

Welcome!

MH I

- goal is to sample from *target density*:

$$\pi(x) \propto \exp[-H(x)/\beta]$$

- the above form is known as the Boltzman form of a distribution
- $H(x)$ is called the fitness or energy function
- β is called the temperature
- EXAMPLE: target density for $X \sim \text{Normal}_1(\mu, \sigma^2)$:

$$\begin{aligned}\pi(x) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right] \\ &\propto \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]\end{aligned}$$

here $H(x) = \frac{1}{2\sigma^2}(x-\mu)^2$ and $\beta = 1$

MH II

- going to use a *proposal distribution pdf* to generate “guesses” or “proposals” for the draws from the target:

$$T(x, \cdot)$$

- EXAMPLE: given x , Normal proposal pdf for $Y \sim \mathbf{Normal}_1(x, \tau^2)$:

$$T(x, \cdot) \equiv \mathbf{Normal}_1(x, \tau^2; \cdot)$$

MH III

- going to have to evaluate the *proposal density pdf*:

$$T(x, y)$$

- EXAMPLE: Normal proposal density:

$$\begin{aligned} T(x, y) &\equiv \text{Normal}_1(x, \tau^2; y) \\ &\propto \exp \left[-\frac{1}{2\tau^2} (y - x)^2 \right] \end{aligned}$$

MH IV

- *acceptance probability* used in Metropolis-Hastings algorithm:

$$\alpha(x, y) = \min \left\{ 1, \frac{\pi(y)T(y, x)}{\pi(x)T(x, y)} \right\}$$

- if proposal is symmetric, i.e., if $T(x, y) = T(y, x)$ then we have:

$$\alpha(x, y) = \min \left\{ 1, \frac{\pi(y)}{\pi(x)} \right\}$$

- if $\pi(y) \geq \pi(x)$ then $\alpha(x, y) = 1$
- if $\pi(y) < \pi(x)$ then $\alpha(x, y) < 1$
- aside: if proposal is symmetric then the algorithm is called Metropolis algorithm
- note since we deal with ratios above its enough to know $\pi(\cdot)$ and $T(\cdot, \cdot)$ up to a proportionality constant

MH V

- the MH algorithm (for N -many iterations):
 1. initialize: set $t = 0$ and get a starting value $x^{(t)}$
 2. propose: generate y from $T(x^{(t)}, \cdot)$
 3. eval: evaluate acceptance probability $\alpha(x^{(t)}, y)$
 4. move: generate u from **Uniform**(0, 1) and set

$$x^{(t+1)} = \begin{cases} y & \text{if } u \leq \alpha(x^{(t)}, y) \\ x^{(t)} & \text{otherwise} \end{cases}$$

5. if $t \geq N$ stop otherwise set $t = t + 1$ and go to step 2
- aside: its enough to compute $\alpha(\cdot, \cdot)$ without the “min part” because $u \leq 1$, what???

MH VI

- process the samples: $\{x^{(t)} : t = 0, 1, \dots, N\}$
 - discard some initial samples, say, $\lfloor N/10 \rfloor$ is the “burn-in” period, for notational ease, reindex the rest as: $\{x^{(t)} : t = 1, 2, \dots, M\}$
 - use the rest for inference
 - EXAMPLE: to estimate the mean of the target density use the estimator:

$$\frac{1}{M} \sum_{t=1}^M x^{(t)}$$

MH VII

- a simple example:
- set up:
 - target: $X \sim \text{Normal}_1(\mu, \sigma^2)$
 - proposal: $Y \sim \text{Normal}_1(x, \tau^2)$
- so we have:
 - $\pi(x) \propto \exp \left[-\frac{1}{2\sigma^2} (x - \mu)^2 \right]$
 - $T(x, \cdot) \equiv \text{Normal}_1(x, \tau^2; \cdot)$
 - $T(x, y) \propto \exp \left[-\frac{1}{2\tau^2} (y - x)^2 \right]$, note its symmetric!
 - $\alpha(x, y) = \min \left\{ 1, \frac{\pi(y)}{\pi(x)} \right\} = \min \left\{ 1, \exp \left[-\frac{1}{2\sigma^2} (y - \mu)^2 + \frac{1}{2\sigma^2} (x - \mu)^2 \right] \right\}$
- important aside: all the above expressions are nice and fine but while implementing do all your computations in log-scale

