Solving the Least-Squares Problem Using Gradients

Objectives

- Introduce gradients of linear and quadratic functions of $\underline{w}$
- Use gradients to solve least-squares problem
- Show solution is a minimizer
- Introduce projection matrices

The Least-Squares Problem

$$\min_{\underline{w}} \|\underline{A}\,\underline{w} - \underline{d}\|_2^2$$

$\underline{A}$: N features; $\underline{w}$: P model parameters; $\underline{d}$: N labels

Note: $\|\underline{z}\|_2^2 = \underline{z}^T\underline{z}$ $(\underline{A}\underline{B})^T = \underline{B}^T\underline{A}^T$

$$(\underline{A}\,\underline{w} - \underline{d})^T(\underline{A}\,\underline{w} - \underline{d}) \Rightarrow$$

$$\min_{\underline{w}} \underbrace{\underline{w}^T\underline{A}^T\underline{A}\,\underline{w} - \underline{w}^T\underline{A}^T\underline{d} - \underline{d}^T\underline{A}\,\underline{w} + \underline{d}^T\underline{d}}_{f(\underline{w})}$$

$f(\underline{w})$: quadratic in $\underline{w}$

Scalar problem: $\min_{w} \underbrace{a^2w^2 - 2\beta w + \gamma^2}_{g(w)}$ set $\frac{d}{dw}g(w) = 0$

$2a^2w - 2\beta = 0 \Rightarrow w = a^{-2}\beta$ $a^2 > 0 \Rightarrow$ concave up, min

Gradients: differentiate $f(\underline{w})$ with respect to vector $\underline{w}$

$$\frac{\partial}{\partial \underline{w}} f(\underline{w}) = \underline{0}$$

Gradients

$$\nabla_{\underline{w}} f(\underline{w}) = \left[\frac{\partial}{\partial w_1} f(\underline{w}) \; \frac{\partial}{\partial w_2} f(\underline{w}) \cdots \frac{\partial}{\partial w_p} f(\underline{w}) \right]^T$$

Suppose $f(\underline{w}) = \underline{w}^T \underline{h} = \underline{h}^T \underline{w} = \sum_{i=1}^{p} w_i h_i$ $\quad \frac{\partial}{\partial w_j} f(\underline{w}) = h_j$

$$\nabla_{\underline{w}} f(\underline{w}) = [h_1 \; h_2 \ldots h_p]^T = \underline{h}$$

Suppose $f(\underline{w}) = \underline{w}^T \underline{Q} \underline{w}$ Can show $\nabla_{\underline{w}} f(\underline{w}) = \underline{Q}^T \underline{w} + \underline{Q}\underline{w}$

symmetric case: $\underline{Q} = \underline{Q}^T \Rightarrow \nabla_{\underline{w}} f(\underline{w}) = 2\underline{Q}\underline{w}$

If $\underline{Q} = \underline{A}^T \underline{A} \Rightarrow (\underline{A}^T \underline{A})^T = \underline{A}^T \underline{A}$ symmetric

Solution: $\nabla_{\underline{w}} \left(\underline{w}^T \underline{A}^T \underline{A} \underline{w} - \underline{w}^T \underline{A}^T \underline{d} - \underline{d}^T \underline{A} \underline{w} + \underline{d}^T \underline{d} \right) = \underline{0}$

$$2 \underline{A}^T \underline{A} w - \underline{A}^T \underline{d} - \underline{A}^T \underline{d} = \underline{0} \Rightarrow \underline{w} = (\underline{A}^T \underline{A})^{-1} \underline{A}^T \underline{d}$$

Solution Attributes

$\underline{w} = (\underline{A}^T\underline{A})^{-1}\underline{A}^T\underline{d}$ $\quad \underline{A}: N \times P,\ P \leq N,\ \text{rank}\,\underline{A} = P \Rightarrow (\underline{A}^T\underline{A})^{-1}$ exists

Minimizer? $\quad f(\underline{w}) = \underline{w}^T\underline{A}^T\underline{A}\underline{w} - \underline{w}^T\underline{A}^T\underline{d} - \underline{d}^T\underline{A}\underline{w} + \underline{d}^T\underline{d}$

$$f(\underline{w}) = \left(\underline{w} - (\underline{A}^T\underline{A})^{-1}\underline{A}^T\underline{d}\right)^T\underline{A}^T\underline{A}\left(\underline{w} - (\underline{A}^T\underline{A})^{-1}\underline{A}^T\underline{d}\right) + \underline{d}^T\underline{d} - \underline{d}^T\underline{A}(\underline{A}^T\underline{A})^{-1}\underline{A}^T\underline{d}$$

$$= \underline{w}^T\underline{A}^T\underline{A}\underline{w} - \underline{w}^T\cancel{\underline{A}^T\underline{A}}\cancel{(\underline{A}^T\underline{A})^{-1}}\underline{A}^T\underline{d} - \underline{d}^T\underline{A}\cancel{(\underline{A}^T\underline{A})^{-1}}\cancel{\underline{A}^T\underline{A}}\underline{w} + \underline{d}^T\underline{A}\cancel{(\underline{A}^T\underline{A})^{-1}}\cancel{\underline{A}^T\underline{A}}(\underline{A}^T\underline{A})^{-1}\underline{A}^T\underline{d}$$
$$+ \underline{d}^T\underline{d} - \underline{d}^T\underline{A}(\underline{A}^T\underline{A})^{-1}\underline{A}^T\underline{d}$$

