# Optimizations for statistical applications I

Stat 580: Statistical Computing

- Theme: Black - White
- Printable version

# Motivating example: likelihood inference

- Let $x_1, \ldots, x_n$ be an iid sample from $f(x|\boldsymbol{\theta}^*)$ where the true parameter value $\boldsymbol{\theta}^*$ is unknown.

- The likelihood function is
$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})$$
where the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ is the maximizer of $L(\boldsymbol{\theta})$.

- Usually it is easier to work with the log likelihood $l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$.

- Typically, maximization of $l(\boldsymbol{\theta})$ is done by solving $l'(\boldsymbol{\theta})$, the score function.

# Motivating example: likelihood inference

- For any $\boldsymbol{\theta}$

$$E_{\boldsymbol{\theta}}\{l'(\boldsymbol{\theta})\} = 0$$

$$E_{\boldsymbol{\theta}}\{l'(\boldsymbol{\theta})l'(\boldsymbol{\theta})^T\} = -E_{\boldsymbol{\theta}}\{l''(\boldsymbol{\theta})\},$$

where $E_{\boldsymbol{\theta}}$ is expectation w.r.t. $f(x|\boldsymbol{\theta})$.

- Fisher Information: $I(\boldsymbol{\theta}) = -E_{\boldsymbol{\theta}}\{l''(\boldsymbol{\theta})\}$
- Observed Fisher Information: $-l''(\boldsymbol{\theta})$

# Motivating example: likelihood inference

- If dim($\boldsymbol{\theta}$) = 1, $I(\boldsymbol{\theta})$ is a nonnegative number. If $\dim(\boldsymbol{\theta}) > 1$, $I(\boldsymbol{\theta})$ is a nonnegative definite matrix.

- $I(\boldsymbol{\theta})$ sets the limit on how accurate an unbiased estimate of $\boldsymbol{\theta}$ can be.

- As $n \to \infty$, $\sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}^*) \Rightarrow \mathcal{N}(0, nI(\boldsymbol{\theta}^*)^{-1})$

# Working with Derivatives

- Suppose $g(\boldsymbol{x})$ is a differentiable function, where $\boldsymbol{x} = (x_1, \ldots, x_n)$.

- To find its (local) maximum or minimum, one method is to solve the equation $g'(\boldsymbol{x}) = 0$, where
$$g'(\boldsymbol{x}) = \left( \frac{\partial g}{\partial x_1}, \ldots, \frac{\partial g}{\partial x_n} \right)^T$$
That is, this is equivalent to solving $f(\boldsymbol{x}) = 0$ where $f = g'$.

# Univariate Case: Newton's Method

- A fast approach for solving $f(x) = 0$:

1. Start with an initial estimate $x_0$.
2. For $t = 0, 1, \ldots,$ compute $x_{t+1} = x_t + h_t$, where
$$h_t = \frac{-f(x_t)}{f'(x_t)}$$
3. Continue until convergence:
$$\frac{|x_{t+1} - x_t|}{|x_t + \Delta|} < \varepsilon$$
where $\Delta$ is small (e.g., 0.00005).

# Univariate Case: Newton's Method

- Also known as Newton-Raphson method
- Need to specify $x_0$
- If $f(x) = 0$ has multiple roots, end result will depend on $x_0$
- Iteration cannot continue if $f'(x_t) = 0$. Try a different initial value if this happens.

# Why does it work?

- Let $x^0$ be the true solution and $\tilde{x}$ be an approximation of $x^0$.

- Taylor expansion:

  $$f(x) = f(\tilde{x}) + (x - \tilde{x})f'(\tilde{x}) + \frac{(x - \tilde{x})^2}{2} f''(\hat{x}),$$

  where $\hat{x}$ lies between $x$ and $\tilde{x}$.

- Since $f(x^0) = 0$, we have

  $$0 = f(\tilde{x}) + (x^0 - \tilde{x})f'(\tilde{x}) + \frac{(x^0 - \tilde{x})^2}{2} f''(\hat{x})$$

- If $x^0$ and $\tilde{x}$ are close, the last term can be ignored:

  $$0 \approx f(\tilde{x}) + (x^0 - \tilde{x})f'(\tilde{x}) \quad \Rightarrow \quad x^0 \approx \tilde{x} - \frac{f(\tilde{x})}{f'(\tilde{x})}$$

# Optimization with Newton's method

- Can be applied to optimize $g$ by applying Newton's method to $f = g'$.
- Both $g'$ (gradient) and $g''$ (Hessian) are needed.
- Computation of $g''$ could be difficult, especially for multi-dimensional function. Many variants of Newton's method avoid the computation of $g''$.