EM I

- goal is to find the Maximum Likelihood Estimator (MLE) or the Maximum A Posterior (MAP) Estimator
- involves two steps:
 - the Expectation step or the E-step
 - the Maximization step or the M-step

EM II

- set up:
 - data: $\mathbf{y} := (y_1, y_2, \dots, y_n)$
 - parameter of interest: θ
 - “nuisance” parameter or “missing” data: \mathbf{z}
- E-step:

$$Q(\theta | \theta^{(t)}) := \begin{cases} E_{\theta^{(t)}} [\log p(\theta, \mathbf{z} | \mathbf{y})] = \int \log p(\theta, \mathbf{z} | \mathbf{y}) p(\mathbf{z} | \theta^{(t)}, \mathbf{y}) d\mathbf{z} & \text{for MAP} \\ E_{\theta^{(t)}} [\log p(\mathbf{z}, \mathbf{y} | \theta)] = \int \log p(\mathbf{z}, \mathbf{y} | \theta) p(\mathbf{z} | \theta^{(t)}, \mathbf{y}) d\mathbf{z} & \text{for MLE} \end{cases}$$

- M-step:

$$\theta^{(t+1)} := \arg \max_{\theta} Q(\theta | \theta^{(t)})$$

EM III

- the EM algorithm with ϵ -close stopping:
 1. initialize: set $t = 0$ and get a starting value $\theta^{(t)}$
 2. E-step: get $Q(\theta | \theta^{(t)})$
 3. M-step: get $\theta^{(t+1)} = \arg \max_{\theta} Q(\theta | \theta^{(t)})$
 4. if $\|\theta^{(t+1)} - \theta^{(t)}\| \leq \epsilon$ stop otherwise set $t = t + 1$ and go to step 2
- in some easy cases you could combine the E-step and the M-step if you have a closed form expression for $Q(\theta | \theta^{(t)})$ and (hence) for $\arg \max_{\theta} Q(\theta | \theta^{(t)})$

EM IV

- EXAMPLE: we want MAP estimator of μ from (with σ^2 unknown):
 - $y_i \sim \text{Normal}_1(\mu, \sigma^2)$, $i = 1, 2, \dots, n$
 - $\mu \sim \text{Normal}_1(\mu_0, \tau_0^2)$
 - $p(\log \sigma) \propto 1$
- so we have:
 - data: $\mathbf{y} := (y_1, y_2, \dots, y_n)$
 - parameter of interest: $\theta = \mu$
 - “nuisance” parameter: $z = \sigma^2$

EM V

- we observe:

$$\begin{aligned}\log p(\theta, \mathbf{z} \mid \mathbf{y}) &= \log p(\mu, \sigma^2 \mid \mathbf{y}) \\ &= \text{const} - \frac{1}{2\tau_0^2} (\mu - \mu_0)^2 - (n+1) \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2\end{aligned}$$

- we also note:

$$\begin{aligned}p(\mathbf{z} \mid \theta^{(t)}, \mathbf{y}) &= p(\sigma^2 \mid \mu^{(t)}, \mathbf{y}) \\ &\equiv \text{Inv} - \chi^2 \left(n, \frac{1}{n} \sum_{i=1}^n (y_i - \mu^{(t)})^2 \right)\end{aligned}$$

EM VI

- E-step: only compute the expectations of the terms which involve θ because other terms are not useful in the M-step
- so we note:

$$\begin{aligned}
 Q(\theta | \theta^{(t)}) &= Q(\mu | \mu^{(t)}) \\
 &= \text{const} - \frac{1}{2\tau_0^2} (\mu - \mu_0)^2 - E_{\mu^{(t)}} \left[\frac{1}{2\sigma^2} \right] \sum_{i=1}^n (y_i - \mu)^2 \\
 &= \text{const} - \frac{1}{2\tau_0^2} (\mu - \mu_0)^2 - \frac{1}{2} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - \mu^{(t)})^2 \right\}^{-1} \sum_{i=1}^n (y_i - \mu)^2
 \end{aligned}$$

- we are ignoring the followin for the mentioned reason

$$-(n+1)E_{\mu^{(t)}}[\log \sigma]$$

EM VII

- M-step: note $Q(\theta | \theta^{(t)}) = Q(\mu | \mu^{(t)})$ is a quadratic in μ and hence easy to maximize
- taking derivatives once (and then twice) one can show

$$\begin{aligned}\theta^{(t+1)} &:= \arg \max_{\theta} Q(\theta | \theta^{(t)}) \\ &= \arg \max_{\mu} Q(\mu | \mu^{(t)}) \\ &= \frac{\frac{1}{\tau_0^2} \mu_0 + \frac{n}{\frac{1}{n} \sum_{i=1}^n (y_i - \mu^{(t)})^2} \bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\frac{1}{n} \sum_{i=1}^n (y_i - \mu^{(t)})^2}} \\ &= \mu^{(t+1)}\end{aligned}$$

