MATH 565: Lecture 11 (02/17/2026)

Today: * Newton's update for nonquadratic J
* line search in Newton
* Newton in regression, L_2-SVM

Recall Newton's method update: $\bar{w} \leftarrow \bar{w} - H^{-1} \nabla J$

Newton's method works perfectly for quadratic J. What about other J?

Consider $J(w) = -w^3 + 4w^2 + 1$

Consider Newton update at $w_0 = 2$ $\left(J(\bar{w}_0) = 9\right)$.

$w^{0} = w_0$ — iteration index

$$\nabla J = -3w^2 + 8w \Rightarrow \nabla J(w_0) = 4$$

$$HJ = -6w + 8 \Rightarrow HJ(w_0) = -4$$

$\rightarrow$ $H < 0$ here! So, Newton's method pushes w (from $w_0 = 2$) to $w^1 = 3$, which is a local max for the quadratic approximation.

$$w^1 \leftarrow w^0 - H^{-1} \nabla J = 2 - \left(-\frac{1}{4}\right) \cdot 4 = 3$$

$$J(w^1) = -27 + 36 + 1 = 10$$

So, the Newton update increases J!

Quadratic approximation of $J(w)$ at $w_0 = 2$

$$J(w) = J(w_0) + (w-2)4 + \frac{1}{2}(w-2)^2(-4)$$

$$= 9 + 4w - 8 - 2(w^2 - 4w + 4)$$

$$= -2w^2 + 12w - 7$$

this parabola is opening down ($\cap$)

$\rightarrow$ $w = 3$ is a local maximum for this quadratic approximation

Line Search

Modify Newton update to

$$\bar{w}^{k+1} \leftarrow \bar{w}^{k} - \alpha H_k^{-1} \nabla J_k$$

accept, i.e., update $\bar{w}$ as described above, only if $J(\bar{w}^{k+1}) < J(\bar{w}^{k})$.

else, change α, or start over from a different $\bar{w}_0$.

Example

$$J(w) = w^2 - \ln(w)$$

(log barrier function) used in interior point methods. Keeps the algorithm from getting too close to $w=0$.

generalization: $J(\bar{w}) = \sum_{i=1}^{d} J_i(w_i)$

where $J_i(w_i) = w_i^2 - \ln(w_i)$

$$HJ = \begin{bmatrix} \ddots & & 0 \\ & 2+\frac{1}{w_i^2} & \\ 0 & & \ddots \end{bmatrix}$$

$$\nabla J = 2w - \frac{1}{w}$$

$$HJ = 2 + \frac{1}{w^2}$$

$\Rightarrow$ J has unique global minimum at $w^* = \frac{1}{\sqrt{2}} = 0.707$.

Starting @ $w_0 = 2$, apply Newton's update:

$$\nabla J = \frac{7}{2} \; (=3.5)$$

$$HJ = \frac{9}{4} \; (=2.25)$$

$$w \leftarrow w_0 - \left(\frac{4}{9}\right)\cdot\left(\frac{7}{2}\right) = \frac{4}{9} \approx 0.44 \; \left(< \frac{1}{\sqrt{2}}\right)$$

$w_0 = 2$, $\frac{4}{9} = H^{-1}$, $\frac{7}{2} = \nabla$

We've overshot (passed over) the global minimum!

With $\alpha \approx 0.831$,

$$w' = w_0 - \alpha H^{-1} \nabla \approx 0.707.$$

Can adapt various line search selection approaches introduced for gradient descent for Newton's method as well.

Newton's Method in ML

Newton in Regression

$$J = \frac{1}{2}\|D\bar{w} - \bar{y}\|^2 = \frac{1}{2}(D\bar{w} - \bar{y})^T (D\bar{w} - \bar{y})$$

quadratic! So, Newton should work in one step

$$D = \begin{array}{c|c} & 1 \; 2 \; \cdots \; d \\ \hline \begin{matrix} 1 \\ 2 \\ \vdots \\ n \end{matrix} & \begin{bmatrix} \\ \text{---} \; \bar{x}_i^T \; \text{---} \\ \\ \end{bmatrix} \end{array} \quad n \times d$$

$$\nabla J = D^T D \bar{w} - D^T \bar{y}$$

$$HJ = D^T D$$

$$HJ = \sum_{i=1}^{n} \bar{x}_i \bar{x}_i^T \quad \text{(Sum of outer products)}$$

$$= \sum_{i=1}^{n} HJ_i$$

Hessian of J = sum of point-specific Hessians

Newton update

$$\bar{w} \leftarrow \bar{w} - \underbrace{H^{-1}\nabla J}$$

$$\bar{w} \leftarrow \bar{w} - (D^T D)^{-1}(D^T D \bar{w} - D^T \bar{y})$$

$$\nabla J = \bar{0} \Rightarrow \bar{w} = (D^T D)^{-1} D^T \bar{y}$$

$$= \bar{w} - \bar{w} + (D^T D)^{-1} D^T \bar{y}$$

$$\bar{w} = (D^T D)^{-1} D^T \bar{y} \longrightarrow \text{optimal solution in a single update.}$$

