# Least Squares

Suppose that we are given a linear system, $Ax=b$, where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$. Assume that $m > n$. Moreover, $\operatorname{rank} A = n$. (full column rank)

In this case, the system is usually inconsistent, that is, there's no $x$ satisfying $Ax=b$.

One way of finding an approximate solution is to consider the residual vector $r = Ax-b \in \mathbb{R}^m$ and to the minimize the norm square of it. This leads to the least squares problem:

$$(LS) \quad \underset{x\in\mathbb{R}^n}{\text{minimize}} \; \|Ax-b\|_2^2 .$$

Note that $f(x) = \|Ax-b\|_2^2 = (Ax-b)^T(Ax-b)$

$$= x^TA^TAx - x^TA^Tb - b^TAx + b^Tb$$

$$= x^TA^TAx - 2(A^Tb)^Tx + b^Tb.$$

Hence, $f$ is a quadratic function of $n$ variables.

$\nabla f(x) = 2A^TAx - 2A^Tb$ and $\nabla^2 f(x) = 2A^TT.$

Let's check if $A^TA$ is psd (even pd):

Note that for any vector $v \in \mathbb{R}^n$, we have

$$v^TA^TAv = (Av)^T(Av) = \|Av\|_2^2 \geq 0.$$

Hence $A^TA$ is positive semidefinite. Then, any stationary point for (LS) is a global minimum point.

Now, rank $A = n$ implies that the columns of $A$, say $a_1, \dots, a_n \in \mathbb{R}^m$, are linearly independent. Moreover for any vector $v \in \mathbb{R}^n$ we have

$$Av = v_1 a_1 + \dots + v_n a_n \in \mathbb{R}^m.$$

By linear independence we know that $Av = 0$ if and only if $v = 0 \in \mathbb{R}^n$.

Then, $v^TA^TAv = \|Av\|_2^2 = 0$ only for $v = 0$, and it's strictly positive for any other vector $v \in \mathbb{R}^n$.

This means that $A^TA$ is positive definite. In particular, it has an inverse. This implies that $\nabla f(x) = 0 \iff$

$$A^TAx = A^Tb \iff (A^TA)^{-1}A^TAx = (A^TA)^{-1}A^Tb$$

$$\iff \hat{x} = (A^TA)^{-1}A^Tb.$$

Note that if we had $m=n$ & $A$ invertible, then this would simplify as

$$\hat{x} = A^{-1}\underbrace{(A^T)^{-1}A^T}_{I}b = A^{-1}b$$

as expected.

Ex: (Data fitting)

Suppose we are given a set of data points $(s_i, t_i)$ for $i=1,\dots,m$ where $s_i \in \mathbb{R}^n$ and $t_i \in \mathbb{R}$. We want to fit this data to a linear form, that is, we look for a vector $x \in \mathbb{R}^n$ s.t. $s_i^Tx = t_i$, $i=1,\dots,m$.

In general, $m$ is much larger than $n$ and the system is over-determined.

Then, one can use the least squares approach.

Let $S = \begin{bmatrix} s_1, \dots, s_m \end{bmatrix} \in \mathbb{R}^{n \times m}$, $t = (t_1, \dots, t_m)^T$.

We want to minimize $\| S^T x - t \|_2^2$.

$\hat{x} = (SS^T)^{-1} St$ is the least square solution.

Ex: (Nonlinear data fitting)

Let $(x_i, y_i)$, $i = 1, \dots, m$ be data points such that $s_i, t_i \in \mathbb{R}, \forall i$.

Assume now, we want to fit data points in a polynomial form, that is,

$$\text{(*)} \quad y_i \simeq a_p x_i^p + a_{p-1} x_i^{p-1} + \dots + a_1 x_i + a_0, \quad i = 1, \dots, m.$$

for some numbers $a_0, \dots, a_p \in \mathbb{R}$. The idea is to find the best $a \in \mathbb{R}^{p+1}$ such that the residual is minimized.

Let $s_i = (x_i^p \;\; x_i^{p-1}, \dots, x_i, 1)^T \in \mathbb{R}^{p+1}$

We can write (*) as $y_i \simeq s_i^T a$.

We obtain a very similar case as in the previous example. Let $S = [s_1, \ldots, s_m] \in \mathbb{R}^{(p+1) \times m}$.

$$y = (y_1, \ldots, y_m)^T \in \mathbb{R}^m$$

We want to minimize $\| S^T a - y \|_2^2$, which is again a least square problem.

