MSCA 37016
Advanced Linear Algebra for Machine Learning

Lecture 3

Danny Ng

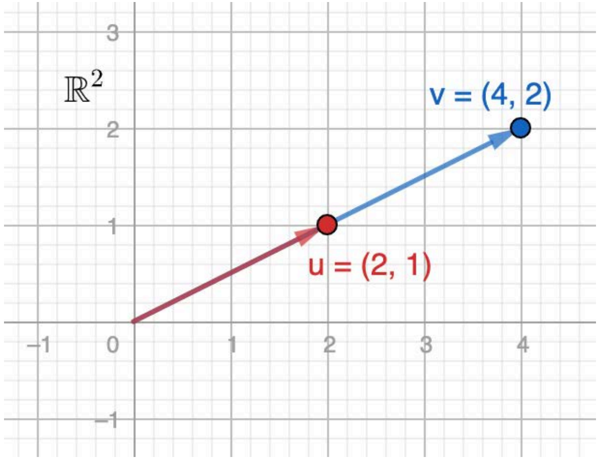
L2-Norm Minimization

- If $\vec{v}$ and $\vec{u} \in \mathbb{R}^d$ are parallel, then it is easy to find a scaling s of $\vec{u}$ so that

$$\vec{v} = s\vec{u}$$

e.g.

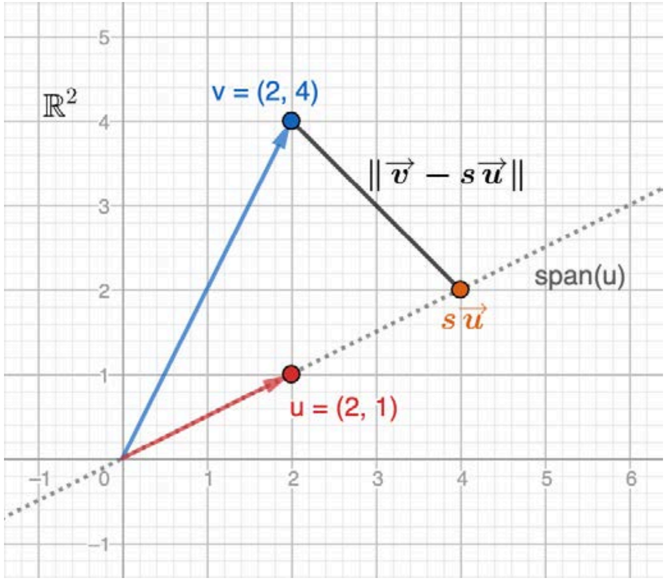
$$\vec{v} = (4, 2) \qquad \vec{u} = (2, 1) \qquad s = 2$$



- **Question:** If $\vec{v}$ and $\vec{u}$ are not parallel, then how should we scale $\vec{u}$ so that

$$\vec{v} \approx s\vec{u}$$

are (not equal but) as close as possible?



L2-Norm Minimization Geometry

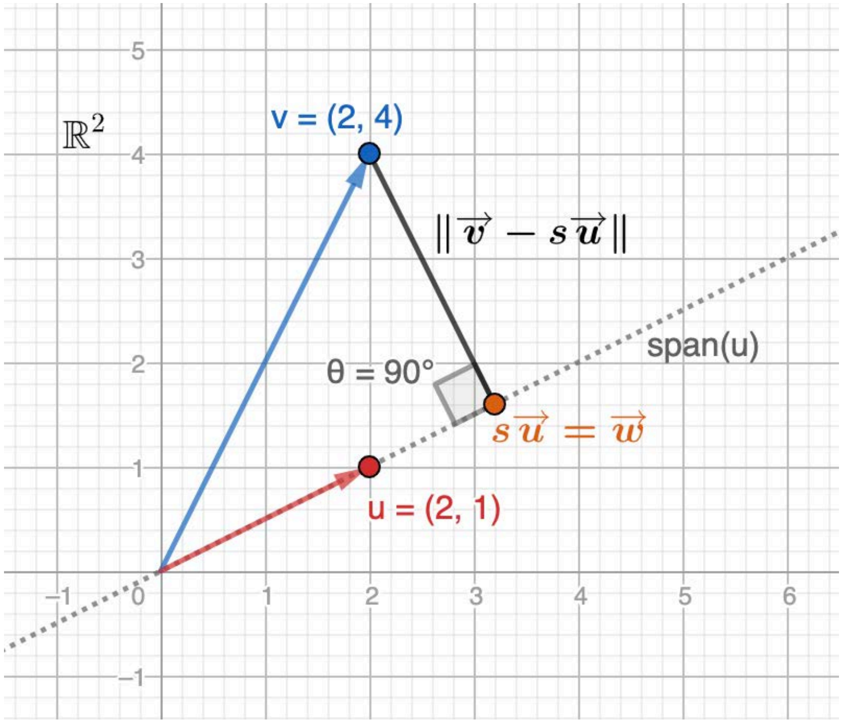
- Mathematically, find scaling s to

$$\min_{s} ||\vec{v} - s\vec{u}||$$

- **Geometric perspective:** The point $s\vec{u}$ on the line of $\vec{u}$ that is closest to $\vec{v}$ must be at a 90° angle

