Bayesian Methods for Machine Learning

Andrey de Aguiar Salvi

December 2020

1 Class 1

Three principles:

- use prior knowledge
- choose the answer that explains the observations the most
- avoid making extra assumptions

1.1 Variable Independence

$$P(X,Y) = P(X)P(Y) \tag{1}$$

1.2 Conditional Probability

$$P(X|Y) = \frac{P(X,Y)}{Y(P)} \tag{2}$$

where $P(X,Y)$ = joint probability and $P(Y)$ = marginal probability.

1.3 Chain Rule

$$P(X,Y) = P(X|Y)P(Y) \tag{3}$$

$$P(X,Y,Z) = P(X|Y,Z)P(Y|Z)P(Z) \tag{4}$$

$$P(X_1,...,X_N) = \prod_{n=1}^{N} P(X_i|X1,...,X_{i-1}) \tag{5}$$

1.4 Marginalization

$$p(X) = \int_{-\inf}^{in} p(X,Y)dY \tag{6}$$

1.5 Bayes Theorem

$$P(\theta|X) = \frac{P(X,\theta)}{P(X)} = \frac{P(X|\theta)P(\theta)}{P(X)} \tag{7}$$

where $P(\theta|X)$ = posterior probability, $P(X|\theta)P(\theta)$ = Likelihood, and $P(X)$ = Evidence.

2 Class 2

2.1 Statistic Approaches

- **Frequentist**:
  - deterministic
  - θ is fixed, X is random
  - work whether data points is higher than the parameters - $|X| >> |\theta|$
  - Train models with Maximum Likelihood:

$$\hat{\theta} = arg\max_{\theta} P(X|\theta) \tag{8}$$

    to maximize the probability of data given the parameters

- **Bayesing**:
  - subjective
  - θ is random, X is fixed (given a set of forces θ, tossing a coin always gives the same result X)
  - work with data on any size - $|X|$
  - Train models with Na"(i)ve Bayes to maximize the probability of the parameters given the data

2.2 On-line learning

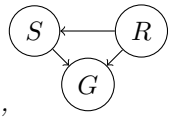
Use the current mini-batch (posterior) to update parameters, and then use it as prior in the new mini-batch.

3 Class 3

3.1 Bayesian Net

Is not a bayesian neural network. The **Nodes** are random variables and the **Edges** are the direct impact. **Model**: is the joint probability over all probabilites

$$P(X_1, ..., X_N) = \prod_{n=1}^{N} P(X_i|P_a(X_i)) \tag{9}$$

where $P_a(X_i)$ is the probability of the parent nodes from the Bayesian Net. *E.g.,* 

where R is father from G and S, S is parent from G, and the equation bellow is the respective bayesian probability

$$P(S, R, G) = P(G|S, R)P(S|R)P(R) \tag{10}$$

4 Class 5

4.1 Univariate Normal Distribution

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{11}$$

4.2 Multivariate Normal Distribution

$$\mathcal{N}(x|\mu, \Sigma^2) = \frac{1}{\sqrt{2\pi\Sigma^2}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^-1(x-\mu)} \tag{12}$$

4.3 Linear Regression

$$L(w) = \sum_{i=1}^{N} (w^T x_i - y_i)^2 = ||w^T X - y|| \rightarrow \min_w \tag{13}$$

$$\hat{w} = arg\min_w L(w) \tag{14}$$

where L is the loss from the bayesian net, w are the weights and X are the data, both are parents of y (target)

$$P(w, y|X) = P(y|X, w)P(w) \tag{15}$$

$$P(y|w, X) = \mathcal{N}(y|w^T X, \sigma^2 \mathcal{I}) \tag{16}$$

$$P(w) = \mathcal{N}(w|0, \gamma^2 \mathcal{I}) \tag{17}$$

$$P(w|y, x) = \frac{P(y, w|X)}{P(y|X)} \rightarrow \max_w \tag{18}$$

$$P(y, w|x) = \frac{P(y|x, w)}{P(w)} \rightarrow \max_w \tag{19}$$

5 Class 6 - Maximum a Posteriori

To learn a distribution

$$\theta_{MP} = arg\max_\theta P(\theta|X) = arg\max_\theta \frac{P(\theta|X)P(\theta)}{P(X)} = arg\max_\theta P(\theta|X)P(\theta) \tag{20}$$

