MAT 653: Statistical Simulation

Instructor: Dr. Wei Li Scribe: Jiangyu Yu

Oct 25th, 2021

EM Algorithm (Deterministic Optimization)

Suppose we have n observables $x_1, x_2, ..., x_n \overset{i.i.d}{\sim} g(x|\theta)$. Let $\boldsymbol{x} = \{x_i\}_{i=1}^n$. Our goal is to compute:

$$\hat{\theta} = \arg\max L(\theta|\boldsymbol{x}) = \prod_{i=1}^{n} g(x_i|\theta),$$

Where $L(\theta|\boldsymbol{x}) = f(\boldsymbol{x}|\theta)$ is the likelihood function.

Denote $L^c(\theta|\boldsymbol{x}, \boldsymbol{z}) = f(\boldsymbol{x}, \boldsymbol{z}|\theta)$, called the complete data likelihood function.

Augment the data with $\boldsymbol{z}$ which are unobservable, so

$$(\boldsymbol{x}, \boldsymbol{z}) \sim f(\boldsymbol{x}, \boldsymbol{z}|\theta).$$

The conditional distribution of $\boldsymbol{z}$ given the observables $\boldsymbol{x}$ is

$$f(\boldsymbol{z}|\boldsymbol{x}, \theta) = \frac{f(\boldsymbol{x}, \boldsymbol{z}|\theta)}{f(\boldsymbol{x}|\theta)}.$$

Use this result on the likelihood function:

$$\begin{aligned} logL(\theta|\boldsymbol{x}) &= logf(\boldsymbol{x}|\theta) \\ &= log\frac{f(\boldsymbol{x}, \boldsymbol{z}|\theta)}{f(\boldsymbol{z}|\boldsymbol{x}, \theta)} \\ &= log(f(\boldsymbol{x}, \boldsymbol{z}|\theta)) - log(f(\boldsymbol{z}|\boldsymbol{x}, \theta)). \end{aligned}$$

We use the notations:

$$\begin{aligned} E_g(f(\boldsymbol{x})) &= \int f(\boldsymbol{x})g(\boldsymbol{x})d\boldsymbol{x} \\ E_{\boldsymbol{z}|\boldsymbol{x}}(h(\boldsymbol{z}, \boldsymbol{x})) &= \int h(\boldsymbol{z}, \boldsymbol{x})f(\boldsymbol{z}|\boldsymbol{x})d\boldsymbol{z}. \end{aligned}$$

Let $\theta^{(0)}$ as our initial guess of the parameter, and take conditional expectation of $\boldsymbol{z}$ given $\boldsymbol{x}$, that is, the integral is taken with respect to $f(\boldsymbol{z}|\boldsymbol{x}, \theta^{(0)})$, on both sides:

$$\begin{aligned} E_{\theta^{(0)}}(log(L(\theta|\boldsymbol{x}))) = log(L(\theta|\boldsymbol{x})) &= E_{\theta^{(0)}}[log f(\boldsymbol{x},\boldsymbol{z}|\theta)|\boldsymbol{x}] - E_{\theta^{(0)}}[log f(\boldsymbol{z}|\boldsymbol{x},\theta)|\boldsymbol{x}] \\ &= Q(\theta|\theta^{(0)},\boldsymbol{x}) - K(\theta|\theta^{(0)},\boldsymbol{x}), \end{aligned}$$

where we take $Q(\theta|\theta^{(0)},\boldsymbol{x}) = E_{\theta^{(0)}}[log f(\boldsymbol{x},\boldsymbol{z}|\theta)|\boldsymbol{x}]$. It turns out for any candidate $\theta^{'}$ for next iterate,

$$K(\theta^{'}|\theta^{0},\boldsymbol{x}) \leq K(\theta^{(0)}|\theta^{(0)},\boldsymbol{x}).$$