- **Vector space perspective:** Find the vector $\vec{w}$ in the vector space $\text{span}(\vec{u})$ that is closest to the given vector $\vec{v}$

$$\min_{\vec{w} \in \text{span}(\vec{u})} ||\vec{v} - \vec{w}||$$



Orthogonal Projection

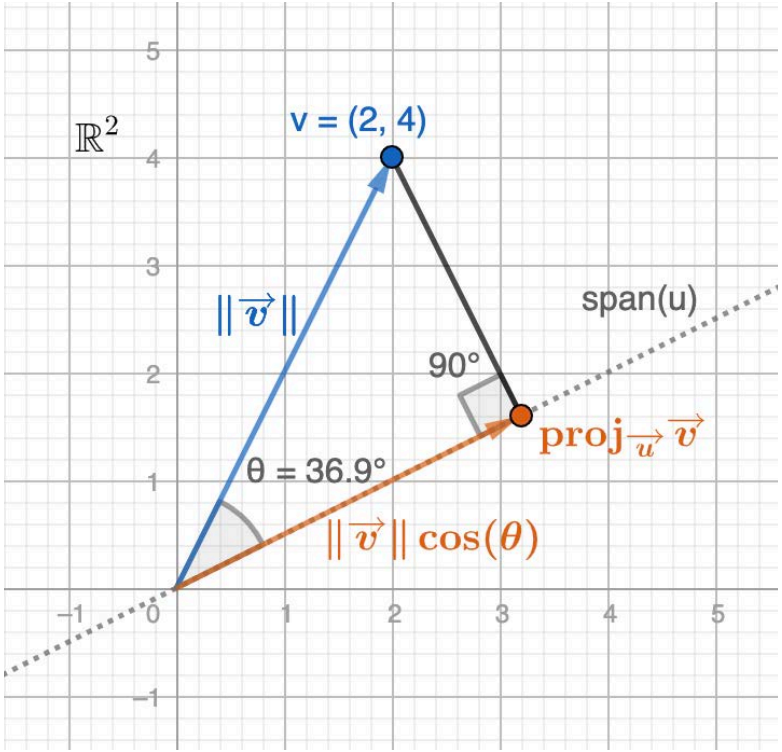
- **Question:** How to project $\vec{v}$ "orthogonally" onto the line of $\vec{u}$?
- It suffice to specify the length and direction of the projected vector

$$\mathrm{proj}_{\vec{u}}\vec{v} = \overbrace{(\|\vec{v}\|\cos\theta)}^{\text{length}}\ \overbrace{\hat{u}}^{\text{direction}}$$

$$= (\|\vec{v}\|\cos\theta)\frac{\vec{u}}{\|\vec{u}\|}$$

$$= \left(\frac{\vec{v}\cdot\vec{u}}{\|\vec{u}\|^2}\right)\vec{u}$$

$$= \underbrace{\left(\frac{\vec{v}\cdot\vec{u}}{\vec{u}\cdot\vec{u}}\right)}_{\text{formula for scaling } s}\vec{u}$$



Orthogonal Projection Examples

e.g.

$$\vec{v} = (3, 5) \quad \text{and} \quad \vec{u} = (2, 1)$$

$$\begin{aligned}\text{proj}_{\vec{u}}\vec{v} &= \left(\frac{\vec{v}\cdot\vec{u}}{\|\vec{u}\|^2}\right)\vec{u} \\ &= \left(\frac{3\times 2 + 5\times 1}{2^2 + 1^2}\right)\vec{u} \\ &= \frac{11}{5}\vec{u} \\ &= \left(\frac{22}{5}, \frac{11}{5}\right)\end{aligned}$$

e.g.

$$\vec{v} = (3, 5) \quad \text{and} \quad \vec{u} = (6, 3)$$

$$\begin{aligned}\text{proj}_{\vec{u}}\vec{v} &= \left(\frac{\vec{v}\cdot\vec{u}}{\|\vec{u}\|^2}\right)\vec{u} \\ &= \left(\frac{3\times 6 + 5\times 3}{6^2 + 3^2}\right)\vec{u} \\ &= \frac{11}{15}\vec{u} \\ &= \left(\frac{22}{5}, \frac{11}{5}\right)\end{aligned}$$

- If $\vec{u}$ is 3x longer, then the scaling s is 1/3 smaller but the projected vector $\text{proj}_{\vec{u}}\vec{v}$ remains the same!

