Expectation-Maximization

Nuno Vasconcelos

ECE Department, UCSD

Expectation-maximization

- we have seen that EM is a framework for ML estimation with missing data
- i.e. problems where we have, two types of random variables
  - *X* observed random variable
  - *Z* hidden random variable
- goal:
  - given iid sample $D = \{x_1, \ldots, x_n\}$
  - find parameters Ψ^* that maximize likelihood with respect to *D*

$$
\begin{aligned}
\Psi^* &= \arg\max_{\Psi} P_{\mathbf{X}}(\mathcal{D}; \Psi) \\
&= \arg\max_{\Psi} \int P_{\mathbf{X}|Z}(\mathcal{D}|z; \Psi) P_Z(z; \Psi) dz
\end{aligned}
$$

Expectation-maximization

- the set

$$D = \{x_1, \ldots, x_n\}$$

  is called the incomplete data

- the set

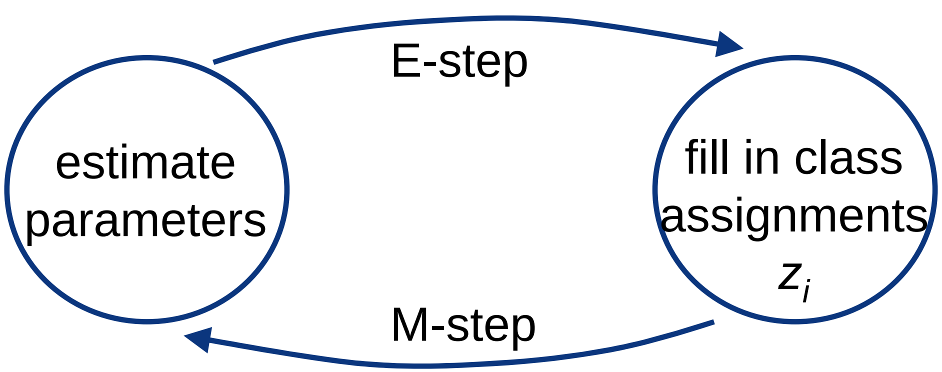
$$D_c = \{(x_1,z_1), \ldots, (x_n,z_n)\}$$

  is called the complete data

- we never get to see it, otherwise the problem would be trivial (standard ML)
- EM solves the problem by iterating between two steps

Expectation-maximization

- the basic idea is quite simple
  1. start with an initial parameter estimate $\Psi^{(0)}$
  2. **E-step:** given current parameters $\Psi^{(i)}$ and observations in D, "guess" what the values of the z_i are
  3. **M-step:** with the new z_i, we have a complete data problem, solve this problem for the parameters, i.e. compute $\Psi^{(i+1)}$
  4. go to 2.
- this can be summarized as



The Q function

- main idea: don't know what complete data likelihood is, but can compute its expected value given observed data
- this is the Q function

$$Q(\Psi; \Psi^{(n)}) = E_{Z|\mathbf{X};\Psi^{(i)}}\left[\log P_{\mathbf{X},Z}(\mathcal{D}, \{z_1, \ldots, z_N\}; \Psi)|\mathcal{D}\right]$$

- and is a bit tricky:
  - it is the expected value of likelihood with respect to complete data (joint X and Z)
  - given that we observed incomplete data (X)
  - note that the likelihood is a function of Ψ (the parameters that we want to determine)
  - but to compute the expected value we need to use the parameter values from the previous iteration (because we need a distribution for Z|X)

Expectation-maximization

- E-step:
  - given estimates $\Psi^{(n)} = \{\Psi^{(n)}_1, \ldots, \Psi^{(n)}_C\}$
  - compute expected log-likelihood of complete data

$$Q(\Psi; \Psi^{(n)}) = E_{Z|\mathbf{X};\Psi^{(n)}}\left[\log P_{\mathbf{X},Z}(\mathcal{D}, \{z_1, \ldots, z_N\}; \Psi)|\mathcal{D}\right]$$

- M-step:
  - find parameter set that maximizes this expected log-likelihood

$$\Psi^{(n+1)} = \arg\max_{\Psi} Q(\Psi; \Psi^{(n)})$$

- let's make this more concrete by looking at a toy example

Example

- toy model: *X* iid, *Z* iid, $X_i \sim N(\mu,1)$, $Z_i \sim \lambda e^{-\lambda z}$, *X* independent of *Z*
- $Q(\Psi;\Psi^{(n)}) = E_{Z|\mathbf{X};\Psi^{(n)}}\left[\log P_{\mathbf{X},Z}(\mathcal{D},\{z_1,\ldots,z_N\};\Psi)|\mathcal{D}\right]$

