Gradient Descent Solutions to Least-Square problems

Iterative solution methods play an important role

Features / labels: x_i, d_i, $i = 1, 2, \cdots, N$

Classifier on model error: $e^2 = \sum_{i=1}^{N} (\underbrace{x_i^T w}_{\text{prediction}} - \overbrace{d_i}^{\text{actual}})^2$

Formulate as a matrix problem

$$A = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix} \quad d = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{bmatrix} \qquad e^2 = \|Aw - d\|_2^2$$

In general, we are interested in the regularized least-squares problems:

$$\underset{w}{\operatorname{argmin}} \ \|Aw - d\|_2^2 + \lambda r(w)$$

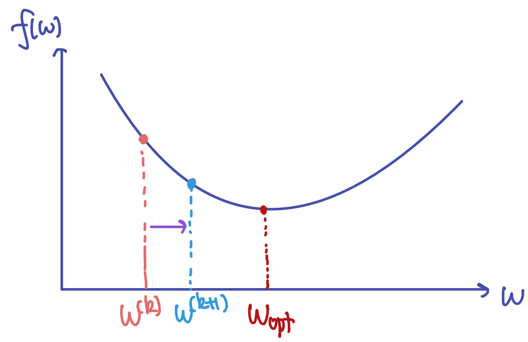
Why consider an iterative solution method?

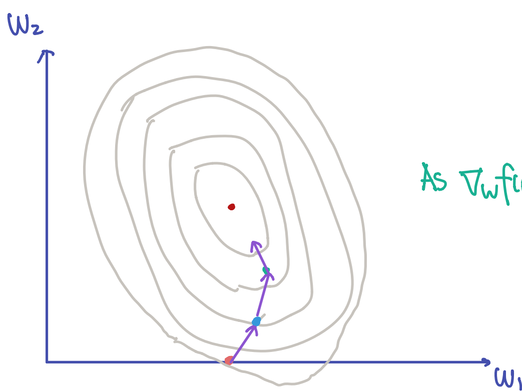
$\left.\begin{array}{l} \text{1. computational cost } (A^TA)^{-1} \\ \text{2. closed form solution maybe unavailable} \\ \text{3. adapt } w \text{ to new features / labels} \end{array}\right\}$ develop iterative approach

Gradient descent finds the minimum

$$f(w) = \|Aw - d\|_2^2$$

$$w^{(k+1)} = w^{(k)} - \underbrace{\tau'}_{\text{step size}} \underbrace{\nabla_w f(w)}_{\text{gradient}} \qquad (\tau' > 0)$$





As $\nabla_w f(w)$ decreases, we are taking smaller steps

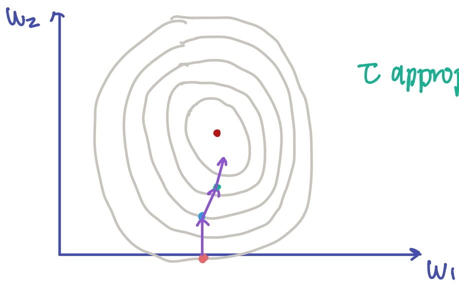
$$f(w) = (Aw - d)^T (Aw - d)$$
$$= w^T A^T A w - 2 w^T A^T d + d^T d$$

$$\nabla_w f(w) = 2A^T A w - 2A^T d$$
$$= 2A^T (Aw - d)$$

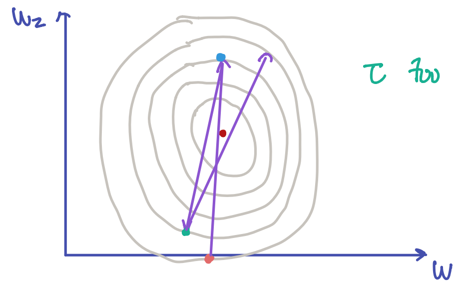
this 2 is absorbed in τ

$$w^{(k+1)} = w^{(k)} - \tau A^T (A w^{(k)} - d)$$ (landweber iteration)

Convergence behavior depends on τ



τ appropriate



τ too big

τ too small: slow convergence
τ too big: no convergence. unstable!