# The Gradient Method

Consider the unconstrained minimization problems: $\left( \underset{x \in \mathbb{R}^n}{\text{minimize}} \; f(x) \right)$

$f: \mathbb{R}^n \to \mathbb{R}$, continuously differentiable

We may be able to solve it by finding the stationary points (after showing existence) and selecting the minimum point among them. However, this is not always possible : (1) $\nabla f(x) = 0$ may be difficult to solve
(2) $\nabla f(x) = 0$ may have inf.'y many solutions...

There are also iterative methods:

- Pick $x_0 \in \mathbb{R}^n$ (arbitrarily, or a guess...)
- Pick a "descent direction" $d_k$ ?
- Find a stepsize $t_k$ s.t. $f(x_k + t_k d_k) < f(x_k)$ ← ?
- $x_{k+1} \leftarrow x_k + t_k d_k$
- Stop if a stopping criteria is satisfied. ← ?

Defn: $f: \mathbb{R}^n \to \mathbb{R}$, continuously differentiable. A vector $0 \neq d \in \mathbb{R}^n$ is a descent direction of $f$ at $x$ if the directional derivative of $f$ in direction $d$ at $x$ is negative, that is,

$$f'(x; d) = \lim_{t \downarrow 0} \frac{f(x+td) - f(x)}{t} < 0.$$

* Note that $f'(x; d) = \nabla f(x)^T d$.

For gradient method, the negative of the gradient is selected as the decent direction, that is, $d_k = -\nabla f(x_k)$.

Note that $f'(x_k; d_k) = \nabla f(x_k)^T(-\nabla f(x_k)) = -\|\nabla f(x_k)\|^2 < 0$.

Indeed, this is the "steepest" descent direction, that is, over all $d \in \mathbb{R}^n$ with $\|d\| = 1$, $-\nabla f(x_k)$ yields the minimum directional derivative

$$f'(x_k, -\nabla f(x_k)) = \min_{d \in \mathbb{R}^n} \{ f'(x_k, d) \mid \|d\| = 1 \}.$$

- How to select the stepsize?

  - constant stepsize: $t_k = \bar{t} \quad \forall k$

    easy to apply but how to choose? If too small, then may converge slow. If not small, may not satisfy $f(x_k + t_k d_k) < f(x_k)$.

  - exact line search: minimize $f$ along the ray $x_k + t \cdot d_k$

    i.e. solve $\displaystyle \min_{t \geq 0}$ minimize $f(x_k + t d_k)$

    This may be difficult (may not be possible) to solve...

  - backtracking: start with an initial guess $t_k = s$, iterate $t_k \leftarrow \beta t_k$ for some $0 < \beta < 1$.

    until we have $f(x_k) - f(x_k + t_k d_k) \geq -\alpha t_k \nabla f(x_k)^T d_k$ for some $\alpha \in (0,1)$.

- Stopping condition For a predetermined error $\varepsilon > 0$, if $\|\nabla f(x_{k+1})\| \leq \varepsilon$, then STOP. Return $x_{k+1}$.

**Ex:** Exact line search for quadratic functions. Let $A \in \mathbb{R}^{n\times n}$ be p.d. $b \in \mathbb{R}^n$, $c \in \mathbb{R}$. Consider $f(x) = x^T A x + 2b^T x + c$

Let $d \in \mathbb{R}^n$ be a descent direction (may not be the steepest descent) at a point $x$. We want to find the stepsize by exact line search.

$$\begin{pmatrix} \text{minimize} & f(x+td) \\ \text{s.t.} & t \geq 0 \end{pmatrix}$$

The decision variable is "$t$". Let's write the objective functions, as a function of $t$:

$$g(t) := f(x+td) = (x+td)^T A (x+td) + 2b^T(x+td) + c$$

$$= x^T A x + t x^T A d + t d^T A x + t^2 d^T A d + 2b^T x + 2t b^T d + c$$

$$= t^2 (d^T A d) + 2t(x^T A d) + 2t b^T d + x^T A x + 2b^T x + c$$

$$= (d^T A d) t^2 + 2(x^T A d + b^T d) t + x^T A x + 2b^T x + c$$

(a quadratic function of $t$)

$$g'(t) = 2(d^T A d) t + 2(x^T A d + b^T d)$$

$$= 2(d^T A d) t + 2 d^T (A x + b)$$

(as $A^T = A$)

$$\Rightarrow t^* = \frac{-d^T(Ax+b)}{d^T A d}.$$

Note that $\nabla f(x) = 2Ax + 2b = 2(Ax+b)$.

Hence, $t^* = \dfrac{-d^T \nabla f(x)}{2 d^T A d}$.