$$\begin{aligned}
&= E_{Z|\mathbf{X};\Psi^{(n)}}\left[-\sum_k \frac{(x_k-\mu)^2}{2} - \frac{N}{2}\log 2\pi - \lambda\sum_k z_k + N\log\lambda|\mathcal{D}\right] \\
&= -\sum_k \frac{(x_k-\mu)^2}{2} - \frac{N}{2}\log 2\pi - \lambda\sum_k E_{Z|\mathbf{X};\Psi^{(n)}}[z_k|x_k] + N\log\lambda \\
&= -\sum_k \frac{(x_k-\mu)^2}{2} - \frac{N}{2}\log 2\pi - \lambda\sum_k E_{Z_k;\Psi^{(n)}}[z_k] + N\log\lambda \\
&= -\sum_k \frac{(x_k-\mu)^2}{2} - \frac{N}{2}\log 2\pi - N\lambda E_{Z;\Psi^{(n)}}[z] + N\log\lambda \\
&= -\sum_k \frac{(x_k-\mu)^2}{2} - \frac{N}{2}\log 2\pi - N\frac{\lambda}{\lambda^{(n)}} + N\log\lambda
\end{aligned}$$

Example

- $\Psi^{(n+1)} = \arg\max_{\Psi} Q(\Psi; \Psi^{(n)})$

- $$Q(\Psi; \Psi^{(n)}) = -\sum_k \frac{(x_k - \mu)^2}{2} - \frac{N}{2}\log 2\pi - N\frac{\lambda}{\lambda^{(n)}} + N\log\lambda$$

$$\frac{\partial Q}{\partial \mu} = 0 \Leftrightarrow \boxed{\mu^{(n+1)} = \frac{1}{n}\sum_k x_k} \qquad \frac{\partial Q}{\partial \lambda} = 0 \Leftrightarrow \boxed{\lambda^{(n+1)} = \lambda^{(n)}}$$

- this makes sense:
  - since hidden variables *Z* are independent of observed *X*
  - ML estimate of μ is always the same: the sample mean, no dependence on z_i
  - ML estimate of λ is always the initial estimate $\lambda^{(0)}$: since the observations are independent of the z_i we have no information on what λ should be, other than initial guess.
- note that model does not make sense, not EM solution

EM for mixtures

- we have also seen a more serious example
- ML estimation of the parameters of a mixture

$$P_{\mathbf{X}}(\mathbf{x};\Psi) = \sum_{c=1}^{C} P_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}|c;\Psi_c)\pi_c$$

- we noted that the right way to represent Z is to use a binary vector of size equal to the # of classes

$$\mathbf{z} \in \{\mathbf{e}_1,\ldots,\mathbf{e}_C\} \qquad \mathrm{e}_j = \begin{bmatrix} 0 & \\ \vdots & \\ 1 & (j^{th} position) \\ \vdots & \\ 0 & \end{bmatrix}$$

- in which case complete data log-likelihood is linear on z_{ij}

$$\log P_{\mathbf{X},Z}(\mathcal{D},\{\mathbf{z}_1,\ldots,\mathbf{z}_n\};\Psi) = \sum_{i,j} z_{ij} \log\left[P_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}_i|\mathbf{e}_j,\Psi)\pi_j\right]$$

EM for mixtures

- the Q function becomes

$$\begin{aligned} Q(\Psi;\Psi^{(n)}) &= E_{Z|\mathbf{X};\Psi^{(n)}}\left[\log P_{\mathbf{X},Z}(\mathcal{D},\{z_1,\ldots,z_N\};\Psi)|\mathcal{D}\right] \\ &= \sum_{i,j} E_{Z|\mathbf{X};\Psi^{(n)}}[z_{ij}|\mathcal{D}]\log\left[P_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}_i|\mathbf{e}_j,\Psi)\pi_j\right] \end{aligned}$$

- i.e. to compute it we only need to find

$$E_{Z|\mathbf{X};\Psi^{(n)}}[z_{ij}|\mathcal{D}],\ \forall i,j$$

- and since z_{ij} is binary and only depends on x_i

$$E_{\mathbf{Z}|\mathbf{X};\Psi^{(n)}}[z_{ij}|\mathcal{D}] = P_{\mathbf{Z}|\mathbf{X}}(z_{ij}=1|\mathbf{x}_i;\Psi^{(n)}) = P_{\mathbf{Z}|\mathbf{X}}(\mathbf{e}_j|\mathbf{x}_i;\Psi^{(n)})$$

- the E-step reduces to computing the posterior probability of each point under each class!

