Gradient Descent Solutions to Least-Square Problems

Objectives

- explain need for iterative algorithms
- derive gradient descent algorithm
- consider impact of step size on convergence
- introduce notion of convex functions

Iterative solution methods play an important role

Features/labels: $\underline{x}_i$, d_i, $i = 1, 2, \dots N$

Classifier or model error: $e^2 = \sum_{i=1}^{N} (\underline{x}_i^T \underline{w} - d_i)^2$

$$\underline{A} = \begin{bmatrix} \underline{x}_1^T \\ \underline{x}_2^T \\ \vdots \\ \underline{x}_N^T \end{bmatrix} \quad \underline{d} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{bmatrix} \quad e^2 = \|\underline{A}\,\underline{w} - \underline{d}\|_2^2$$

Regularized least squares: $\arg\min_{\underline{w}} \|\underline{A}\,\underline{w} - \underline{d}\|_2^2 + \lambda r(\underline{w})$

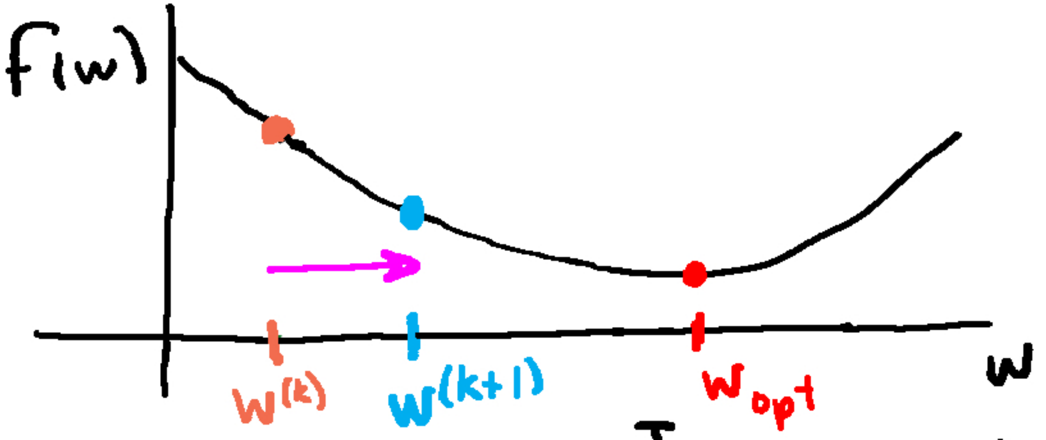
$$\left.\begin{array}{l} \text{1. Computational cost } (\underline{A}^T\underline{A})^{-1} \\ \text{2. Closed form solution may be unavailable} \\ \text{3. Adapt } \underline{w} \text{ to new features/labels} \end{array}\right\} \text{develop iterative approach}$$

