Warmup: Suppose $f(w) = 3w^2 - 5w + 2$

what value of w minimizes f?

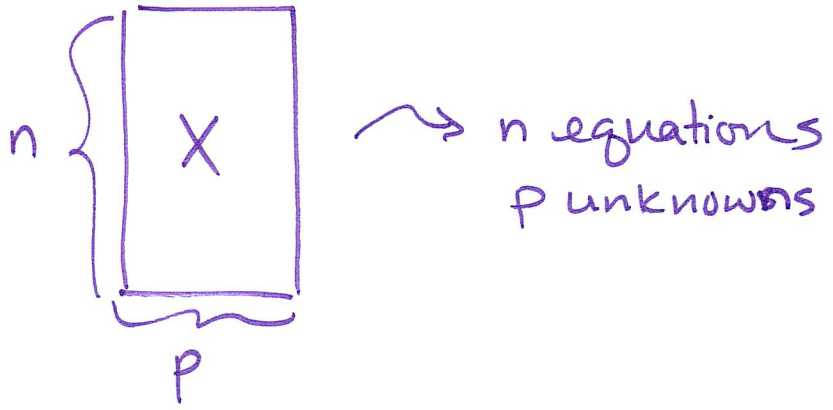
$$\frac{df}{dw} = 6w - 5 = 0$$

$$w = 5/6$$

Lecture 5: Least Squares

given:
vector of labels $y \in \mathbb{R}^n$

matrix of features $X \in \mathbb{R}^{n \times p}$



want:
vector of weights $\underline{w} \in \mathbb{R}^p$

assume:
$n \geq p$, $\text{Rank}(X) = p$ (have p linearly independent columns)

if $\underline{y} = X\underline{w}$, then have system of n linear equations

i^{th} equation $y_i = w_1 x_{i1} + w_2 x_{i2} + \cdots + w_p x_{ip}$

$$= \sum_{j=1}^{p} w_j x_{ij} = \langle \underline{w}, \underline{x}_{\cdot i} \rangle$$

$\underline{x}_{\cdot i}$ ← i^{th} row of X

In general, $\underline{y} \neq X\underline{w}$ for any $\underline{w}$

(because modeling errors, noise)

Define residual $r_i = y_i - \langle \underline{w}, \underline{x}_{\cdot i} \rangle$

Goal: find $\underline{w}$ to minimize $\sum_{i=1}^{n} |r_i|^2$ (sum of squared residuals / errors)

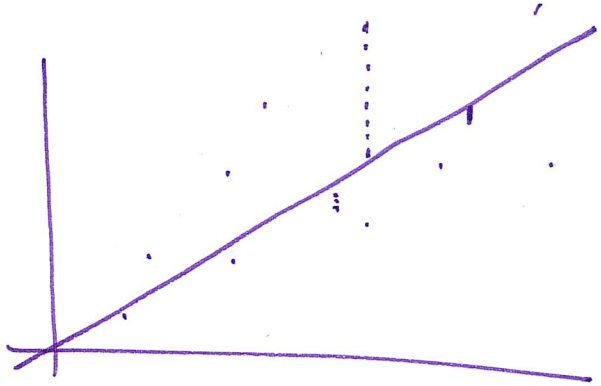
Why sum of squared errors?

a) magnify effect of large errors

b) make math easy (can compute derivatives)

c) nice geometric interpretation

d) coincides w/ modeling $\underline{y} = X\underline{w} + \underline{\varepsilon}$, $\underline{\varepsilon}$ = Gaussian noise (not this course)