To see this,that is, for any $\theta^{'}$:

$$\begin{aligned} E_{\theta^{(m)}}(\log f(\boldsymbol{z}|\boldsymbol{x},\theta^{'})|\boldsymbol{x}) &\leq E_{\theta^{((m))}}(\log f(\boldsymbol{z}|\boldsymbol{x},\theta^{(m)})|\boldsymbol{x}) \\ &= \int \log f(\boldsymbol{z}|\boldsymbol{x},\theta^{(m)}) f(\boldsymbol{z}|x,\theta^{(m)}) d\boldsymbol{z}. \end{aligned}$$

Call $g(\boldsymbol{z}) = f(\boldsymbol{z}|\boldsymbol{x},\theta^{'})$, $h(\boldsymbol{z}) = f(\boldsymbol{z}|\boldsymbol{x},\theta^{(m)})$. It suffices to show

$$E_h\Big[\log\frac{h(z)}{g(z)}\Big] \geq 0.$$

In the above inequality, we use Jesen's inequality:

$$\begin{aligned} LHS &= \int log(\frac{h(z)}{g(z)}h(z))dz \\ &= -\int log(\frac{g(z)}{h(z)}h(z))dz \\ &\geq -log\int\frac{g(z)}{h(z)}h(z)dz = 0. \end{aligned}$$

So to maximize $log(L(\theta|\boldsymbol{x})$ over θ, it suffices to just maximize $Q(\theta|\theta^{(0)},\boldsymbol{x})$ over θ. By maximizing $Q(\theta|\theta^{(0)},\boldsymbol{x})$ over θ, one obtain the maximizer $\theta^{(1)}$ as the next iterate; we then by maximizing $Q(\theta|\theta^{(1)},\boldsymbol{x})$ over θ, obtaining the next iterate $\theta^{(2)}$– the process can keep go on.

EM algorithm

Based on the result, we have two main steps for EM algorithm: at step m,

(1) E Step: compute $Q(\theta|\theta^{(m)},\boldsymbol{x})$ as a function of θ and $\theta^{(m)}$.

(2) M Step: $\theta^{(m+1)} = \arg\max_{\theta} Q(\theta|\theta^{(m)},\boldsymbol{x})$.

Remark

(1) EM algorithm only generates the limit point of $\theta^{(m)}$ that is a stationary point of the objective function $log(L(\theta|\boldsymbol{x})$. In practice, you'll try different starting values of $\theta^{(0)}$.

(2) Notice that, for $h(x) = E[H(x, Z)]$ where the expectation is taken wrt to the random variable Z.

$$\begin{aligned}
\max_x h(x) &= \max_x E[H(x, Z)] \\
&= \max_x E[H(X, Z)|X = x] \\
&= \max_x \int H(x, z) f(z|x) dz,
\end{aligned}$$

we can use Monte Carlo to approximate the objective function:

$$\frac{1}{m} \sum_{i=1}^{m} H(x, Z_i) \to \int H(x, z) f(z|x) dz,$$

where $Z_i \overset{i.i.d}{\sim} f(z|x)$.

If we approximate Q function by this idea, this then gives the so-called Monte-Carlo EM:

$$\hat{Q}(\theta|\theta^{(m)}, \boldsymbol{x}) = \frac{1}{T} \sum_{j=1}^{T} log[L^c(\theta|\boldsymbol{x}, \boldsymbol{z}_j)].$$

where $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T$ is an i.i.d. random sample generated from $f(\boldsymbol{z}|\theta^{(m)}, \boldsymbol{x})$.

(3) We may not need to find the exact maximizer in the process . Instead, sometimes we just find $\theta^{(m+1)}$ that can improve upon the value of Q at the current $\theta^{(m)}$, that is,

$$Q(\theta^{(m+1)}|\theta^{(m)}, \boldsymbol{x}) \geq Q(\theta^{(m)}|\theta^{(m)}, \boldsymbol{z}),$$

we called that “generalized EM Algorithm".