سوال ۱ آ)

تابع log-likelihood به صورت زیر است:

$$l(\pi,\mu,\sigma) = \ln P(X|\pi,\mu,\sigma) = \sum_{n=1}^{N} \ln P(x^n|\pi,\mu,\sigma)$$

(تعداد داده‌ها)

$$= \sum_{n=1}^{N} \ln \sum_{z^n=1}^{K} \underbrace{P(x^n|z^n;\mu,\sigma)P(z^n|\pi)}$$

(تعداد کلاس‌ها)

$$= \sum_{n=1}^{N} \log \sum_{j=1}^{K} P(x^n, z^n=j|\theta)$$

$\theta:(\pi,\mu,\sigma)$

(چون $P(x) = \sum_z P(x|z)P(z)$)

عبارت داخل سیگمای سمت راست را در یک q_j ضرب و تقسیم می‌کنیم:

$$l(X,\theta) = \sum_{n=1}^{N} \log \sum_{j=1}^{K} q_j \frac{P(x^n, z^n=j|\theta)}{q_j}$$

نامساوی jensen: اگر f تابعی محدب باشد و $P_1, P_2, \dots, P_n$ اعدادی بین 0 و یک باشند که جمعشان 1 است،

$$f\left(\sum_{i=1}^{n} P_i x_i\right) \geq \sum_{i=1}^{n} P_i f(x_i) = E[f(x)]$$

در نتیجه:

$$\sum_{n=1}^{N} \log \sum_{j=1}^{K} q_j \frac{P(x^n, z^n=j|\theta)}{q_j} \geq \overbrace{\sum_{n=1}^{N}\sum_{j=1}^{K} q_j \log \frac{P(x^n, z^n=j|\theta)}{q_j}}^{Q}$$

حال یک حدس خوب برای q_j، $P(z^n=j|x^n,\theta^{old})$ است (θ^{old}: مقادیر فعلی پارامترها)

چرا؟ چون عبارت جلوی log می‌شود $\frac{P(x^n, z^n|\theta)}{P(z^n=j|x^n,\theta^{old})}$ که همان $P(x^n|z^n=j,\theta^{old})$ است.

در آن صورت نامساوی به مساوی تبدیل می‌شود.

پس Q به صورت زیر می‌شود:

$$Q(\theta) = \sum_{n=1}^{N}\sum_{j=1}^{K} P(z^n=j|x^n,\theta^{old}) \log\left(P(x^n, z^n=j|\theta)\right)$$

ادامه‌ی سؤال ۱) آ)

$$Q(\theta) = E_{p(z^n \mid x^n, \theta^{old})}\left[\log P(x^n, z^n \mid \theta)\right]$$

(قسمتی که به θ وابسته نبود حذف شده است)

حال یک متغیر γ_k^n برای داده‌ی nام و کلاس kام تعریف می‌کنیم

$$\gamma_k^n = P(z^n = k \mid x^n) \overset{\text{Bayes}}{=} \frac{P(x^n \mid z^n = k)\, P(z^n = k)}{P(x^n)}$$

$$= \frac{\overbrace{P(z^n = k)}^{\pi_k}\, \overbrace{P(x^n \mid z^n = k)}^{\mathcal{N}(x^n, \mu_k, \sigma_k)}}{\sum_{j=1}^{K} P(z^n = j)\, P(x^n \mid z^n = j)}$$

$$= \frac{\pi_k\, \mathcal{N}(x^n, \mu_k, \sigma_k)}{\sum_{j=1}^{K} \pi_j\, \mathcal{N}(x^n, \mu_k, \sigma_k)}$$

که $\mathcal{N}$ همان pdf توزیع نرمال است.

حال γ_k^n را در فرمول تابع Q جایگذاری می‌کنیم:

$$E_{p(z^n \mid x^n)}\left[\sum_{n=1}^{N} \log\left(P(x^n, z^n \mid \theta)\right)\right]$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_k^n \left(\log\Big(\underbrace{P(z^n = k \mid \theta)}_{\pi_k}\Big) + \log\Big(\underbrace{P(x^n \mid z^n = k, \theta)}_{\mathcal{N}(x^n, \mu_k, \sigma_k)}\Big)\right)$$

$$\boxed{= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_k^n \log(\pi_k) + \sum_{k=1}^{K} \sum_{n=1}^{N} \gamma_k^N \log\left(\mathcal{N}(x^n \mid \mu_k, \sigma_k)\right)}$$

حال این تابع را در که optimize می‌کنیم (یعنی π_kها و μ_kها و σ_k را طوری انتخاب می‌کنیم که آن را optimize کند.)

سوال ۱)

قسمت 2) اثبات غیر نزولی بودن Q

می دانیم در مرحله‌ی M، پارامترها با فرمول زیر آپدیت می شوند:

$$\mu_k^* = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_k^n x^n$$

$$\sigma_k^* = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_k^n (x - \mu_k)^2$$

$$\pi_k^* = \frac{N_k}{N}$$

به صورتیکه $N_k = \sum_{n=1}^{N} \gamma_k^n$ تعریف شده است.