Expectation-maximization

- and the EM algorithm reduces to
  1. E-step: Q function

  $$h_{ij} = P_{\mathbf{Z}|\mathbf{X}}(\mathbf{e}_j|\mathbf{x}_i;\Psi^{(n)})$$

  $$Q(\Psi;\Psi^{(n)}) = \sum_{i,j} h_{ij} \log\left[P_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}_i|\mathbf{e}_j,\Psi)\pi_j\right]$$

  2. M-step: solve the maximization, deriving a closed-form solution if there is one

  $$\Psi^{(n+1)} = \arg\max_{\Psi} \sum_{ij} h_{ij} \log\left[P_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}_i|\mathbf{e}_j,\Psi)\pi_j\right]$$

  under whatever constraints need to be considered, e.g.

  $$\sum_j \pi_j = 1$$

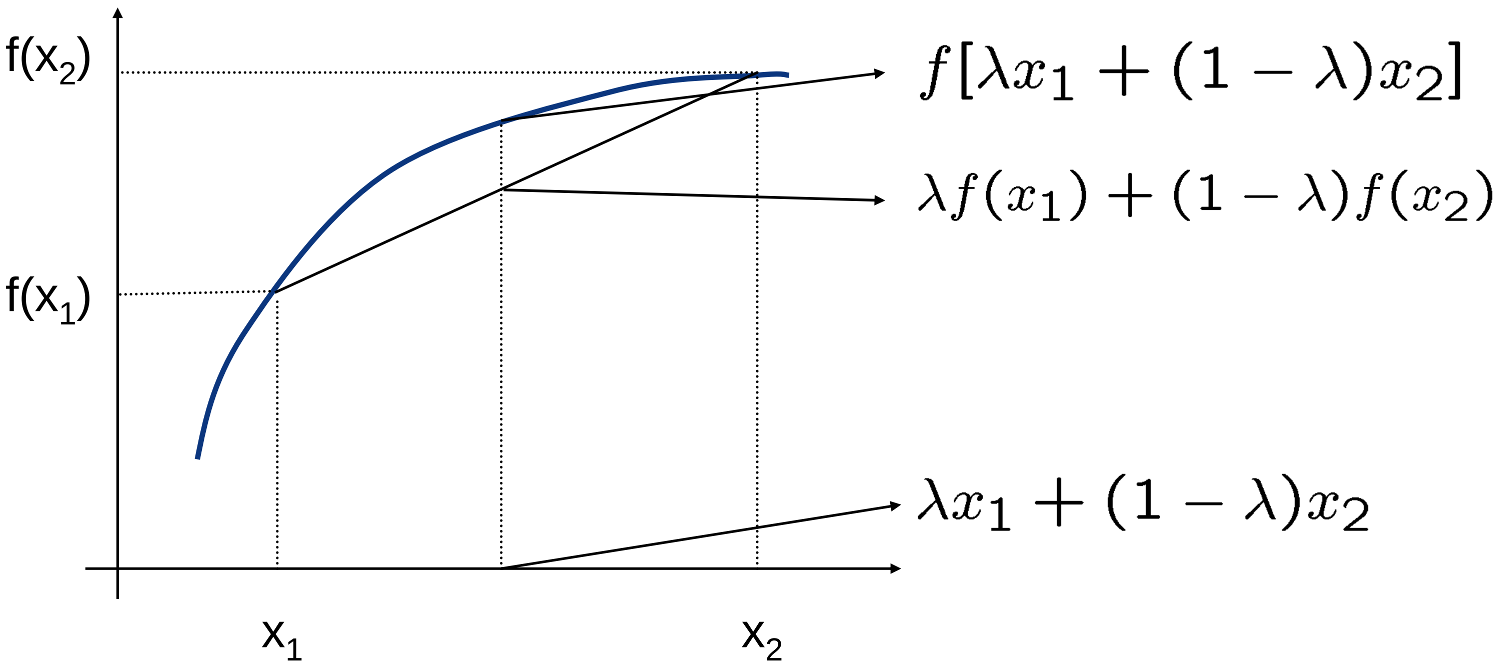
Convergence of EM

- so far we have shown that EM
  - makes intuitive sense
  - leads to intuitive update equations
- the obvious question is: "how do we know that it converges to something useful?"
- it turns out that the proof is frustratingly simple
  - "it takes longer to understand what each term means than to do the proof itself"
- the only tool that we really need is Jensen's inequality
- since this is such a useful inequality, let's go over it in some detail

Concave functions

- a function *f(x)* is concave in *(a,b)* if for all x_1, x_2 in *(a,b)* and λ in *[0,1]*

$$f[\lambda x_1 + (1-\lambda)x_2] \geq \lambda f(x_1) + (1-\lambda) f(x_2)$$



Jensen's inequality

- if *f(x)* is concave and X a random variable then

$$E[f(x)] \leq f(E[x])$$

- the proof is easy for discrete distributions, where it can be done by induction

1. assume *X* has two states with probability p_1, p_2. If f is concave, by definition

$$\begin{aligned} E[f(x)] &= p_1 f(x_1) + p_2 f(x_2) \\ &\leq f[p_1 x_1 + p_2 x_2] = f(E[x]) \end{aligned}$$

