Geometry of the Squared-Error Surface

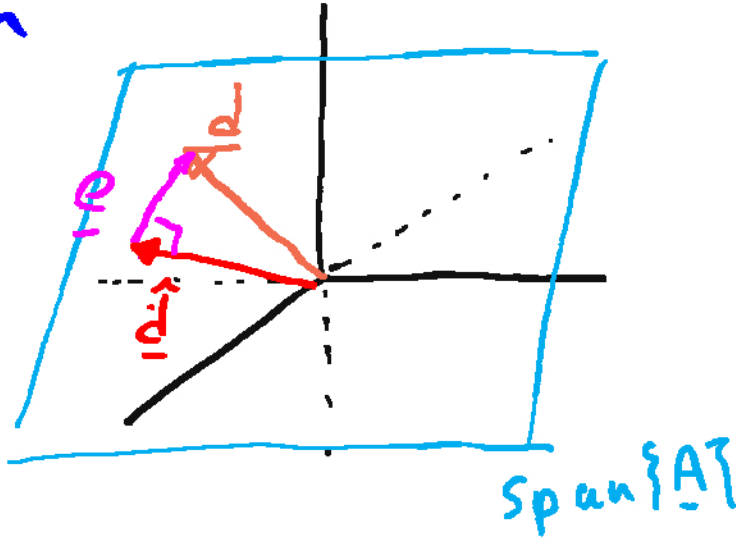
Objectives

- visualize squared-error cost function $f(\underline{w})$

- special cases

- general case

Squared Error Cost Function



$$\min_{\underline{w}} \|\underline{A}\,\underline{w} - \underline{d}\|_2^2 \Rightarrow \min_{\underline{w}} f(\underline{w})$$

N features, P parameters, N labels

- geometry of error $\underline{e} = \underline{d} - \underline{A}\,\underline{w}$
  N-dimensional space

- geometry of $f(\underline{w})$ in P-dimensional space

★ $$f(\underline{w}) = (\underline{w} - \underline{w}_0)^T \underline{A}^T \underline{A} (\underline{w} - \underline{w}_0) + \underline{d}^T \underline{P}_{A_\perp} \underline{d}$$

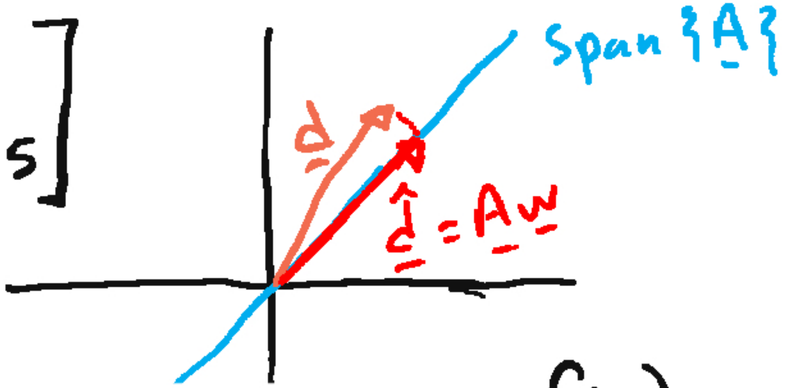
$$\underline{w}_0 = (\underline{A}^T \underline{A})^{-1} \underline{A}^T \underline{d} \qquad \underline{P}_{A_\perp} = \underline{I} - \underbrace{\underline{A}(\underline{A}^T\underline{A})^{-1}\underline{A}^T}_{P_A}$$

$$\underline{A}^T\underline{A} > 0 \Rightarrow f(\underline{w}) \geq f(\underline{w}_0) = \underline{d}^T \underline{P}_{A_\perp} \underline{d}$$

↓ The result when the first terms are 0 $(w = w_0)$

Example: $\underline{A} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $\underline{d} = \begin{bmatrix} 1 \\ 1.5 \end{bmatrix}$