ثابت می کنیم که مقادیر سمت راست، همان مقادیر اکسترمم مطلق تابع Q برحسب μ، σ و π هستند. در اینصورت چون طبق تعریف اکسترمم،

$$\underbrace{Q(\pi', \mu', \sigma') \leq Q(\pi^*, \mu^*, \sigma^*)}_{}$$

$\forall \pi', \mu', \sigma'$ Q مرحله‌ی بعدی

در اینصورت Q مرحله‌ی بعدی حتماً از Q مرحله‌ی قبلی بیشتر خواهد بود.

$$Q(\theta | \theta^{old}) = \sum_{k=1}^{K} \sum_{n=1}^{N} \gamma_k^n \log \pi_k + \sum_{k=1}^{K} \sum_{n=1}^{N} \gamma_k^n \log \left[\frac{1}{\sqrt{2\pi}\,\sigma_k} e^{-\frac{1}{2}\left(\frac{x - \mu_k}{\sigma_k}\right)^2} \right]$$

$$\mathcal{L}(Q(\pi)) = \sum_{k=1}^{K} \sum_{n=1}^{N} \gamma_k^n \log \pi_k - \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) \quad , \quad \sum_{k=1}^{K} \pi_k = 1$$

λ ضریب لاگرانژ

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = \left(\sum_{n=1}^{N} \gamma_k^n \right) \times \frac{1}{\pi_k} - \lambda \times 1 = 0 \rightarrow \pi_k = \frac{1}{\lambda} \times \sum_{n=1}^{N} \gamma_k^n$$

جایگذاری در Constraint: $\sum_{k=1}^{K} \pi_k = 1 \longrightarrow \sum_{k=1}^{K} \left(\frac{1}{\lambda} \sum_{n=1}^{N} \gamma_k^n \right) = 1 \longrightarrow \boxed{\lambda = \sum_{k=1}^{K} \sum_{n=1}^{N} \gamma_k^n = \sum_{n=1}^{N} \underbrace{\sum_{k=1}^{K} \gamma_k^n}_{1} = N}$

$$\boxed{\pi_k = \frac{\sum_{n=1}^{N} \gamma_k^N}{N}}$$

برای اینکه بدانیم مقدار بهینه ی بدست آمده برای π_k مطلق argmax، Q برحسب π_k است و نه argmax نسبی، مشتق دوم Q برحسب π_k را می گیریم

$$\frac{\partial^2 Q}{(\partial \pi_k)^2} = \underbrace{\left(\sum_{n=1}^{N} \gamma_k^n\right)}_{} \times -\frac{1}{\pi_k^2} \leq 0$$

هر γ_k^n مثبت است
پس این مقدار مثبت است

پس چون تقعر تابع Q برحسب π_k همواره منفی است، مقدار بدست آمده برای π_k در صفحه‌ی قبل، ماکزیمم مطلق است.

$$\frac{\partial Q}{\partial \mu_k} = \sum_{n=1}^{N} \gamma_k^N \times \left(+\frac{1}{\cancel{2}}\right) \times \left(+\frac{1}{\sigma_k}\right) \times \cancel{2} \left(\frac{x^n - \mu_k}{\sigma_k}\right)$$

$$= \frac{1}{\sigma_k^2} \sum_{n=1}^{N} \gamma_k^N \left(x^n - \mu_k\right)$$

$$\frac{\partial Q}{\partial \mu_k} = 0 \rightarrow \frac{1}{\sigma_k^2} \sum_{n=1}^{N} \gamma_k^n \left(x^n - \mu_k\right) = 0$$

$$\rightarrow \sum_{n=1}^{N} \gamma_k^n \times x^n - \sum_{n=1}^{N} \gamma_k^n \times \underbrace{\mu_k}_{\text{مستقل از n}} = 0$$

$$\rightarrow \boxed{\mu_k^* = \frac{\sum_{n=1}^{N} \gamma_k^n x^n}{\sum_{n=1}^{N} \gamma^n}}$$

حالا از مشتق دوم Q نسبت به μ_k را بگیریم:

$$\frac{\partial^2 Q}{(\partial \mu_k)^2} = \frac{1}{\sigma_k^2} \times \sum_{n=1}^{N} \gamma_k^n (-1) = -\underbrace{\frac{\sum_{n=1}^{N} \gamma_k^n}{\sigma_k^2}}_{>0} \leq 0$$

پس Q برحسب μ_k هم محدب اکید است، پس نقطه‌ی بهینه ی بدست آمده برای μ_k، argmax مطلق Q است.

