Machine Learning 2

Information Theory
- Relative Entropy
- revisited

Patrick Forré

Relative Entropy / Kullback-Leibler Divergence

- Let $q(x)$ and $p(x)$ be two distribution for the same observation space.

- Then the Relative Entropy or Kullback-Leibler Divergence is defined:

$$\mathrm{KL}\left(q(X)\|p(X)\right) := \mathbb{E}_{q(X)}\left[\log \frac{q(X)}{p(X)}\right]$$

- Fundamental Inequality:

$$\mathrm{KL}\left(q(X)\|p(X)\right) \geq 0,$$

with equality if and only if $q(x) = p(x)$ for almost all values x.

Chain Rule and Data Processing Inequality for KL

- Let $q(x, z)$ and $p(x, z)$ two joint distribution for the same product space.

- Then we have the Chain Rule:

$$\mathrm{KL}\left(q(X,Z)\|p(X,Z)\right) = KL\left(q(X)\|p(X)\right) + \mathbb{E}_{q(x)}\left[KL\left(q(Z|X)\|p(Z|X)\right)\right]$$

≥ 0 $\quad$ ≥ 0

Proof:

$$\log\frac{q(z|x)q(x)}{p(z|x)p(x)} = \log\frac{q(z|x)}{p(z|x)} + \log\frac{q(x)}{p(x)}$$

- From this we get the Data Processing Inequality:

$$\mathrm{KL}\left(q(X,Z)\|p(X,Z)\right) \geq KL\left(q(x)\|p(x)\right)$$

with equality if and only if $q(z\,|\,x) = p(z\,|\,x)$ for almost all values x and all z.

Principle of Minimal Relative Entropy

- Let $q(x)$ be an underlying 'true' distribution of interest and $\{p_\theta(x) \mid \theta \in \Theta\}$ a statistical model for $q(x)$.

- The Principle of Minimal Relative Entropy says one should try to minimize:

$$\theta^* \in \arg\min_\theta \mathrm{KL}\left(q(X) \| p_\theta(X)\right)$$

- We have seen that if one replaces the (unknown) $q(x)$ with an empirical distribution $\hat{q}(x)$ one arrives at the learning framework of Maximum Likelihood Estimation.

Machine Learning 2

Variational Inference
- Ideas and Principles

Patrick Forré

Remark: Approximate Inference

- If $q(x)$ is the distribution / marginal of interest and d a divergence measure quantifying the difference between $q(x)$ and a model distribution $p_\theta(x)$,
- then minimizing such a divergence (e.g. $d = \mathrm{KL}$):

$$\arg\min_\theta d\left(q(X) \| p_\theta(X)\right)$$

between $q(x)$ and an approximating model $p_\theta(x)$ can be seen as a general approach for approximate inference.

- Small values for $d\left(q(X) \| p_\theta(X)\right)$ means good approximation.

