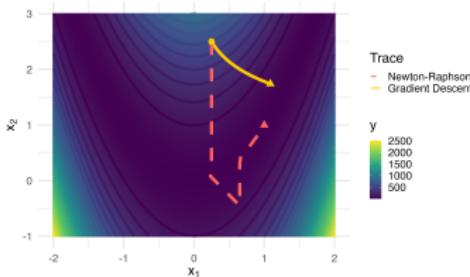


Optimization in Machine Learning

Second order methods Fisher Scoring



Learning goals

- Fisher Scoring
- Newton-Raphson vs. Fisher scoring
- Logistic regression



RECAP OF NEWTON'S METHOD

- Second-order Taylor expansion of log-likelihood around the current iterate $\theta^{[t]}$:

$$\ell(\theta) \approx \ell(\theta^{[t]}) + \nabla \ell(\theta^{[t]})^\top (\theta - \theta^{[t]}) + \frac{1}{2} (\theta - \theta^{[t]})^\top [\nabla^2 \ell(\theta^{[t]})] (\theta - \theta^{[t]})$$

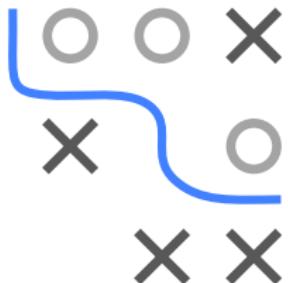
- We then differentiate w.r.t. θ and set the gradient to zero:

$$\nabla \ell(\theta^{[t]}) + [\nabla^2 \ell(\theta^{[t]})] (\theta - \theta^{[t]}) = \mathbf{0}$$

- Solving for $\theta^{(t)}$ yields the pure Newton-Raphson update:

$$\theta^{[t+1]} = \theta^{[t]} + [-\nabla^2 \ell(\theta^{[t]})]^{-1} \nabla \ell(\theta^{[t]})$$

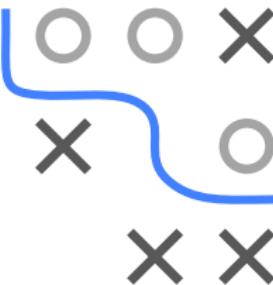
- Potential stability issue:** pure Newton-Raphson updates do not always converge. Its quadratic convergence rate is “local” in the sense that it requires starting close to a solution.



FISHER SCORING

- Fisher's scoring method replaces the negative *observed Hessian* $-\nabla^2 \ell(\theta)$ by the Fisher information matrix, i.e., the variance of $\nabla \ell(\theta)$, which, under weak regularity conditions, equals the negative *expected Hessian*

$$\mathbb{E}[\nabla \ell(\theta) \nabla \ell(\theta)^\top] = \mathbb{E}[-\nabla^2 \ell(\theta)],$$

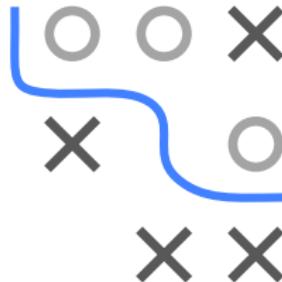


and is positive semi-definite under exchangeability of expectation and differentiation

- **NB:** it can be shown that $\mathbb{E}[\nabla \ell(\theta)] = \mathbf{0}$, which provides the expression of the variance of $\nabla \ell(\theta)$ as the expected outer product of the gradients
- Therefore the Fisher scoring iterates are given by

$$\theta^{[t+1]} = \theta^{[t]} + \mathbb{E}[-\nabla^2 \ell(\theta^{[t]})]^{-1} \nabla \ell(\theta^{[t]})$$

NEWTON-RAPHSO VS. FISHER SCORING



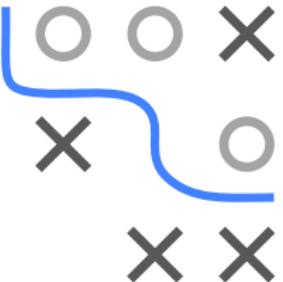
Aspect	Newton-Raphson	Fisher scoring
Second-order Matrix	Exact negative Hessian matrix	Fisher information matrix
Curvature	Exact	Approximated
Computational Cost	Higher	Lower (often has a simpler structure)
Convergence	Fast but potentially unstable	Slower but more stable
Positive Definite	Not guaranteed	Yes with Fisher information
Use Case	General non-linear optimization	Likelihood-based models, especially GLMs

