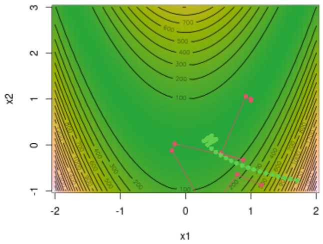
Optimization in Machine Learning

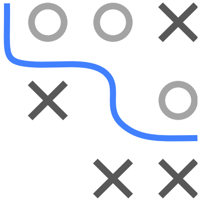
Second order methods
Fisher Scoring



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

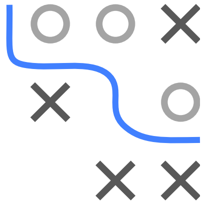
$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \mathbb{E}[-\nabla^2 \ell(\boldsymbol{\theta}^{(t)})]^{-1} \nabla \ell(\boldsymbol{\theta}^{(t)})$$

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

We want to minimize the following risk

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

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.

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

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