Lecture 13

Bound optimization

General idea:

Goal: maximize some function $\ell(\vec{\theta})$ wrt $\vec{\theta}$.

→ Construct a surrogate function $Q(\vec{\theta}, \vec{\theta}^t)$ ($\vec{\theta}^t$: current prm values)

s.t. $\begin{cases} \ell(\vec{\theta}) \geq Q(\vec{\theta}, \vec{\theta}^t) & \text{and} \\ \ell(\vec{\theta}^t) = Q(\vec{\theta}^t, \vec{\theta}^t) . \end{cases}$

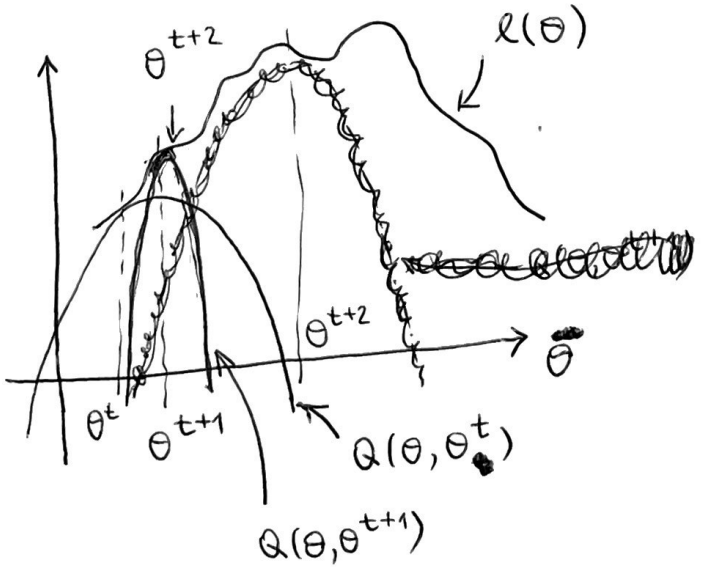
→ Find $\vec{\theta}^{t+1} = \underset{\vec{\theta}}{\operatorname{argmax}}\, Q(\vec{\theta}, \vec{\theta}^t)$

Then $\ell(\vec{\theta}^{t+1}) \geq Q(\vec{\theta}^{t+1}, \vec{\theta}^t) \geq Q(\vec{\theta}^t, \vec{\theta}^t) = \ell(\vec{\theta}^t)$

($\geq$: by construction; $\geq$: by argmax)

$\ell(\vec{\theta}^{t+1}) \geq \ell(\vec{\theta}^t)$ monotonically at each step

Ex.



quadratic Q-function

[Expectation maximization (EM) algorithm]

$\begin{cases} \vec{y}_n = \text{observable data} \quad (n=1,\dots,N) \\ \vec{z}_n = \text{hidden variables} \end{cases}$

alternation between

→ E-step (expectation step) ⇒ estimate hidden vars

→ M-step (maximization step) ⇒ compute MLE or MAP parameter estimates given the hidden vars & the observed data

---

Maximize log-likelihood:

$$\ell(\vec{\theta}) = \sum_{n=1}^{N} \log p(\vec{y}_n|\vec{\theta}) = \sum_{n=1}^{N} \log\left(\sum_{\vec{z}_n} p(\vec{y}_n, \vec{z}_n|\vec{\theta})\right)$$

Consider $\{q_n(\vec{z}_n)\}_{n=1}^{N}$ – a set of some prob. distributions over hidden vars $\vec{z}_n$: $\sum_{\vec{z}_n} q_n(\vec{z}_n)=1$

Then $\ell(\vec{\theta}) = \sum_n \log\left(\sum_{\vec{z}_n} q_n(\vec{z}_n) \frac{p(\vec{y}_n, \vec{z}_n|\vec{\theta})}{q_n(\vec{z}_n)}\right) \geq$

$$\geq \sum_{n,\vec{z}_n} q_n(\vec{z}_n) \log \frac{p(\vec{y}_n, \vec{z}_n|\vec{\theta})}{q_n(\vec{z}_n)} \equiv \underbrace{\sum_n \mathcal{J}(\vec{\theta}, q_n)}_{\text{evidence lower bound (ELBO)}}$$