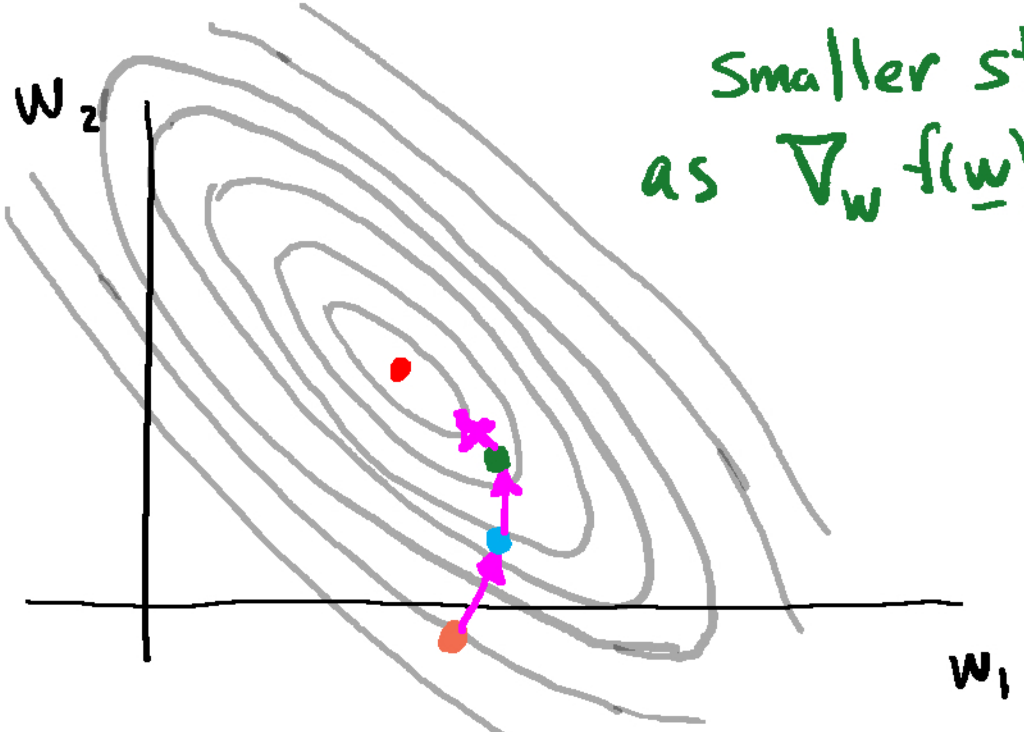
Gradient descent finds the minimum

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

$$\underline{w}^{(k+1)} = \underline{w}^{(k)} - \tau' \nabla_{\underline{w}} f(\underline{w})$$

$(\tau' > 0)$ step size gradient





smaller steps as $\nabla_{\underline{w}} f(\underline{w}) \downarrow$

$$f(\underline{w}) = (\underline{A}\underline{w} - \underline{d})^T(\underline{A}\underline{w} - \underline{d})$$

$$= \underline{w}^T\underline{A}^T\underline{A}\underline{w} - 2\underline{w}^T\underline{A}^T\underline{d} + \underline{d}^T\underline{d}$$

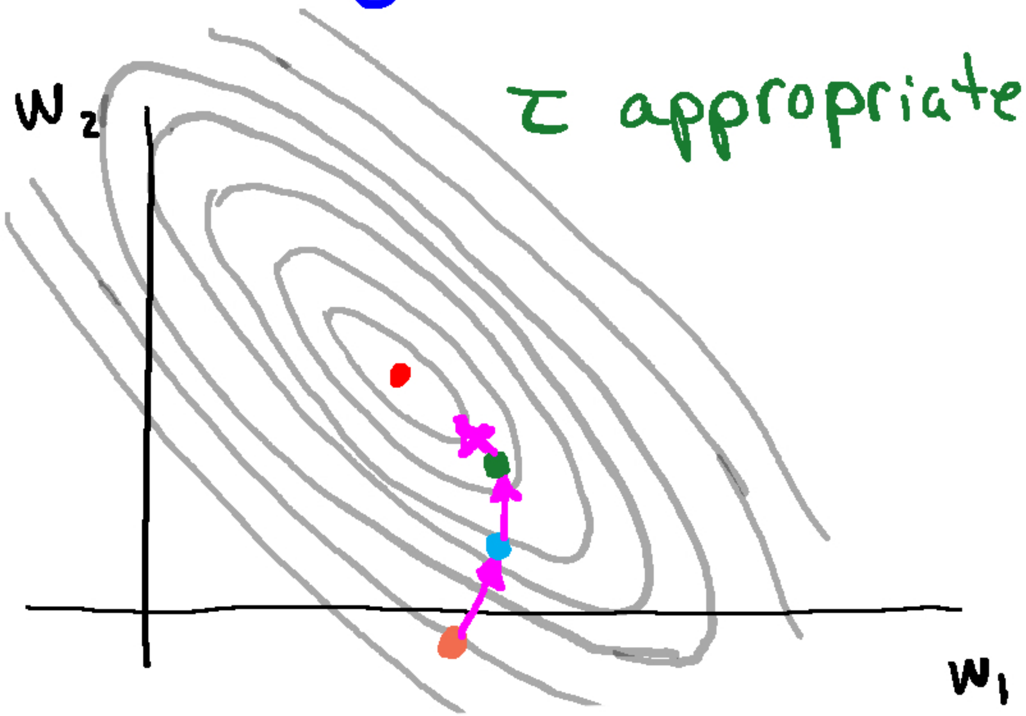
$$\nabla_w f(\underline{w}) = 2\underline{A}^T\underline{A}\underline{w} - 2\underline{A}^T\underline{d}$$