Residual of Orthogonal Projection

- The **residual** of an orthogonal projection is

$$\vec{e} = \vec{v} - \text{proj}_{\vec{u}}\vec{v}$$

- Residual norm is the minimum value of the optimization

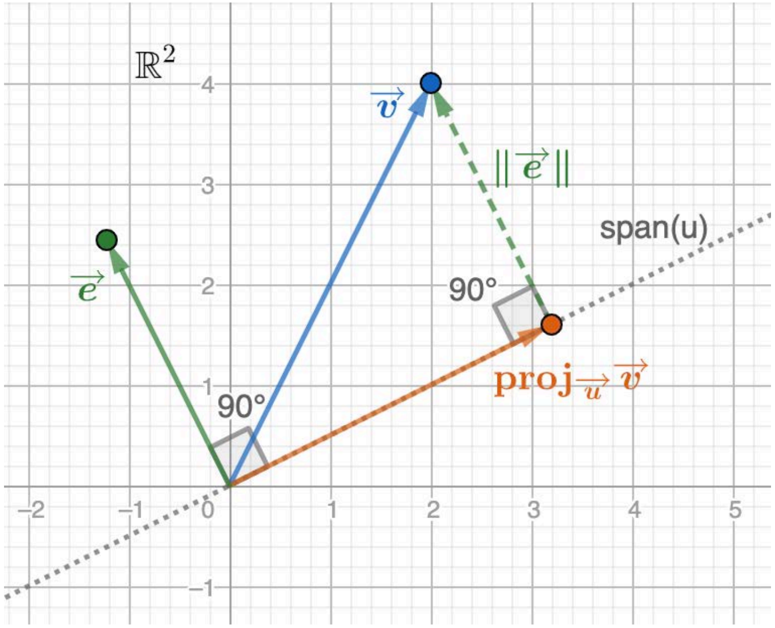
$$\|\vec{e}\| = \min_{s}\|\vec{v} - s\vec{u}\|$$

- Geometrically, $\|\vec{e}\|$ is the **distance** between the point $\vec{v}$ and the line of $\vec{u}$
- By construction,

$$\vec{e} \perp \text{proj}_{\vec{u}}\vec{v}$$

- $\vec{v}$ is decomposed into sum of 2 orthogonal components

$$\vec{v} = (\text{proj}_{\vec{u}}\vec{v}) + \vec{e}$$



Simple Linear Regression

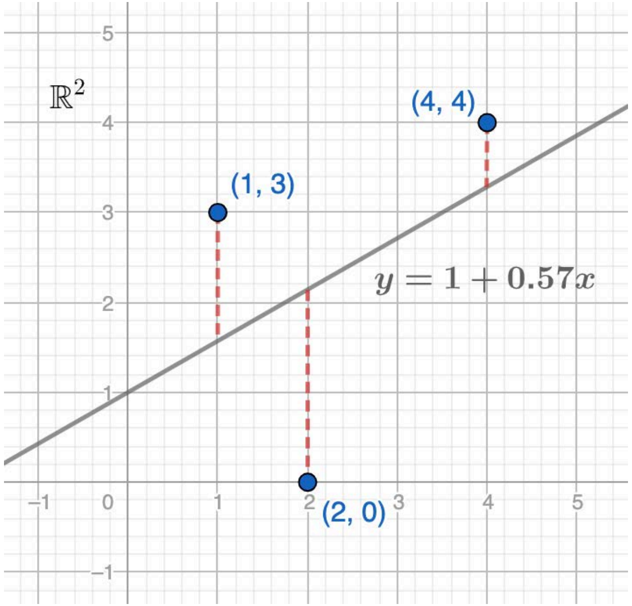
- Observed data: $(x_1, y_1), \ldots, (x_n, y_n)$
- Model equation:

$$y_i = \overbrace{\beta_0 + \beta_1 x_i}^{\text{line}} + \epsilon_i \quad \text{for } i = 1, \ldots, n$$

intercept slope noise term

- Find optimal β_0^* and β_1^* such that the line fits the data best

$$y_i \approx \beta_0^* + \beta_1^* x_i \quad \text{for } i = 1, \ldots, n$$



- **Least squares minimization**: Find optimal β_0^* and β_1^* to

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} \overbrace{|y_i - (\beta_0 + \beta_1 x_i)|}^{\text{vertical discrepancy}}{}^2$$

Vector Formulation

- Write the response, intercept, and explanatory data as vectors in $\mathbb{R}^n$

$$\vec{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \vec{\mathbf{1}} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \quad \vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

$$\sum_{i=1}^{n} |y_i - (\beta_0 + \beta_1 x_i)|^2$$

$$= \left\| \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_0 \end{pmatrix} - \beta_1 \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \right\|^2$$

$$= \left\| \vec{y} - \left(\beta_0 \vec{\mathbf{1}} + \beta_1 \vec{x}\right) \right\|^2$$

- Recall L_2-norm $\|\vec{v}\|^2 = \sum_{i=1}^{d} |v_i|^2$

Use Orthogonal Projection?