We eliminate the denominator from the last equation, once time it don't have θ Problems:

- it is not variant to reparametrization. Ex learning about a gaussian will be not usefull in sigmoid(gaussian)
- strange "loss function"
- we do not have the posteriori of θ
- can't compute credible intervals

6 Class 7 - Conjugate Distribution

It is a way to avoid computing the evidence ($P(X)$), which is costly. In the Bayes Probability

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \tag{21}$$

in the likelihood, $P(X|\theta)$ is fixed by the model, $P(X)$ is fixed by the data, and $P(\theta)$ is our own choice. The prior $P(\theta)$ is conjugate to the likelihood if the prior and the posterior $P(X|\theta)$ lie in the same family distributions.

*E.g.,*if the prior is $P(X|\theta) = \mathcal{N}(x|\mu, \sigma^2)$ and the posterior is $P(\theta) = \mathcal{N}(\theta|m, s^2)$, thus the conjugate of posterior $P(\theta|X)$ is $\mathcal{N}(\theta|a, b^2)$.

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} = \frac{\mathcal{N}(x|\theta, 1)\mathcal{N}(\theta|0, 1)}{P(X)} \tag{22}$$

$$P(\theta|X) \propto e^{-\frac{1}{2}(X-\theta)^2} e^{-\frac{1}{2}\theta^2} \tag{23}$$

$$P(\theta|X) \propto e^{-(\theta - \frac{X}{2})^2} \tag{24}$$

$$P(\theta|X) = \mathcal{N}(\theta|\frac{X}{2}, \frac{X}{2}) \tag{25}$$

7 Class 8

7.1 Gamma Distribution

$$\Gamma(\gamma|a,b) = \frac{b^a}{\Gamma(a)}\gamma^{a-1}e^{-b\gamma} \tag{26}$$

where

- $\gamma, a, b > 0$
- $\Gamma(n) = (n-1)!$
- the expectation, or mean, $\mathbb{E}[\gamma] = \frac{a}{b}$
- $Mode[\gamma] = \frac{a-1}{b}$
- $Var[\gamma] = \frac{a}{b^2}$

7.2 Precision

$$\gamma = \frac{1}{\sigma^2} \tag{27}$$

If we replace the variance in the Normal Distribution to the inverse of the Precision, we get

$$\mathcal{N}(x|\mu,\gamma^{-1}) = \frac{\sqrt{\gamma}}{\sqrt{2\pi}}e^{-\gamma\frac{(x-\mu)^2}{2}} \tag{28}$$

thus, the conjugate prior in respect to the precision is

$$\mathcal{N}(x|\mu,\gamma^{-1}) \propto \gamma^{\frac{1}{2}}e^{-b\gamma} \tag{29}$$

$$P(\gamma) \propto \gamma^{a-1}e^{-b\gamma} \tag{30}$$

$$P(\gamma|X) = \Gamma(\gamma|a,b) \tag{31}$$

then, the prior is

$$P(\gamma|X) = \Gamma(\gamma|a,b) \propto \gamma^{a-1}e^{-b\gamma} \tag{32}$$

the posterior is

$$P(\gamma|X) \propto P(X|\gamma)P(\gamma) \tag{33}$$

dropping out all the constants, we have

$$P(\gamma|X) \propto \left(\gamma^{\frac{1}{2}}e^{-\gamma\frac{(X-\theta)^2}{2}}\right)\left(\gamma^{a-1}e^{-b\gamma}\right) \tag{34}$$

re-arranging the terms, we get

$$P(\gamma|X) \propto \gamma^{\frac{1}{2}+a-1}e^{-\gamma\frac{(X-\theta)^2}{2}}e^{-\gamma(b+\frac{(X-\mu)^2}{2})} \tag{35}$$

finally

$$P(\gamma|X) = \Gamma(a+\frac{1}{2}, b+\frac{(X-\mu)^2}{2}) \tag{36}$$

8 Class 9

8.1 Beta Distribution

$$\beta(X|a,b) = \frac{1}{\beta(a,b)}X^{a-1}(1-X)^{b-1} \tag{37}$$

where:

- $X \in [0,1]$
- $a, b > 0$

- $$\beta(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \tag{38}$$
- $$\mathbb{E}[X] = \frac{a}{a+b} \tag{39}$$
- $$Mode[X] = \frac{a-1}{a+b-2} \tag{40}$$
- $$Var[X] = \frac{ab}{(a+b)^2(a+b-1)} \tag{41}$$

*E.g.,*to model a distribution with mean 0.8 and standard deviation of 0.1, thus $\mathbb{E}[X] = 0.8$ and $Var[X] = 0.1^2$. Consequently, a Beta distribution with $a = 12$ and $b = 3$ model it.

8.2 Bernoulli Distribution

The Beta distribution is the conjugate of the Bernoulli likelihood. *E.g.,*a Bernoulli likelihood from a dataset

$$P(X|\theta) = \theta^{N_1}(1-\theta)^{N_0} \tag{42}$$

where N_1 is the number of ones in X and N_0 the number of zeros. Thus, the Beta distribution is

$$P(\theta) = \beta(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1} \tag{43}$$

Multiplying the likelihood by the prior

$$P(\theta|X) \propto P(X|\theta)P(\theta) \tag{44}$$

using the before, but replacing the terms by the equivalent equations above, we have

$$P(\theta|X) \propto \theta^{N_1}(1-\theta)^{N_0}\beta(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1} \tag{45}$$

and rearranging the terms, we have

$$P(\theta|X) \propto \theta^{N_1+a-1}(1-\theta)^{N_0+b-1} \tag{46}$$

and finally, recognizing the equation before as a Beta distribution, we have

$$P(\theta|X) = \beta(N_1+a, N_0+b) \tag{47}$$

8.3 Posteriors

- Pros
  - Exact posterior
  - Easy for on-line learning. *E.g.,*$P(\theta|X) = \beta(N_1+a, N_2+b)$
- Cons
  - In some cases, the conjugate prior may be inadequate

9 Class 10 - Latent Variable

It is a hidden variable that you never observe. *E.g.,*creating a dataset of a job interview, measuring the GPA, IQ, School degree, and Phone, from a first stage of the interview, and a last attribute Onsite performance, from the second interview. Traditional ML models will suffer with the missing data. Even more, probably there are some correlation between the variables, and ponder each combination will increase the number of attributes. Thus, we can link all these attributes with a latent variable, called in this example as Intelligence. We can model the problem as:

$$P(X_1, X_2, X_3, X_4, X_5) = \sum_{i=1}^{N} P(X_1, X_2, X_3, X_4, X_5|I)P(I) \tag{48}$$

where I is the latent variable. We can also simplify the problem with

$$P(X_1, X_2, X_3, X_4, X_5) = \sum_{i=1}^{N} P(X_1|I)P(X_2|I)P(X_3|I)P(X_4|I)P(X_5|I)P(I) \tag{49}$$

which breaks the table in 5 fewer tables, reduce the model complexity and improves the flexibility of the model.

10 Class 12 - Gaussian Mixture Model (GMM)

It is a model of soft clustering with N gaussian's can be described as

$$P(X|\theta) = \pi_1\mathcal{N}(\mu_1, \Sigma_1) + \pi_2\mathcal{N}(\mu_2, \Sigma_2) + ... + \pi_N\mathcal{N}(\mu_N, \Sigma_N) \tag{50}$$

where the θ weight matrix is

$$\theta = \{\pi_1, \pi_2, ..., \pi_N, \mu_1, \mu_2, ..., \mu_N, \Sigma_1, \Sigma_2, ..., \Sigma_N\} \tag{51}$$

To goal of the train is

$$\max_{\theta} P(X|\theta) = \prod_{i=1}^{N} P(X_1|\theta) = \prod_{i=1}^{N} (pi_1\mathcal{N}(\mu_1, \Sigma_1) + ...) \tag{52}$$

By definition, the covariance matrix $\Sigma_k \succ 0$, once time we need to compute Σ^{-1}. Otherwise, we will have divisions by zero.

