# EE445 Mod2-Lec1: Introduction to Least Squares

References:

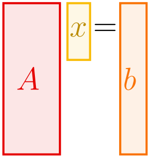
- [VMLS]: Chapter 12

# Least Squares Set-up

- Linear Regression ([VMLS, Ch. 2.3]) is the simplest form of machine learning out there.
- Consider an $m \times n$ matrix $A$—i.e., $A \in \mathbb{R}^{m\times n}$—and vectors $b \in \mathbb{R}^m$ and $x \in \mathbb{R}^n$
- **Goal**: Find a solution to $Ax = b$—that is, find $x$ such that $Ax = b$
- **ML Intepretation**:
  - $A$ is a matrix of training data—i.e., $m$ is the number of samples, and $n$ is the number of 'features'
  - $m$–dimensional vector $b$ contains 'target values' or observations of real world phenomena
  - $n$-dimensional vector $x$ is a set of feature weights

# Overdetermined System of Equations→Least Squares Opt

- **Goal**: Find a solution to $Ax = b$—that is, find $x$ such that $Ax = b$
- However, typically $A$ is a 'tall' matrix or what we call an 'over-determined' system—i.e., there are more equations ($m$) than variables to choose ($n$).

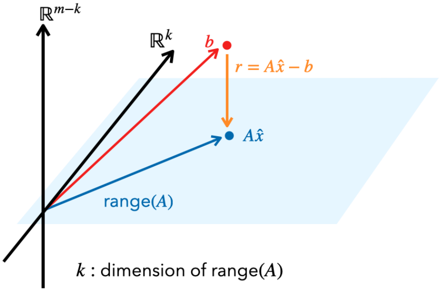



- There is often not an exact solution → formulate an **optimization problem** to find as *close* a solution as possible—i.e., an *least squares approximate solution*

# Least Squares Optimization Problem

- Least squares optimization problem:

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2$$




- Components of the problem:
  - **decision variable**: $x \in \mathbb{R}^n$
  - **data**: $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$
  - **objective**: $\|Ax - b\|_2^2$
- vector of residuals $r \in \mathbb{R}^m$: let $\hat{x}$ be the solution to the least squares opt problem.

$$r := A\hat{x} - b$$

# Example Applications: Advertising Purchases

- Consider $m$ demographic groups (audiences) that we want to advertise to, with a target number of 'impressions' or views for each group, $b$
- To reach these groups, we purchase advertising in $n$ different channels (e.g., different web publishers, radio, print,...), in amounts that given as a vector $x \in \mathbb{R}^n$.
- The matrix $A \in \mathbb{R}^{m \times n}$ specifies the number of impressions in each group per dollar spending in the channels—i.e., entry $a_{ij}$ is the number of impressions in group $i$ per dollar spent on advertising in channel $j$.
  - The $j$–th column of $A$ gives the effectiveness or reach (in impressions per dollar) for channel $j$.
  - The $i$–th row of $A$ shows which media demographic group $i$ is exposed to.
- **Goal**: find $x$ such that $\|Ax - b\|_2^2$ is as small as possible (minimized)

# Other Examples

- Stock market prediction:
- Weather forecasting:
- Predicting impact of GPA/SAT scores on college admissions
- Predicting/forecasting housing prices as a function of size, location, etc.

# Combing back to the optimization problem

- Any vector $\hat{x}$ satisfying the following is a solution (i.e., a least squares approximate solution):
$$\|A\hat{x} - b\|_2^2 \leq \|Ax - b\|_2^2 \quad \text{for all} \quad (\forall)\ x \in \mathbb{R}^n$$
- Importantly, it need not be the case that $A\hat{x} = b$!
- **Regression**: We say that $\hat{x}$ is the result of *regressing* the vector $b$ onto the columns of $A$.