- To minimize the quantity

$$\min_{\beta_0, \beta_1} \left\| \vec{y} - \left(\beta_0 \vec{\mathbf{1}} + \beta_1 \vec{x}\right) \right\|$$

- **Geometric perspective:** Project $\vec{y}$ orthogonally onto both $\vec{\mathbf{1}}$ and $\vec{x}$ simultaneously to obtain the optimal scaling β_0^* and β_1^*

- **Vector space perspective:** Find the vector y^* from $\operatorname{span}\left(\vec{\mathbf{1}}, \vec{x}\right)$ that is closest to the given vector $\vec{y}$

$$\min_{y^* \in \operatorname{span}(\vec{\mathbf{1}}, \vec{x})} \|\vec{y} - y^*\|$$

Linear Regression Geometry

- Least squares approximation is

$$\min_{\beta_0,\beta_1} \|\vec{y} - (\beta_0 \vec{\mathbf{1}} + \beta_1 \vec{x})\|$$

$$\min_{y^* \in \text{span}(\vec{\mathbf{1}},\vec{x})} \|\vec{y} - y^*\|$$

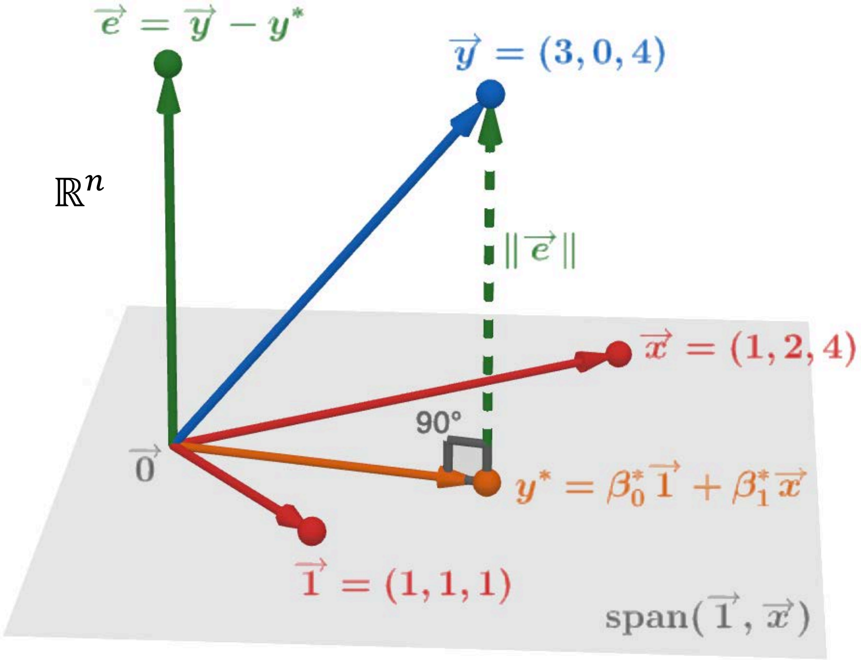
- Project $\vec{y}$ orthogonally onto the subspace spanned by $\vec{\mathbf{1}}$ and $\vec{x}$

$$\vec{y}^* = \text{proj}_{\text{span}(\vec{\mathbf{1}},\vec{x})}\, \vec{y} = \beta_0^* \vec{\mathbf{1}} + \beta_1^* \vec{x}$$

- Residual vector

$$\vec{e} = \vec{y} - \vec{y}^*$$

is perpendicular to the fitted value / prediction vector $\vec{y}^*$



Multiple Linear Regression

- n data points but $p > 1$ explanatory variables
- Each i^{th} observed data:

$$y_i \in \mathbb{R} \qquad \vec{x}_i = (x_{i1}, \dots, x_{ip}) \in \mathbb{R}^p$$

- Model equation:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

for $i = 1, \dots, n$

Matrix Form

- Rewrite together as

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \approx \beta_0 + \beta_1 \begin{pmatrix} x_{11} \\ \vdots \\ x_{n1} \end{pmatrix} + \cdots + \beta_p \begin{pmatrix} x_{1p} \\ \vdots \\ x_{np} \end{pmatrix}$$

$$= \underbrace{\begin{pmatrix} | & x_{11} & \dots & x_{1p} \\ \vec{\mathbf{1}} & \vdots & & \vdots \\ | & x_{n1} & \dots & x_{np} \end{pmatrix}}_{\textbf{design matrix } X \in \mathbb{R}^{n \times (1+p)}} \times \underbrace{\begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}}_{\vec{\beta} \in \mathbb{R}^{1+p}}$$