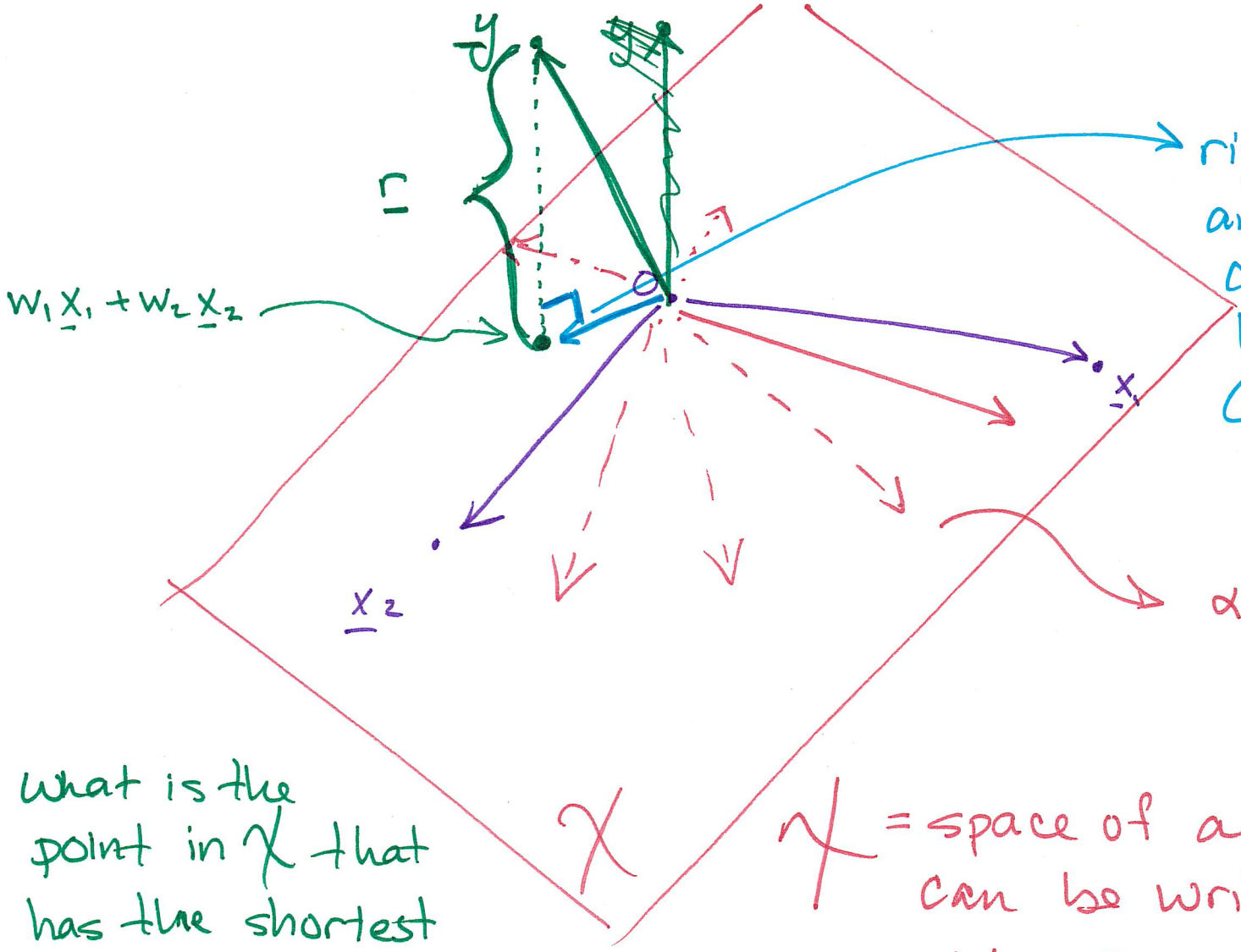


First, Geometry.

$p = 2, \quad n = 3$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \approx \begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \end{bmatrix} w_1 + \begin{bmatrix} x_{12} \\ x_{22} \\ x_{32} \end{bmatrix} w_2 \quad , \qquad \underline{r} = \underline{y} - w_1 \underline{x}_1 - w_2 \underline{x}_2$$

↑ 1st col of X



right angle because any other angle would correspond to a longer distance (bigger r)