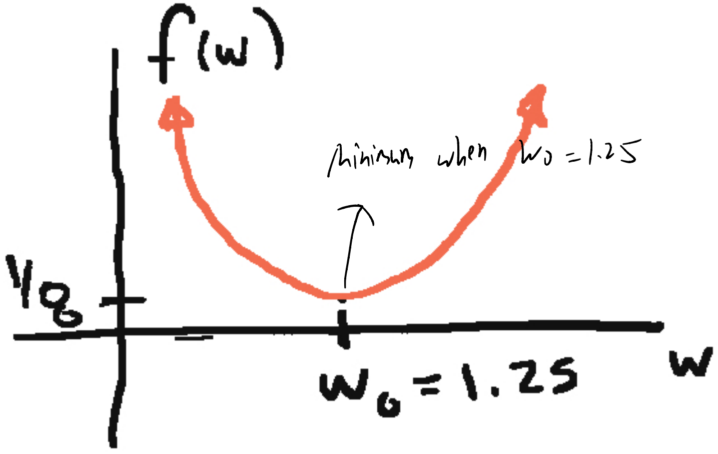
$$w_0 = (\underline{A}^T\underline{A})^{-1}\underline{A}^T\underline{d}$$
$$= \tfrac{1}{2}(2.5) = 1.25$$



$$f(\underline{w}) = (\underline{w}-\underline{w}_0)^T\underline{A}^T\underline{A}(\underline{w}-\underline{w}_0) + \underline{d}^T\underline{P}_{A\perp}\underline{d}$$

(Sum of squared errors)

$$= 2(w-w_0)^2 + \frac{1}{8}$$ → Cost of function



Example: $\underline{A} = \begin{bmatrix} 3 & 0 \\ 0 & 5 \\ -4 & 0 \end{bmatrix}$, $\underline{d} = \begin{bmatrix} 0 \\ 5 \\ 12.5 \end{bmatrix}$

$\underline{A}^T\underline{A} = \begin{bmatrix} 25 & 0 \\ 0 & 25 \end{bmatrix}$ $\quad \underline{w}_0 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$

$$f(\underline{w}) = 25(\underline{w}-\underline{w}_0)^T(\underline{w}-\underline{w}_0) + \underline{d}^T\underline{P}_{A\perp}\underline{d}$$
$$= 25(w_1+2)^2 + 25(w_2-1)^2 + \underline{d}^T\underline{P}_{A\perp}\underline{d}$$

contours of constant $f(\underline{w})$

$$25(w_1+2)^2 + 25(w_2-1)^2 = c^2$$

circle of radius $c/5$

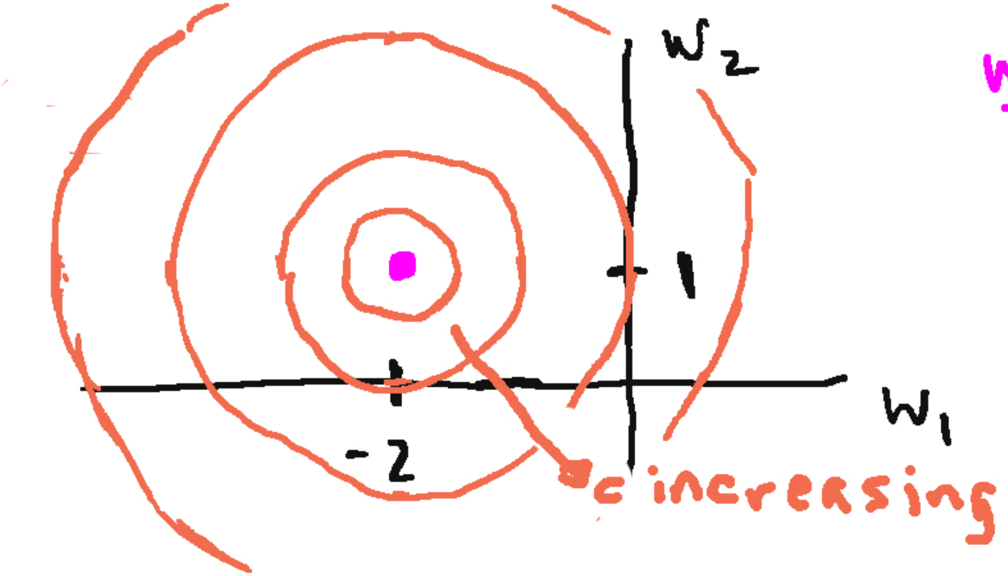
centered at $(-2,1)$

Example (con't):