Remark: Latent Variable Model vs Bayesian Setting

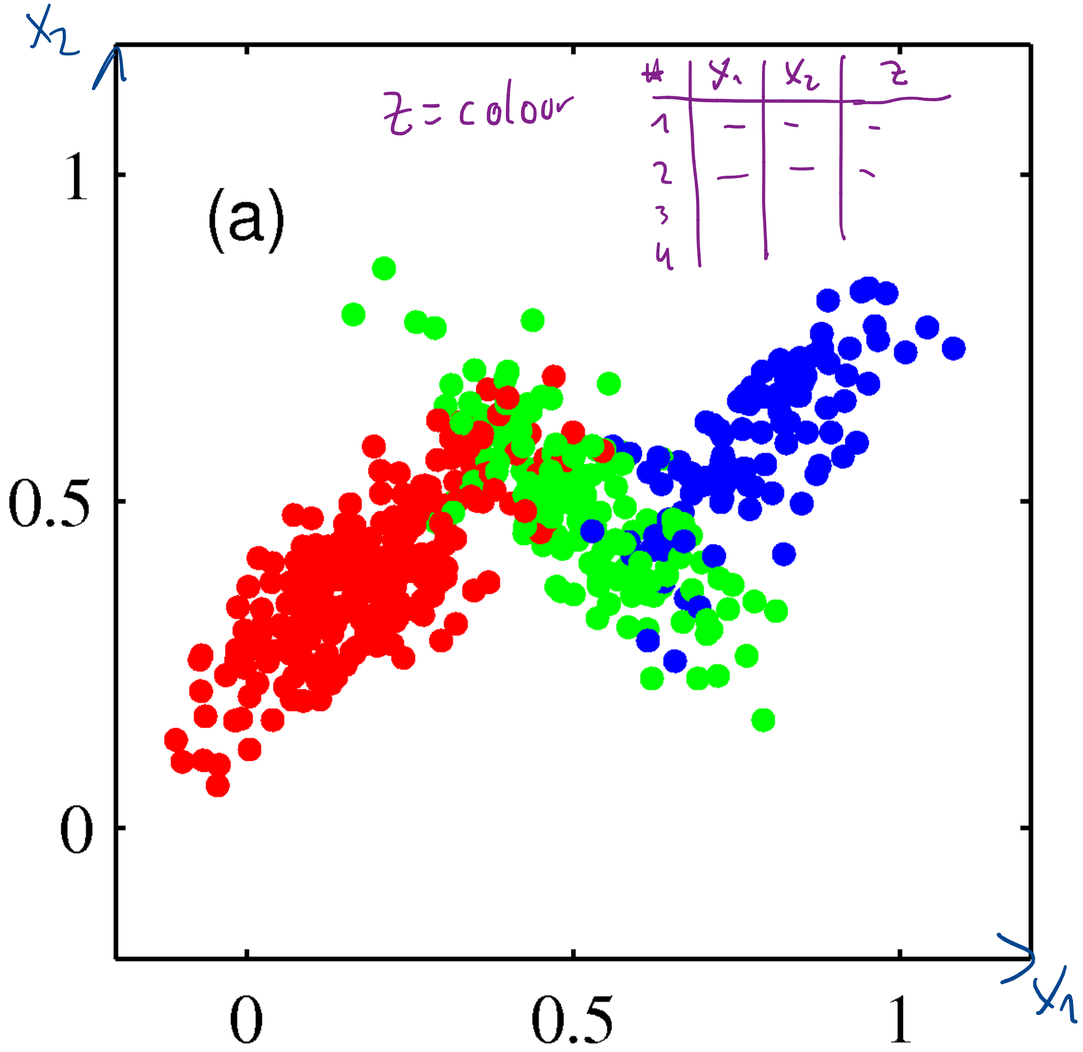
- Consider joint distribution $p_\theta(x, z) = p_{\theta_2}(x \mid z) \cdot p_{\theta_1}(z)$ and data $\hat{X}$.

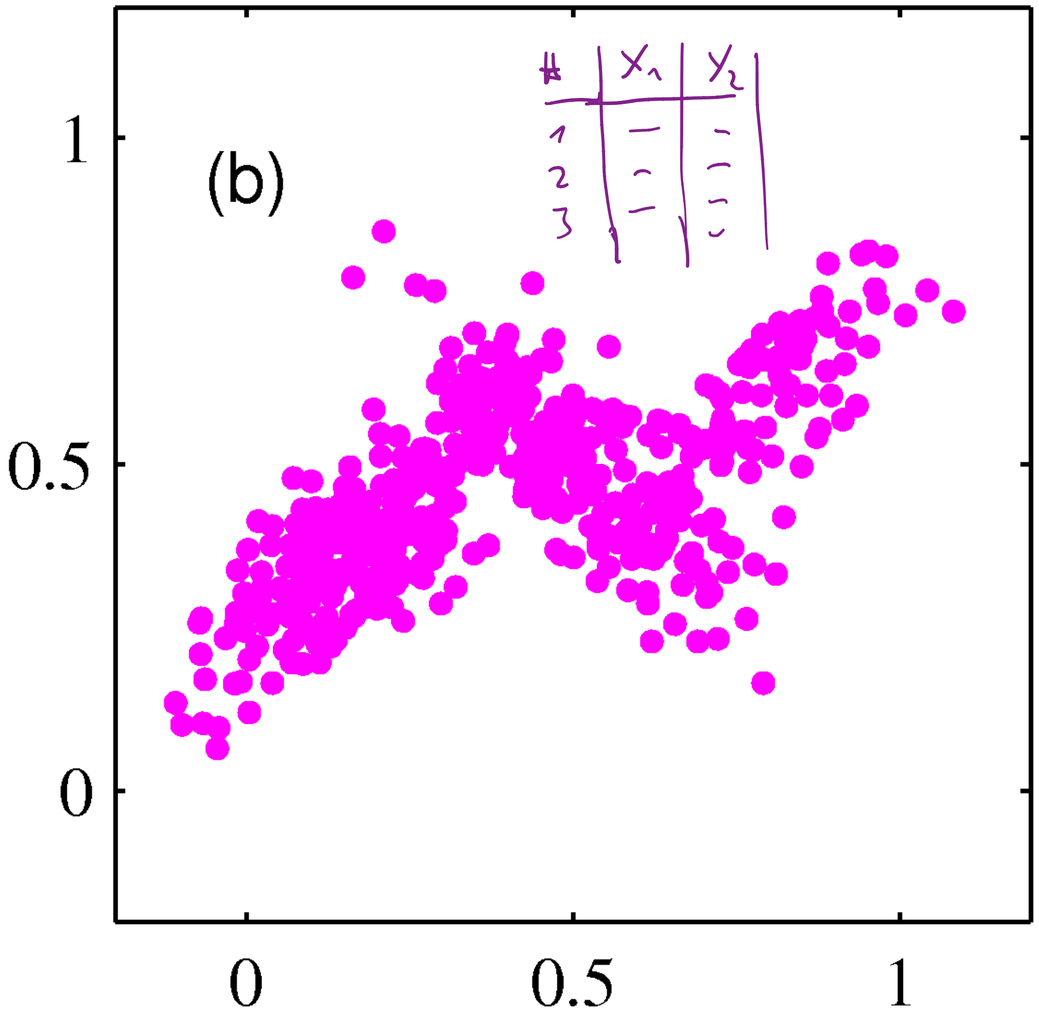
- Latent Variable Model
- z: latent variables
- θ: parameters
- $p_{\theta_1}(z)$: marginal latent distribution
- $p_\theta(\hat{X})$: likelihood
- $p_{\theta_2}(x \mid z)$: latent conditional distribution
- $p_\theta(z \mid x)$: conditional distribution
- Fitting θ to $\hat{X}$: Maximum Likelihood Estimation