$\alpha \underline{x}_1 + \beta \underline{x}_2$ = weighted combo x_1 and x_2

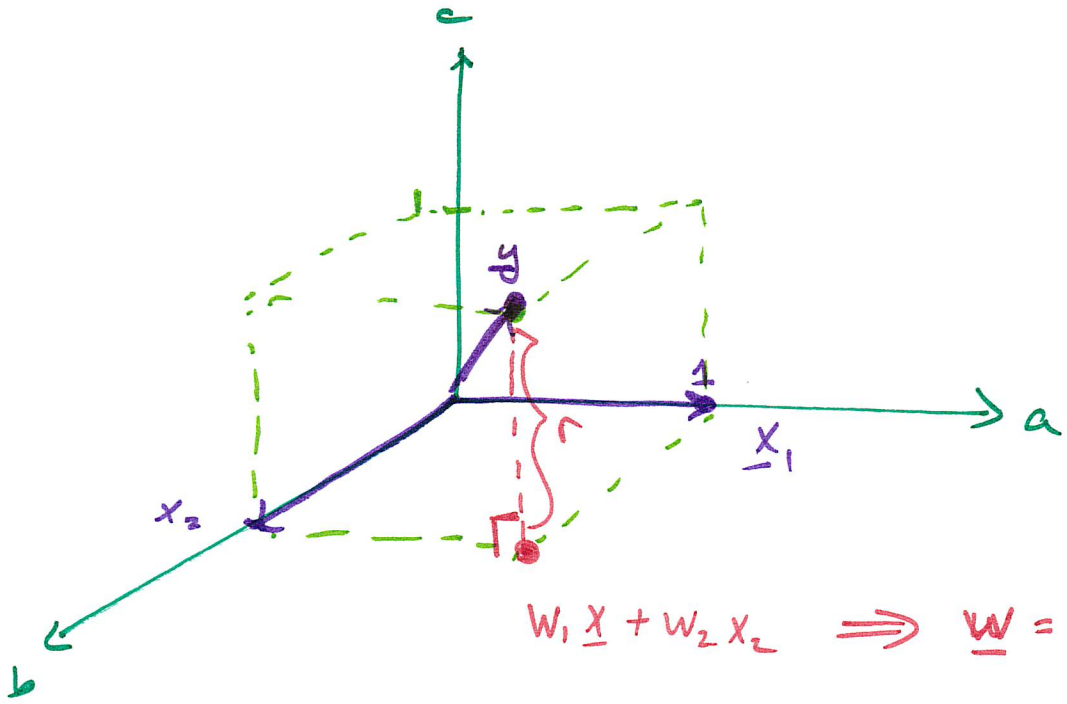
What is the point in χ that has the shortest distance to y?

χ = space of all vectors that can be written as $\alpha \underline{x}_1 + \beta \underline{x}_2$ for some $\alpha, \beta \in \mathbb{R}$

= span of cols of X

(y may not live in this space)

ex) $\underline{x}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}_{c}^{b}{}^{a}$, $\underline{x}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$, $\underline{y} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$



$w_1 \underline{x} + w_2 x_2 \Rightarrow \underline{w} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $X\underline{w} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$

Algebraic approach:

goal: choose $\underline{w}$ to minimize $\sum_{i=1}^{n} r_i^2$

$$\hat{\underline{w}} = \arg\min_{\underline{w}} \sum_{i=1}^{n} r_i^2$$

↑ argument that minimizes objective

$$= \arg\min_{\underline{w}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} w_j x_{ij} \right)^2$$

$$= \arg\min_{\underline{w}} \| \underline{r} \|_2^2$$

$$= \arg\min_{\underline{w}} \underbrace{\| \underline{y} - X \underline{w} \|_2^2}_{f(\underline{w})}$$

$$\left(\sum_i a_i^2 \right)^{1/2} = \|a\|_2$$

Approach 1: use geometry.

know $\hat{r} = y - X\hat{w}$ is orthogonal / perpendicular to plane / span of cols of X.

i^{th} col of X

$X_i^T \hat{r} = 0$ for each i

two vectors u, v are orthogonal if $\langle u, v \rangle = u^T v = 0$

$\hat{r} \in \mathbb{R}^n$

$X_i \in \mathbb{R}^n$ (col of X)

$X_{i\cdot} \in \mathbb{R}^p$ (row of X)

$$X^T \hat{r} = 0$$

$$\Rightarrow X^T(y - X\hat{w}) = 0$$

$\Rightarrow$ $\hat{w}$ is solution to linear system of eqns $X^T y = X^T X \hat{w}$

Matrix inverse:

for a square matrix A, it's inverse A^{-1} is a square matrix satisfies

$$AA^{-1} = A^{-1}A = I = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}$$

e.g. $A = \begin{bmatrix} 1/4 & 0 \\ 0 & 2 \end{bmatrix} \Rightarrow A^{-1} = \begin{bmatrix} 4 & 0 \\ 0 & 1/2 \end{bmatrix}$

$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \Rightarrow A^{-1} = \begin{bmatrix} -2 & 1 \\ 3/2 & -1/2 \end{bmatrix}$

not all matrices have inverses:

$A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \Rightarrow$ no inverse

Back to Least squares:

$\hat{w}$ satisfies $X^T y = X^T X \hat{w}$

so if $X^T X$ is invertible ($(X^T X)^{-1}$ exists)

then $\hat{w} = (X^T X)^{-1} X^T y$

Approach 2: use vector calculus

if we have a vector w and a function of it $f(w)$

then the **gradient** (vector version of derivative) is

$$\nabla_w f(w) = \begin{bmatrix} df(w)/dw_1 \\ df(w)/dw_2 \\ \vdots \\ df(w)/dw_r \end{bmatrix} = \text{vector same size as } w$$

ex. $f(\underline{w}) = \langle \underline{a}, \underline{w} \rangle = \underline{w}^T \underline{a} = \sum_{i=1}^{p} w_i a_i$

$$\nabla_{\underline{w}} \; \underline{w}^T \underline{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \underline{a}$$

(analog $f(w) = aw$, $\frac{df}{dw} = a$)

$\nabla_{\underline{w}} \; \underline{w}^T A \underline{w} = A\underline{w} + A^T \underline{w}$; if A is symmetric ($A = A^T$), then

$\uparrow$ pxp matrix

$$\nabla_{\underline{w}} \; w^T A w = 2Aw$$

(analog $f(w) = aw^2$, $\frac{df}{dw} = 2aw$)

Least squares:

$$\hat{w} = \arg\min_{w} \underbrace{\|y - Xw\|_2^2}_{f(w)}$$

$$\|a\|_2^2 = a^T a$$

$$f(w) = (y - Xw)^T (y - Xw)$$

$$= y^T y - \underbrace{y^T X w}_{= w^T X^T y} - w^T X^T y + w^T X^T X w$$

$$= y^T y - 2 w^T \underbrace{X^T y}_{a} + w^T \underbrace{X^T X}_{A} w$$

$X^T X$ is symmetric for any X

$$\nabla_w f(w) = 0 - 2X^T y + 2X^T X w = 0$$

$\hat{w}$ solves $X^T y = X^T X w$

$\hat{w} = (X^T X)^{-1} X^T y$ if $(X^T X)^{-1}$ exists.

Does $(X^TX)^{-1}$ exist?

Thm 3.10: If columns of X are linearly independent, then X^TX is invertible / full rank / non-singular and there exists a least squares solution

$$\underline{\hat{w}} = (X^TX)^{-1}X^T\underline{y}$$

Proof:

X^TX is symmetric, so X^TX is invertible if it is positive definite — i.e. for all $\underline{z} \neq 0$,

$$\underline{z}^TX^TX\underline{z} > 0$$

Let $\underline{\tilde{z}} = X\underline{z}$. then $\underbrace{\underline{z}^TX^T}_{\underline{\tilde{z}}^T}\underbrace{X\underline{z}}_{\underline{\tilde{z}}} = \underline{\tilde{z}}^T\underline{\tilde{z}} = \|\underline{\tilde{z}}\|_2^2$

Recall:

$\|\tilde{z}\|_2 \geq 0$ for all $\tilde{z}$, and $\|\tilde{z}\|_2$ if and only if $\tilde{z} = 0$

now $\tilde{z} = X z$. can $\tilde{z} = 0$ when $z \neq 0$?

no. because $\tilde{z}$ = weighted sum of cols of X
we assumed cols of X were linearly independent
therefore $\tilde{z} = 0$ implies $z = 0$

thus $z^T X^T X z > 0$ for all $z \neq 0$.

$\Rightarrow$ $X^T X$ is invertible.

Is $\hat{w}$ the Least squares solution?

if it is, then for any (other) w

$$\|y - X\hat{w}\|_2^2 \leq \|y - Xw\|_2^2 \iff \|\hat{r}\|_2^2 \leq \|r\|_2^2$$

Let $\hat{r} = y - X\hat{w}$, ⇒

$$r = y - Xw = \underbrace{y - X\hat{w}}_{\hat{r}} + X\hat{w} - Xw$$

$$r = \hat{r} + X(\hat{w} - w)$$

$$\|r\|_2^2 = r^T r = (\hat{r} + X(\hat{w}-w))^T (\hat{r} + X(\hat{w}-w))$$

$$= \hat{r}^T\hat{r} + \underbrace{\hat{r}^T X}_{=0}(\hat{w}-w) + (\hat{w}-w)^T\underbrace{X^T\hat{r}}_{=0} + \underbrace{(\hat{w}-w)^T X^T X(\hat{w}-w)}_{\geq 0 \text{ b/c } X^TX \text{ p.d.}}$$

$$\geq \|\hat{r}\|^2 + 0 + 0 \quad \star$$

$\|r\|^2 \geq \|\hat{r}\|^2$, and only equal if $\hat{w} = w \Rightarrow \hat{w} = (X^TX)^{-1}X^Ty$ is LS soln