# Example 1

To maximize $g(x) = \frac{\log x}{1+x}$, first find

$$f(x) = g'(x) = \frac{1 + \frac{1}{x} - \log x}{(1+x)^2}$$

$$f'(x) = g''(x) = \frac{-(3 + \frac{4}{x} + \frac{1}{x^2} - 2\log x)}{(1+x)^3}$$

Therefore,

$$h_t = \frac{(x_t + 1)(1 + \frac{1}{x_t} - \log x_t)}{3 + \frac{4}{x_t} + \frac{1}{x_t^2} - 2\log x_t}$$

# Example 1

A simpler formula: note that solving $f(x) = 0$ is the same as solving $1 + \frac{1}{x} - \log(x) = 0$. Treat $1 + \frac{1}{x} - \log(x)$ as a new function.

Then,

$$h_t = x_t - \frac{x_t^2 \log x_t}{1 + x_t} \qquad \Rightarrow \qquad x_{t+1} = 2x_t - \frac{x_t^2 \log x_t}{1 + x_t}.$$

# Example 2

To maximize log likelihood $l(\theta)$:

$$\theta_{t+1} = \theta_t - \frac{l(\theta_t)}{l''(\theta_t)}$$

Consider the model with shift $p(x|\theta) = p(x - \theta)$. Given observations $x_1, \ldots, x_n$ iid $\sim p(x|\theta)$,

$$l(\theta) = \sum_{i=1}^{n} \log p(x_i - \theta)$$

$$l'(\theta) = -\sum_{i=1}^{n} \frac{p'(x_i - \theta)}{p(x_i - \theta)}$$

$$l''(\theta) = \sum_{i=1}^{n} \frac{p''(x_i - \theta)}{p(x_i - \theta)} - \sum_{i=1}^{n} \left\{ \frac{p'(x_i - \theta)}{p(x_i - \theta)} \right\}^2$$

# Secant Method

- Approximate $f'(x)$ by $\frac{f(x_t)-f(x_{t-1})}{x_t-x_{t-1}}$.
- The Newton's method becomes the secant method:
$$x_{t+1} = x_t - \frac{f(x_t)(x_t - x_{t-1})}{f(x_t) - f(x_{t-1})}.$$
We need to specify $x_0$ and $x_1$ to begin the iterations.
- Remember that $f' = g''$

# Fisher Scoring

- Another variant of Newton's method
- Specific for MLE
- Replace the Hessian $l''(\theta)$ by its expectation, i.e., Fisher information:

$$\theta_{t+1} = \theta_t + \frac{l'(\theta_t)}{I(\theta_t)}.$$

- In practice, use Fisher Scoring at the beginning to make rapid improvement, then Newton's method for refinement near the end.

# Example

Continuing from $p(x|\theta) = p(x-\theta)$, using Fisher Scoring. We need to compute $I(\theta) = -E_\theta\{l''(\theta)\}$.

$$
\begin{aligned}
I(\theta) &= -nE_\theta\left[\frac{p''(x-\theta)}{p(x-\theta)} - \left\{\frac{p'(x-\theta)}{p(x-\theta)}\right\}^2\right] \\
&= -n\int\left[\frac{p''(x-\theta)}{p(x-\theta)} - \left\{\frac{p'(x-\theta)}{p(x-\theta)}\right\}^2\right]p(x-\theta)dx \\
&= -n\int p''(x-\theta)dx + n\int\frac{\{p'(x-\theta)\}^2}{p(x-\theta)}dx \\
&= -n\frac{d^2}{d\theta^2}\int p(x-\theta)dx + n\int\frac{p'(x)^2}{p(x)}dx \\
&= -n\frac{d^2}{d\theta^2}1 + n\int\frac{p'(x)^2}{p(x)}dx \\
&= 0 + n\int\frac{p'(x)^2}{p(x)}dx
\end{aligned}
$$

# Multivariate Case: Newton's Method

- Now, $g$ is a function of $\boldsymbol{x} = (x_1, \ldots, x_p)^T$.

- Generalization is straight forward. To maximize or minimize $g(\boldsymbol{x})$, use
$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \{g''(\boldsymbol{x}_t)\}^{-1} g'(\boldsymbol{x}_t)$$
where
  - $g''(\boldsymbol{x})$ is a $p \times p$ matrix with $(i, j)$-th element as $\frac{\partial^2 g(\boldsymbol{x})}{\partial x_i \partial x_j}$
  - $g'(\boldsymbol{x}) = \left[\frac{\partial g(\boldsymbol{x})}{\partial x_1}, \ldots, \frac{\partial g(\boldsymbol{x})}{\partial x_p}\right]^T$, a $p \times 1$ vector.
  - note: need to compute the inverse of $g''(\boldsymbol{x}_t)$.