# Column Interpretation

$$A = \begin{bmatrix} | & \cdots & | \\ a_1 & \cdots & a_n \\ | & \cdots & | \end{bmatrix}, \quad a_i \in \mathbb{R}^m$$

- Least squares problem is equivalent to finding a linear combination of the columns that is closest to $b \in \mathbb{R}^m$:

$$\|Ax - b\|_2^2 = \|x_1 \cdot a_1 + \cdots + x_n \cdot a_n - b\|_2^2$$

  where $x_i \cdot a_i$ is element-wise multiplication of the vector $a_i$ by the scalar $x_i$
- For a solution $\hat{x}$, we have that $A\hat{x} = \hat{x}_1 \cdot a_1 + \cdots + \hat{x}_n \cdot a_n$
- $A\hat{x}$ is the *closest* (in Euclidean distance) to $b \in \mathbb{R}^m$ among all linear combinations of vectors $a_1, \ldots, a_n \in \mathbb{R}^m$

# Row Interpretation

$$A = \begin{bmatrix} — & \tilde{a}_1^\top & — \\ \vdots & \dots & \vdots \\ — & \tilde{a}_m^\top & — \end{bmatrix}, \quad \tilde{a}_i \in \mathbb{R}^n$$

- Recall that $r = Ax - b$ is the residual vector
- The components of $r$ are then given by $r_i = \tilde{a}_i^\top x - b_i, \quad i = 1, \dots, m$
- The objective can be rewritten as

$$\|Ax - b\|_2^2 = (\tilde{a}_1^\top x - b_1)^2 + \cdots + (\tilde{a}_m^\top x - b_m)^2$$

# Example

$$A = \begin{bmatrix} 2 & 0 \\ -1 & 1 \\ 0 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \quad m = 3,\ n = 2$$

- $Ax = b \implies \{2x_1 = 1,\ -x_1 + x_2 = 0,\ 2x_2 = -1\}$ a system that has no solution
- Least squares problem: using row interpretation we have

$$\min_{x_1, x_2} \{(2x_1 - 1)^2 + (-x_1 + x_2)^2 + (2x_2 + 1)^2\}$$

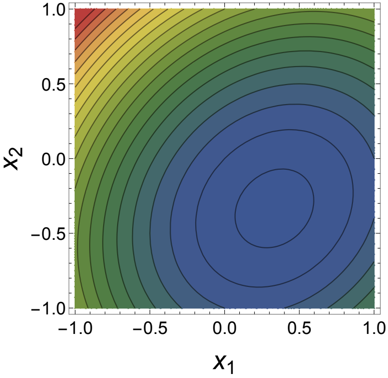
# Aside: Finding Minima via Calculus [VMLS, App. C]

- **Calculus**: to find $\min_x f(x)$, we set $\frac{d}{dx} f(x) = 0$ and find $x^*$ that solves the equation, and check that $\frac{d^2}{dx^2} f(x)\Big|_{x=x^*} > 0$
- **Multivariable Case:**
  - $\nabla f(x) = 0 \iff \{\frac{\partial}{\partial x_i} f(x) = 0, \ i = 1, \dots, n\}$
  - Hessian: $\nabla^2 f(x)|_{x=x^*} > 0$ where

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \quad \text{and} \quad \nabla^2 f(x)|_{x=x^*} > 0 \iff \text{eigenvalues positive}$$

# Example Continued

$$\min_{x_1,x_2}\{(2x_1-1)^2+(-x_1+x_2)^2+(2x_2+1)^2\}$$




$$\begin{bmatrix} \frac{\partial}{\partial x_1} f(x) \\ \frac{\partial}{\partial x_2} f(x) \end{bmatrix} = \begin{bmatrix} 4(2x_1-1)-2(-x_1+x_2) \\ 2(-x_1+x_2)+4(2x_2+1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\implies \begin{bmatrix} 10x_1 - 2x_2 \\ 2x_1 - 10x_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \end{bmatrix} \implies \hat{x} = \begin{bmatrix} \frac{1}{3} \\ -\frac{1}{3} \end{bmatrix}$$

- Observe: $A\hat{x} \neq b$.
- Indeed, $r = A\hat{x} - b = (-\frac{1}{3}, -\frac{2}{3}, \frac{1}{3})$ and $\|A\hat{x} - b\|_2^2 = \frac{2}{3}$

# Least Squares Solution via Calculus

**Assumption [A1]:** The columns of $A$ are linearly independent—i.e., $\sum_{i=1}^{n} c_i a_i = 0 \iff c_i = 0 \quad \forall i = 1, \ldots, n$

- Any minimizer $\hat{x}$ of $f(x) = \|Ax - b\|_2^2$ must satisfy

$$\frac{\partial f}{\partial x_i}(\hat{x}) = 0, \quad i = 1, \ldots, n \iff \nabla f(\hat{x}) = 0$$