Recall that we obtained the same solution using gradient descent (GD) approaches. $\nabla J = \bar{0}$ gave $\bar{w} = (D^T D)^{-1} D^T \bar{y}$.

Newton in SVM

We look at L_2-SVM. (The hinge loss J is not smooth)

$$J = J_{L_2\text{-svm}}(\bar{\omega}) = \frac{1}{2}\sum_{i=1}^{n}\max\{0, (1-y_i(\bar{\omega}^T\bar{x}_i))\}^2$$

$$= \sum_{i=1}^{n} J_i \quad \text{for} \quad J_i = \frac{1}{2}\max\{0, 1-y_i(\bar{\omega}^T\bar{x}_i)\}^2.$$

Here, $J_i = f_i(z) = \frac{1}{2}\max\{0, 1-y_i z\}^2$ for $z = \bar{\omega}^T\bar{x}_i$.

$$\frac{\partial f_i(z)}{\partial z} = -y_i \max\{0, 1-y_i z\}$$

$$\Rightarrow \nabla J_i = \frac{\partial J_i}{\partial \bar{\omega}} = \underbrace{-y_i \max\{0, 1-y_i(\bar{\omega}^T\bar{x}_i)\}}_{\text{scalar}}\bar{x}_i$$

If we were using least squares classification (instead of L_2-SVM), we get $\nabla J_{LS} = -y_i(1-y_i(\bar{\omega}^T\bar{x}_i))\bar{x}_i$

We can rewrite ∇J_i using the indicator function $\delta(\cdot)$:

$$\nabla J_i = \underbrace{(\bar{\omega}^T\bar{x}_i - y_i)\,\delta(1-y_i(\bar{\omega}^T\bar{x}_i) > 0)}_{\text{scalar}}\cdot\bar{x}_i$$

$$\Rightarrow \nabla J(\bar{\omega}) = \sum_{i=1}^{n}\nabla J_i$$

$$= D^T\Delta_{\bar{\omega}}(D\bar{\omega} - \bar{y})$$

where $\Delta_{\bar{\omega}} = [\mathrm{diag}(\delta(1-y_i(\bar{\omega}^T\bar{x}_i) > 0))]$

↳ diagonal matrix whose i^{th} entry is $\delta(1-y_i(\bar{\omega}^T\bar{x}_i) > 0)$.

<u>Hessian?</u>
$$HJ = \sum_{i=1}^{n} H_i = \sum_{i=1}^{n} HJ_i$$

Note: $$\nabla J_i = \underbrace{(\bar{w}^T \bar{x}_i - y_i)\,\delta\left(1 - y_i(\bar{w}^T\bar{x}_i) > 0\right)}_{\text{scalar}} \cdot \bar{x}_i$$

$$= s_i(\bar{w}) \cdot \bar{x}_i \quad \text{where } s_i(\bar{w}) = -y_i \max\{0, 1 - y_i(\bar{w}^T\bar{x}_i)\}$$

$$\Rightarrow \quad HJ_i = \bar{x}_i \left[\frac{\partial s_i}{\partial \bar{w}}\right]^T$$

$$= \bar{x}_i \left(y_i^2\, \delta\left(1 - y_i(\bar{w}^T\bar{x}_i) > 0\right)\right) \bar{x}_i^T \qquad y_i^2 = 1$$

$$= \delta\left(1 - y_i(\bar{w}^T\bar{x}_i) > 0\right) \bar{x}_i \bar{x}_i^T$$

$$\Rightarrow \quad HJ = \sum_{i=1}^{n} HJ_i = \sum_{i=1}^{n} \delta\left(1 - y_i(\bar{w}^T\bar{x}_i) > 0\right) \bar{x}_i \bar{x}_i^T$$

For J_{LS}, we get $H = D^T D$.

Here, for L_2-SVM, we get
$$HJ_{L_2\text{-SVM}} = D^T \Delta_{\bar{w}} D$$

where $\Delta_{\bar{w}} = \left[\operatorname{diag}\left(\delta\left(1 - y_i(\bar{w}^T\bar{x}_i) > 0\right)\right)\right]$.