2. assume that the inequality holds for all random variables of n states, i.e.

$$\sum_{i=1}^{n} p_i f(x_i) \;\leq\; f\left(\sum_{i=1}^{n} p_i x_i\right)$$

Jensen's inequality

- assume $\sum_{i=1}^{n} p_i f(x_i) \;\leq\; f\left(\sum_{i=1}^{n} p_i x_i\right)$
- then for a r.v. with n+1 states

$$
\begin{aligned}
E[f(x)] &= \sum_{i=1}^{n+1} p_i f(x_i) = \sum_{i=1}^{n} p_i f(x_i) + p_{n+1} f(x_{n+1}) \\
&= (1 - p_{n+1}) \sum_{i=1}^{n} \frac{p_i}{1 - p_{n+1}} f(x_i) + p_{n+1} f(x_{n+1}) \\
&\leq (1 - p_{n+1}) f\left(\sum_{i=1}^{n} \frac{p_i}{1 - p_{n+1}} x_i\right) + p_{n+1} f(x_{n+1})
\end{aligned}
$$

and from the definition of concavity

$$
E[f(x)] \;\leq\; f\left((1 - p_{n+1}) \sum_{i=1}^{n} \frac{p_i}{1 - p_{n+1}} x_i + p_{n+1} x_{n+1}\right)
$$

Jensen's inequality

- $$E[f(x)] \leq f\left((1-p_{n+1})\sum_{i=1}^{n}\frac{p_i}{1-p_{n+1}}x_i + p_{n+1}x_{n+1}\right)$$
  $$= f\left(\sum_{i=1}^{n+1} p_i x_i\right) = f(E[x])$$
- in summary:
  - inequality holds for r.v. with two states
  - given that it holds for n states it also holds for n+1 states
  - hence, by induction, it follows that for all discrete distributions and concave f(.)

$$\boxed{E[f(x)] \leq f(E[x])}$$

- the result generalizes for the continuous case, but the proof is more complicated

EM convergence

- we are now ready to show that EM converges
- recall: the goal is to maximize $\log P_{\mathbf{X}}(\mathcal{D};\Psi)$
- using

$$P_{\mathbf{X},\mathbf{Z}}(\mathcal{D},\mathbf{z};\Psi) = P_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathcal{D};\Psi)P_{\mathbf{X}}(\mathcal{D};\Psi)$$

- this can be written as

$$\log P_{\mathbf{X}}(\mathcal{D};\Psi) = \log P_{\mathbf{X},\mathbf{Z}}(\mathcal{D},\mathbf{z};\Psi) - \log P_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathcal{D};\Psi)$$

- taking expectations on both sides and using the fact that the LHS does not depend on *Z*

$$\begin{aligned}\log P_{\mathbf{X}}(\mathcal{D};\Psi) &= E_{\mathbf{Z}|\mathbf{X};\Psi^{(n)}}[\log P_{\mathbf{X},\mathbf{Z}}(\mathcal{D},\mathbf{z};\Psi)|\mathcal{D}] \\ &- E_{\mathbf{Z}|\mathbf{X};\Psi^{(n)}}[\log P_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathcal{D};\Psi)|\mathcal{D}]\end{aligned}$$

EM convergence

- and plugging in the definition of the Q function

$$\begin{aligned}\log P_{\mathbf{X}}(\mathcal{D};\Psi) &= E_{\mathbf{Z}|\mathbf{X};\Psi^{(n)}}[\log P_{\mathbf{X},\mathbf{Z}}(\mathcal{D},\mathbf{z};\Psi)|\mathcal{D}] \\ &- E_{\mathbf{Z}|\mathbf{X};\Psi^{(n)}}[\log P_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathcal{D};\Psi)|\mathcal{D}] \\ &= Q(\Psi|\Psi^{(n)}) + H(\Psi|\Psi^{(n)})\end{aligned}$$

- where we have also introduced

$$\begin{aligned}H(\Psi|\Psi^{(n)}) &= -E_{\mathbf{Z}|\mathbf{X};\Psi^{(n)}}[\log P_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathcal{D};\Psi)|\mathcal{D}] \\ &= -\int P_{\mathbf{Z}|\mathbf{X};\Psi^{(n)}}(\mathbf{z}|\mathcal{D};\Psi^{(n)}) \log P_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathcal{D};\Psi) d\mathbf{z}\end{aligned}$$

EM convergence

- the key to proving convergence is this equation

$$\log P_{\mathbf{X}}(\mathcal{D};\Psi) \;=\; Q(\Psi|\Psi^{(i)}) + H(\Psi|\Psi^{(i)})$$

