CS 524: Introduction to Optimization
Lecture 28 : Least Squares Problems

Michael Ferris

Computer Sciences Department
University of Wisconsin-Madison

November 8, 2023

Review of linear equations

System of m linear equations in n unknowns:

$$\begin{array}{c} a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n = b_2 \\ \vdots \quad \vdots \quad \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n = b_m \end{array} \iff \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$$

Compact representation: $Ax = b$. Only three possibilities:

1. exactly one solution (e.g. $x_1 + x_2 = 3$ and $x_1 - x_2 = 1$)
2. infinitely many solutions (e.g. $x_1 + x_2 = 0$)
3. no solutions (e.g. $x_1 + x_2 = 1$ and $x_1 + x_2 = 2$)

Review of linear equations

- The set of solutions of $Ax = b$ is an **affine subspace**.
- If $m > n$, there is (usually but not always) no solution. This is the case where A is **tall** (overdetermined).
  - Can we find x so that $Ax \approx b$?
  - One possibility is to use **least squares**.
- If $m < n$, there are infinitely many solutions. This is the case where A is **wide** (underdetermined).
  - Among all solutions to $Ax = b$, which one should we pick?
  - One possibility is to use **regularization**.

In this lecture, we will discuss **least squares**.

Least squares

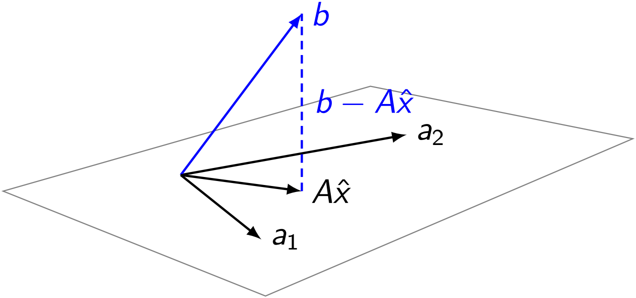
- Typical case of interest: $m > n$ (overdetermined). If there is no solution to $Ax = b$ we try instead to have $Ax \approx b$.
- The least-squares approach: make Euclidean norm $\|Ax - b\|$ as small as possible.
- Equivalently: make $\|Ax - b\|^2$ as small as possible.

Standard form:

$$\boxed{\underset{x}{\text{minimize}} \quad \left\|Ax - b\right\|^2}$$

It's an unconstrained optimization problem.

Geometry of LS



- The set of points $\{Ax\}$ is a subspace.
- We want to find $\hat{x}$ such that $A\hat{x}$ is closest to b.
- Insight: $(b - A\hat{x})$ must be orthogonal to the subspace.
- i.e. $0 = (Ax)^T(b - A\hat{x}) = x^T(A^Tb - A^TA\hat{x})$ for all x
- Since this holds for all x, the normal equations are satisfied:

$$A^TA\hat{x} = A^Tb$$

- Alternatively: $\nabla_x \|Ax - b\|^2 = 0$!

Normal equations

Theorem: If $\hat{x}$ satisfies the normal equations, then $\hat{x}$ is a solution to the least-squares optimization problem

$$\underset{x}{\text{minimize}} \quad \|Ax - b\|^2$$

Proof: Suppose $A^{\mathsf{T}}A\,\hat{x} = A^{\mathsf{T}}b$. Let x be any other point.

$$\begin{aligned}
\|Ax - b\|^2 &= \|A(x - \hat{x}) + (A\hat{x} - b)\|^2 \\
&= \|A(x - \hat{x})\|^2 + \|A\hat{x} - b\|^2 + 2(x - \hat{x})^{\mathsf{T}}A^{\mathsf{T}}(A\hat{x} - b) \\
&= \|A(x - \hat{x})\|^2 + \|A\hat{x} - b\|^2 \\
&\geq \|A\hat{x} - b\|^2
\end{aligned}$$

Vector norms

We want to solve $Ax = b$, but there is no solution. Define the residual to be the quantity $r := b - Ax$. We can't make it zero, so instead we try to make it small. Many options!