Regularized $0 < \tau < \frac{2}{\|A\|_{op}^2}$ for convergence

Recall: $\|A\|_{op} = \|A\|_2 = \sigma_{max}(A)$

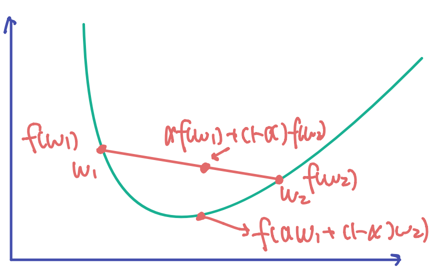
Convergence: $f(w^{(k+1)}) < f(w^{(k)})$, we want to make sure that cost decreases

$$\|Aw^{(k+1)} - d\|_2^2 < \|Aw - d\|_2^2$$

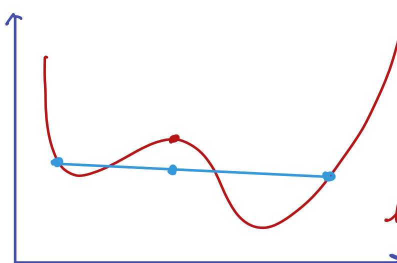
Notes - guaranteed convergence for $0 < \tau < \frac{2}{\|A\|_{op}^2}$

$$w^{(0)} = 0, \quad w^{(k+1)} = w^{(k)} - \tau A^T(Aw^{(k)} - d) \longrightarrow (A^TA)^{-1}A^Td$$

Gradient descent is effective for convex cost functions



convex



local minima

not convex

$$f(\alpha w_1 + (1-\alpha) w_2) \leq \alpha f(w_1) + (1-\alpha) f(w_2), \quad 0 < \alpha < 1, \text{ all } w_1 > w_2 \; ; \quad \frac{d^2}{dw^2} f(w) \geq 0$$

Multidimensional case:

$$[H(w)]_{ij} = \frac{\partial^2}{\partial w_i \partial w_j} f(w), \quad H(w) \succeq 0$$

Proof: Bounds on step size for Guaranteed Convergence

$$f(w) = \|Aw - d\|_2^2$$

$$w^{(k+1)} = w^{(k)} - \tau A^T(Aw^{(k)} - d),\ \forall k \quad , \quad \tau > 0 \qquad \text{often, } w^{(0)} = 0$$

$$f(w^{(k+1)}) = \|Aw^{(k+1)} - d\|_2^2 < f(w) = \|Aw^{(k)} - d\|_2^2$$

$$f(w^{(k+1)}) = \|A[w^{(k)} - \tau A^T(Aw^{(k)} - d)] - d\|_2^2$$

$$= \|Aw^{(k)} - \tau AA^T(Aw^{(k)} - d) - d\|_2^2$$

$$= \|(Aw^{(k)} - d) - \tau AA^T(Aw^{(k)} - d)\|_2^2$$

let $c = Aw^{(k)} - d$, let $e = \tau(AA^T(Aw^{(k)} - d))$

$$f(w^{(k+1)}) = \|c - e\|_2^2 = (c-e)^T(c-e) = \|c\|_2^2 + \|e\|_2^2 - 2e^Tc$$

$$f(w^{(k+1)}) = \|Aw^{(k)} - d\|_2^2 + \|\tau AA^T(Aw^{(k)} - d)\|_2^2 - 2[\tau AA^T(Aw^{(k)} - d)]^T(Aw^{(k)} - d)$$

$$= f(w^k) + \tau^2\|AA^T(Aw^{(k)} - d)\|_2^2 - 2\tau[(Aw^{(k)} - d)^T A][A^T(Aw^{(k)} - d)]$$

let $v = A^T(Aw^{(k)} - d)$

$$f(w^{(k+1)}) = f(w^k) + \tau^2\|Av\|_2^2 - 2\tau v^Tv$$

$$f(w^{(k+1)}) - f(w^{(k)}) = q(\tau) = \tau^2\|Av\|_2^2 - 2\tau v^Tv,$$

since wts $f(w^{(k+1)}) < f(w^{(k)}) \rightarrow f(w^{(k+1)}) - f(w^{(k)}) < 0$

$$\max_g \|Xg\|_2 < \|X\|_{op}\|g\|_2$$

$$\tau^2\|Av\|_2^2 \leq \tau^2\|A\|_{op}^2\|v\|_2^2$$

$$-2\tau v^Tv = -2\tau\|v\|_2^2$$

so. $q(\tau) \leq \tau^2\|A\|_{op}^2\|v\|_2^2 - 2\tau\|v\|_2^2$

$$q(\tau) \leq (\tau\|A\|_{op}^2 - 2)\underbrace{\tau\|v\|_2^2}_{} \geq 0, > 0 \text{ if } v \neq 0$$

if wts $q(t) < 0$, then

$$\tau \|A\|_{op}^2 - 2 < 0$$

$$\tau < \frac{2}{\|A\|_{op}^2}$$