Lecture 12: Optimization + EM Lecture # 1

- Overview: Today we review basic optimization algorithms and the (vanilla) EM algorithm. In the rest of the module we will look at more advanced (and more useful) variants of the EM algorithm.

- First, note that most optimization problems (e.g., finding MLE's, MAP's, even CI's) can be reduced to finding the root of an equation i.e., solving for:

$$g(x) = 0.$$

  i.e., to maximize $f(x)$, we can just solve $g(x) = f'(x) = 0$. Lets review some basic root-finding algorithms: bisection, Newton-Raphson and Fisher Scoring.

- **Bisection**:
  - Let $g : \mathbb{R} \mapsto \mathbb{R}$ be a continuous function on $[a, b]$ s.t. $g(a) \cdot g(b) < 0$.
  - Intermediate Value Theorem $\Rightarrow g(x) = 0$for some $x \in (a, b)$.
  - Let $l = a$, $u = b$ be lower and upper boundaries and fix a precision $\epsilon > 0$. The algorithm is as follows:

```
converged = False
while (!converged):
 c := (l + u)/2
 if |g(c)| < epsilon:
   converged = True #solved
 else:
   if (g(l)  g(c) < 0):
     u = c
   else: # g(c) x g(u) < 0
     l = c
return c
```

    What is this doing graphically? Pro's and con's of the algorithm? Interval length?

- **Newton-Raphson:** Arguably the most famous root-finding/optimization algorithm. Iterative algorithm to solve for x_* where $g(x_*) = 0$. Let x_t be the current guess for x_* then update via:

$$x_{t+1} = x_t + \eta_t.$$

  How to select η_t? Taylor expansion of g, if η_t is small:

$$g(x_{t+1}) = g(x_t + \eta_t) \approx g(x_t) + \eta_t g'(x_t) + O(\eta_t^2)$$

  Solving for $g(x_{t+1}) = 0$ gives $\eta_t = -g(x_t)/g'(x_t)$ i.e.,

$$x_{t+1} = x_t - \frac{g(x_t)}{g'(x_t + 1)}.$$

  When *maximizing* $l(\theta)$ this becomes:

$$x_{t+1} = x_t - \frac{l'(x_t)}{l''(x_t + 1)}.$$

For multivariate $g : \mathbb{R}^p \mapsto \mathbb{R}^p$ we obtain:

$$\vec{x}_{t+1} = \vec{x}_t - [\nabla g(\vec{x}_t)]^{-1} g(\vec{x}_t).$$

e.g., to *maximize* $\ell : \mathbb{R}^p \mapsto \mathbb{R}$:

$$\vec{x}_{t+1} = \vec{x}_t - \left[\nabla \nabla^T \ell(\vec{x}_t)\right]^{-1} \nabla \ell(\vec{x}_t).$$

Quadratic convergence:

$$\lim_{t \to \infty} \frac{|x_{t+1} - x_*|}{|x_t - x_*|^2} = c < \infty.$$

Pros and cons?

- **Scoring:** Here we introduce a special modification of Newton-Raphson tailored to statistical applications. Recall the NR update to maximize ℓ:

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

The scoring algorithm instead uses:

$$\theta_{t+1} = \theta_t + \frac{\ell'(\theta_t)}{\mathcal{I}(\theta_t)},$$

where:

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

is the expected Fisher information. Multivariate version:

$$\theta_{t+1} = \theta_t + \mathcal{I}^{-1}(\theta_t)\ell'(\theta_t)$$

Pros and cons?

- Examples...

- **The EM Algorithm:** While NR and scoring are useful for many problems, there are many more where complications arise. For example, the likelihood may involve integrals with no analytic form e.g., Generalized Linear Mixed Models. Simple nested Binomial GLMM with Normal random effects:

$$\begin{aligned} \eta_{ij} &= x_{ij}^T \beta + z_i^T \gamma_i, \\ Y_{ij} | \gamma_i &\sim \text{Bin}(n_{ij}, g^{-1}(\eta_{ij})), \\ \gamma_i &\sim N\left(0, \Sigma^{-1}\right). \end{aligned}$$

What is the likelihood here?

Enter the EM algorithm: designed for maximizing likelihoods (or posterior distributions) in the presence of 'missing data'. As we will see, it turns out that 'missing data' is defined very loosely, and in reality there does not need to be any actual 'data' that is missing.

The algorithm is based upon the Q-function, the expected complete-data log-likelihood:

$$Q(\theta|\theta^{(t)}) = \mathbb{E}\left[l(Y_{com}|\theta)|Y_{obs}, \theta = \theta^{(t)}\right] \tag{1}$$

The algorithm works by selecting an initial $\theta^{(0)}$, setting $t = 0$ and iteratively computing:

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta|\theta^{(t)})$$

It can be shown that $\lim_{t\to\infty} \theta^{(t)} = \theta_*$ where $\ell'(\theta_*) = 0$ i.e., the algorithm converges to a (possibly local) mode of the log-likelihood function.

- Example: Probit Regression.
- Important implementation notes.
- Proof of convergence.
- Extension to finding MAP estimators.

- The EM alphabet soup. Next few lectures, we will some of these variants...
  - ECM, MCEM, MCECM, MCMCEM, MCMCECM, AECM, PXEM, IEM

These variants are designed to address various complications that arise in practice. Some are designed to improve the rate of convergence, some to improve tractability, some to approximate quantities that cannot be computed analytically.