- In matrix form, the gradient is

$$\nabla f(x) = 2A^\top (Ax - b)$$ [VMLS, page 184]

# Let's verify

- Least squares objective in summation form:

$$f(x) = \|Ax - b\|_2^2 = \sum_{i=1}^{m} \left( \sum_{j=1}^{n} a_{ij} x_j - b_i \right)^2$$

- Let $v = \nabla f(x) \in \mathbb{R}^n$ where $v_\ell = \frac{\partial}{\partial x_\ell} f(x)$—i.e.,

$$\begin{aligned} v_\ell = \frac{\partial f}{\partial x_\ell}(x) &= \sum_{i=1}^{m} 2 \left( \sum_{j=1}^{n} a_{ij} x_j - b_i \right) a_{i\ell} \\ &= \sum_{i=1}^{m} 2(A^\top)_{\ell i}(Ax - b)_i = (2A^\top(Ax - b))_\ell \end{aligned}$$

# Least Squares Solution via Calculus Continued

- Any minimizer $\hat{x}$ of $f(x) = \|Ax - b\|_2^2$ must satisfy

$$\nabla f(\hat{x}) = 2A^\top(A\hat{x} - b) = 0 \iff A^\top A\hat{x} = A^\top b \quad \text{[normal equations]}$$

- **Gram matrix**: $A^\top A$ has entries which are the inner products of the columns of $A$
- **[A1]** $\implies$ $A^\top A$ is invertible [VMLS, §11.5,pg. 214]
- Hence, $\hat{x} = (A^\top A)^{-1}A^\top b$ is the *only* solution of the normal equations
- **Pseudo-inverse**: $A^\dagger := (A^\top A)^{-1}A^\top$ is a left inverse of $A$
- $\hat{x} = A^\dagger b$ solves $Ax = b$ **if** the set of equations has a solution otherwise it is said to be the least squares approximate solution.

# Direct Verification of the Solution

- Let's check via direct verification: we will show that for any $x \neq \hat{x} = A^\dagger b$ we have the estimate
$$\|A\hat{x} - b\|_2^2 < \|Ax - b\|_2^2$$
- Indeed,
$$\begin{aligned}\|Ax - b\|_2^2 &= \|(Ax - A\hat{x}) + (A\hat{x} - b)^2\|_2^2 \\ &= \|A(x - \hat{x})\|_2^2 + \|A\hat{x} - b\|_2^2 + 2(x - \hat{x})^\top A^\top (A\hat{x} - b)\end{aligned}$$
since $\|u + v\|_2^2 = (u + v)^\top (u + v) = \|u\|_2^2 + \|v\|_2^2 + 2u^\top v$
- **Claim**: $(x - \hat{x})^\top A^\top (A\hat{x} - b) = 0$
**proof**: since $(A^\top A)\hat{x} = A^\top b$ [normal equations], we have
$$(x - \hat{x})^\top A^\top (A\hat{x} - b) = (x - \hat{x})^\top (A^\top A\hat{x} - A^\top b) = 0$$

# Direct Verification of the Solution

- we know that $(x-\hat{x})^\top A^\top (A\hat{x}-b)=0$
- Coming back to the expression for the objective, we have

$$\|Ax-b\|_2^2 = \underbrace{\|A(x-\hat{x})\|_2^2}_{\geq 0} + \|A\hat{x}-b\|_2^2$$

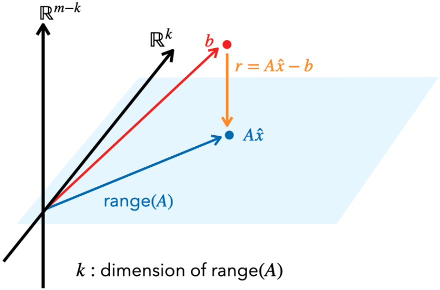
- Hence, we deduce

$$\|A\hat{x}-b\|_2^2 \leq \|Ax-b\|_2^2$$

- **Row form of solution**: sometimes its useful to express the solution as

$$\hat{x} = A^\dagger b = (A^\top A)^{-1} A^\top b = \left(\sum_{i=1}^m \tilde{a}_i \tilde{a}_i^\top\right)^{-1} \left(\sum_{i=1}^m b_i \tilde{a}_i\right)$$