11 Class 13 - Training a GMM

Assuming that our data points X was generated by a latent variable t. Thus, the probability of a point to belongs to the class c given the parameters θ is

$$P(t = c|\theta) = \pi_c \tag{53}$$

and

$$P(X|t = c, \theta) = \mathcal{N}(X, \mu_c, \Sigma_c) \tag{54}$$

Marginalizing the latent variable, we have

$$P(X|\theta) = \sum_{c=1}^{N} P(X|t = c, \theta)P(t = c|\theta) \tag{55}$$

which is the summation of the likelihood times the prior, and is exactly the same as the first equation, but ignoring the latent variable.

Training a GMM, if we have hard-labels of the points, it is a easy problem: we just need to compute the mean and standard deviation of each cluster. In the opposite side, if we have the parameters (consequently the gaussians), thus we can estimate the source ($P(t = 1|X, \theta) = \frac{P(X|t=1,\theta)P(t=1|\theta)}{Z}$), which is the joint probability (the likelihood times the prior). The problem of train a GMM is a Chicken and Egg problem:

- Need gaussian parameters (θ) to estimate the source (soft-labels)
- Need sources to estimate the gaussian parameters

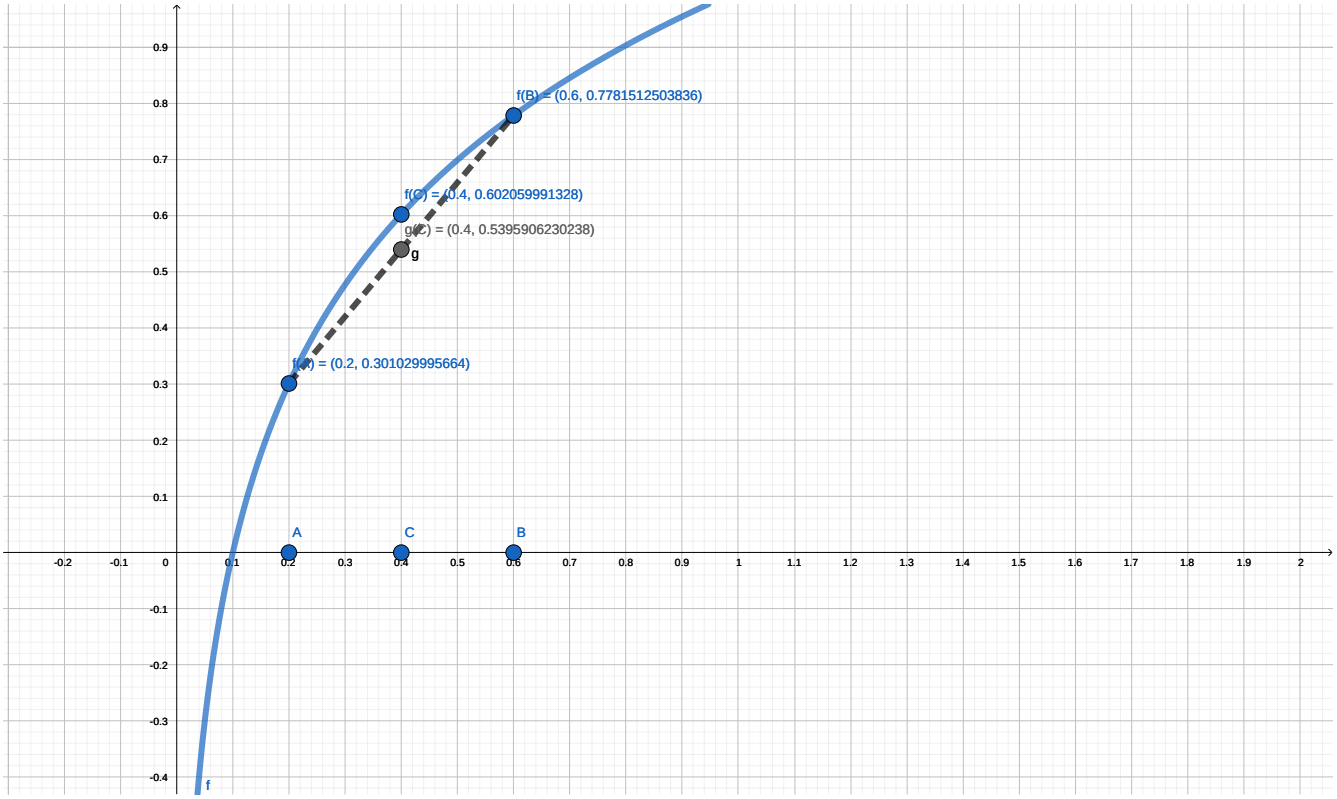


Figure 1: Showing the inequality from the equation. The equation from the right inequality side, which generates the straight segment, we call as G(x).

12 Class 14 - EM GMM

1. Start with N randomly placed gaussians parameters θ
2. Until convergence:
   (a) For each point X_i, compute $P(t = c|X_i, \theta)$
   (b) Update gaussian parameters θ to fit points assigned to them

- EM can train GMM faster than Stochastic Gradient Descent
- EM suffers from local maxima (the exact solution is NP-Hard)

13 Class 15

13.1 Concavity

A function is concave if its second derivative is negative, or, if

$$f(\alpha a + (1-\alpha)b) \geq \alpha f(a) + (1-\alpha)f(b) \tag{56}$$

where $0 \leq \alpha \leq 1$. The Figure 1 shows an *e.g.,*