# Multivariate Case: Newton's Method

- If $g''(\boldsymbol{x}_t)$ is near-singular, try replacing it with $M_t = g''(\boldsymbol{x}_t) + \alpha I$, where $\alpha$ is as small as possible (increase it until $M_t$ is stable).
- When $p$ is big, to speed up the process, only update $g''(\boldsymbol{x}_t)$ every second iteration. That is, every other time, we are using $g''(\boldsymbol{x}_t)$ and otherwise, we use $g''(\boldsymbol{x}_{t-1})$.

# Multivariate Case: Fisher Scoring

- Use $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + I(\boldsymbol{\theta}_t)^{-1} l'(\boldsymbol{\theta}_t)$.

# Multivariate Case: Other Newton-like Methods

- Computing $g''(\boldsymbol{x})$ or $\{g''(\boldsymbol{x})\}^{-1}$ could be hard.

- The idea is to replace $g''(\boldsymbol{x})$ by some easily computable matrix, say $M(\boldsymbol{x})$. Then:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - M(\boldsymbol{x}_t)^{-1} g'(\boldsymbol{x}_t).$$

# Multivariate Case: Steepest Ascent Method

- Set $M(\boldsymbol{x}_t) = -\alpha_t^{-1} I_p$, where $I_p$ is the $p \times p$ identity matrix and $\alpha_t > 0$ is the step size which can shrink to ensure ascent. Then:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \alpha_t g'(\boldsymbol{x}_t).$$

- If at step $t$, the original step turns out to be downhill; i.e., if $g(\boldsymbol{x}_{t+1}) < g(\boldsymbol{x})$, the updating can be backtracked by halving $\alpha_t$.

- Also known as steepest descent (for minimization).

# Gauss-Newton Method

- Want to maximize $g(\boldsymbol{\theta}) = -\sum_{i=1}^{n}\{y_i - f_i(\boldsymbol{\theta})\}^2$ where each $f_i(\boldsymbol{\theta})$ is differentiable.

- First consider linear regression:

$$y_i = \boldsymbol{x}_i^T\boldsymbol{\theta} + \varepsilon, \qquad i = 1, \ldots, n$$

where $f_i(\boldsymbol{\theta}) = \boldsymbol{x}_i^T\boldsymbol{\theta}$ in this case.

- The least squares estimator of $\boldsymbol{\theta}$ maximizes $g(\boldsymbol{\theta})$ with $f_i(\hat{\boldsymbol{\theta}}) = \boldsymbol{x}_i^T\hat{\boldsymbol{\theta}}$.

- $\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ and $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$.

- Gauss-Newton uses a similar idea for nonlinear $f_i(\boldsymbol{\theta})$.

# Gauss-Newton Method

- Let $\boldsymbol{\theta}^*$ be the unknown maximizer of $g(\boldsymbol{\theta})$.
- Consider $h(\boldsymbol{u}) = -\sum_{i=1}^{n}\{y_i - f_i(\boldsymbol{\theta} - \boldsymbol{u})\}^2$.
- $h(\boldsymbol{u})$ is maximized by $\boldsymbol{u}^* = \boldsymbol{\theta}^* - \boldsymbol{\theta}$ ($\boldsymbol{u}^*$ is unknown).
- If $\boldsymbol{\theta}$ is near $\boldsymbol{\theta}^*$, $\boldsymbol{u}^* \approx 0$ and by Taylor expansion of $h(\boldsymbol{u})$, $\boldsymbol{u}^*$ should be close to the maximizer of
$$-\sum_{i=1}^{n}\{y_i - f_i(\boldsymbol{\theta}) - f_i'(\boldsymbol{\theta})^T\boldsymbol{u}\}^2.$$
Now, this resembles the classical linear regression problem with $y_i = y_i - f_i(\boldsymbol{\theta}), \boldsymbol{x}_i^T = f_i'(\boldsymbol{\theta})^T$, and $\boldsymbol{\theta} = \boldsymbol{u}$.

# Gauss-Newton Method

- We have

$$\boldsymbol{u}^* = \boldsymbol{\theta}^* - \boldsymbol{\theta} \approx (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{Z}$$

where

$$\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{\theta}) = \{f_1'(\boldsymbol{\theta}), \ldots, f_n'(\boldsymbol{\theta})\}^T$$

and

$$\boldsymbol{Z} = \boldsymbol{Z}(\boldsymbol{\theta}) = \{y_1 - f_1(\boldsymbol{\theta}), \ldots, y_n - f_n(\boldsymbol{\theta})\}^T.$$

- The updating formula is

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + (\boldsymbol{A}_t^T\boldsymbol{A}_t)^{-1}\boldsymbol{A}_t^T\boldsymbol{Z}_t$$

where $\boldsymbol{A}_t = \boldsymbol{A}(\boldsymbol{\theta}_t)$ and $\boldsymbol{Z}_t = \boldsymbol{Z}(\boldsymbol{\theta}_t)$.