- Rewrite the minimization in matrix form

$$\boxed{\min_{\beta_0, \dots, \beta_p} \sum_{i=1}^{n} \left| y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \right|^2 = \min_{\vec{\beta}} \left\| \vec{y} - X\vec{\beta} \right\|^2}$$

Matrix Calculus

- **Question:** How to solve the least squares problem for $\vec{\beta}^*$ algebraically?

$$\min_{\vec{\beta}} \left\|\vec{y} - X\vec{\beta}\right\|^2$$

- **Step 1:** Rewrite as

$$\begin{aligned} f(\vec{\beta}) &= \left\|\vec{y} - X\vec{\beta}\right\|^2 \\ &= \left(\vec{y} - X\vec{\beta}\right)^T \left(\vec{y} - X\vec{\beta}\right) \\ &= \left(\vec{y}^T - \vec{\beta}^T X^T\right)\left(\vec{y} - X\vec{\beta}\right) \\ &= \underbrace{\vec{\beta}^T X^T X\vec{\beta} - 2\vec{\beta}^T X^T \vec{y} + \vec{y}^T \vec{y}}_{\text{quadratic function in variable } \vec{\beta}} \end{aligned}$$

- **Step 2:** To minimize $f(\vec{\beta})$, use the first-order condition

$$\frac{\partial}{\partial \vec{\beta}} f(\vec{\beta}) = \vec{0}$$

In 1-dimensional case,

$$\frac{d}{dx}(ax^2 + bx + c) = 2ax + b$$

In high-dimensional case, use matrix calculus

$$\begin{aligned} &\frac{\partial}{\partial \vec{\beta}}\left(\vec{\beta}^T X^T X\vec{\beta} - 2\vec{\beta}^T X^T \vec{y} + \vec{y}^T \vec{y}\right) \\ &= 2X^T X\vec{\beta} - 2X^T \vec{y} \end{aligned}$$

Normal Equation

- **Step 3:** A solution $\vec{\beta}^*$ of

$$\min_{\vec{\beta}} \left\|\vec{y} - X\vec{\beta}\right\|^2$$

is a solution of the **normal equation**

$$X^T X \vec{\beta} = X^T \vec{y}$$

$$\underbrace{\underset{(1+p)\times n}{X^T}\;\underset{n\times(1+p)}{X}}_{A \in \mathbb{R}^{(1+p)\times(1+p)}}\;\underset{(1+p)\times 1}{\vec{\beta}} = \underbrace{\underset{(1+p)\times n}{X^T}\;\underset{n\times 1}{\vec{y}}}_{\vec{b} \in \mathbb{R}^{1+p}}$$

- This is a system of $1 + p$ linear equations in $1 + p$ unknown variables $\vec{\beta}$

Formula for $\vec{\beta}^*$

- If $X^T X \in \mathbb{R}^{(1+p)\times(1+p)}$ is invertible, then the linear regression has unique **fitted model parameters**

$$\vec{\beta}^* = (X^T X)^{-1} X^T \vec{y}$$

- Fitted values / predictions

$$\begin{aligned}\vec{y}^* &= \operatorname{proj}_{\operatorname{range}(X)} \vec{y} \\ &= X\vec{\beta}^* \\ &= \underbrace{X(X^T X)^{-1} X^T}_{\text{subspace projection formula}} \vec{y}\end{aligned}$$

- **Question:** But when is $X^T X$ invertible?

Full Rank Design Matrix

- **Fact:** Two matrices $A \in \mathbb{R}^{r \times c}$ and $A^T A \in \mathbb{R}^{c \times c}$ always have the same rank

- The matrix

$$X^T X \in \mathbb{R}^{(1+p)\times(1+p)}$$

is invertible iff it has full rank $1 + p$ iff the design matrix

$$X \in \mathbb{R}^{n\times(1+p)}$$

also has rank $1 + p$

- In this case, no explanatory variable is redundant i.e. no multicollinearity

Model Identifiability

- A multiple linear regression with p explanatory variables will have unique fitted model parameter

$$\vec{\beta}^* = (X^T X)^{-1} X^T \vec{y}$$

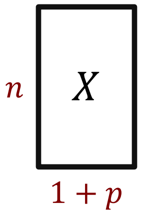
if and only if the $1 + p$ column vectors of the design matrix

$$X = \begin{pmatrix} | & x_{11} & \cdots & x_{1p} \\ \vec{\mathbf{1}} & \vdots & & \vdots \\ | & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

are linearly independent

Large Data Set $n \gg p$

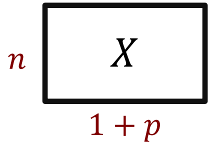
- In practice, we usually have a lot of data but only a few explanatory variables i.e. $n \gg p$
- The design matrix is "tall"