- minimize the largest component (a.k.a. the ∞-norm)

$$\|r\|_\infty = \max_i |r_i|$$

- minimize the sum of absolute values (a.k.a. the 1-norm)

$$\|r\|_1 = |r_1| + |r_2| + \cdots + |r_m|$$

- minimize the Euclidean norm (a.k.a. the 2-norm)

$$\|r\|_2^2 = \|r\|^2 = r_1^2 + r_2^2 + \cdots + r_m^2$$

We can think of these functions as loss functions applied to the residual r

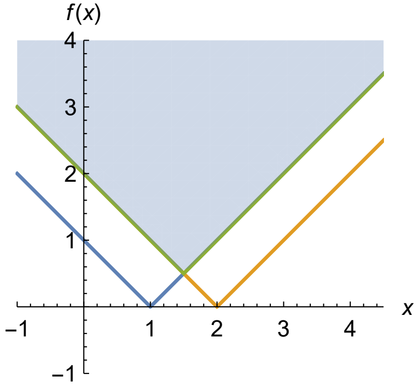
Example 1: L_∞ (see 28leastsq.ipynb)

Find x so that $Ax = \begin{bmatrix} 1 \\ 1 \end{bmatrix} x = \begin{bmatrix} x \\ x \end{bmatrix}$ is close to $b = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

Minimize largest component:

$$\min_x \max\{|x-1|, |x-2|\}$$

Optimum is at $x = 1.5$.



Easy: reformulate as an LP

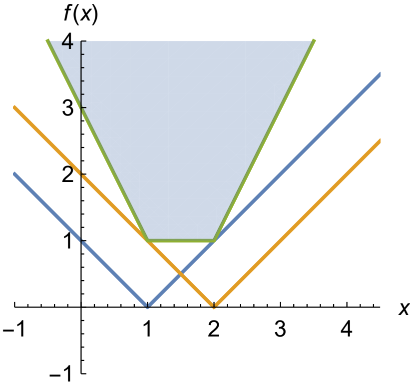
Example 2: L_1 (see `28leastsq.ipynb`)

Example: find $\begin{bmatrix} x \\ x \end{bmatrix}$ that is closest to $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$.

Minimize sum of components:

$$\min_x \ |x-1| + |x-2|$$

Optimum is any $1 \leq x \leq 2$.



Easy: reformulate as an LP

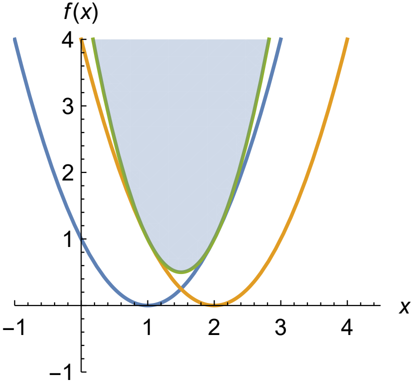
Example 3: least squares (see `28leastsq.ipynb`)

Example: find $\begin{bmatrix} x \\ x \end{bmatrix}$ that is closest to $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$.

Minimize sum of squares:

$$\min_x \, (x-1)^2 + (x-2)^2$$

Optimum is at $x = 1.5$.



Is this a convex QP?

Extension: k largest terms

Look at k-norm defined as sum of largest k absolute value terms.

$$\|x\|_\infty = \|x\|_{[1]} \leq \ldots \leq \|x\|_{[n]} = \|x\|_1$$

and its dual norm

$$\|\cdot\|_{[k]} = \max\{\frac{1}{k}\|\cdot\|_1, \|\cdot\|_\infty\}$$

$$\|x\|_{[k]} = \max_\alpha x^T \alpha \text{ s.t. } -1 \leq \alpha_i \leq 1, \sum_i |\alpha_i| \leq k$$

which is equal (by lp duality) to:

$$\begin{aligned} &\min kw + \textstyle\sum_i p_i + q_i \\ &\text{ s.t. } p_i - q_i + u_i - v_i = x_i, -u_i - v_i + w = 0, p, q, u, v, w \geq 0 \end{aligned}$$

Parametric regression (see `28leastsq.ipynb`)

We are given noisy data points (z_i, y_i) (the training set).