- note, in particular, that

$$
\begin{aligned}
\log P_{\mathbf{X}}(\mathcal{D};\Psi^{(n+1)}) - \log P_{\mathbf{X}}(\mathcal{D};\Psi^{(n)}) = \\
= \;& Q(\Psi^{(n+1)}|\Psi^{(n)}) + H(\Psi^{(n+1)}|\Psi^{(n)}) \\
& -[Q(\Psi^{(n)}|\Psi^{(n)}) + H(\Psi^{(n)}|\Psi^{(n)})] \\
= \;& Q(\Psi^{(n+1)}|\Psi^{(n)}) - Q(\Psi^{(n)}|\Psi^{(n)}) \\
& +H(\Psi^{(n+1)}|\Psi^{(n)}) - H(\Psi^{(n)}|\Psi^{(n)})
\end{aligned}
$$

EM convergence

- but, by definition of the M-step

$$\Psi^{(n+1)} = \arg\max_{\Psi} Q(\Psi|\Psi^{(n)})$$

- it follows that

$$\boxed{Q(\Psi^{(n+1)}|\Psi^{(n)}) \geq Q(\Psi^{(n)}|\Psi^{(n)})}$$

- and since

$$\log P_{\mathbf{X}}(\mathcal{D};\Psi^{(n+1)}) - \log P_{\mathbf{X}}(\mathcal{D};\Psi^{(n)}) =$$
$$= Q(\Psi^{(n+1)}|\Psi^{(n)}) - Q(\Psi^{(n)}|\Psi^{(n)})$$
$$+ H(\Psi^{(n+1)}|\Psi^{(n)}) - H(\Psi^{(n)}|\Psi^{(n)})$$

we have

$$\log P_{\mathbf{X}}(\mathcal{D};\Psi^{(n+1)}) \geq \log P_{\mathbf{X}}(\mathcal{D};\Psi^{(n)})$$

EM convergence

- we have

$$\log P_{\mathbf{X}}(\mathcal{D}; \Psi^{(n+1)}) \geq \log P_{\mathbf{X}}(\mathcal{D}; \Psi^{(n)})$$

- if

$$\boxed{H(\Psi^{(n+1)}|\Psi^{(n)}) \quad \geq \quad H(\Psi^{(n)}|\Psi^{(n)})}$$

- but, from

$$H(\Psi|\Psi^{(n)}) = -E_{\mathbf{Z}|\mathbf{X};\Psi^{(n)}}[\log P_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathcal{D}; \Psi)|\mathcal{D}]$$

we have

$$\begin{aligned} & H(\Psi^{(n+1)}|\Psi^{(n)}) - H(\Psi^{(n)}|\Psi^{(n)}) \\ = \quad & -E_{\mathbf{Z}|\mathbf{X};\Psi^{(n)}}\left[\log \frac{P_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathcal{D}; \Psi^{(n+1)})}{P_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathcal{D}; \Psi^{(n)})}|\mathcal{D}\right] \end{aligned}$$

EM convergence

- and, since the log is a concave function, by Jensen's $E[f(x)] \le f(E[x])$

- $H(\Psi^{(n+1)}|\Psi^{(n)}) - H(\Psi^{(n)}|\Psi^{(n)})$

$$
\begin{aligned}
&= -E_{\mathbf{Z}|\mathbf{X};\Psi^{(n)}}\left[\log \frac{P_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathcal{D};\Psi^{(n+1)})}{P_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathcal{D};\Psi^{(n)})}|\mathcal{D}\right] \\
&\ge -\log E_{\mathbf{Z}|\mathbf{X};\Psi^{(n)}}\left[\frac{P_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathcal{D};\Psi^{(n+1)})}{P_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathcal{D};\Psi^{(n)})}|\mathcal{D}\right] \\
&= -\log \int P_{\mathbf{Z}|\mathbf{X};\Psi^{(n)}}(\mathbf{z}|\mathcal{D};\Psi^{(n)}) \frac{P_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathcal{D};\Psi^{(n+1)})}{P_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathcal{D};\Psi^{(n)})} d\mathbf{z} \\
&= -\log 1 = 0
\end{aligned}
$$

EM convergence

- this shows that

$$\log P_{\mathbf{X}}(\mathcal{D}; \Psi^{(n+1)}) \geq \log P_{\mathbf{X}}(\mathcal{D}; \Psi^{(n)})$$

- i.e. the log-likelihood of the incomplete data can only increase from iteration to iteration
- hence the algorithm converges
- note that there is no guarantee of convergence to a global minimum, only local

Geometric interpretation

- one can also derive a geometric interpretation from

$$\log P_{\mathbf{X}}(\mathcal{D};\Psi) \;=\; Q(\Psi|\Psi^{(n)}) + H(\Psi|\Psi^{(n)})$$