# Orthogonality Principle

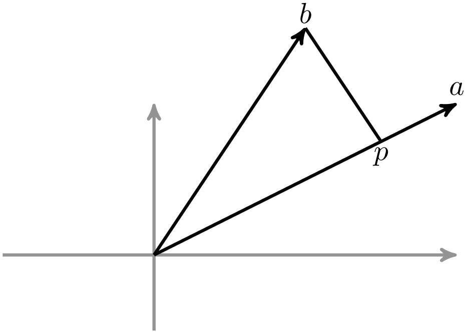



- $A\hat{x}$ is the linear combination of columns of $A$ closest to $b$
- Residual $r = A\hat{x} - b$ satisfies the so orthogonality principle:

$$(Az) \perp r \quad \forall\ z \in \mathbb{R}^n$$

- **Why?**

# Let's Look at the Vector case



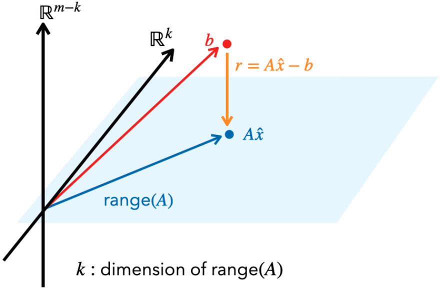

- Since $p$ lies along the vector $a$, we know that $p = xa$ for some $x$
- Also, $a$ is perpendicular to $r = b - xa$—i.e.,

$$a^\top(b - xa) = 0 \implies xa^\top a = a^\top b$$

$$x = \frac{a^\top b}{a^\top a} \text{ and } p = ax = a\frac{a^\top b}{a^\top a}$$

- projection matrix: $P = a(a^\top a)^{-1}a^\top$

# Orthogonality Principle




- $A\hat{x}$ is the linear combination of columns of $A$ closest to $b$
- Residual $r = A\hat{x} - b$ satisfies the so orthogonality principle:

$$(Az) \perp r \quad \forall\, z \in \mathbb{R}^n$$

- **Why?**

- First, [normal equations] $\iff A^\top(A\hat{x} - b) = 0$
- Hence, for any $z \in \mathbb{R}^n$, we have

$$(Az)^\top r = (Az)^\top(A\hat{x} - b) = z^\top A^\top(A\hat{x} - b) = 0$$

# Projection

- The least squares solution is $\hat{x} = (A^\top A)^{-1}A^\top b$ and the prediction is $\hat{y} = A\hat{x}$
- $P = A(A^\top A)^{-1}A^\top$ is a projection matrix: it projects on to the subspace formed by the columns of $A$
  - $P$ is a projection matrix if $P^2 = P$
- Orthogonal decomposition: $b = b_{\mathcal{R}(A)} + b_{\mathcal{R}(A)^\perp}$ where $b_{\mathcal{R}(A)} = A\hat{x}$ and $b_{\mathcal{R}(A)^\perp} = r = A\hat{x} - b$

# More Examples

Consider

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix}$$

Find the least squares approximate solution to $Ax = b$.

**Solution.** First

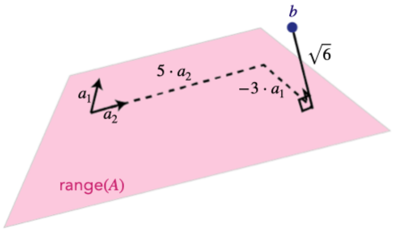
$$A^\top A = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 5 & 3 \\ 3 & 3 \end{bmatrix} \quad \text{and} \quad A^\top b = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 6 \end{bmatrix}$$

$$(A^\top A)^{-1} = \frac{1}{15-9} \begin{bmatrix} 3 & -3 \\ -3 & 5 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{5}{6} \end{bmatrix} \implies \hat{x} = \underbrace{\begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{5}{6} \end{bmatrix}}_{(A^\top A)^{-1}} \begin{bmatrix} 0 \\ 6 \end{bmatrix} = \begin{bmatrix} -3 \\ 5 \end{bmatrix}$$

# Example Continued

- The solution minimizes the distance from $A\hat{x}$ to $b$:

$$\|b - A\hat{x}\|_2^2 = \left\| \begin{bmatrix} 6 \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 5 \\ 2 \\ -1 \end{bmatrix} \right\|_2^2 = \left\| \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} \right\|_2^2$$



# Numerical Examples

see `Mod2-N1.ipynb`