13.2 Jensen's inequality

Generalizing the concavity for any point, we have

$$f(\mathbb{E}_{p(t)t}) \geq \mathbb{E}_{p(t)} f(t) \tag{57}$$

13.3 Kullback-Leibler Divergence

It is a way to measure the difference between two probabilistic functions.

$$\mathcal{KL}(p||q) = \int q(x) \log \frac{q(x)}{p(x)} dx \tag{58}$$

Properties:

1.
$$\mathcal{KL}(p||q) \neq \mathcal{KL}(q||p) \tag{59}$$

2.
$$\mathcal{KL}(q||q) = 0 \tag{60}$$

3.
$$\mathcal{KL}(p||q) > 1 \tag{61}$$

Proof

$$-\mathcal{KL}(p||q) = \mathbb{E}_q \left(-\log \frac{q}{p} \right) = \mathbb{E}_q \left(\log \frac{p}{q} \right) \tag{62}$$

$$\leq \log \left(\mathbb{E}_q \frac{p}{q} \right) = \log \int q(x) \frac{p(x)}{q(x)} dx = 0 \tag{63}$$

14 Class 16 - Expectation Maximization

Using the log for mathematical conveniences, we want to

$$\max_\theta \log(P(X|\theta)) = \log \left(\prod_{i=1}^{N} p(x_i|\theta) \right) \tag{64}$$

with the log properties, we have

$$\max_\theta \log(P(X|\theta)) = \log \left(\sum_{i=1}^{N} \log(p(x_i|\theta)) \right) \tag{65}$$

The probability $\log(p(x|\theta))$ is

$$\log(p(x|\theta)) = \log \left(\sum_{i=1}^{N} \log(p(x_i|\theta)) \right) \tag{66}$$

which we can change the marginal likelihood of the data object x_i by the definition, resulting in

$$= \sum_{i=1}^{N} \log \left(\sum_{c=1}^{C} p(x_i, t_i = c|\theta) \right) \tag{67}$$

with the Jensen's inequality, we have

$$= \sum_{i=1}^{N} \log \left(\sum_{c=1}^{C} p(x_i, t_i = c|\theta) \right) \geq \mathcal{L}(\theta) \tag{68}$$

which means that instead of maximize the original marginal log likelihood, we will maximize a lower bound instead, which is more easy to maximize. Multiplying a dividing a term by the same value, we don't change the function. So, for convenience, we have

$$= \sum_{i=1}^{N} \log \left(\sum_{c=1}^{C} \frac{q(t_i = c)}{q(t_i = c)} p(x_i, t_i = c|\theta) \right) \tag{69}$$