ادامه‌ی سؤال ۱) قسمت ۲) اثبات بیشینه بودن Q

$$\frac{\partial Q}{\partial \sigma_k} = \frac{1}{\cancel{\sqrt{2\pi}}} \times \left(\frac{-1}{\sigma_k^2}\right) \times \cancel{\sqrt{2\pi}}\, \sigma_k \times \sum_{n=1}^{N} \gamma_k^n$$

$$+ \sum_{n=1}^{N} \gamma_k^n \times \left(-\frac{1}{\cancel{2}}\right) \times (x^n - \mu_k) \times \frac{-1}{\sigma^2} \times \cancel{2} \left(\frac{x^n - \mu_k}{\sigma_k}\right)$$

$$\frac{\partial Q}{\partial \sigma_k} = -\frac{\sum_{n=1}^{N} \gamma_k^n}{\sigma_k} + \frac{1}{\sigma_k^3} \sum_{n=1}^{N} \gamma_k^n (x - \mu_k)^2 = 0$$

از آنجایی که $\sigma_k = 0$ جزو دامنه‌ی تابع نیست،

$$\rightarrow -\sum_{n=1}^{N} \gamma_k^n + \frac{1}{\sigma_k^2} \sum_{n=1}^{N} \gamma_k^n (x - \mu)^2 = 0$$

$$\rightarrow \boxed{\sigma_k^2 = \frac{\sum_{n=1}^{N} \gamma_k^n (x - \mu)^2}{\sum_{n=1}^{N} \gamma_k^n}}$$

حالا اگر مشتق دوم بگیریم، با در حالت قبل کمی فرق دارد:

$$\frac{\partial^2 Q}{(\partial \sigma_k)^2} = \frac{-3}{\sigma_k^{\cancel{4}\,2}} \sum_{n=1}^{N} \gamma_k^n (x - \mu_k)^2 + \frac{\sum_{n=1}^{N} \gamma_k^n}{\cancel{\sigma_k^2}} = 0$$

$$\frac{-3}{\sigma_k^2} \times \sum_{n=1}^{N} \gamma_k^n (x - \mu_k)^2 + \sum_{n=1}^{N} \gamma_k^n = 0$$

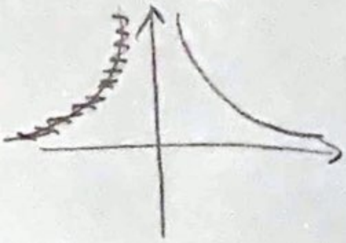
در اینجا مشتق دوم یک ریشه مثبت دارد. پس با مشتق دوم نمی‌توان اثبات کرد که نقطه‌ی بدست آمده، بیشینه‌ی مطلق است و نه نسبی. اما با رسم شکل تابع مشتق می‌توان این کار را کرد. دقت می‌کنیم که $\frac{\partial Q}{\partial \sigma_k}$ به فرم $\frac{1}{\sigma_k}\left(-A + \frac{B}{\sigma_k^2}\right)$ است. و برای انحراف معیار ما فقط σهای بزرگتر از صفر را قابل قبول می‌دانیم (حتی در حالت ماتریسی هم همینطور است)

ادامه‌ی سؤال ۱) قسمت 2) اثبات غیر نزولی بودن Q

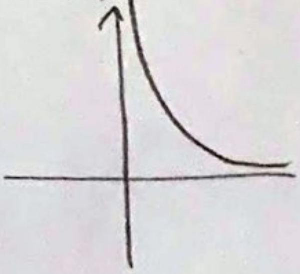
$\sigma_k > 0$

رسم تابع $\frac{\partial Q}{\partial \sigma_k} = \frac{1}{\sigma_k}\left(-A + \frac{B}{\sigma_k^2}\right)$ که $0 \leq B , A$

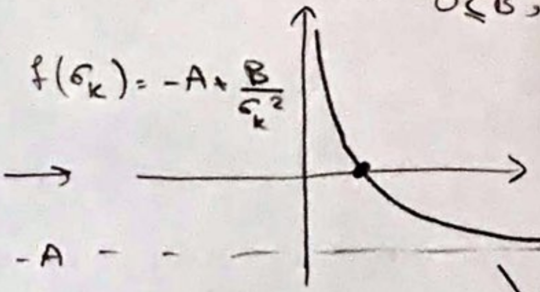
$f(\sigma_k) = \frac{1}{\sigma_k^2}$



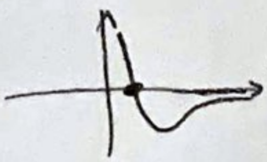
$\rightarrow$ $f(\sigma_k) = \frac{B}{\sigma_k^2}$



$\rightarrow$ $f(\sigma_k) = -A + \frac{B}{\sigma_k^2}$



حالا اگر $\frac{1}{\sigma_k}$ را در تابع فوق ضرب کنیم، با توجه به اینکه σ_k همواره بزرگتر از 0 است، درست است که شکل تابع تغییر می‌کند (در واقع به صورت



در می‌آید) اما ریشه‌ی جدیدی تولید نمی‌شود. پس ریشه‌ی مشتق، اکسترمم مطلق است و چون مشتق قبل از آن مثبت و بعد از آن منفی است، پس مقدار Q در σ_k ی بدست آمده ماکزیمم مطلق است نه مینیمم.

پس چون مقداردهی σ و μ و π در M-step به argmax مطلق Q است، پس Q مرحله‌ی جدید نمی‌تواند از Q مرحله‌ی قبلی کمتر باشد.