- The few column vectors in $\mathbb{R}^n$ are likely to be linearly independent. So X has full rank $1 + p$
- Unique fitted parameter $\vec{\beta}^*$

High Dimensional Data $n \ll p$

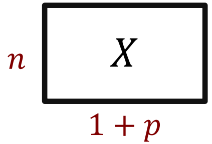
- Sometimes, the explanatory variables are high-dimensional (e.g. genomics with few patients) i.e. $n \ll p$
- The design matrix is "wide"



- It can only have up to rank $n < 1 + p$
- No unique fitted parameter $\vec{\beta}^*$
- Such linear regression model might predict well but lack model explainability

Overfitting $n \leq p$

- If we have more explanatory variables than the number of data i.e. $n \leq p$



the $1 + p$ column vectors in $\mathbb{R}^n$ are likely to span $\mathbb{R}^n$ i.e.

$$\text{range}(X) = \mathbb{R}^n$$

- Model fits the data perfectly

$$y^* = \text{proj}_{\text{range}(X)}\, \vec{y} = \vec{y}$$

with no error $\vec{e} = \vec{y} - y^* = \vec{0}$

Weighted Least Squares

- If each data error is weighted differently

$$\sum_{i=1}^{n} w_i \left| y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \right|^2$$

then the normal equation is modified as

$$X^T W X \vec{\beta} = X^T W \vec{y}$$

with diagonal matrix

$$W = \begin{pmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_n \end{pmatrix} \in \mathbb{R}^{n \times n}$$

- Application: “Iteratively reweighted least squares” algorithm to fit GLM model

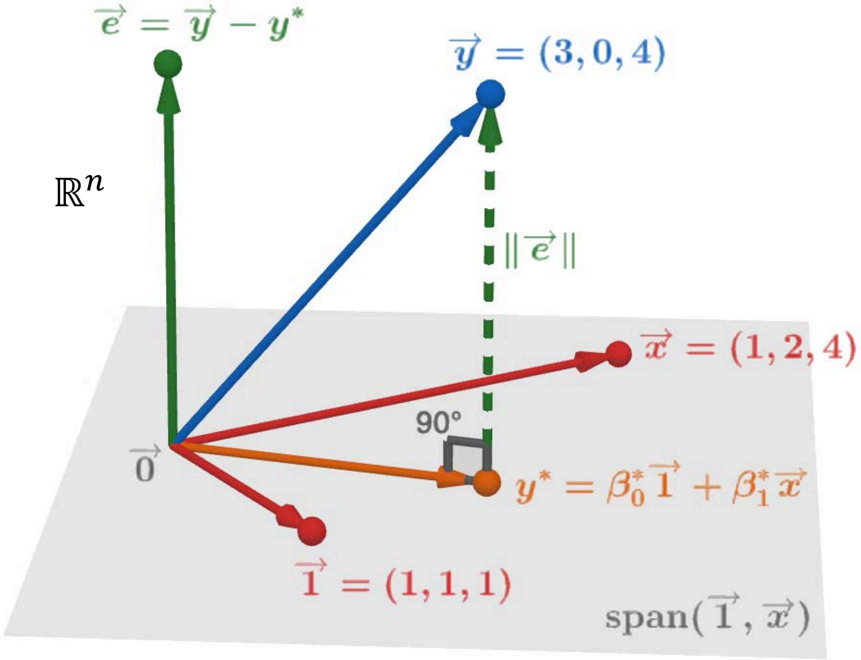
Model Intercept and Sum of Residuals

- If a linear regression model has an intercept term

$$y = \beta_0 + \beta_1 x_1 + \cdots$$

then

1. Design matrix X has a column of $\vec{\mathbf{1}}$ vector
2. Residual vector $\vec{e} \perp \text{range}(X)$ so we have $\vec{e} \perp \vec{\mathbf{1}}$
3. Dot product $\vec{e} \cdot \vec{\mathbf{1}} = 0$
4. Sum of the residuals $\sum_{i=1}^{n} e_i = 0$
5. Model does not over/under-estimate on average