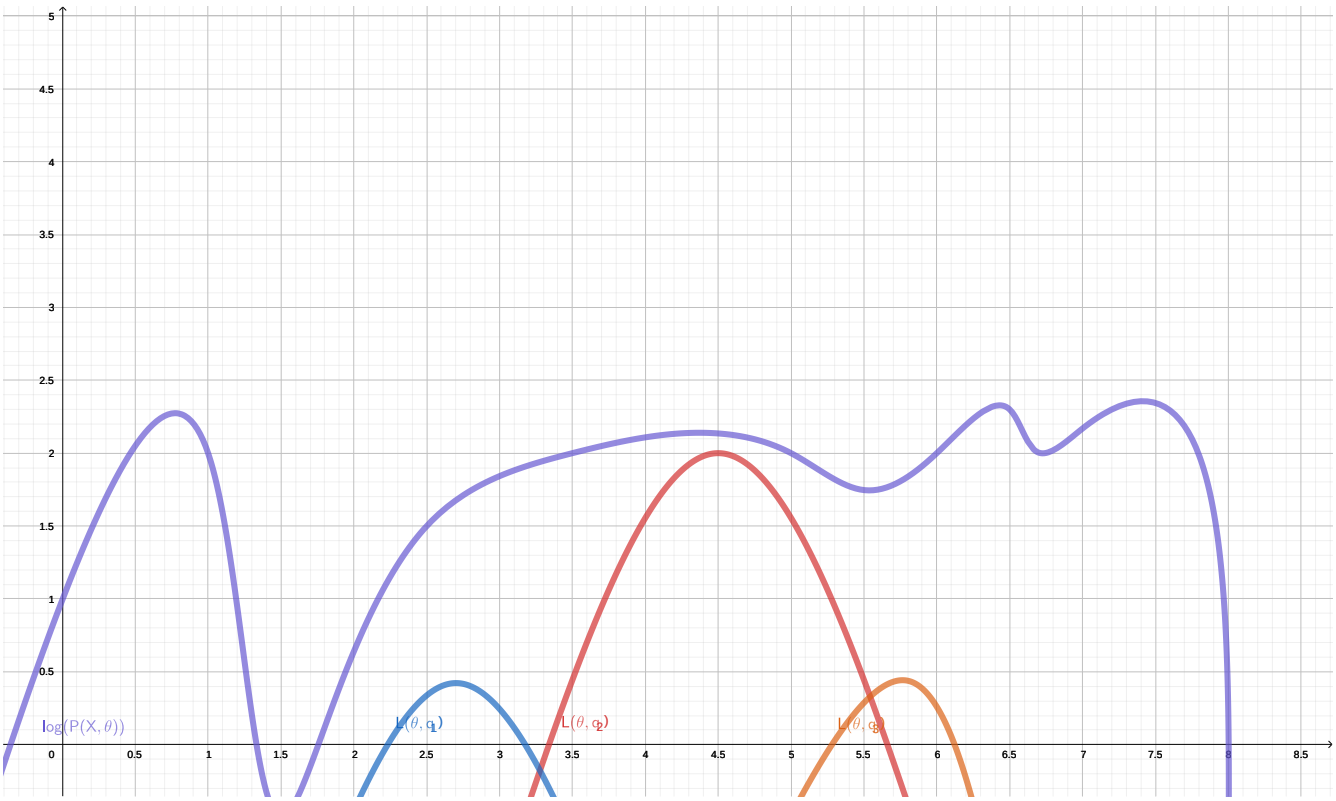


Figure 2: Loss from the General Form of Expectation of Maximization, changing q value. θ is the x-axis.

