Lecture 8: Fisher information

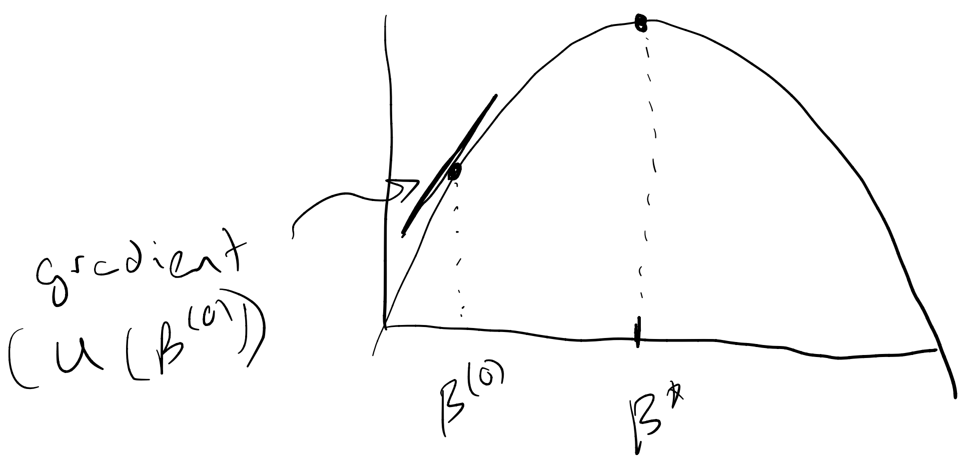
Recap: Newton's method

To find β^* such that $U(\beta^*) = 0$, when there is no closed-form solution we use Newton's method:

- Begin with an initial guess $\beta^{(0)}$
- Iteratively update: $\beta^{(r+1)} = \beta^{(r)} - \mathbf{H}^{-1}(\beta^{(r)})U(\beta^{(r)})$
- Stop when the algorithm converges

Intuition:

log likelihood



Hessian tells us about curvature of the log-likelihood

Some intuition about Hessians

Example: Suppose that $\beta = (\beta_0, \beta_1)^T \in \mathbb{R}^2$, and

$$\ell(\beta) = -\beta_0^2 - 100\beta_1^2$$

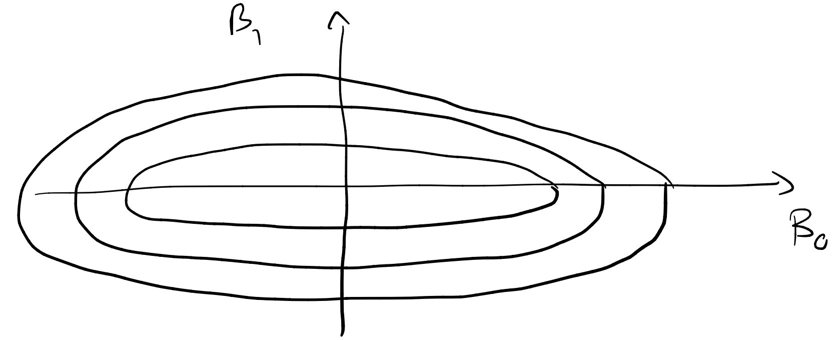
Calculate

$$U(\beta) = \begin{bmatrix} \frac{\partial \ell}{\partial \beta_0} \\ \frac{\partial \ell}{\partial \beta_1} \end{bmatrix} \qquad \mathbf{H}(\beta) = \begin{bmatrix} \frac{\partial^2 \ell}{\partial \beta_0^2} & \frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 \ell}{\partial \beta_1^2} \end{bmatrix}$$

$$= \begin{bmatrix} -2\beta_0 \\ -200\beta_1 \end{bmatrix} \qquad = \begin{bmatrix} -2 & 0 \\ 0 & -200 \end{bmatrix}$$

$$\ell(\beta) = -\beta_0^2 - 100\beta_1^2$$

Level curves



(a small change in β_1 makes a bigger difference to $\ell(\beta)$ than a small change in β_0)

$$H(\beta) = -\begin{bmatrix} 2 & 0 \\ 0 & 200 \end{bmatrix}$$

$$H^{-1}(\beta) = -\begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{200} \end{bmatrix}$$

Suppose start with $\beta^{(0)} = \begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix}$

$$\beta^{(1)} = \begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix} + \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{200} \end{bmatrix} \begin{bmatrix} -2(0.1) \\ -200(0.1) \end{bmatrix}$$

$$= \begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix} - \begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Contrast: gradient ascent / gradient descent

$\beta^{(r)}$ current guess $\qquad X^TWX$

step size

new guess: $\beta^{(r+1)} = \beta^{(r)} + \alpha\, U(\beta^{(r)})$

Previous example:

$$\begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix} + \alpha \begin{bmatrix} -2(0.1) \\ -200(0.1) \end{bmatrix}$$