Linear Transformation

- A **linear transformation** / function

$$f : \mathbb{R}^d \mapsto \mathbb{R}^n$$

satisfies

1. Additivity: For any $\vec{u}, \vec{v} \in \mathbb{R}^d$

$$f(\vec{u} + \vec{v}) = f(\vec{u}) + f(\vec{v})$$

2. Homogeneity: For any $\vec{v} \in \mathbb{R}^d$ and $a \in \mathbb{R}$

$$f(a\vec{v}) = af(\vec{v})$$

- Notation: Write $T(\vec{v})$ instead of $f(\vec{v})$

Examples of Not Linear

- Here are some functions $f : \mathbb{R} \mapsto \mathbb{R}$ that are not linear transformation

"parabolic curve"

e.g. $f(x) = x^2$

$$f(5 \times 2) = (5 \times 2)^2 = 100$$

$$5f(2) = 5(2)^2 = 20$$

"line not through origin"

e.g. $f(x) = 2x + 1$

$$f(3 + 4) = 2(3 + 4) + 1 = 15$$

$$f(3) + f(4) = 2(3) + 1 + 2(4) + 1 = 16$$

Linear Transformation = Matrix-Vector Multiplication

e.g. Let $T : \mathbb{R}^2 \mapsto \mathbb{R}^2$ be

$$T(x, y) = (y, x)$$

i.e. swapping the 2 coordinates

- Check it is a linear transformation i.e. additivity and homogeneity
- Can view T as a matrix multiplication

$$T(x, y) = \begin{pmatrix} 0x + 1y \\ 1x + 0y \end{pmatrix}$$

$$= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

"permutation matrix"

- Let $A \in \mathbb{R}^{n \times d}$ be a matrix. The matrix-vector multiplication

$$T(\vec{v}) = A\vec{v}$$

is a linear transformation $T : \mathbb{R}^d \mapsto \mathbb{R}^n$

- Moreover, any linear transformation $T : \mathbb{R}^d \mapsto \mathbb{R}^n$ must be a matrix-vector multiplication by some matrix $A \in \mathbb{R}^{n \times d}$

e.g. $T(x, y, z) = (ax + by + cz, dx + ey + fz)$

$$A = \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix}$$

- $T: \mathbb{R}^d \mapsto \mathbb{R}^d$ is an **invertible** function when $A \in \mathbb{R}^{d \times d}$ is an invertible matrix

Linear Transformation Properties

- If $T : \mathbb{R}^d \mapsto \mathbb{R}^n$ is a linear transformation:

  1. Preserve origin
  $$T(\vec{0}) = \vec{0}$$
  2. For any $\vec{v} \in \mathbb{R}^d$
  $$T(-\vec{v}) = -T(\vec{v})$$
  3. For any $\vec{u}, \vec{v} \in \mathbb{R}^d$
  $$T(\vec{u} - \vec{v}) = T(\vec{u}) - T(\vec{v})$$
  4. Preserve linear combination
  $$\begin{aligned} T(a_1\vec{v}_1 + \cdots + a_k\vec{v}_k) \\ = a_1T(\vec{v}_1) + \cdots + a_kT(\vec{v}_k) \end{aligned}$$

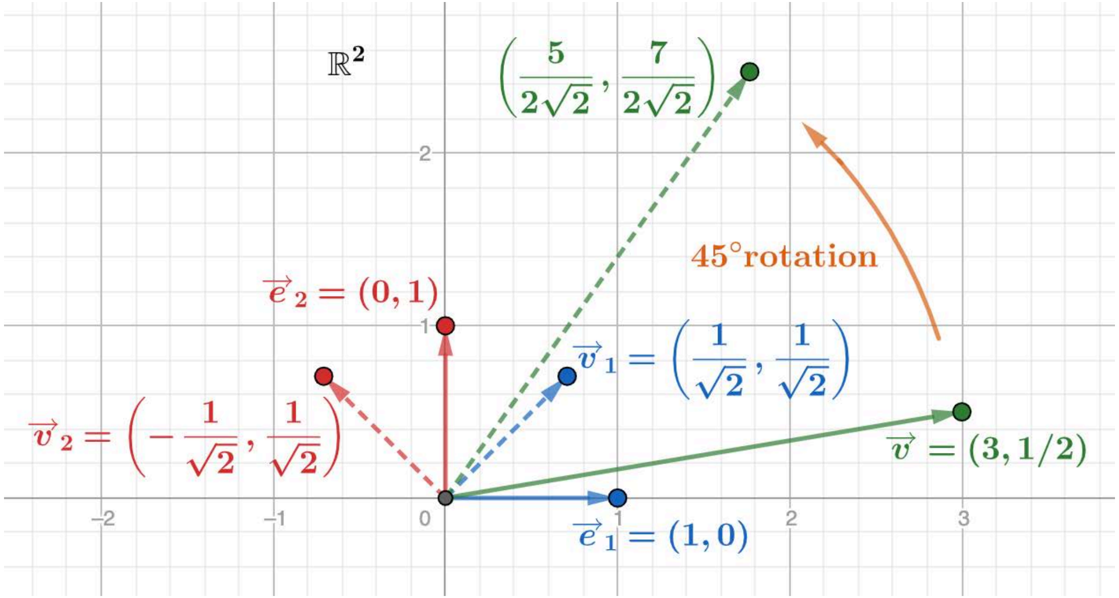
- If we know how T transforms each of
$$\vec{v}_i \mapsto T(\vec{v}_i)$$
then we already know how it transforms any
$$\vec{v} = a_1\vec{v}_1 + \cdots + a_k\vec{v}_k$$