rewriting, what we have is the Jensen's inequality, in this equation

$$\log\left(\sum_c \alpha_c \upsilon_c\right) = \sum_c \left(\log \alpha_c(\upsilon_c)\right) \tag{70}$$

applying the logarithm properties from the Jensen's inequality, we rebuild the function to

$$\geq \sum_{i=1}^{N}\sum_{c=1}^{C}\left(q(t_i = c)\log\frac{p(x_i, t_i = c|\theta}{q(t_i = c)}\right) \tag{71}$$

$$= \mathcal{L}(\theta, q) \tag{72}$$

Graphically, we are changing the loss as we change the value of q, as in Figure 2. In practice, fixing q we have a new loss $\mathcal{L}$, as the blue curve in figure, which tends to have an global optimum point similar to a local optimum point in the original curve, the purple. In the next q step, we have another curve. Repeating iteratively, these variational lower-bound curves tends to lead θ to the global optimal point from the original curve.

Summarizing:

- $\log p(X|\theta) \geq \mathcal{L}(\theta, q)$ for any q, where $\mathcal{L}(\theta, q)$ is the variational lower bound
- **E-step**

$$q^{k+1} = arg\max_q \mathcal{L}(\theta^k, q) \tag{73}$$

- **M-step**

$$\theta^{k+1} = arg\max_\theta \mathcal{L}(\theta, q^{k+1}) \tag{74}$$

15 Class 17 - E Step

Fixing θ and maximizing $\mathcal{L}$, which is a log likelihood, changing q, which is a distribution

$$\max_{q} \mathcal{L}(\theta^k, q) \tag{75}$$

as in Figure 2, we want to minimize the gap between the purple curve, the real log likelihood $\log P(X|\theta)$, and some another curve, which is fixed the current lower bound at step k. We can describe the gap as

$$gap = \log P(X|\theta) - \mathcal{L}(\theta^k, q) \tag{76}$$

which is equal to

$$= \sum_{i=1}^{N} \log P(X_i|\theta) - \sum_{i=1}^{N}\sum_{c=1}^{C} q(t_i = c) \log \frac{P(X_i, t_i = c|\theta)}{q(t_i = c)} \tag{77}$$

rearranging the summations, we have

$$= \sum_{i=1}^{N} \left(\log P(X_i|\theta) * \sum_{c=1}^{N} q(t_i = c) - \sum_{c=1}^{C} q(t_i = c) \log \frac{P(X_i, t_i = c|\theta)}{q(t_i = c)} \right) \tag{78}$$

but $\sum_{c=1}^{N} q(t_i = c)$ is the summation of the probabilities of all the classes, which is always 1. As the two inner summations have this element, we can rearrange the equation, giving

$$= \sum_{i=1}^{N}\sum_{c=1}^{N} q(t_i = c) \left(\log P(X_i|\theta) - \log \frac{P(X_i, t_i = c|\theta)}{q(t_i = c)} \right) \tag{79}$$

and using the logarithm properties of division of terms, we have

$$= \sum_{i=1}^{N}\sum_{c=1}^{N} q(t_i = c) \left(\log \frac{P(X_i|\theta) q(t_i = c)}{P(X_i, t_i = c|\theta)} \right) \tag{80}$$

by some bayesian rules, $P(X_i, t_i = c|\theta) = P(t_i = c|X_i, \theta) P(X_i|\theta)$, replacing the terms, we can simplify the $P(X_i|\theta)$ in the numerator and denominator, giving

$$= \sum_{i=1}^{N}\sum_{c=1}^{N} q(t_i = c) \left(\log \frac{q(t_i = c)}{P(t_i = c|X_i, \theta)} \right) \tag{81}$$

where the inner summation is exactly the Kullback-Leibler Divergence $\mathcal{KL}\left(q(t_i||P(t_i|X_i, \theta)\right)$. Our final equation is

$$gap = \sum_{i=1}^{N} \mathcal{KL}\left(q(t_i||P(t_i|X_i, \theta)\right) \tag{82}$$

As we want to maximize the lower bound $\mathcal{L}$, we want to minimize the difference gived by the first gap equation, which means minimize the summation of the Kullback-Leibler Divergence.

16 Class 18 - M Step