(probably converge more slowly than Newton's method, b/c we don't use information about second derivative)

Trade-off:

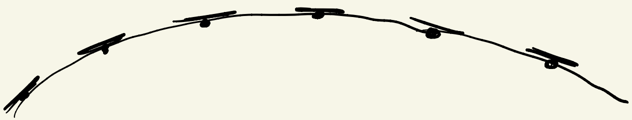
gradient ascent/descent:
- more steps
- doesn't require matrix inversion of second derivatives

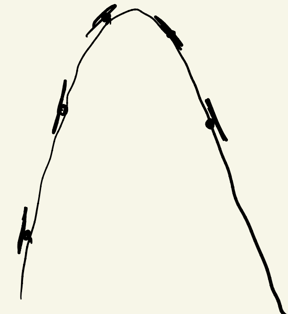
Newton's method
- fewer steps
- steps are more computationally expensive

One more perspective

harder to maximize

easier to maximize





slopes $u(\beta)$ are close to 0

slopes $u(\beta)$ are far from 0

$\Rightarrow$ suggest look @ variability in $u(\beta)$ (as a function of observed data)

Logistic regression: $u(\beta) = X^T(Y-p)$

X design matrix

Y vector of responses

$p =$ vector of probabilities

$$
\begin{aligned}
Var(u(\beta) \mid X) &= Var(X^T(Y-p) \mid X) \\
&= X^T \, Var(Y-p \mid X) \, X \\
&= X^T \, Var(Y \mid X) \, X \\
&= X^T \begin{bmatrix} p_1(1-p_1) & p_2(1-p_2) & \cdots & p_n(1-p) \end{bmatrix} X = -H(\beta)
\end{aligned}
$$

Fisher information

<u>Def</u>: Let $\ell(\theta | Y)$ be a log likelihood, and $U(\theta) = \frac{\partial \ell}{\partial \theta}$

The <u>Fisher information</u> is

$$\underset{\sim}{I}(\theta) = Var(U(\theta) \mid \theta)$$

i.e. the variance of $U(\theta)$, given θ is the true parameter

$Var(Y+c) = Var(Y)$

$Var(Y_i) = p(1-p)$

Example: Bernoulli sample

Suppose that $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Bernoulli}(p)$.

$$L(p|Y) = p^{\sum_i Y_i}(1-p)^{n-\sum_i Y_i}$$

$$\ell(p|Y) = \left(\sum_i Y_i\right)\log p + \left(n - \sum_i Y_i\right)\log(1-p_i)$$

$$u(p) = \frac{\partial}{\partial p}\ell(p|Y) = \frac{\sum_i Y_i}{p} - \frac{\left(n - \sum_i Y_i\right)}{1-p}$$

$$Var(u(p)|p) = Var\left(\frac{\sum_i Y_i}{p} - \frac{\left(n-\sum_i Y_i\right)}{(1-p)} \,\Big|\, p\right)$$

$$= Var\left(\frac{\left(\sum_i Y_i\right)(1-p) - \left(n - \sum_i Y_i\right)p}{p(1-p)} \,\Big|\, p\right)$$

$$= Var\left(\frac{\sum_i Y_i}{p(1-p)} \,\Big|\, p\right) = \frac{1}{p^2(1-p)^2}\sum_i Var(Y_i) = \frac{np(1-p)}{p^2(1-p)^2} = \frac{n}{p(1-p)}$$

$$\mathcal{I}(p) = \frac{n}{p(1-p)} \qquad \mathbb{E}\left[-\frac{\partial^2}{\partial p^2} \ell(p \mid Y) \right] = \frac{n}{p(1-p)}$$

$$\hat{p} = \frac{1}{n} \sum_i Y_i$$

$$\mathrm{Var}(\hat{p}) = \frac{1}{n^2} \cdot n p(1-p) = \frac{p(1-p)}{n} = \mathcal{I}^{-1}(p)$$

Linear regression: $\mathrm{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} = \mathcal{I}^{-1}(\beta)$

Logistic regression: $\mathrm{Var}(\hat{\beta}) = (X^T W X)^{-1} = \mathcal{I}^{-1}(\beta)$

$$W = \begin{bmatrix} p_1(1-p_1) & & & \\ & p_2(1-p_2) & & \\ & & \ddots & \\ & & & p_n(1-p_n) \end{bmatrix}$$

Properties

Under certain regularity conditions, we have:

① $\mathbb{E}[\, u(\theta) \mid \theta \,] = 0$

② $\mathcal{I}(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \ell(\theta \mid Y) \;\middle|\; \theta \right]$

Example: Bernoulli sample

Suppose that $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Bernoulli}(p_i)$.