- Bayesian Setting
- z: parameters
- θ: hyper-parameters
- $p_{\theta_1}(z)$: prior distributions
- $p_\theta(\hat{X})$: evidence
- $p_{\theta_2}(x \mid z)$: likelihood
- $p_\theta(z \mid x)$: posterior distribution
- Fitting θ to $\hat{X}$: Empirical Bayes (max. evidence)

- Names and interpretations blur, but the same math applies!

Example - Latent Variable Model - Clustering





Two approaches: MLE + VI

- Let $q(x)$ be the 'true' distribution of interest and $\{p_\theta(x,z) = p_{\theta_2}(x\,|\,z) \cdot p_{\theta_1}(z)\,|\,\theta \in \Theta\}$ a (generative) latent variable model for $q(x)$.
- To compare $q(x)$ to $p_\theta(x,z)$ we have two approaches:

1.) • marginalize $p_\theta(x,z)$ to $p_\theta(x)$ and minimize: $\min_\theta \mathrm{KL}\left(q(X)\|p_\theta(X)\right)$ → MLE

2.) • introduce a variational family of inference distributions: $\{q_w(z\,|\,x)\,|\,w \in \mathscr{W}\}$ → VI
(meant to approximate $p_\theta(z\,|\,x)$) and minimize:

$$\min_{w,\theta} \mathrm{KL}\left(q_w(X,Z)\|p_\theta(X,Z)\right)$$

- Replacing expectation $\mathbb{E}_{q(X)}$ with an empirical mean $\mathbb{E}_{\hat{q}(X)}$ leads in 1st case to MLE and in the 2nd case to the learning framework of Variational Inference.

Maximum Likelihood Estimation vs Variational Inference

- By the chain rule:

$$\mathbb{E}_{q(X)}\left[\mathrm{KL}\left(q_w(Z\,|\,X)\|p_\theta(Z\,|\,X)\right)\right] + \mathrm{KL}\left(q(X)\|p_\theta(X)\right) = \mathrm{KL}\left(q_w(X,Z)\|p_\theta(X,Z)\right)$$

  we see that Variational Inference degenerates to MLE if the inference model is flexible enough to minimize the left hand term.

- We have the inequality for w, θ:

$$0 \leq \mathrm{KL}\left(q(X)\|p_\theta(X)\right) \leq \mathrm{KL}\left(q_w(X,Z)\|p_\theta(X,Z)\right)$$

  with (rhs) equality iff $q_w(z\,|\,x) = p_\theta(z\,|\,x)$ for all values x, z.