↑ Jensen's inequality

E-step

Now, consider

$$\tilde{\mathcal{I}}(\vec{\theta}, q_n) = \sum_{\vec{z}_n} q_n(\vec{z}_n) \log \frac{p(\vec{z}_n | \vec{y}_n, \vec{\theta})\, p(\vec{y}_n | \vec{\theta})}{q_n(\vec{z}_n)} =$$

$$= \sum_{\vec{z}_n} q_n(\vec{z}_n) \log p(\vec{y}_n | \vec{\theta}) + \sum_{\vec{z}_n} q_n(\vec{z}_n) \log \frac{p(\vec{z}_n | \vec{y}_n, \vec{\theta})}{q_n(\vec{z}_n)} \quad (**)$$

$$(**) = \log p(\vec{y}_n | \vec{\theta}) - \underbrace{D_{KL}(q_n(\vec{z}_n) \,||\, p(\vec{z}_n | \vec{y}_n, \vec{\theta}))}_{\geq 0}.$$

$\uparrow$

$\sum_{\vec{z}_n} q_n(\vec{z}_n) = 1$

$\swarrow$ tight lower bound

If $q_n(\vec{z}_n) = p(\vec{z}_n | \vec{y}_n, \vec{\theta})$,

$$\tilde{\mathcal{I}}(\vec{\theta}, q_n) = \log p(\vec{y}_n | \vec{\theta}) \quad (*)$$

Finally, define $Q(\vec{\theta}, \vec{\theta}^t) = \sum_n \tilde{\mathcal{I}}(\vec{\theta}, \underbrace{p(\vec{z}_n | \vec{y}_n, \vec{\theta}^t)}_{q_n(\vec{z}_n)})$

Indeed, by Eq. $(*)$ $Q(\vec{\theta}^t, \vec{\theta}^t) = \sum_n \log p(\vec{y}_n | \vec{\theta}^t)$

$= \ell(\vec{\theta}^t)$.

Moreover, $Q(\vec{\theta}, \vec{\theta}^t) \leq \ell(\vec{\theta})$ due to the ≥ 0 D_{KL} term.

M-step

Need to maximize $\sum_n \mathcal{I}(\vec{\theta}, p(\vec{z}_n | \vec{y}_n, \vec{\theta}^t))$ wrt $\vec{\theta}$. Since $-\sum_{\vec{z}_n} q_n(\vec{z}_n) \log q_n(\vec{z}_n) =$

$= H(q_n)$ does not depend on $\vec{\theta}$, it suffices to maximize

$$\sum_{n, \vec{z}_n} q_n(\vec{z}_n) \log p(\vec{y}_n, \vec{z}_n | \vec{\theta}) \quad \text{to get } \vec{\theta}^{t+1}.$$

$\uparrow$ expectation value wrt $q_n(\vec{z}_n)$

Ex. gaussian mixture model (GMM)

$$p(\vec{y} | \vec{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\vec{y} | \vec{\mu}_k, \Sigma_k) \qquad \pi_k \geq 0, \ \sum_k \pi_k = 1$$

application: clustering of $\vec{y}_n \in \mathbb{R}^D$

Associate each data point $\vec{y}_n$ with a hidden var $z_n \in \{1, \dots, K\}$ which labels clusters:

$$\underbrace{p(z_n = k | \vec{y}_n, \vec{\theta})}_{\text{responsibility of cluster } k \text{ for datapoint } n} = \frac{p(z_n = k | \vec{\theta}) \, p(\vec{y}_n | z_n = k, \vec{\theta})}{\sum_{k'=1}^{K} p(z_n = k' | \vec{\theta}) \, p(\vec{y}_n | z_n = k', \vec{\theta})}$$

E-step

$$r_{nk}^t = \frac{\pi_k^t \, p(\vec{y}_n | \vec{\theta}_k^t)}{\sum_{k'} \pi_{k'}^t \, p(\vec{y}_n, \vec{\theta}_{k'}^t)}, \text{ where}$$

$\vec{\theta}_k$ = prms describing (mixture component) k.

