normal equations")

Remark 0.0.1. The normal equation tend to be ill-conditioned in practice, so although this is a good way to solve the LS problem in theory, there are superior numerical algorithms.

One involves matrix factorizations like the QR and SVD. I encourage you to find out more!

Example 0.0.2. *See "least squares examples" note.*

A special case of the linear LS problem is the problem of linear regression (here I mean we have only two parameters and we seek a line of best fit).

Given data points $(x_1, y_1), \ldots, (x_m, y_m)$, we seek a, b to minimize

$$\begin{aligned} &\sum_{i=1}^{m}(y_i - (ax_i + b))^2 && (7)\\ &= \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} - \begin{bmatrix} ax_1 + b \\ \vdots \\ ax_m + b \end{bmatrix}^2 && (8)\\ &= \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} - \begin{bmatrix} x_1 & 1 \\ & \vdots \\ x_m & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}^2 && (9)\\ &= ||\vec{y} - A\vec{w}||^2 && (10) \end{aligned}$$

with $\vec{w} = \begin{bmatrix} a \\ b \end{bmatrix}$ and $A = \begin{bmatrix} x_1 & 1 \\ & \vdots \\ x_m & 1 \end{bmatrix}$

The normal equations constitute a system of two equations in two variables, with coefficient matrix

$$A^T A = \begin{bmatrix} x_1 & \dots & x_m \\ 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} x_1 & & 1 \\ & \vdots & \\ x_m & & 1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^m x_i^2 & \sum_{i=1}^m x_i \\ \sum_{i=1}^m x_i & m \end{bmatrix} \tag{11}$$

The RHS $A^T \vec{y}$ is given by

$$A^T \vec{y} = \begin{bmatrix} x_1 & \dots & x_m \\ 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^m x_i y_i \\ \sum_{i=1}^m y_i \end{bmatrix} \tag{12}$$

Therefore the optimal a, b satisfy

$$a \sum_{i=1}^m x_i^2 + b \sum_{i=1}^m x_i = \sum_{i=1}^m x_i y_i, \tag{13}$$

$$a \sum_{i=1}^m x_i + bm = \sum_{i=1}^m y_i \tag{14}$$

To conclude, we illustrate why this problem is called the "Least Square" problem.

By definition, we seek a, b to minimize the sum of $(y_i - (ax_i + b))^2$. This quantity can be interpreted as the area of a square with side length given by the residual $y_i - (ax_i + b)$, as in the diagram below. The problem is then to find (a, b) such that the square are (simultaneously) as small as possible.