Empirical versions of the KL divergencies

- $-\mathrm{KL}\left(q(X)\|p_\theta(X)\right) - H(q(X)) =$ $= -\mathbb{E}_{q(x)}\left[\log \frac{q(x)}{p_\theta(x)}\right] - \mathbb{E}_{q(x)}\left[-\log q(x)\right]$

$= \mathbb{E}_{q(x)}\left[\log p_\theta(x)\right]$

$\overset{LLN}{\approx} \mathbb{E}_{\hat{q}(x)}\left[\log p_\theta(x)\right] = \frac{1}{N}\log p_\theta(\hat{x})$

evidence / log-likelihood

VI

- $-\mathrm{KL}\left(q_w(X,Z)\|p_\theta(X,Z)\right) - H(q(X))$ $\approx \frac{1}{N}\mathbb{E}_{q_w(z|\hat{x})}\left[\log \frac{p_\theta(z,\hat{x})}{q_w(z|\hat{x})}\right]$

evidence-lower bound (ELBO)

$= -\mathbb{E}_{q_w(z|x)q(x)}\left[\log \frac{q_w(z|x)q(x)}{p_\theta(z,x)}\right] - \mathbb{E}_{q(x)}\left[-\log q(x)\right]$

$= +\mathbb{E}_{q_w(z|x)q(x)}\left[\log \frac{p_\theta(z,x)}{q_w(z|x)}\right] - \mathbb{E}_{q_w(z|x)q(x)}\left[\log q(x)\right] + \mathbb{E}_{q(x)}\left[\log q(x)\right]$

$= \mathbb{E}_{q(x)}\,\mathbb{E}_{q_w(z|x)}\left[\log \frac{p_\theta(z,x)}{q_w(z|x)}\right] \overset{LLN}{\approx} \mathbb{E}_{\hat{q}(x)}\,\mathbb{E}_{q_w(z|x)}\left[\log \frac{p_\theta(z,x)}{q_w(z|x)}\right] = \frac{1}{N}\mathbb{E}_{q_w(z|\hat{x})}\left[\log \frac{p_\theta(z,\hat{x})}{q_w(z|\hat{x})}\right]$

Empirical versions: Log-Likelihood and Evidence Lower Bound

- Log-Likelihood: $L_{\hat{X}}(\theta) = \log p_\theta(\hat{X}) =$

- ELBO: $L_{\hat{X}}(w, \theta) =$

Fundamental Lemma of Variational Inference

- Let $p_\theta(z, x)$ be a latent variable model and $\hat{X}$ data and $q_w(z \mid x)$ be an inference model (meant to approximate $p_\theta(z \mid \hat{X})$).
- Then the evidence lower bound is a lower bound to the evidence (aka log-likelihood):

$$\log p_\theta(\hat{X}) = L_{\hat{X}}(\theta) \geq L_{\hat{X}}(w, \theta) = \mathbb{E}_{q_w(Z|\hat{X})}\left[\log \frac{p_\theta(Z, \hat{X})}{q_w(Z \mid \hat{X})}\right]$$

ELBO

- Equality holds iff $q_w(z \mid \hat{X}) = p_\theta(z \mid \hat{X})$ for all z.
- Proof: Arguments from before or directly with Jennsen's inequality.

Proof

$$\log p(\hat{x}) = \log \int \frac{p(z,\hat{x})}{q(z|\hat{x})} \cdot q(z|\hat{x})\, dz$$

$$= \log \mathbb{E}_{q(z|\hat{x})}\left[\frac{p(z,\hat{x})}{q(z|\hat{x})}\right]$$

$$\geq \mathbb{E}_{q(z|\hat{x})}\left[\log \frac{p(z,\hat{x})}{q(z|\hat{x})}\right] = \text{ELBO}$$

□

The Framework of Variational Inference (VI)

- Let $q(x)$ be an underlying 'true' distribution of interest and $\hat{X}$ data sampled from it.
  Let $\{p_\theta(x,z) \mid \theta \in \Theta\}$ be a latent variable model for $q(x)$ (or a Bayesian setting).
- Instead of maximizing the evidence/log-likelihood $\log p_\theta(\hat{X})$ w.r.t. θ,
- in Variational Inference (VI) one introduces a variational family of inference distributions $\{q_w(z \mid x) \mid w \in \mathscr{W}\}$ and Maximizes the Evidence Lower BOund (ELBO):

$$L_{\hat{X}}(w,\theta) := \mathbb{E}_{q_w(Z|\hat{X})}\left[\log \frac{p_\theta(Z,\hat{X})}{q_w(Z \mid \hat{X})}\right]$$

  w.r.t. both parameters w, θ.
- VI was justified either by the Minimal Relative Entropy Principle applied to the joint distributions plus the empirical approximation of $\mathbb{E}_{q(X)}$ via the law of large numbers, or via the Maximum Likelihood Estimation and its (log) approximation by the ELBO from below.