- by noting that

$$H(\Psi|\Psi^{(n)}) = -E_{\mathbf{Z}|\mathbf{X};\Psi^{(n)}}[\log P_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathcal{D};\Psi)|\mathcal{D}]$$

$$= \; -\int P_{\mathbf{Z}|\mathbf{X};\Psi^{(n)}}(\mathbf{z}|\mathcal{D};\Psi^{(n)}) \log P_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathcal{D};\Psi) d\mathbf{z}$$

- is of the form

$$H(\Psi|\Psi^{(n)}) = -\int p_n(\mathbf{z}) \log p(\mathbf{z}) d\mathbf{z}$$

$$= \; \int p_n(\mathbf{z}) \log \frac{p_n(\mathbf{z})}{p(\mathbf{z})} d\mathbf{z} - \int p_n(\mathbf{z}) \log p_n(\mathbf{z}) d\mathbf{z}$$

Geometric interpretation

- is of the form

$$H(\Psi|\Psi^{(n)}) = -\int p_n(\mathbf{z}) \log p(\mathbf{z}) d\mathbf{z}$$

$$= \int p_n(\mathbf{z}) \log \frac{p_n(\mathbf{z})}{p(\mathbf{z})} d\mathbf{z} - \int p_n(\mathbf{z}) \log p_n(\mathbf{z}) d\mathbf{z}$$

$$= KL[p_n||p] + H[p_n]$$

- where *KL[p||q]* is the Kullback-Leibler divergence between *p* and *q*, and *H[p]* the entropy of *p*
- it can be shown that these two quantities are never negative, from which $H(\Psi|\Psi^{(n)}) \geq 0$ and
- since

$$\log P_{\mathbf{X}}(\mathcal{D};\Psi) = Q(\Psi|\Psi^{(n)}) + H(\Psi|\Psi^{(n)})$$

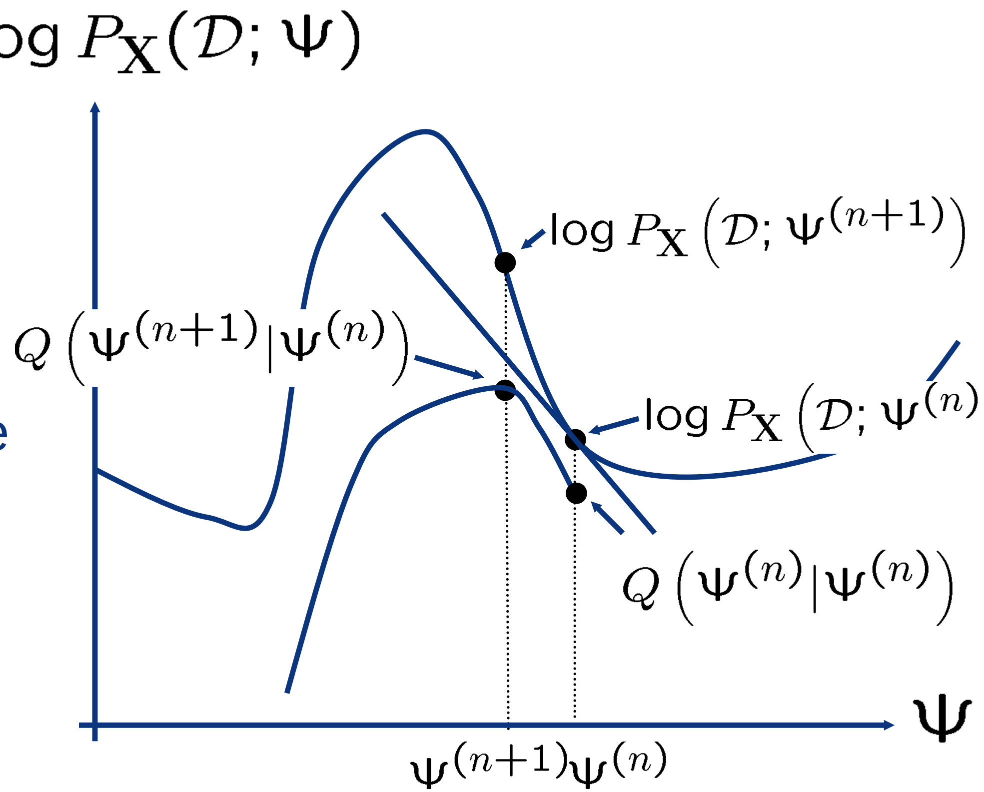
Geometric interpretation

- we have

$$\log P_{\mathbf{X}}(\mathcal{D};\Psi) \;\geq\; Q(\Psi|\Psi^{(n)})$$

- which means that the Q function is a lower bound to the log-likelihood of the observed data

- this allows an interpretation of the EM steps as
  - E-step: lower-bound the observed log-likelihood
  - M-step: maximize the lower bound



Geometric interpretation

- consider next the difference between cost and bound

$$\log P_{\mathbf{X}}(\mathcal{D};\Psi) \;-\; Q(\Psi|\Psi^{(n)}) = H(\Psi|\Psi^{(n)})$$

