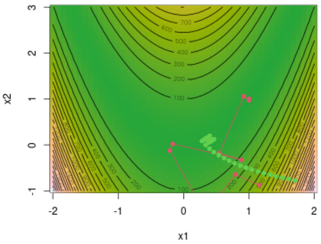
Optimization in Machine Learning

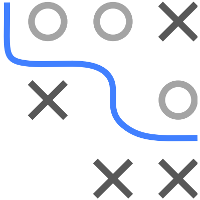
Second order methods
Fisher Scoring



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 $\boldsymbol{\theta}^{(t)}$:

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

We then differentiate w.r.t. $\boldsymbol{\theta}$ and set the gradient to zero:

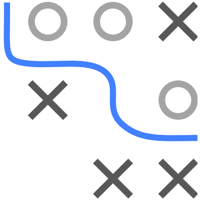
$$\nabla\ell(\boldsymbol{\theta}^{(t)}) + [\nabla^2\ell(\boldsymbol{\theta}^{(t)})](\boldsymbol{\theta}-\boldsymbol{\theta}^{(t)}) = \mathbf{0}$$

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

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

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* $-\nabla^2 \ell(\boldsymbol{\theta})$ by the Fisher information matrix, i.e., the variance of $\nabla \ell(\boldsymbol{\theta})$, which, under weak regularity conditions, equals the negative *expected Hessian*

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

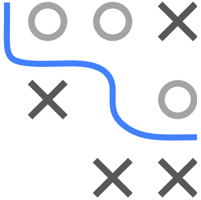
and is positive semi-definite under exchangeability of expectation and differentiation.

NB: it can be shown that $\mathbb{E}[\nabla \ell(\boldsymbol{\theta})] = \mathbf{0}$, which provides the expression of the variance of $\nabla \ell(\boldsymbol{\theta})$ as the expected outer product of the gradients.

Therefore the Fisher scoring iterates are given by

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

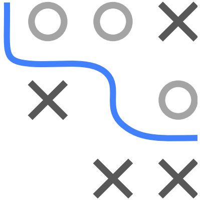
NEWTON-RAPHSON VS. FISHER SCORING



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 $\left(\mathbf{x}^{(i)}, y^{(i)}\right) \in \mathbb{R}^{p+1} \times \{0, 1\}$, $y^{(i)} | \, \mathbf{x}^{(i)} \sim \mathit{Bernoulli}(\pi^{(i)})$.

We want to minimize the following risk

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

with respect to $\boldsymbol{\theta}$, where the probabilistic classifier $\pi^{(i)} = \pi\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right) = s\left(f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)$, the sigmoid function $s(f) = \frac{1}{1+\exp(-f)}$ and the score $f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right) = \boldsymbol{\theta}^\top \mathbf{x}$.

NB: Note that $\frac{\partial}{\partial f} s(f) = s(f)(1 - s(f))$ and $\frac{\partial f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} = \left(\mathbf{x}^{(i)}\right)^\top$.

For more details we refer to the i2ml lecture.

LOGISTIC REGRESSION / 2

Partial derivative of empirical risk using chain rule:

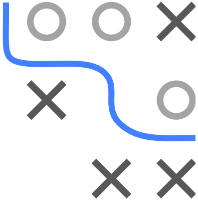
$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) &= -\sum_{i=1}^{n} \frac{\partial}{\partial \pi^{(i)}} \left(y^{(i)} \log(\pi^{(i)}) + (1 - y^{(i)}) \log(1 - \pi^{(i)}) \right) \frac{\partial \pi^{(i)}}{\partial \boldsymbol{\theta}} \\
&=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

where $\mathbf{X} = \left(\mathbf{x}^{(1)\top}, \ldots, \mathbf{x}^{(n)\top} \right)^\top \in \mathbb{R}^{n \times (p+1)}, \mathbf{y} = \left(y^{(1)}, \ldots, y^{(n)} \right)^\top,$

$\pi(\mathbf{X} |\ \boldsymbol{\theta}) = \left(\pi^{(1)}, \ldots, \pi^{(n)} \right)^\top \in \mathbb{R}^n.$

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

LOGISTIC REGRESSION / 3



The Hessian of logistic regression:

$$\nabla^2_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}} = \frac{\partial^2}{\partial \boldsymbol{\theta}^\top \partial \boldsymbol{\theta}} \mathcal{R}_{\text{emp}} = \frac{\partial}{\partial \boldsymbol{\theta}^\top} \sum_{i=1}^{n} \left(\pi^{(i)} - y^{(i)}\right) \left(\mathbf{x}^{(i)}\right)^\top$$

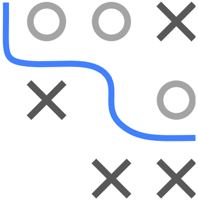
$$= \sum_{i=1}^{n} \mathbf{x}^{(i)} \left(\pi^{(i)} \left(1 - \pi^{(i)}\right)\right) \left(\mathbf{x}^{(i)}\right)^\top$$

$$= \mathbf{X}^\top \mathbf{D} \mathbf{X}$$

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

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

LOGISTIC REGRESSION / 4



We now have

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

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

Newton-Raphson:

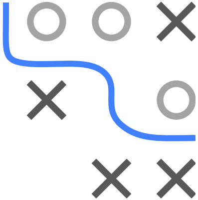
$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - [\mathbf{X}^\top \mathbf{D} \mathbf{X}]^{-1} \nabla_{\boldsymbol{\theta}^{(t)}} \mathcal{R}_{\text{emp}}$$

Fisher scoring:

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

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. $y^{(i)} \sim P(y^{(i)}|\mathbf{x}^{(i)}, \boldsymbol{\theta})$ coincides with $\nabla^2_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ 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 pos. 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 $g(\mu) = \eta = \mathbf{x}^\top\theta$.

One can show that mean $\mu = \mu(\mathbf{x}) = b'(\delta) = g^{-1}(\eta)$, variance $Var(y|\mathbf{x}) = a(\phi)b''(\delta)$, 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 / 2

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:

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

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

G is a diagonal matrix with element $g'(\mu^{(i)})$.

GENERALIZED LINEAR MODELS / 3

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

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

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

GENERALIZED LINEAR MODELS / 4

Fisher scoring:

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

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.