Machine Learning 2

Variational Inference
- Expectation-Maximization Algorithm

Patrick Forré

Expectation-Maximization Algorithm

- Let $q(x)$ be an underlying 'true' distribution of interest and $\hat{X}$ data sampled from it. Let $\{p_\theta(x,z) \mid \theta \in \Theta\}$ be a latent variable model for $q(x)$ and $\{q_w(z \mid x) \mid w \in \mathscr{W}\}$ a variational family of inference distributions.
- To maximize the ELBO (VI): $L_{\hat{X}}(w,\theta) := \mathbb{E}_{q_w(Z|\hat{X})}\left[\log \frac{p_\theta(Z,\hat{X})}{q_w(Z \mid \hat{X})}\right]$ we alternate:
  - E-step: $w^{(t+1)} := \operatorname{argmax}_w L_{\hat{x}}(w, \theta^{(t)}) + \lambda \cdot (\text{constraints})$
  - M-step: $\theta^{(t+1)} := \operatorname{argmax}_\theta L_{\hat{x}}(w^{(t+1)}, \theta) + \tilde{\lambda} \cdot (\text{constraints})$
- In practice we also need to include Lagrange multiplier in each optimization step to ensure constraints (e.g. normalization of probabilities).
- Note that the ELBO can only improve in every step, so the EM-algorithm converges.

Simplifying the steps

- E-step:

$$w^{(t+1)} := \underset{w}{\operatorname{argmax}}\, L_{\hat{x}}(w, \theta^{(t)}) = \underset{w}{\operatorname{argmax}}\, \mathbb{E}_{q_w(z|\hat{x})}\left[\log \frac{p_{\theta^{(t)}}(z, \hat{x})}{q_w(z|\hat{x})}\right]$$
$$= \underset{w}{\operatorname{argmin}}\, KL(q_w(z|\hat{x}) \,\|\, p_{\theta^{(t)}}(z|\hat{x}))$$

- M-step:

$$\theta^{(t+1)} := \underset{\theta}{\operatorname{argmax}}\, L_{\hat{x}}(w^{(t+1)}, \theta) = \underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_{q_{w^{(t+1)}}(z|\hat{x})}\left[\log \frac{p_\theta(z, \hat{x})}{q_{w^{(t+1)}}(z, \hat{x})}\right]$$
$$= \underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_{q_{w^{(t+1)}}(z|\hat{x})}\left[\log p_\theta(z, \hat{x})\right]$$

The EM-Algorithm with best possible E-steps

- E-step: M-step:
  - $w^{(t+1)} := \operatorname{argmin}_w KL(q_w(Z|\hat{X}) \| p_{\theta^{(t)}}(Z|\hat{X}))$ $\theta^{(t+1)} := \operatorname{argmax}_\theta \mathbb{E}_{q_{w^{(t+1)}}(Z|\hat{X})}[\log p_\theta(Z, \hat{X})]$
- The best possible E-step occurs when we can arrange $w^{(t+1)}$ such that:
  - $q_{w^{(t+1)}}(Z|\hat{X}) = p_{\theta^{(t)}}(Z|\hat{X})$
  - $\mathbb{E}_{q_{w^{(t+1)}}(Z|\hat{X})}[\log p_\theta(Z, \hat{X})] = \mathbb{E}_{p_{\theta^{(t)}}(Z|\hat{X})}[\log p_\theta(Z, \hat{X})]$ implying: $=: Q(\theta, \theta^{(t)})$
- So the objective in the M-step after a best possible E-step would be:
  - $\theta^{(t+1)} := \operatorname{argmax}_\theta Q(\theta, \theta^{(t)})$

The EM-Algorithm with best possible E-steps

- The EM-Algorithm in this case reduces to alternating:
  - E-step: Evaluate: $Q(\theta, \theta^{(t)}) := \mathbb{E}_{p_{\theta^{(t)}}(z|\hat{x})}[\log p_\theta(z, \hat{x})]$
  - M-step: Compute: $\theta^{(t+1)} := \arg\max_\theta Q(\theta, \theta^{(t)}) \quad + \lambda(\text{constraints})$
- In the M-step don't forget Lagrange multipliers if there are constraints.
- The M-step can sometimes be solved by putting gradients to 0 or using gradient methods for optimization.