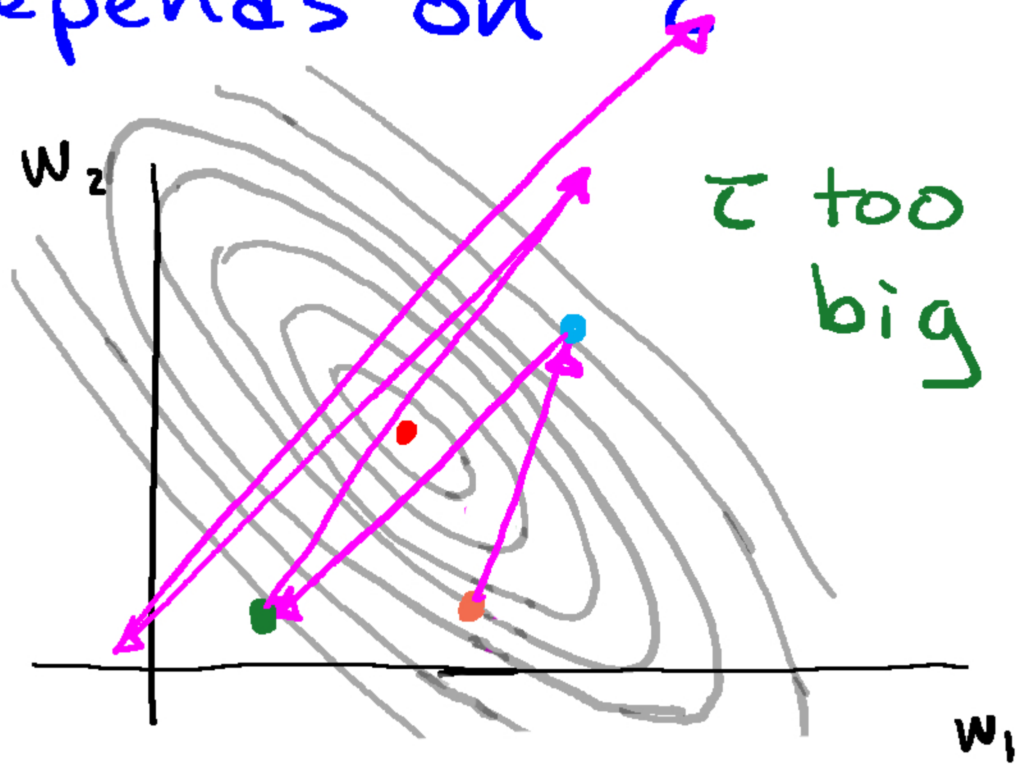
$$= 2\underline{A}^T(\underline{A}\underline{w} - \underline{d})$$

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

(Landweber iteration)

Convergence behavior depends on τ





τ too small: slow convergence

τ too big: no convergence
unstable!

Require $0 < \tau < 2/\|\underline{A}\|_{op}^2$ for convergence

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

Convergence: $f(\underline{w}^{(k+1)}) < f(\underline{w}^{(k)})$ cost decreases as k increases