- which can be written as

$$\boxed{H(\Psi|\Psi^{(n)}) \;=\; KL[p_n||p] + H[p_n]}$$

with

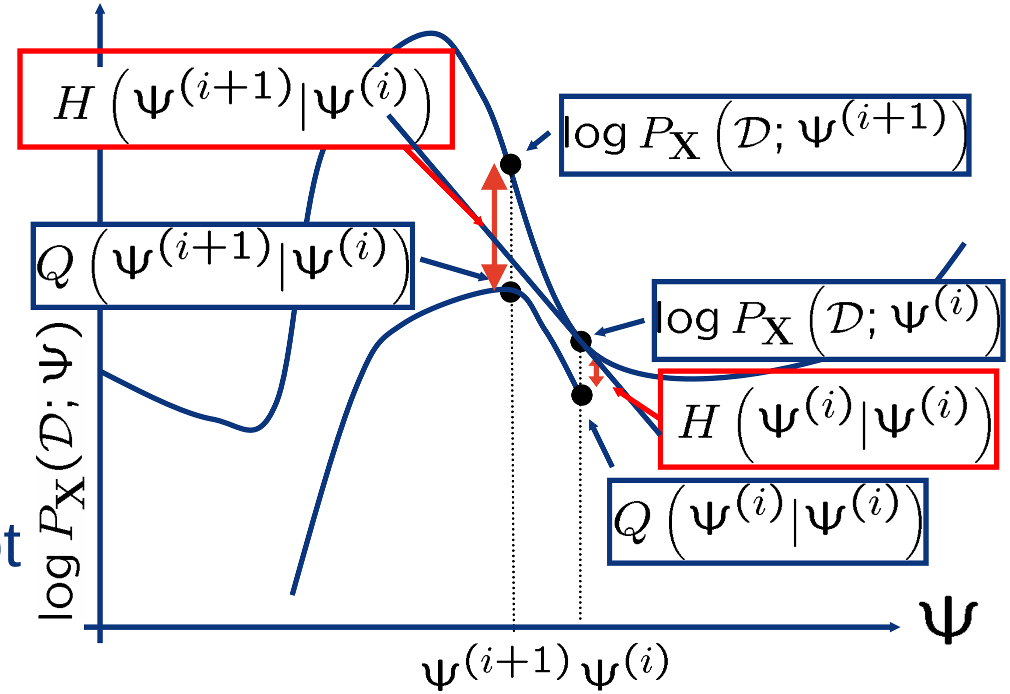
$$p_n(\mathbf{z}) = P_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathcal{D};\Psi^{(n)}) \qquad p(\mathbf{z}) = P_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathcal{D};\Psi)$$

- hence

$$\begin{aligned} H(\Psi^{(n+1)}|\Psi^{(n)}) - H(\Psi^{(n)}|\Psi^{(n)}) &= \\ &= KL[p_n||p_{n+1}] + H[p_n] - KL[p_n||p_n] - H[p_n] \\ &= KL[p_n||p_{n+1}] \geq 0 \end{aligned}$$

Geometric interpretation

- note that since
  - by definition of M-step: $Q(\Psi^{(n+1)}|\Psi^{(n)}) \;\geq\; Q(\Psi^{(n)}|\Psi^{(n)})$
  - by non-negativity of KL: $H(\Psi^{(n+1)}|\Psi^{(n)}) \geq H(\Psi^{(n)}|\Psi^{(n)})$
- it follows that $\log P_{\mathbf{X}}\left(\mathcal{D};\Psi^{(n+1)}\right) \geq \log P_{\mathbf{X}}\left(\mathcal{D};\Psi^{(n)}\right)$
- EM converges without need for step sizes
- this is not the case for gradient ascent which uses the linear approximation
- if we move too far, there will be overshoot



Extensions

- note that in the proof we have really only used the fact that

$$Q(\Psi^{(n+1)}|\Psi^{(n)}) \geq Q(\Psi^{(n)}|\Psi^{(n)})$$

- this means that
  - in M-step we do not necessarily need to maximize the Q-function
  - any step that increases it is sufficient
- Generalized EM-algorithm
  - E-step: compute

$$Q(\Psi|\Psi^{(n)}) = E_{\mathbf{Z}|\mathbf{X};\Psi^{(n)}}[\log P_{\mathbf{X}|\mathbf{Z}}(\mathcal{D},\mathbf{z};\Psi)|\mathcal{D}]$$