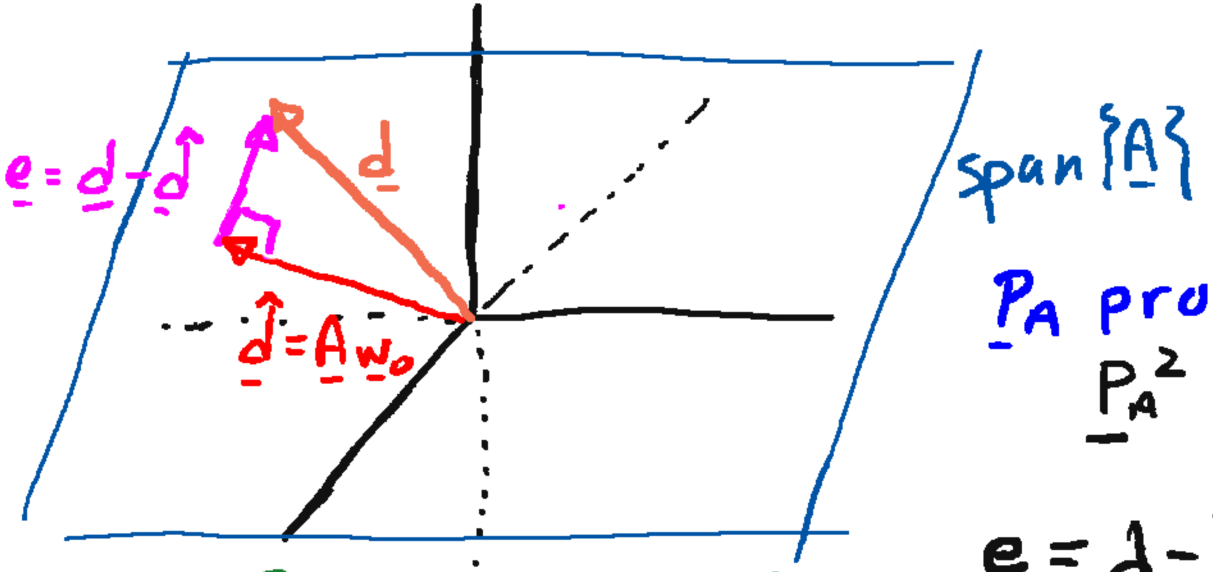
$$= \underline{w}^T\underline{A}^T\underline{A}\underline{w} - \underline{w}^T\underline{A}^T\underline{d} - \underline{d}^T\underline{A}\underline{w} + \cancel{\underline{d}^T\underline{A}(\underline{A}^T\underline{A})^{-1}\underline{A}^T\underline{d}} + \underline{d}^T\underline{d} - \cancel{\underline{d}^T\underline{A}(\underline{A}^T\underline{A})^{-1}\underline{A}^T\underline{d}} \checkmark$$

So $\quad f(\underline{w}) = \underline{z}^T(\underline{w})\underline{A}^T\underline{A}\,\underline{z}(\underline{w}) + \underline{d}^T\underline{d} - \underline{d}^T\underline{A}(\underline{A}^T\underline{A})^{-1}\underline{A}^T\underline{d}$

$\underline{z}(\underline{w}) = \underline{w} - \underline{w}_0\ ;\ \underline{w}_0 = (\underline{A}^T\underline{A})^{-1}\underline{A}^T\underline{d}$

$$\boxed{\begin{array}{l}\underline{A}^T\underline{A} > 0 \Rightarrow \min\limits_{\underline{w}} f(\underline{w}) \text{ when } \underline{w} = \underline{w}_0 \\ \min\limits_{\underline{w}} f(\underline{w}) = \underline{d}^T\underline{d} - \underline{d}^T\underline{A}(\underline{A}^T\underline{A})^{-1}\underline{A}^T\underline{d}\end{array}}$$

Projection and the Pythagorean Theorem



$$\hat{\underline{d}} = \underline{A}(\underline{A}^T\underline{A})^{-1}\underline{A}^T\underline{d} = \underline{P}_A\,\underline{d}$$

$$\underline{P}_A = \underline{A}(\underline{A}^T\underline{A})^{-1}\underline{A}^T$$ "projection matrix"

$\underline{P}_A$ projects $\underline{d}$ onto span{$\underline{A}$}

$$\underline{P}_A^2 = \underline{A}\cancel{(\underline{A}^T\underline{A})^{-1}}\cancel{\underline{A}^T\underline{A}}(\underline{A}^T\underline{A})^{-1}\underline{A}^T = \underline{A}(\underline{A}^T\underline{A})^{-1}\underline{A}^T$$
$$= \underline{P}_A$$

$$\underline{e} = \underline{d} - \hat{\underline{d}} = (\underline{I} - \underline{P}_A)\,\underline{d} = \underline{P}_{A\perp}\underline{d}$$

$\underline{P}_{A\perp} = \underline{I} - \underline{P}_A$ projects onto space $\perp$ to span{$\underline{A}$}

$$\underline{P}_{A\perp}^2 = (\underline{I} - \underline{P}_A)(\underline{I} - \underline{P}_A) = \underline{I} - 2\underline{P}_A + \underline{P}_A^2$$
$$= \underline{I} - \underline{P}_A = \underline{P}_{A\perp}$$

Right triangle

$$\|\underline{d}\|_2^2 = \|\underline{e}\|_2^2 + \|\hat{\underline{d}}\|_2^2$$

$$\Rightarrow \|\underline{e}\|_2^2 = \|\underline{d}\|_2^2 - \|\hat{\underline{d}}\|_2^2$$
$$= \underline{d}^T\underline{d} - \underline{d}^T\underline{P}_A\underline{d}$$
$$= \underline{d}^T(\underline{I} - \underline{P}_A)\,\underline{d}$$