- In many cases Newton-Raphson and Fisher scoring are equivalent (see below)

LOGISTIC REGRESSION

- The goal of logistic regression is to predict a binary event
- Given n observations $(\mathbf{x}^{(i)}, y^{(i)}) \in \mathbb{R}^{p+1} \times \{0, 1\}$,
 $y^{(i)} | \mathbf{x}^{(i)} \sim \text{Bernoulli}(\pi^{(i)})$
- We want to minimize the following risk

$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = - \sum_{i=1}^n y^{(i)} \log(\pi^{(i)}) + (1 - y^{(i)}) \log(1 - \pi^{(i)})$$



with respect to $\boldsymbol{\theta}$, where the probabilistic classifier

$\pi^{(i)} = \pi(\mathbf{x}^{(i)} | \boldsymbol{\theta}) = s(f(\mathbf{x}^{(i)} | \boldsymbol{\theta}))$, the sigmoid function

$s(f) = \frac{1}{1+\exp(-f)}$ and the score $f(\mathbf{x}^{(i)} | \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{x}$

- **NB:** Note that $\frac{\partial}{\partial f} s(f) = s(f)(1 - s(f))$ and $\frac{\partial f(\mathbf{x}^{(i)} | \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = (\mathbf{x}^{(i)})^\top$
- For more details we refer to the [i2ml](#) lecture

LOGISTIC REGRESSION

- Partial derivative of empirical risk using chain rule:

$$\begin{aligned}\frac{\partial}{\partial \theta} \mathcal{R}_{\text{emp}}(\theta) &= - \sum_{i=1}^n \frac{\partial}{\partial \pi^{(i)}} (y^{(i)} \log(\pi^{(i)}) + (1 - y^{(i)}) \log(1 - \pi^{(i)})) \frac{\partial \pi^{(i)}}{\partial \theta} \\ &= - \sum_{i=1}^n \left(\frac{y^{(i)}}{\pi^{(i)}} - \frac{1 - y^{(i)}}{1 - \pi^{(i)}} \right) \frac{\partial s(f(\mathbf{x}^{(i)} | \theta))}{\partial f(\mathbf{x}^{(i)} | \theta)} \frac{\partial f(\mathbf{x}^{(i)} | \theta)}{\partial \theta} \\ &= \sum_{i=1}^n (\pi^{(i)} - y^{(i)}) (\mathbf{x}^{(i)})^\top \\ &= (\pi(\mathbf{X} | \theta) - \mathbf{y})^\top \mathbf{X}\end{aligned}$$



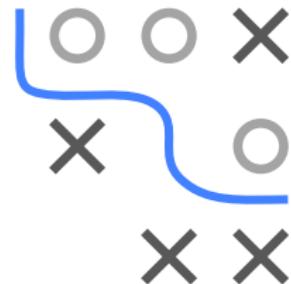
where $\mathbf{X} = (\mathbf{x}^{(1)^\top}, \dots, \mathbf{x}^{(n)^\top})^\top \in \mathbb{R}^{n \times (p+1)}$, $\mathbf{y} = (y^{(1)}, \dots, y^{(n)})^\top$,
 $\pi(\mathbf{X} | \theta) = (\pi^{(1)}, \dots, \pi^{(n)})^\top \in \mathbb{R}^n$

- $\nabla_\theta \mathcal{R}_{\text{emp}} = \left(\frac{\partial}{\partial \theta} \mathcal{R}_{\text{emp}} \right)^\top$

