General EM

- X = observed
- Z = latent
- Goal: Maximize$_\theta$ $p(X|\theta) = \sum_z p(X, z|\theta)$
- Decompose $\log(p(x|\theta))$:

$$\log p(x|\theta) = \mathcal{L}(q,\theta) + KL(q\|p) \qquad \text{where } q \text{ is any p.m.f.}$$

$$\mathcal{L}(q,\theta) = \sum_z q(z) \log\left(\frac{p(x,z|\theta)}{q(z)}\right) \quad \text{ELBO}$$

$$KL(q\|p) = -\sum_z q(z) \log\left(\frac{p(z|x,\theta)}{q(z)}\right) \quad \text{KL-Divergence}$$

↑ this holds for *any* distribution over z ($q(z)$) (not just the true distribution of z)

proof:

$$\mathcal{L}(q,\theta) = \sum_z q(z)\left(\log \frac{p(x,z|\theta)}{q(z)}\right) = \sum_z q(z) \log\left(\frac{p(z|x,\theta)\, p(x|\theta)}{q(z)}\right)$$

$$= \sum_z q(z) \log\left(\frac{p(z|x,\theta)}{q(z)}\right) + \underbrace{\sum_z q(z)}_{=1} \log p(x|\theta)$$

$$= -KL(q\|p) + \log p(x|\theta) \quad //$$

- Note: $KL(q\|p)$ is always ≥ 0

proof:

$$KL(q\|p) = E\left[-\log\left(\frac{p(z|x,\theta)}{q(z)}\right)\right]$$

$-\log$ is convex

$\Rightarrow$ jensen's inequality $\geq -\log E\left[\frac{p(z|x,\theta)}{q(z)}\right)$

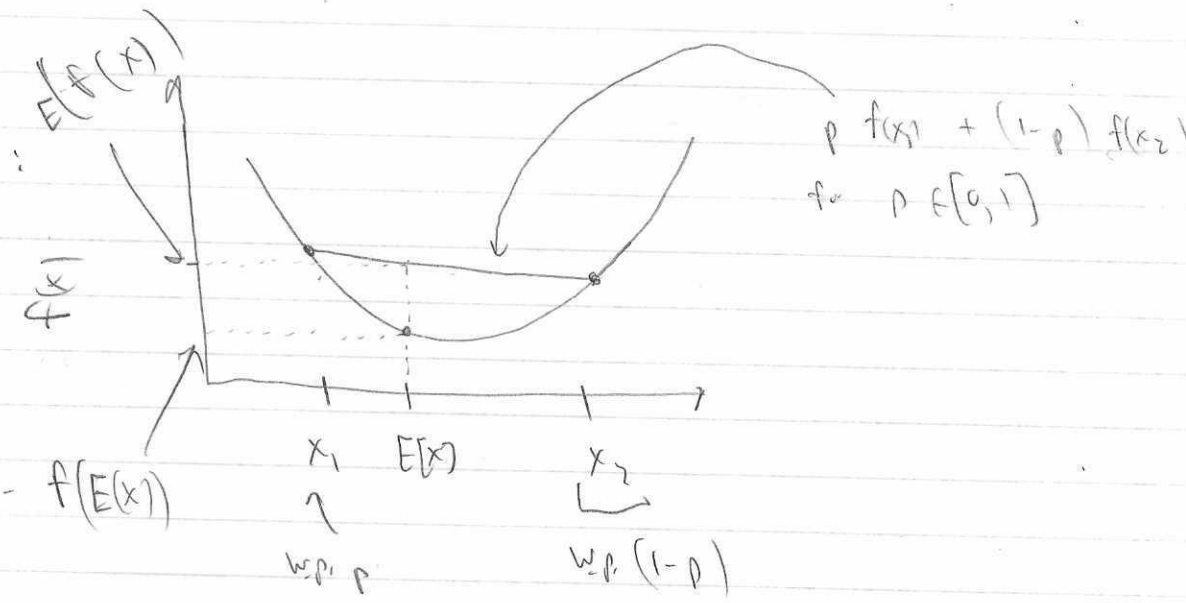
$$= -\log\left(\sum q(z)\frac{p(z|x,\theta)}{q(z)}\right)$$

$$= -\log\left(\underbrace{\sum p(z|x,\theta)}_{=1}\right)$$

$$= -\log(1)$$

$$= 0$$

- Jensen's:



so $E(f(x)) \geq f[E(x)]$ for $f()$ convex

$f() = -\log()$ here

- So $\mathcal{L}(q, \theta)$ is a lower bound on $\log p(x|\theta)$

$$\mathcal{L}(q, \theta) \leq \log p(x|\theta)$$

- EM Algorithm:

$$\log p(x|\theta) = \mathcal{L}(q, \theta) + \overbrace{KL(q||p)}^{\geq 0}$$

Step 1: Maximize $\mathcal{L}(q, \theta)$ over q, holding θ fixed

↳ during this step, $p(x|\theta)$ is constant, so $\mathcal{L}(q, \theta)$ is maximized when $KL(q||p)$ is 0

$$KL(q||p) = -\sum_z q(z) \log \left[\frac{p(z|x, \theta)}{q(z)} \right]$$

↑ if this equals $p(z|x, \theta)$, then

$$= -\sum_z q(z) \log(1) = 0$$

So E-step: set $q(z) = p(z|x, \theta)$

↳ posterior of $z|x$

Step 2: Maximize $\mathcal{L}(q, \theta)$ over θ holding q fixed

↑ this increases the lower bound on $\log p(x|\theta)$

↑ As you change θ, $q(\cdot)$ becomes no longer equal to posterior, so KL ↑

$$\log p(x|\theta) = \underbrace{\mathcal{L}(q, \theta)}_{\uparrow} + \underbrace{KL(q \| p)}_{\uparrow}$$

$\Rightarrow \log p(x|\theta)$ ↑

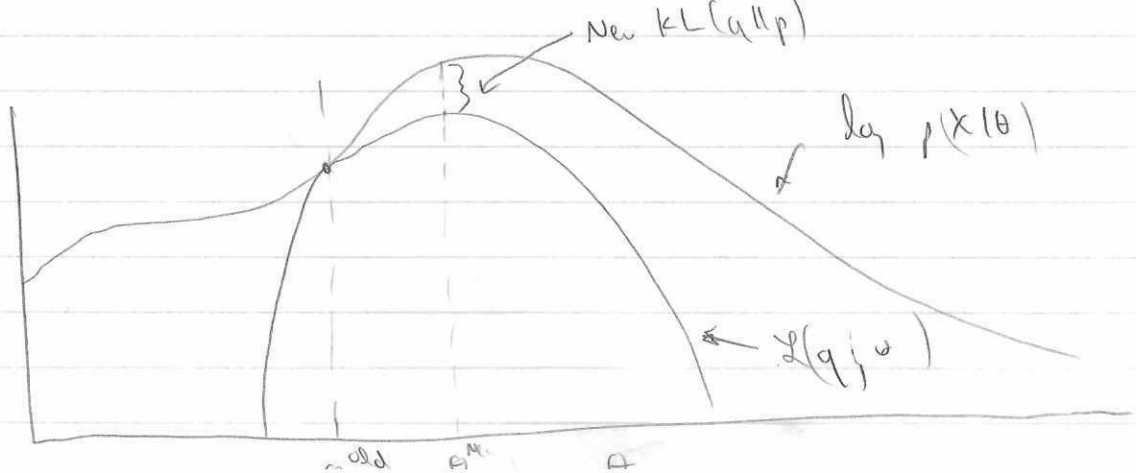
$$\max_{\theta} \mathcal{L}(q, \theta)$$

$$= \max_{\theta} \sum_{z} q(z) \log \left(\frac{p(x, z|\theta)}{q(z)} \right)$$

$$= \max_{\theta} \sum_{z} q(z) \log p(x, z|\theta) + \text{constant}$$

$$= \max_{\theta} E\left[\log p(x, z|\theta) \right]$$

↑ expectation taken over current posterior at z



• HW: Implement EM algorithm from Li (2011)

Inputs: A, a matrix A[i,h] = genotype log-likelihood for individual i at dosage h at a SNP

pi0, a k-vector containing initial values of prior genotype probabilities. Default rep($\frac{1}{k+1}$, k+1)

nite, an integer; # iterations of EM, default 1

tol, stopping criterion. Default 0.001

Function

$$\pi_h^{(new)} = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi_h^{old} (a_{ih}) \wedge e^{A[i,h]}}{\sum_{h=0}^{k} \pi_h^{old} (a_{ih}) \wedge e^{A[i,h]}}$$

Objective function: $\sum_{i=1}^{n} \log \left[\sum_{h=0}^{k} a_{ih} \pi_h \right]$

Stop when objective differs by less than tol in adjacent iterations.

Return

B, a matrix, B[i,h] = log-posterior probability for individual i having dosage h at a SNP

$$b_{ih} = \frac{\pi_h a_{ih}}{\sum_{h=0}^{k} \pi_h a_{ih}}$$