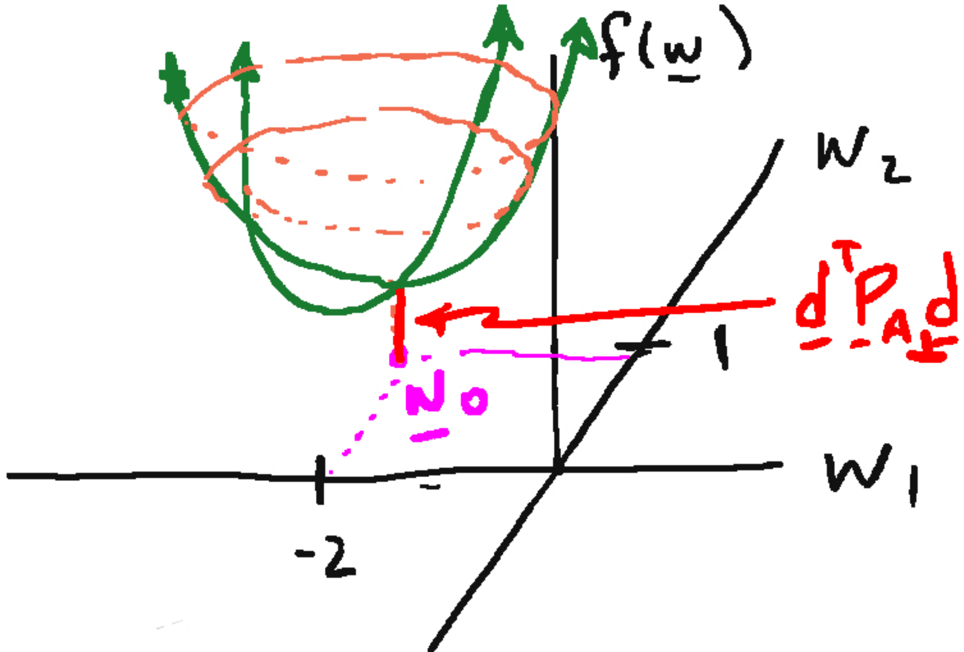
$$f(\underline{w}) = 25(w_1+2)^2 + 25(w_2-1)^2 + \underline{d}^T \underline{P}_A^{\perp} \underline{d}$$

$$\underline{w}_0 = \begin{bmatrix} -2 \\ 1 \end{bmatrix} \qquad 25\left[(w_1+2)^2 + (w_2-1)^2\right] = c^2$$

circular contours



Cross sections at fixed w_1 or w_2 are parabolas



Bowl shaped surface

- $f(\underline{w})$ = const are circles
- parabolic in each coordinate ($\underline{w}$)

Example: $\underline{A}^T\underline{A} = \begin{bmatrix} \lambda_1^2 & 0 \\ 0 & \lambda_2^2 \end{bmatrix}$, $\lambda_1^2 > \lambda_2^2$