- Conceptually, linear combination "passes through" linear transformation intact

Linear Transformation By Orthogonal Matrix

e.g. Let $T : \mathbb{R}^2 \mapsto \mathbb{R}^2$ be $T(\vec{v}) = Q\vec{v}$ with orthogonal matrix

$$Q = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$$



- Transform standard basis

$$Q\vec{e}_1 = Q\begin{pmatrix}1\\0\end{pmatrix} = \begin{pmatrix}1/\sqrt{2}\\1/\sqrt{2}\end{pmatrix}$$

$$Q\vec{e}_2 = Q\begin{pmatrix}0\\1\end{pmatrix} = \begin{pmatrix}-1/\sqrt{2}\\1/\sqrt{2}\end{pmatrix}$$

- By linearity

$$\begin{aligned} Q\begin{pmatrix}3\\1/2\end{pmatrix} &= Q(3\vec{e}_1 + 1/2\vec{e}_2) \\ &= 3(Q\vec{e}_1) + 1/2(Q\vec{e}_2) \\ &= \begin{pmatrix}5/2\sqrt{2}\\7/2\sqrt{2}\end{pmatrix} \end{aligned}$$

Orthogonal Transformation = Rigid = Rotation or Reflection

- Linear transformation $T(\vec{v}) = Q\vec{v}$ by orthogonal matrix

$$Q = \underbrace{\begin{pmatrix} | & & | \\ \vec{v}_1 & \dots & \vec{v}_d \\ | & & | \end{pmatrix}}_{\text{orthonormal basis}} \in \mathbb{R}^{d\times d}$$

- It transforms standard basis $\{\vec{e}_1, \dots, \vec{e}_d\}$ to the orthonormal basis $\{\vec{v}_1, \dots, \vec{v}_d\}$

$$\vec{e}_i \mapsto T(\vec{e}_i) = Q\vec{e}_i = \vec{v}_i$$

i.e. **change of coordinate system**

- **Fact:** Orthogonal transformation $\vec{x} \mapsto Q\vec{x}$ preserves dot product

$$\begin{aligned} T(\vec{u}) \cdot T(\vec{v}) &= (Q\vec{u}) \cdot (Q\vec{v}) \\ &= \vec{u}^T Q^T Q \vec{v} = \vec{u} \cdot \vec{v} \end{aligned}$$

Hence it preserves norm $\|Q\vec{v}\| = \|\vec{v}\|$, distance, and angle too

- Such a **rigid** transformation must be either a rotation or a reflection

- Orthogonal transformation $\vec{x} \mapsto Q\vec{x}$ is either a **rotation** or a **reflection**

- Application: Rotate data points in principal component analysis PCA

Matrix Multiplication = Composition of Linear Transformations

- Let two linear transformations

$$T_1 : \mathbb{R}^d \mapsto \mathbb{R}^k \quad \text{and} \quad T_2 : \mathbb{R}^k \mapsto \mathbb{R}^n$$
$$A_1 \in \mathbb{R}^{k \times d} \qquad\qquad A_2 \in \mathbb{R}^{n \times k}$$

- **Fact:** Their composition $T = T_1 \circ T_2$

$$T : \mathbb{R}^d \mapsto \mathbb{R}^k \mapsto \mathbb{R}^n$$

$$\vec{v} \mapsto T_1(\vec{v}) \mapsto \underbrace{T_2(T_1(\vec{v}))}_{(T_1 \circ T_2)(\vec{v})}$$

is also a linear transformation

- **Question:** What matrix A corresponds to the linear transformation T?

- By matrix multiplication, we have

$$\vec{v} \mapsto A_1\vec{v} \mapsto A_2(A_1\vec{v}) = (A_2 \times A_1)(\vec{v})$$

Thus the matrix corresponds to T must be

$$A = A_2 \times A_1 \in \mathbb{R}^{n \times d}$$

- Matrix multiplication is defined in such a (complicated) way just to make the math of linear transformation composition works!

- In general, composition is not commutative

$$T_1 \circ T_2 \neq T_1 \circ T_2 \quad \text{so} \quad A_2A_1 \neq A_1A_2$$