Learn with Hidden Variables

$$p(d, h \mid \underline{\lambda}) = \frac{1}{Z[\underline{\lambda}]} e^{\underline{\lambda} \cdot \underline{\Phi}(d,h)}$$

d - observed variable
h - hidden variable

$$P(d \mid \underline{\lambda}) = \sum_h P(d, h \mid \underline{\lambda})$$

Maximum Likelihood (ML) estimate of $\underline{\lambda}$

$$\hat{\underline{\lambda}} = \underset{\underline{\lambda}}{\text{ARG MAX}}\ P(d \mid \underline{\lambda}) = \underset{\underline{\lambda}}{\text{ARG MIN}}\ -\log P(d \mid \underline{\lambda})$$

Claim: minimizing $-\log P(d \mid \underline{\lambda})$ with respect to $\underline{\lambda}$ is equivalent to minimizing

$$F[\underline{\lambda}, q] = -\log P(d \mid \underline{\lambda}) + \sum_h q(h) \log \frac{q(h)}{P(h \mid d, \underline{\lambda})}$$

with respect to $\underline{\lambda}$ and $q(h)$, where $q(h)$ is a probability distribution on the hidden variables h - i.e. $q(h) \geq 0$, for all h, and $\sum_h q(h) = 1$.

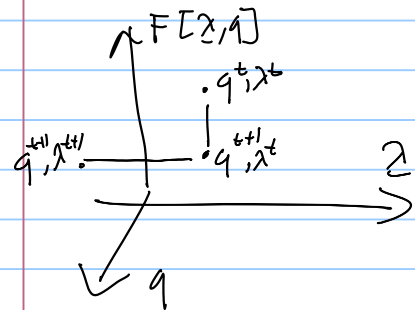
Proof. $\sum_h q(h) \log \frac{q(h)}{P(h \mid d, \underline{\lambda})}$ is the Kullback-Leibler divergence.

It ≥ 0, with $= 0$ only if $q(h) = P(h \mid d, \underline{\lambda})$

So to minimize $F[\underline{\lambda}, q]$ you can minimize w.r.t. $q()$ to set $q(h) = P(h \mid d, \underline{\lambda})$, then you have to minimize $-\log P(d \mid \underline{\lambda})$ w.r.t. $\underline{\lambda}$, which is the original problem.

$F[\underline{\lambda}, q]$ can be rewritten as

$$F[\underline{\lambda}, q] = \sum_h q(h) \log q(h) - \sum_h q(h) \log P(d, h \mid \underline{\lambda})$$



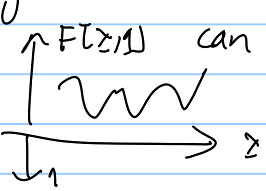
Try to minimize $F[\underline{\lambda}, q]$ by Coordinate descent:

(i) Fix $\underline{\lambda}^t$, minimize $F[\underline{\lambda}, q]$ w.r.t. q to get q^{t+1}

(ii) Fix q^{t+1}, minimize $F[\underline{\lambda}, q]$ w.r.t. $\underline{\lambda}$ to get $\underline{\lambda}^{t+1}$

repeat.

Each step is guaranteed to reduce $F[\underline{\lambda}, q]$. So the algorithm will converge to a local, or global, minimum.



can have local minima, so no guarantee that the algorithm will reach a global minimum.
- i.e. EM may not converge to the ML estimate of λ.

(1) Minimize $F[\underline{\lambda}^t, q]$ w.r.t q

gives $\quad q^{t+1}(h) = P(h|d,\lambda^t) = \dfrac{P(h,d|\lambda^t)}{\sum_h P(h,d|\lambda^t)}$

requires the ability to calculate $\sum_h P(h,d|\lambda^t)$ may be difficult

(2) Minimize $F[\underline{\lambda}, \underline{q}^t]$ w.r.t. $\underline{\lambda}$.

solve $\quad \dfrac{\partial F(\underline{\lambda}, q^t)}{\partial \underline{\lambda}} = 0, \quad \dfrac{\partial}{\partial \underline{\lambda}} \sum_h q^t(\underline{h}) \{ \underline{\lambda} \cdot \underline{\phi}(h,d) - \log Z[\underline{\lambda}] \},$

$$\sum_h q^t(\underline{h})\, \underline{\phi}(\underline{h},\underline{d}) = \sum_{h,d} \underline{\phi}(h,\underline{d})\, P(\underline{h},d|\underline{\lambda}^{t+1})$$

statistics of $\underline{d}$, with expectation over the hidden variables $\underline{h}$ w.r.t. $q^t(\underline{h})$

expected statistics of the model with parameter $\underline{\lambda}$

(Compare to ML formula for learning a model without hidden variables)

Note: solving this equation is often not easy because it requires computing $\sum_{h,d} \underline{\phi}(h,d) P(h,d|\lambda^{t+1})$

There are hidden variables so we cannot match the data statistics $\underline{\phi}(h,\underline{d})$ to the model statistics $\sum_{h,d} \underline{\phi}(h,d) P(\underline{h},d|\lambda^{t+1})$ — because we do not know $\underline{h}$.

So instead we try to estimate a distribution $q(\underline{h})$ over h.

This is like a chicken and egg problem

→ we estimate $q(h)$ assuming we know λ $\Rightarrow \dfrac{\partial F}{\partial q} = 0$

→ then we estimate λ assuming we know $q(h)$ $\Rightarrow \dfrac{\partial F}{\partial \lambda} = 0$

Extension to multiple data $D = \{d^\mu : \mu = 1 \text{ to } N\}$,

ML minimizes $\quad -\sum_\mu \log P(d^\mu|\lambda)$

$$F[\lambda, \{q^\mu\}] = -\sum_\mu \sum_{h^\mu} q^\mu(h_\mu) \log q^\mu(h_\mu) - \sum_\mu \sum_{h_\mu} q_\mu(h_\mu) \log P(h_\mu, d_\mu|\lambda)$$

w.r.t. λ and $\{q^\mu\}$.

Minimize w.r.t. q_μ $\quad q_\mu^{t+1}(h_\mu) = P(h_\mu|d_\mu, \underline{\lambda}^t) \qquad \mu = 1 \text{ to } N$

Minimize w.r.t. $\underline{\lambda}$ $\quad -\sum_\mu \sum_{h_\mu} q_\mu(h_\mu)\, \underline{\phi}(h_\mu, d_\mu) = N \sum_{h,d} \phi(h,d) \dfrac{e^{\underline{\lambda} \cdot \underline{\phi}(h,d)}}{Z[\lambda]}$