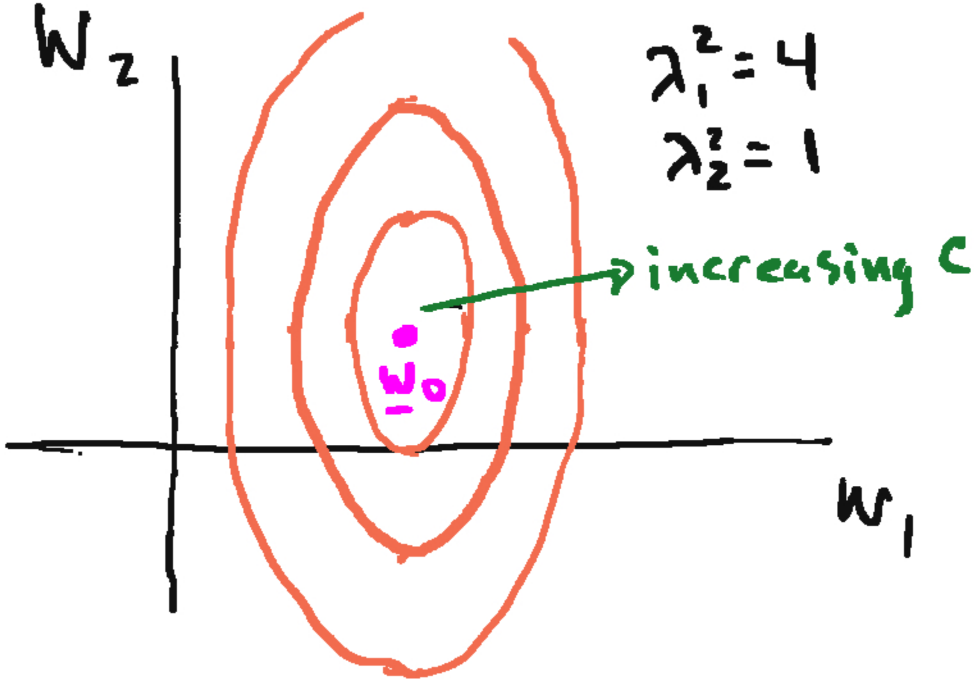
$$f(\underline{w}) = (\underline{w}-\underline{w}_0)^T \underline{A}^T\underline{A}(\underline{w}-\underline{w}_0) + \underline{d}^T \underline{P}_{A\perp}\underline{d}$$

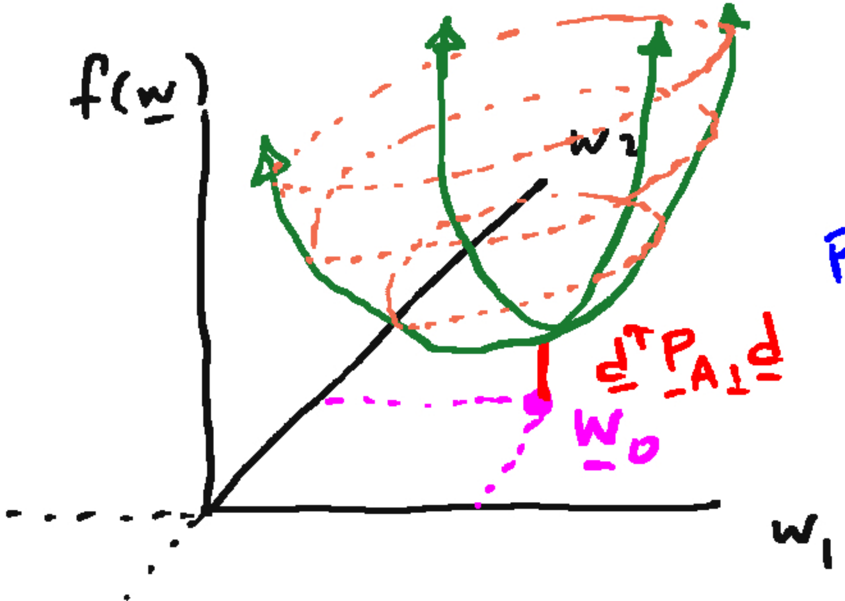
Consider $(\underline{w}-\underline{w}_0)^T \underline{A}^T\underline{A}(\underline{w}-\underline{w}_0) = \text{const}$

$$\lambda_1^2 (w_1 - w_{10})^2 + \lambda_2^2 (w_2 - w_{20})^2 = c^2$$

★ Ellipse: center $\underline{w}_0$, major axis $\frac{2c}{\lambda_2}$ 轴长, minor axis $\frac{2c}{\lambda_1}$ 轴长

Fix w_1 or w_2: $(\underline{w}-\underline{w}_0)^T \underline{A}^T\underline{A}(\underline{w}-\underline{w}_0)$ is a parabola





elliptical cross sections parallel to w_1-w_2 plane

parabolic ⊥ to w_1-w_2 plane

What if $\underline{A}^T\underline{A}$ is not diagonal?