Machine Learning 2

Variational Inference
- Expectation-Maximization Algorithm
- Example: Mixtures of Exponential Families

Patrick Forré

$q(x)$ true distr. $x^{(1)}, \dots x^{(N)}$, iid sample

$$p(z|\pi) = \prod_{k=1}^{K} \pi_k^{z_k}$$

$z_k = \mathbb{1}_k(z)$ $\qquad z \in \{1, \dots, K\}$

$z_k^{(n)} := \mathbb{1}_k(z^{(n)})$ $\qquad z^{(n)} \hookrightarrow x^{(n)}$ $\qquad \hat{X} = [x^{(1)}, \dots x^{(N)}]^T$

$$p(x|z,\eta) = \prod_{k=1}^{K} \left(\exp\left(\eta_k^T T_k(x) - A_k(\eta_k) + B_k(x) \right) \right)^{z_k}$$

$Z = (z_k^{(n)})_{\substack{n=1,\dots N \\ k=1\dots K}}$

$$p(Z, \hat{X}|\pi, \eta) \overset{iid}{=} \prod_{n=1}^{N} \prod_{k=1}^{K} \exp\left[z_k^{(n)} \cdot \left(\log \pi_k + \eta_k^T T_k(x^{(n)}) + B_k(x^{(n)}) - A_k(\eta_k) \right) \right]$$

$$\log p(Z, \hat{X}|\pi, \eta) = \sum_{n,k} z_k^{(n)} \cdot \left[\eta_k^T T_k(x^{(n)}) + B_k(x^{(n)}) - A_k(\eta_k) + \log \pi_k \right]$$

$$\log p(z, x \mid \pi, \eta) = \sum_{n,k} z_k^{(n)} \cdot \left[\eta_k^T T_k(x^{(n)}) + B_k(x^{(n)}) - A_k(\eta_k) + \log \pi_k \right]$$

E-step $\quad \mathbb{E}_{p(z \mid x, \tilde{\pi}, \tilde{\eta})} \left[\log p(z, x \mid \pi, \eta) \right] =: Q\left((\pi, \eta), (\tilde{\pi}, \tilde{\eta})\right)$

$$\gamma_k(x^{(n)}) := \mathbb{E}_{p(z \mid x, \tilde{\pi}, \tilde{\eta})} \left[z_k^{(n)} \right] = p(z^{(n)} = k \mid x^{(n)}, \tilde{\eta}, \tilde{\pi})$$

$$p(z^{(n)}, x^{(n)} \mid \tilde{\pi}, \tilde{\eta}) = \prod_{k=1}^{K} \tilde{\pi}_k^{z_k^{(n)}} \cdot \exp\left(\tilde{\eta}_k^T T_k(x^{(n)}) + B_k(x^{(n)}) - A_k(\eta_k) \right)^{z_k^{(n)}}$$

$$p(x^{(n)} \mid \tilde{\pi}, \tilde{\eta}) = \sum_{k=1}^{K} \tilde{\pi}_k \cdot \exp\left(\tilde{\eta}_k^T T_k(x^{(n)}) + B_k(x^{(n)}) - A_k(\eta_k) \right)$$

$$\rightarrow \gamma_k(x^{(n)}) = \frac{\tilde{\pi}_k \cdot \exp\left(\tilde{\eta}_k^T T_k(x^{(n)}) + B_k(x^{(n)}) - A_k(\tilde{\eta}_k) \right)}{\sum_{j=1}^{K} \tilde{\pi}_j \cdot \exp\left(\tilde{\eta}_j^T T_j(x^{(n)}) + B_j(x^{(n)}) - A_j(\tilde{\eta}_j) \right)}$$