LOGISTIC REGRESSION

- The Hessian of logistic regression:

$$\begin{aligned}\nabla_{\theta}^2 \mathcal{R}_{\text{emp}} &= \frac{\partial^2}{\partial \theta^{\top} \partial \theta} \mathcal{R}_{\text{emp}} = \frac{\partial}{\partial \theta^{\top}} \sum_{i=1}^n \left(\pi^{(i)} - y^{(i)} \right) \left(\mathbf{x}^{(i)} \right)^{\top} \\ &= \sum_{i=1}^n \mathbf{x}^{(i)} \left(\pi^{(i)} \left(1 - \pi^{(i)} \right) \right) \left(\mathbf{x}^{(i)} \right)^{\top} \\ &= \mathbf{X}^{\top} \mathbf{D} \mathbf{X}\end{aligned}$$



where $\mathbf{D} \in \mathbb{R}^{n \times n}$ is a diagonal matrix containing the variances of $y^{(i)}$ on the diagonals

$$\mathbf{D} = \text{diag} \left(\pi^{(1)}(1 - \pi^{(1)}), \dots, \pi^{(n)}(1 - \pi^{(n)}) \right)$$

LOGISTIC REGRESSION

- We now have

$$\nabla_{\theta} \mathcal{R}_{\text{emp}} = \mathbf{X}^{\top} (\pi(\mathbf{X} | \theta) - \mathbf{y})$$

$$\nabla_{\theta}^2 \mathcal{R}_{\text{emp}} = \mathbf{X}^{\top} \mathbf{D} \mathbf{X}$$

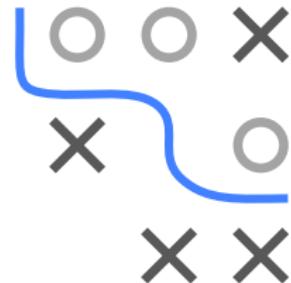
- Newton-Raphson:

$$\theta^{[t+1]} = \theta^{[t]} - [\mathbf{X}^{\top} \mathbf{D} \mathbf{X}]^{-1} \nabla_{\theta^{[t]}} \mathcal{R}_{\text{emp}}$$

- Fisher scoring:

$$\theta^{[t+1]} = \theta^{[t]} - \mathbb{E}[\mathbf{X}^{\top} \mathbf{D} \mathbf{X}]^{-1} \nabla_{\theta^{[t]}} \mathcal{R}_{\text{emp}}$$

- Note that the Hessian does not depend on the $y^{(i)}$ explicitly but only depends on $\mathbb{E}[y^{(i)}] = \pi^{(i)}$. Thus the expectation of the observed Hessian w.r.t. $y^{(i)} \sim P(y^{(i)} | \mathbf{x}^{(i)}, \theta)$ coincides with $\nabla_{\theta}^2 \mathcal{R}_{\text{emp}}(\theta)$ itself



GENERALIZED LINEAR MODELS

- $y|\mathbf{x}$ belongs to an **exponential family** with density:

$$p(y|\delta, \phi) = \exp \left\{ \frac{y\delta - b(\delta)}{a(\phi)} + c(y, \phi) \right\},$$

where δ is the natural parameter and $\phi > 0$ is the dispersion parameter

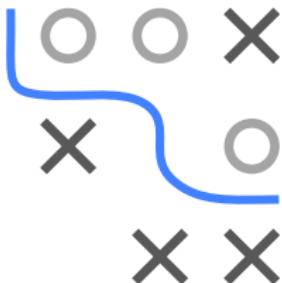
- We often take $a_i(\phi) = \frac{\phi}{w_i}$, with ϕ a positive constant, and w_i is a weight

- Generalized linear models (GLMs) relate the conditional mean

$\mu(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$ of y to a linear predictor η via a strictly increasing link function $g(\mu) = \eta = \mathbf{x}^\top \boldsymbol{\theta}$

- One can show that mean $\mu = \mu(\mathbf{x}) = b'(\delta) = g^{-1}(\eta)$, variance $\text{Var}(y|\mathbf{x}) = a(\phi)b''(\delta)$, where

$$\frac{\partial b(\delta)}{\partial \theta} = \frac{\partial b(\delta)}{\partial \delta} \frac{\partial \delta}{\partial \mu} \frac{\partial \mu}{\partial \eta} \frac{\partial \eta}{\partial \theta} = \mu \frac{1}{b''(\delta)} \frac{1}{g'(\mu)} \mathbf{x}$$