M-step

$$\ell^t(\vec{\theta}) = E\Big[\sum_n \big(\log p(z_n | \vec{\pi}) + \log p(\vec{y}_n | z_n, \vec{\theta}) \big) \Big]$$

(wrt q_n; $\vec{\pi} = (\pi_1, \dots, \pi_K)$) (11)

$$\ominus E\Big[\sum_n \log\Big(\prod_k \pi_k^{z_{nk}} \Big) + \sum_n \log\Big(\prod_k \mathcal{N}(\vec{y}_n | \vec{\mu}_k, \Sigma_k)^{z_{nk}} \Big) \Big]$$

(11)

$\uparrow z_{nk} = \mathbb{I}(z_n = k)$ 'one-hot' encoding

$$= \sum_{n,k} \underbrace{E[z_{nk}]}_{r_{nk}^t} \log \pi_k + \sum_{n,k} E[z_{nk}] \log \mathcal{N}(\vec{y}_n | \vec{\mu}_k, \Sigma_k) =$$

$$= \sum_{n,k} r_{nk}^t \log \pi_k - \frac{1}{2} \sum_{n,k} r_{nk}^t \Big[(\vec{y}_n - \vec{\mu}_k)^T \Sigma_k^{-1} (\vec{y}_n - \vec{\mu}_k) + \log |\Sigma_k| \Big] + \text{const}$$

$\ell^t(\vec{\theta})$ is maximized by

$$\begin{cases} \vec{\mu}_k^{t+1} = \dfrac{\sum_n r_{nk}^t \vec{y}_n}{r_k^t}, & \underbrace{r_k^t = \sum_n r_{nk}^t} \\ \Sigma_k^{t+1} = \dfrac{\sum_n r_{nk}^t (\vec{y}_n - \vec{\mu}_k^{t+1})(\vec{y}_n - \vec{\mu}_k^{t+1})^T}{r_k^t}, & (+) \\ \pi_k^{t+1} = \dfrac{1}{N} \sum_n r_{nk}^t = \dfrac{r_k^t}{N} \end{cases}$$

These prms can be used to estimate r_{nk}^{t+1} in the next E-step, etc.

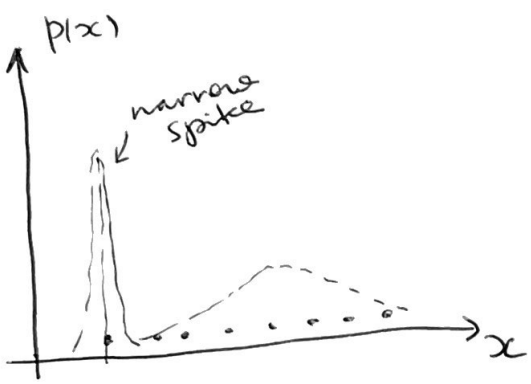
Note: MLE on the GMM might overfit.

Suppose that $\Sigma_k = \sigma_k^2 \mathbb{I}$, $\forall k$.

It's possible to center one component on a single datapoint: $\vec{\mu}_{k'} = \vec{y}_{n'}$ say.

Then $\mathcal{N}(\vec{y}_{n'} | \vec{\mu}_{k'}, \sigma_{k'}^2 \mathbb{I}) = \frac{1}{\sqrt{2\pi\sigma_k^2}}$

This will $\to \infty$ as $\sigma_k \to 0$ $\Rightarrow$ "collapsing variance problem".



$K=2$

Solution: switch from MLE to MAP estimation by introducing priors into Eq. (+); modified M-step.

For example, $\pi_k^{t+1} = \frac{r_k^t + \alpha_k - 1}{N + \sum_k \alpha_k - K}$ ← pseudocounts

Dirichlet prior