EM VIII

- EXAMPLE: we want MLE for mixture proportions, $(\pi_1, \pi_2, \dots, \pi_k)$:
 - we have k -many known densities $f_j(\cdot)$, $j = 1, 2, \dots, k$
 - there are k -many unknown proportions π_j , $j = 1, 2, \dots, k$ with
$$\sum_{j=1}^k \pi_j = 1$$
 - $y_i \sim \sum_{j=1}^k \pi_j f_j(\cdot)$, $i = 1, 2, \dots, n$
- so we have:
 - data: $\mathbf{y} := (y_1, y_2, \dots, y_n)$
 - parameter of interest: $\theta := (\pi_1, \pi_2, \dots, \pi_k)$
 - introduce “missing” data: $\mathbf{z} := (z_1, z_2, \dots, z_n)$ such that
$$[z_i \mid \theta] \sim \text{Multinomial}(1, \theta), \quad i = 1, 2, \dots, n$$
 - note here we need to cook up the “missing data” in such a way that integrating / summing it out gives us back our original model, see next slide

EM IX

- now we can rewrite our model as:
 - $[y_i | z_i = e_j, \theta] \sim f_j(\cdot)$, $i = 1, 2, \dots, n$, $j = 1, 2, \dots, k$
 - $p(z_i = e_j | \theta) = \pi_j$, $i = 1, 2, \dots, n$, $j = 1, 2, \dots, k$
 - here e_j is the j -th canonical vector for $j = 1, 2, \dots, k$ (e.g. $e_1 = (1, 0, 0, \dots, 0)$ etc.)
- check that: $\sum_{\mathbf{z}} p(\mathbf{y}, \mathbf{z} | \theta) = p(\mathbf{y} | \theta)$
- so we have:

$$\log p(\mathbf{z}, \mathbf{y} | \theta) = \sum_{i=1}^n \sum_{j=1}^k z_{ij} \log \{ \pi_j f_j(y_i) \} \quad \text{and}$$

$$p(z_{ij} = 1 | \mathbf{y}, \theta^{(t)}) = p(z_i = e_j | \mathbf{y}, \theta^{(t)})$$

$$= \frac{\pi_j^{(t)} f_j(y_i)}{\sum_{j'=1}^k \pi_{j'}^{(t)} f_{j'}(y_i)} = a_{ij}^{(t)}, \quad \text{say}$$

EM X

- E-step:

$$\begin{aligned} Q(\theta | \theta^{(t)}) &= \sum_{i=1}^n \sum_{j=1}^k E_{\theta^{(t)}}(z_{ij}) \log \{\pi_j f_j(y_i)\} \\ &= \sum_{i=1}^n \sum_{j=1}^k a_{ij}^{(t)} \log \{\pi_j f_j(y_i)\} \end{aligned}$$

- M-step: its a constrained maximization problem with $\sum_{j=1}^k \pi_j = 1$ which gives:

$$\begin{aligned} \theta^{(t+1)} &:= \arg \max_{\theta} Q(\theta | \theta^{(t)}) \\ &= \frac{\sum_{i=1}^n a_{ij}^{(t)}}{\sum_{i=1}^n \sum_{j=1}^k a_{ij}^{(t)}} = \frac{1}{n} \sum_{i=1}^n a_{ij}^{(t)} \end{aligned}$$

EM XI

- the tricky (theoretical) part of the EM algorithm is that many “missing data” schemes may give rise to the same model under consideration but not all are helpful
- EXAMPLE: in the mixture proportions example defining z the following way is not helpful at all (although it satisfies $\sum_z p(\mathbf{y}, z | \theta) = p(\mathbf{y} | \theta)$)

$$p(z_i = j | \theta) = \pi_j, i = 1, 2, \dots, n, j = 1, 2, \dots, k$$

note here z_i is of dimension 1 as opposed to k , as before

- to find the “best” missing data scheme is an art, really
- check out “The Art of Data Augmentation” by David A. van Dyk and Xiao-Li Meng (*go Harvard Stats!*)