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

For logistic regression: $U(\beta) = X^T(Y-p)$

$H(\beta) = -X^TWX$

$W = \text{diag}(p_i(1-p_i))$

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 the score function

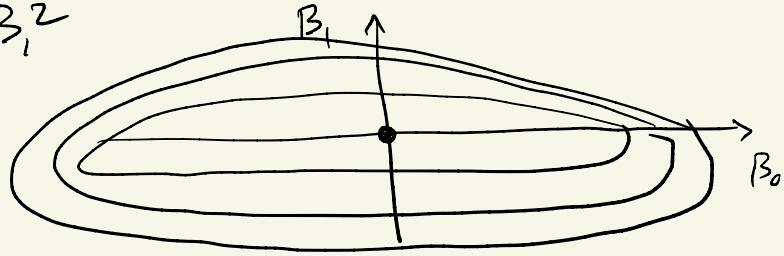
$$U(\beta) = \begin{bmatrix} \frac{\partial \ell}{\partial \beta_0} \\ \frac{\partial \ell}{\partial \beta_1} \end{bmatrix} = \begin{bmatrix} -2\beta_0 \\ -200\beta_1 \end{bmatrix}$$

and the Hessian

$$\mathbf{H}(\beta) = \begin{bmatrix} \frac{\partial^2 \ell}{\partial \beta_0^2} & \frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 \ell}{\partial \beta_1^2} \end{bmatrix} = \begin{bmatrix} -2 & 0 \\ 0 & -200 \end{bmatrix}$$

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

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



captures information about curvature of $\ell(\beta)$

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

amount we move depends on curvature of $\ell(\beta)$

$$\beta^{(r)} = \begin{bmatrix} 0.1 \\ 0.1 \end{bmatrix}$$

move along gradient

$$\beta^{(r+1)} = \beta^{(r)} - H^{-1}(\beta^{(r)})\, U(\beta^{(r)})$$

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

More intuition

which $\ell(\beta)$ is "easier" to maximize?

Harder to maximize
(lots of βs have similar $\ell(\beta)$)

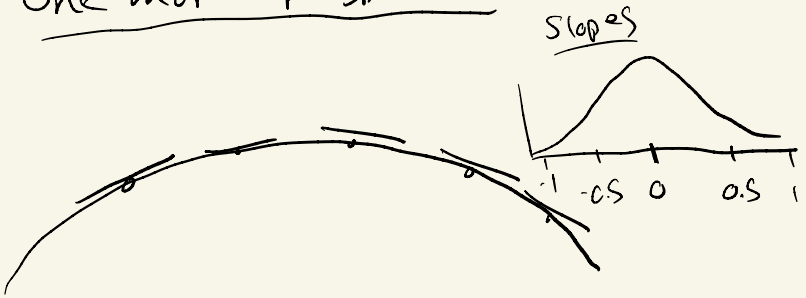
second derivative is close to 0

easier to maximize
(small changes in β =>
large changes in $\ell(\beta)$)

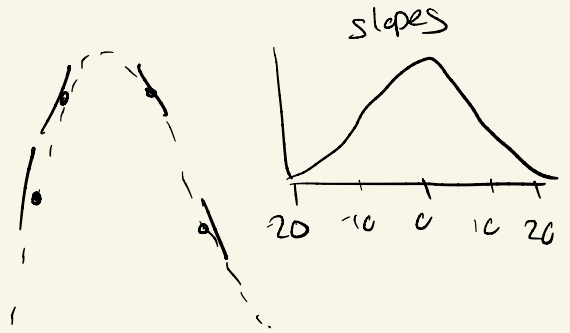
second derivative far from 0

=> Hessian tells us how easy function is to maximize

One more perspective:



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



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

more variability $U(\beta)$

$\Rightarrow$ look @ variability in $U(\beta)$. Larger variance $\Rightarrow$ easier to maximize

Logistic regression:

$$U(\beta) = X^T(Y-p)$$

$$\begin{aligned} \text{var}(U(\beta)) &= X^T \text{Var}(Y-p) X \\ &= X^T \text{Var}(Y) X \\ &= X^T \begin{bmatrix} p_1(1-p_1) & & \\ & p_2(1-p_2) & \\ & & \ddots \; p_n(1-p_n) \end{bmatrix} X \\ &= X^T W X = -H(\beta) \end{aligned}$$

$\text{Var}(Y_i) = p_i(1-p_i)$

Fisher information

Def: Let $\ell(\theta|Y)$ be a log likelihood, and $U(\theta) = \frac{\partial \ell}{\partial \theta}$ the score function. The Fisher information is

$$\mathcal{I}(\theta) = \text{Var}(U(\theta)|\theta)$$

i.e. the variance of the score, if θ is the true parameter

$Y_i \sim \text{Bernoulli}(p_i)$

$\log\left(\frac{p_i}{1-p_i}\right) = \beta^T x_i$

$$\begin{aligned}\mathcal{I}(\beta) &= \text{Var}(U(\beta)) \\ &= \text{Var}(X^T(Y-p)) \\ &= X^T W X\end{aligned}$$

$W = \text{diag}(p_i(1-p_i))$

Example: Bernoulli sample

$\mathrm{Var}(Y_i) = p(1-p)$

$\mathrm{Var}\left(\sum_{i=1}^n Y_i\right) = np(1-p)$ (independent)

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

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

$$\ell(p \mid Y) = \left(\sum_{i=1}^n Y_i\right) \log p + \left(n - \sum_{i=1}^n Y_i\right) \log(1-p)$$

$$u(p) = \frac{\sum_i Y_i}{p} - \frac{\left(n - \sum_{i=1}^n Y_i\right)}{1-p}$$

$$\mathrm{Var}(u(p) \mid p) = \mathrm{Var}\left(\frac{\sum_i Y_i}{p} - \frac{(n - \sum_i Y_i)}{1-p}\right)$$

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

$$= \frac{np(1-p)}{p^2(1-p)^2} = \frac{n}{p(1-p)}$$

$$\hat{\mathcal{I}}(p) = \frac{n}{p(1-p)}$$

Properties

Under certain regularity conditions, we have the following:

① $\mathbb{E}(u(\theta)|\theta) = 0$

(this implies that $\mathcal{I}(\theta) = \text{Var}(u(\theta)|\theta)$
$= \mathbb{E}(u^2(\theta)|\theta)$)

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

(Fisher information captures curvature of log likelihood)

Example: Bernoulli sample

$\mathbb{E}[\sum_i Y_i] =$
$n\,\mathbb{E}[Y_i] = np$

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

$$u(p) = \frac{\sum_i Y_i}{p} - \frac{(n - \sum_i Y_i)}{1-p}$$

1) $\mathbb{E}[u(p)|p] = \mathbb{E}\left[\frac{\sum_i Y_i}{p}\right] - \mathbb{E}\left[\frac{n - \sum_i Y_i}{1-p}\right]$

$$= \frac{np}{p} - \frac{n(1-p)}{1-p} = n - n = 0 \quad \checkmark$$

2) $$\frac{\partial^2 \ell}{\partial p^2} = -\frac{\sum_i Y_i}{p^2} - \frac{(n - \sum_i Y_i)}{(1-p)^2}$$

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

$$= \frac{np}{p^2} + \frac{n(1-p)}{(1-p)^2} = \frac{n}{p} + \frac{n}{1-p} = \frac{n}{p(1-p)} \quad \checkmark$$