☆ Can always write: $\underline{A}^T\underline{A} = \underline{U}\,\underline{\Lambda}^2\underline{U}^T$, $\underline{\Lambda}^2 = \text{diag}\{\lambda_1^2, \lambda_2^2 \ldots \lambda_p^2\}$

columns of $\underline{U}$ are orthonormal: $\underline{U} = [\underline{u}_1, \underline{u}_2 \cdots \underline{u}_p]$

$$\underline{u}_\ell^T \underline{u}_k = \begin{cases} 1 & \ell = k \\ 0 & \ell \neq k \end{cases}$$

Eigendecomposition (more later)

Note: ☆ $\underbrace{\underline{U}^T\underline{U}}_{= \underline{I}} = \begin{bmatrix} \underline{u}_1^T \\ \vdots \\ \underline{u}_p^T \end{bmatrix} \begin{bmatrix} \underline{u}_1 & \cdots & \underline{u}_p \end{bmatrix} = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix} = \underline{I}$

Consider $(\underline{w}-\underline{w}_0)^T \underline{A}^T\underline{A}(\underline{w}-\underline{w}_0) = (\underline{w}-\underline{w}_0)^T \underline{U}\,\underline{\Lambda}^2 \underbrace{\underline{U}^T(\underline{w}-\underline{w}_0)}_{\underline{z}} = \underline{z}^T\underline{\Lambda}^2\underline{z}$

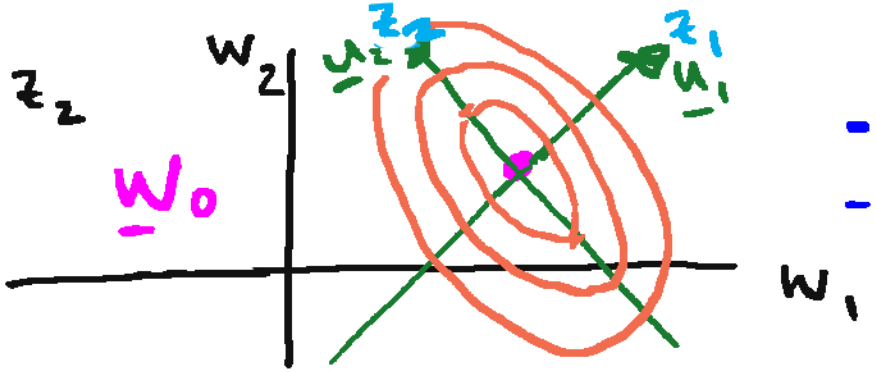
$UZ = u_1 z_1 + \cdots + u_p z_p$

Elliptical contours in $\underline{z}$ major/minor axes

$\underline{w} - \underline{w}_0 = \underline{U}\,\underline{z} = \underline{u}_1 z_1 + \underline{u}_2 z_2$

Bases: $\underline{u}_1, \underline{u}_2$

Coefficients: z_1, z_2



Rotated bowl
- elliptical contours
- parabolic in z_1 and z_2

☆ Just rotated, all other things the same

Summary $f(\underline{w})$:

- bowl shaped surface, concave up
- elliptical constant contours
  - major/minor axis directions eigenvectors $\underline{A}^T\underline{A}$
  - major/minor axis lengths eigenvalues $\underline{A}^T\underline{A}$
- unique bottom (minimum) at $\underline{w}_o$
- Concepts extend to $p > 2$