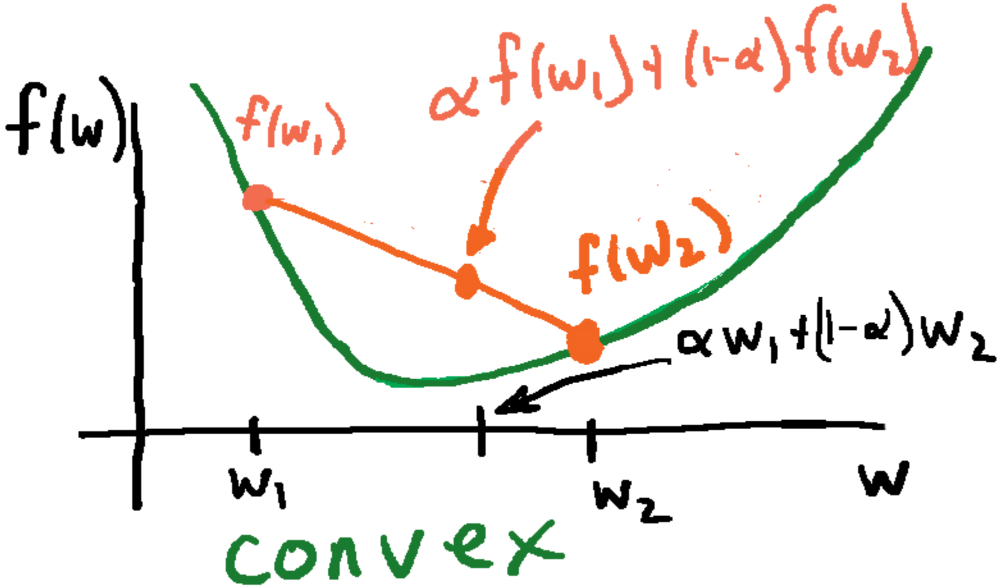
$$\|\underline{A}\,\underline{w}^{(k+1)} - d\|_2^2 < \|\underline{A}\,\underline{w}^{(k)} - d\|_2^2$$

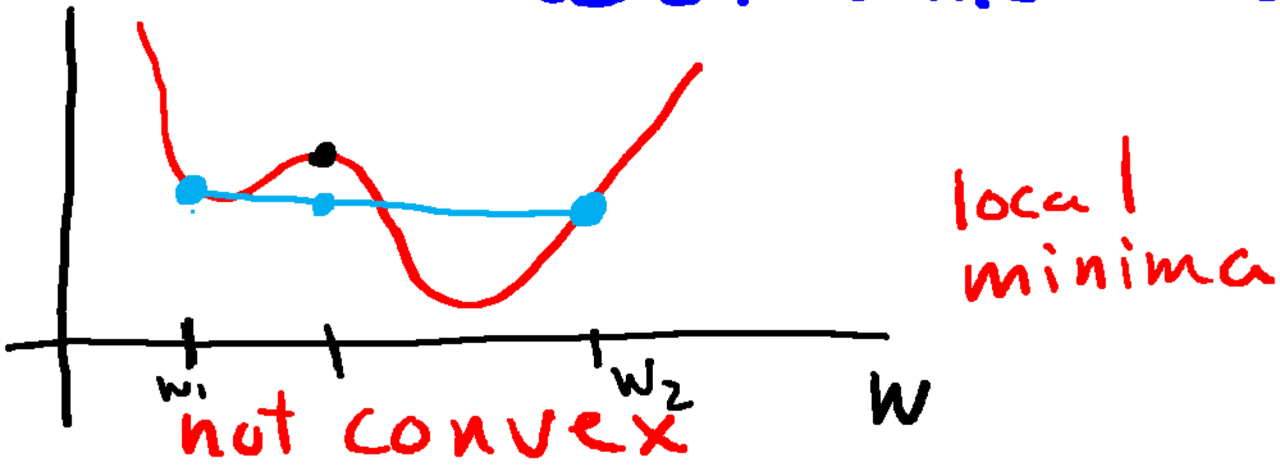
Notes - guaranteed convergence for

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

$$\underline{w}^{(0)} = \underline{0}, \quad \underline{w}^{(k+1)} = \underline{w}^{(k)} - \tau \underline{A}^T(\underline{A}\,\underline{w}^{(k)} - \underline{d}) \xrightarrow[k]{} (\underline{A}^T\underline{A})^{-1}\underline{A}^T\underline{d}$$

Gradient descent is effective for convex cost functions





$$f(\alpha w_1 + (1-\alpha) w_2) \leq \alpha f(w_1) + (1-\alpha) f(w_2); \qquad \frac{d^2}{dw^2} f(w) \geq 0$$

$$0 < \alpha < 1, \text{ all } w_1, w_2$$

Multidimensional case

$$\underline{H}(\underline{w}) \geq 0 \qquad [\underline{H}(\underline{w})]_{ij} = \frac{\partial^2}{\partial w_i \partial w_j} f(\underline{w})$$