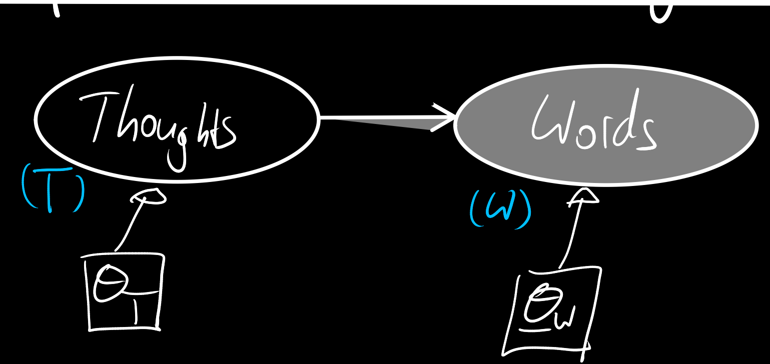
Expectation-Maximization Algorithm-Intro



$T \sim \text{Bernoulli}(\theta_T)$

$W \sim \text{Bernoulli}(\underline{\theta}_W[T])$

$T \in \{\overset{0}{\text{Bad}}, \overset{1}{\text{Good}}\}$

$W \in \{\overset{0}{\text{Bad}}, \overset{1}{\text{Good}}\}$

given Data

$$D = \{(?, 1), (?, 1), (?, 0), \dots\}$$

what are θ_T & $\underline{\theta}_W$

→ Maximum Likelihood under missing data

given Data

$$D = \{(?, w^{[0]}), (?, w^{[1]}), \dots (?, w^{[N-1]})\}$$

actually only data on the words

$$D = \{w^{[0]}, w^{[1]}, \dots, w^{[N-1]}\}$$

[MLE requires Likelihood of data
our model $p(T, W)$ but we only have W]

↳ Marginal Likelihood using $p(W)$

Marginal Likelihood

$(\underline{\theta} = [\theta_T, \underline{\theta}_W^T]^T)$

$$\mathcal{L}(D; \underline{\theta}) \overset{\text{i.i.d.}}{=} \prod_{i=0}^{N-1} p(W = w^{[i]})$$

↳ Marginal: Marginalize the joint

$$p(W) = \sum_T p(T, W)$$

$\left(\text{here: } p(W) = \sum_{t=0}^{1} p(T=t, W)\right)$

[Disclaimer: The EM does not work
for the Bern-Bern Models]

$$\mathcal{L}(D; \underline{\theta}) = \prod_{i=0}^{N-1} \sum_T p(T, W = w^{[i]})$$

Marginal Log-Likelihood

$$\ell(D; \underline{\theta}) = \log(\mathcal{L}(D; \underline{\theta}))$$

$$= \sum_{i=0}^{N-1} \log\left(\sum_T p(T, W = w^{[i]})\right)$$

problem: hard to optimize later-on

How do get the log into the 2nd sum?

[The marginal

$$\sum_T p(T, W = w^{[i]}) = \sum_T q(T) \frac{p(T, W = w^{[i]})}{q(T)}$$

"Say we have another distribution over T, $q(T)$" (→ similar to importance sampling)

$$= \mathbb{E}_{T \sim q(T)}\left[\frac{p(T, W = w^{[i]})}{q(T)}\right]$$

$$\ell(D; \underline{\theta}) = \sum_{i=0}^{N-1} \log \mathbb{E}_{T \sim q(T)}\left[\frac{p(T, W = w^{[i]})}{q(T)}\right]$$

[Jensen's inequality:

$$\log \mathbb{E}[\cdot] \leq \mathbb{E}[\log(\cdot)]$$

⇒ upper bound

okay because we maximize later on anyways]

$$\ell(D; \underline{\theta}) \leq \sum_{i=0}^{N-1} \mathbb{E}_{T \sim q(T)}\left[\log\left(\frac{p(T, W = w^{[i]})}{q(T)}\right)\right]$$

still an upper estimate

$$= \sum_{i=0}^{N-1} \sum_T q(T) \log\left(\frac{p(T, W = w^{[i]})}{q(T)}\right)$$

$$= \sum_{i=0}^{N-1} \sum_T q(T) \cdot \left(\log p(T, W = w^{[i]}) - \log q(T)\right)$$

What is $q(T)$

Maybe as the posterior

$$q(T) = p(T \mid W = w^{[i]})$$

$$\ell(D; \underline{\theta}) \leq \sum_{i=0}^{N-1} \sum_T p(T \mid W = w^{[i]}) \cdot \left(\log p(T, W = w^{[i]}) - \log p(T \mid W = w^{[i]})\right)$$

$$= \sum_{i=0}^{N-1} \sum_T p(T \mid W = w^{[i]}; \underline{\theta}) \cdot \left(\log p(T, W = w^{[i]}; \underline{\theta}) - \log p(T \mid W = w^{[i]}; \underline{\theta})\right)$$

→ chicken-egg problem:

- we need $\underline{\theta}$ to model $q(T)$
- we need $q(T)$ to maximize $\underline{\theta}$

hmm…

💡 iterative algorithm

guess $\underline{\theta}^{[0]}$:

use $\underline{\theta}^{[k]}$ to guess $q(T)$ (E-Step)

use $q(T)$ to update $\underline{\theta}$ to $\underline{\theta}^{[k+1]}$ (M-Step)

$$\ell(D; \underline{\theta}^{[k]}, \underline{\theta}) \leq \sum_{i=0}^{N-1} \sum_T p(T \mid W = w^{[i]}; \underline{\theta}^{[k]}) \cdot \left(\log p(T, W = w^{[i]}; \underline{\theta}) - \log p(T \mid W = w^{[i]}; \underline{\theta}^{[k]})\right)$$

const w.r.t. $\underline{\theta}$

$$\underline{\theta}^{[k+1]} = \underset{\underline{\theta}}{\operatorname{argmax}}\left(\ell(D; \underline{\theta}^{[k]}, \underline{\theta})\right)$$

$$Q(D; \underline{\theta}^{[k]}, \underline{\theta}) = \sum_{i=0}^{N-1} \sum_T \underbrace{p(T \mid W = w^{[i]}; \underline{\theta}^{[k]})}_{\text{responsibility for Mixture Distribution}} \cdot \log p(T, W = w^{[i]}; \underline{\theta})$$

$$\underline{\theta}^{[k+1]} = \underset{\underline{\theta}}{\operatorname{argmax}}\left(Q(D; \underline{\theta}^{[k]}, \underline{\theta})\right)$$

EM-Algorithm:

- randomly initialize $\underline{\theta}^{[0]}$
- until convergence
  - calculate responsibilities $\varrho_t^{[i]} = p(T = t \mid W = w^{[i]}; \underline{\theta}^{[k]})$
  - update parameters by Maximization

$$\underline{\theta}^{[k+1]} = \underset{\underline{\theta}}{\operatorname{argmax}}\left(\sum_{i=0}^{N-1} \sum_{t \in T} \varrho_t^{[i]} \log(p(T = t, W = w^{[i]}; \underline{\theta}))\right)$$

$\underline{\theta}^* = \lim_{k \to \infty} \underline{\theta}^{[k]}$ … the MLE under missing data