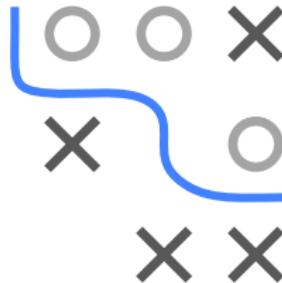


GENERALIZED LINEAR MODELS

- We can estimate δ using MLE with sample $(\mathbf{x}^{(i)}, y^{(i)})$ for $i = 1, \dots, n$
- Take $a^{(i)}(\phi) = \frac{\phi}{w^{(i)}}$, ϕ is a positive constant, we could ignore it since the goal is to maximize the function:

$$\begin{aligned}\nabla \ell_\theta(\delta, \phi) &= \sum_{i=1}^n \frac{w_i(y^{(i)} - \mu^{(i)})}{b''(\delta)g'(\mu^{(i)})} \mathbf{x}^{(i)} \\ &= \sum_{i=1}^n \frac{w^{(i)}(y^{(i)} - \mu^{(i)})g'(\mu^{(i)})}{b''(\delta)[g'(\mu^{(i)})]^2} \mathbf{x}^{(i)} \\ &= \mathbf{X}^\top \mathbf{W} \mathbf{G} (\mathbf{Y} - \boldsymbol{\mu})\end{aligned}$$

- \mathbf{W} is a diagonal matrix with element $\frac{w^{(i)}}{b''(\delta)[g'(\mu^{(i)})]^2}$
- \mathbf{G} is a diagonal matrix with element $g'(\mu^{(i)})$



GENERALIZED LINEAR MODELS



$$\begin{aligned}-\nabla^2 \ell_\theta(\delta, \phi) &= \sum_{i=1}^n \frac{\mathbf{w}^{(i)}}{b''(\delta)[g'(\mu^{(i)})]^2} \mathbf{x}^{(i)} \mathbf{x}^{(i)\top} \\&+ \sum_{i=1}^n \frac{\mathbf{w}^{(i)}(y^{(i)} - \mu^{(i)})(b''(\delta)g''(\mu^{(i)})/g'(\mu^{(i)}))}{[b''(\delta)g'(\mu^{(i)})]^2} \mathbf{x}^{(i)} \mathbf{x}^{(i)\top} \\&+ \sum_{i=1}^n \frac{\mathbf{w}^{(i)}(y^{(i)} - \mu^{(i)})(b'''(\delta)/b''(\delta))}{[b''(\delta)g'(\mu^{(i)})]^2} \mathbf{x}^{(i)} \mathbf{x}^{(i)\top}\end{aligned}$$

$$\mathbb{E}[-\nabla^2 \ell_\theta(\delta, \phi)] = \sum_{i=1}^n \frac{\mathbf{w}^{(i)}}{b''(\delta)[g'(\mu^{(i)})]^2} \mathbf{x}^{(i)} \mathbf{x}^{(i)\top} = \mathbf{X}^\top \mathbf{W} \mathbf{X}$$

- Iteratively Reweighted Least Squares (IRLS) with weights $\frac{\mathbf{w}^{(i)}}{b''(\delta)[g'(\mu^{(i)})]^2}$

GENERALIZED LINEAR MODELS

- Fisher scoring:

$$\begin{aligned}\theta^{[t+1]} &= \theta^{[t]} + (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{G}(\mathbf{Y} - \mu) \\ &= (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \left(\mathbf{G}(\mathbf{Y} - \mu) + \mathbf{X} \theta^{[t]} \right)\end{aligned}$$



- For canonical link where $\eta = \delta (= g(\mu) = \mathbf{x}^\top \theta)$, the second and third term of Hessian cancel each other out and Hessian coincides with Fisher information matrix since

$$\frac{\partial \eta}{\partial \delta} = 1 \Rightarrow b''(\delta) = \frac{1}{g'(\mu^{(i)})} \Rightarrow \frac{b'''(\delta)}{b''(\delta)} = -\frac{g''(\mu^{(i)})}{[g'(\mu^{(i)})]^2}$$

- This will now be a convex problem with Fisher scoring equal to Newton's method
- There are also hybrid algorithms that start out with IRLS which is easier to initialize, and switch over to Newton-Raphson after some iterations