E-step:

$$Q((\pi,\eta),(\tilde{\pi},\tilde{\eta})) := \sum_{n,k} \gamma_k(x^{(n)}) \cdot \left[\eta_k^T T_k(x^{(n)}) + B_k(x^{(n)}) - A_k(\eta_k) + \log \pi_k \right]$$

where

$$\gamma_k(x^{(n)}) := \frac{\tilde{\pi}_k \cdot \exp\left(\tilde{\eta}_k^T T_k(x^{(n)}) + B_k(x^{(n)}) - A_k(\tilde{\eta}_k)\right)}{\sum_{j=1}^{K} \tilde{\pi}_j \cdot \exp\left(\tilde{\eta}_j^T T_j(x^{(n)}) + B_j(x^{(n)}) - A_j(\tilde{\eta}_j)\right)}$$

M-step: $(\pi^*, \eta^*) = \underset{\pi, \eta}{\operatorname{argmax}} \; Q\big((\pi, \eta), (\hat{\pi}, \hat{\eta})\big)$

$$(\pi^*, \eta^*) = \underset{\pi, \eta}{\operatorname{argmax}} \sum_{n,k} \gamma_k(x^{(n)}) \cdot \Big[\eta_k^T T_k(x^{(n)}) + B_k(x^{(n)}) - A_k(\eta_k) + \log \pi_k \Big]$$

$$0 \overset{!}{=} \frac{\partial Q}{\partial \eta_k} = \sum_{n=1}^{N} \gamma_k(x^{(n)}) \cdot \Big[T_k(x^{(n)}) - \underbrace{A_k'(\eta_k)}_{\mathbb{E}[T_k(X) \mid \eta_k]} \Big]$$

$$\rightsquigarrow \quad A_k'(\eta_k) = \frac{1}{N} \sum_{n=1}^{N} \gamma_k(x^{(n)}) \cdot T_k(x^{(n)})$$

$$\rightsquigarrow \quad \eta_k = (A_k')^{-1} \Big(\frac{1}{N} \sum_{n=1}^{N} \gamma_k(x^{(n)}) \cdot T_k(x^{(n)}) \Big)$$

or use samples to approximate $\mathbb{E}[T_k(X) \mid \eta_k]$

Or $\quad \eta_k := \eta_k + \alpha \cdot \Big(\sum_{n=1}^{N} \gamma_k(x^{(n)}) \Big[T_k(x^{(n)}) - A_k'(\eta_k) \Big]$

M-step:

$$Q\left((\eta,\pi),(\tilde{\eta},\tilde{\pi})\right)$$

$$(\pi^*,\eta^*) = \underset{\pi,\eta}{\operatorname{argmax}} \sum_{n,k} \gamma_k(x^{(n)}) \cdot \left[\eta_k^T T_k(x^{(n)}) + B_k(x^{(n)}) - A_k(\eta_k) + \log \pi_k \right]$$

constraints: $\sum_{k=1}^{K} \pi_k = 1$ $\rightsquigarrow$ $\mathcal{L}(\eta,\pi) := Q(\eta,\pi) + \lambda \cdot \left(1 - \sum_{k=1}^{K} \pi_k\right)$

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

$$\Rightarrow \lambda \cdot \pi_k = \sum_{n=1}^{N} \gamma_k(x^{(n)})$$

$$\Rightarrow \lambda = \lambda \cdot 1 = \sum_{k=1}^{K} \lambda \cdot \pi_k = \sum_{k,n} \gamma_k(x^{(n)}) = \sum_{n=1}^{N} \underbrace{\sum_{k} \gamma_k(x^{(n)})}_{1} = N$$

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma_k(x^{(n)})$$

E-step : $\forall k = 1 \dots K$, $\forall n = 1 \dots N$:

$$\gamma_k^{(t+1)}(x^{(n)}) := \frac{\pi_k^{(t)} \cdot \exp\left(\eta_k^{(t)\top} T_k(x^{(n)}) + B_k(x^{(n)}) - A_k(\eta_k^{(t)})\right)}{\sum_{j=1}^{K} \pi_j^{(t)} \cdot \exp\left(\eta_j^{(t)\top} T_j(x^{(n)}) + B_j(x^{(n)}) - A_j(\eta_j^{(t)})\right)}$$

M-step $\forall k = 1 \dots K$:

$$\pi_k^{(t+1)} := \frac{1}{N} \sum_{n=1}^{N} \gamma_k^{(t+1)}(x^{(n)})$$

$$\eta_k^{(t+1)} := (A_k')^{-1}\left(\frac{1}{N} \sum_{n=1}^{N} \gamma_k^{(t+1)}(x^{(n)}) \cdot T_k(x^{(n)})\right)$$