- We suspect they are related by

$$y \approx pz^2 + qz + r =: \phi(z; x)$$

- Find $x = (p, q, r)$, so $\phi(z; x)$ best agrees with the data y.

Writing all the equations:

$$\begin{array}{c} y_1 \approx pz_1^2 + qz_1 + r \\ y_2 \approx pz_2^2 + qz_2 + r \\ \vdots \\ y_m \approx pz_m^2 + qz_m + r \end{array} \quad \Rightarrow \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \approx \begin{bmatrix} z_1^2 & z_1 & 1 \\ z_2^2 & z_2 & 1 \\ \vdots & \vdots & \vdots \\ z_m^2 & z_m & 1 \end{bmatrix} \begin{bmatrix} p \\ q \\ r \end{bmatrix} =: Ax$$

- $a_i^T = (z_i^2, z_i, 1)$ for $i = 1, \ldots, m$
- Curve fitting problem is also called parametric regression

Example: curve-fitting

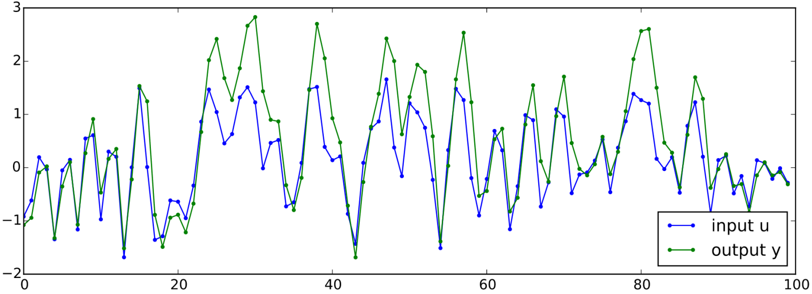
- More complicated: $y \approx pe^z + q\cos(z) - r\sqrt{z} + sz^3$
- Find $x = (p, q, r, s)$ that best agrees with the data
- Writing all the equations:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \approx \begin{bmatrix} e^{z_1} & \cos(z_1) & -\sqrt{z_1} & z_1^3 \\ e^{z_2} & \cos(z_2) & -\sqrt{z_2} & z_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ e^{z_m} & \cos(z_m) & -\sqrt{z_m} & z_m^3 \end{bmatrix} \begin{bmatrix} p \\ q \\ r \\ s \end{bmatrix} =: Ax$$

- $a_i^T = (e_i^z, \cos(z_i), -\sqrt{z_i}, z_i^3)$ for $i = 1, \ldots, m$;
  $\phi(z; x) = pe^z + q\cos(z) - r\sqrt{z} + sz^3$
- Still a linear least squares problem (data is nonlinear)

Moving average model

- We are given a time series of input data $u_1, u_2, \ldots, u_T$ and output data $y_1, y_2, \ldots, y_T$. Example:



- A "moving average" model with window size k assumes each output is a weighted combination of k previous inputs:

$$y_t \approx w_0 u_t + w_1 u_{t-1} + \cdots + w_{k-1} u_{t-(k-1)} \text{ for all } t$$

- find weights $w_0, \ldots, w_{k-1}$ that best agree with the data.

Example: moving average model

- Moving average model:

$$y_t \approx w_0 u_t + w_1 u_{t-1} + w_2 u_{t-2} \text{ for all } t$$

- Writing all the equations (e.g. $k = 3$):

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_T \end{bmatrix} \approx \begin{bmatrix} u_1 & 0 & 0 \\ u_2 & u_1 & 0 \\ u_3 & u_2 & u_1 \\ \vdots & \vdots & \vdots \\ u_T & u_{T-1} & u_{T-2} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}$$

- Solve least squares problem! `28movingAV.ipynb`