  - M-step: pick $\Psi^{(n+1)}$ such that

$$Q(\Psi^{(n+1)}|\Psi^{(n)}) \geq Q(\Psi^{(n)}|\Psi^{(n)})$$

Extensions

- Generalized EM-algorithm
  - E-step: compute

$$Q(\Psi|\Psi^{(n)}) = E_{\mathbf{Z}|\mathbf{X};\Psi^{(n)}}[\log P_{\mathbf{X}|\mathbf{Z}}(\mathcal{D}, \mathbf{z}; \Psi)|\mathcal{D}]$$

  - M-step: pick $\Psi^{(n+1)}$ such that

$$Q(\Psi^{(n+1)}|\Psi^{(n)}) \geq Q(\Psi^{(n)}|\Psi^{(n)})$$

- very useful when M-step is itself non-trivial:
  - e.g. if there is no closed-form solution one has to resort to numerical methods, like gradient ascent
  - can be computationally intensive, lots of iterations per M-step
  - in these cases, it is usually better to just perform a few iterations and move on to the next E-step
  - no point in precisely optimizing M-step if everything is going to change when we compute the new E-step

MAP parameter estimates

- so far we have concentrated on ML estimation
- EM can be equally applied to obtain MAP estimates, with a straightforward extension
- recall that for MAP the goal is

$$\begin{aligned}\Psi^* &= \arg\max_{\Psi} P_{\Psi|\mathbf{X}}(\Psi|\mathcal{D}) \\ &= \arg\max_{\Psi} P_{\mathbf{X}|\Psi}(\mathcal{D}|\Psi) P_{\Psi}(\Psi)\end{aligned}$$

- this is not very different from ML, we just multiply by $P_{\Psi}(\Psi)$
- still a problem of estimation from incomplete data, with

$$P_{\mathbf{X}|\Psi}(\mathcal{D}|\Psi) = \int P_{\mathbf{X}|\mathbf{Z},\Psi}(\mathcal{D}|\mathbf{z},\Psi) P_{\mathbf{Z}|\Psi}(\mathbf{z}|\Psi) d\mathbf{z}$$

MAP parameter estimates

- and there is a complete data posterior

$$P_{\Psi|\mathbf{X},\mathbf{Z}}(\Psi|\mathcal{D},\mathbf{z})$$

- the E step is now to compute

$$\begin{aligned} E_{\mathbf{Z}|\mathbf{X},\Psi}[\log P_{\Psi|\mathbf{X},\mathbf{Z}}(\Psi|\mathcal{D},\mathbf{z})|\mathcal{D},\Psi^{(n)}] &= \\ = \;& E_{\mathbf{Z}|\mathbf{X},\Psi}[\log P_{\mathbf{X},\mathbf{Z}|\Psi}(\mathcal{D},\mathbf{z}|\Psi)|\mathcal{D},\Psi^{(n)}] + \\ & + E_{\mathbf{Z}|\mathbf{X},\Psi}[\log P_{\Psi}(\Psi)|\mathcal{D},\Psi^{(n)}] - \\ & - E_{\mathbf{Z}|\mathbf{X},\Psi}[\log P_{\mathbf{X},\mathbf{Z}}(\mathcal{D},\mathbf{z})|\mathcal{D},\Psi^{(n)}] \\ = \;& Q(\Psi|\Psi^{(n)}) + \log P_{\Psi}(\Psi) - \\ - \;& E_{\mathbf{Z}|\mathbf{X},\Psi}[\log P_{\mathbf{X},\mathbf{Z}}(\mathcal{D},\mathbf{z})|\mathcal{D},\Psi^{(n)}] \end{aligned}$$

- note that the last term does not depend on Ψ
- does not affect M-step, we can drop it

MAP parameter estimates

- hence the E-step does not really change
- E step: compute

$$Q(\Psi|\Psi^{(n)}) \;=\; E_{\mathbf{Z}|\mathbf{X},\Psi}[\log P_{\mathbf{X},\mathbf{Z}|\Psi}(\mathcal{D},\mathbf{z}|\Psi)|\mathcal{D},\Psi^{(n)}]$$

- and the M-step becomes

$$\Psi^{(n+1)} \;=\; \arg\max_{\Psi}\left\{Q(\Psi|\Psi^{(n)}) + \log P_{\Psi}(\Psi)\right\}$$

- this is the MAP-EM algorithm
- note that M-step looks like a standard Bayesian estimate procedure, and typically is
- e.g. for mixtures, it is equivalent to computing Bayesian estimates for each component, under "soft-assignments"

MAP parameter estimates

- in result, the estimates are similar to standard Bayesian estimates, but with
  - each point contributing to the parameters of all components
  - contribution weighted by the assignment probability
- but the important fact is that all the properties of Bayesian estimates still apply
  - conjugate priors
  - interpretation as additional, properly biased data, etc.
- this is a reason why our study of Bayesian estimation with simple models was so important
  - while a Gaussian is a fairly weak model
  - most densities can be approximated by a mixture of Gaussians
  - with EM we can generalize all we did quite easily

Any Questions?