Probabilistic Machine Learning
Lecture 21
Expectation Maximization

Philipp Hennig

06 July 2020

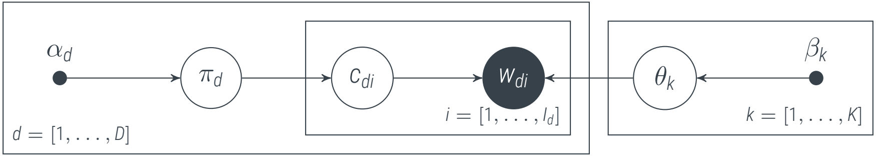
Eberhard Karls Universität Tübingen

Faculty of Science
Department of Computer Science
Chair for the Methods of Machine Learning

#	date	content	Ex	#	date	content	Ex
1	20.04.	Introduction	1	14	09.06.	Generalized Linear Models	
2	21.04.	Reasoning under Uncertainty		15	15.06.	Exponential Families	8
3	27.04.	Continuous Variables	2	16	16.06.	Graphical Models	
4	28.04.	Monte Carlo		17	22.06.	Factor Graphs	9
5	04.05.	Markov Chain Monte Carlo	3	18	23.06.	The Sum-Product Algorithm	
6	05.05.	Gaussian Distributions		19	29.06.	Example: Modelling Topics	10
7	11.05.	Parametric Regression	4	20	30.06.	Mixture Models	
8	12.05.	Learning Representations		21	06.07.	EM	11
9	18.05.	Gaussian Processes	5	22	07.07.	Variational Inference	
10	19.05.	Understanding Kernels		23	13.07.	Fast Variational Inference	12
11	26.05.	Gauss-Markov Models		24	14.07.	Kernel Topic Models	
12	25.05.	An Example for GP Regression	6	25	20.07.	Outlook	
13	08.06.	GP Classification	7	26	21.07.	Revision	

Designing a probabilistic machine learning method:

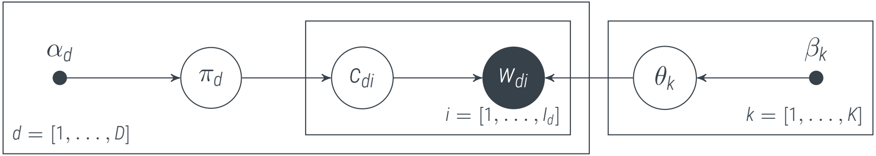
1. get the **data**
   1.1 try to collect as much meta-data as possible
2. build the **model**
   2.1 identify quantities and datastructures; assign names
   2.2 design a generative process (graphical model)
   2.3 assign (conditional) distributions to factors/arrows (use exponential families!)
3. design the **algorithm**
   3.1 consider conditional independence
   3.2 try standard methods for early experiments
   3.3 run unit-tests and sanity-checks
   3.4 identify bottlenecks, find customized approximations and refinements



To draw I_d words $w_{di} \in [1, \dots, V]$ of document $d \in [1, \dots, D]$:

- Draw K topic distributions θ_k over V words from $\quad p(\Theta \mid \boldsymbol{\beta}) = \prod_{k=1}^{K} \mathcal{D}(\theta_k; \beta_k)$
- Draw D document distributions over K topics from $\quad p(\Pi \mid \boldsymbol{\alpha}) = \prod_{d=1}^{D} \mathcal{D}(\pi_d; \alpha_d)$
- Draw topic assignments c_{ik} of word w_{di} from $\quad p(C \mid \Pi) = \prod_{i,d,k} \pi_{dk}^{c_{dik}}$
- Draw word w_{di} from $\quad p(w_{di} = v \mid c_{di}, \Theta) = \prod_k \theta_{kv}^{c_{dik}}$

Useful notation: $n_{dkv} = \#\{i : w_{di} = v, c_{ijk} = 1\}$. Write $n_{dk:} := [n_{dk1}, \dots, n_{dkV}]$ and $n_{dk\cdot} = \sum_v n_{dkv}$, etc.



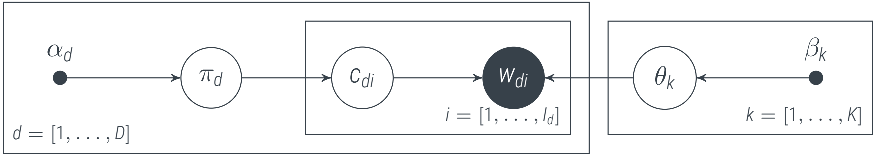
$$p(C, \Pi, \Theta, W) = \left(\prod_{d=1}^{D} \frac{\Gamma(\sum_k \alpha_{dk})}{\prod_k \Gamma(\alpha_{dk})} \prod_{k=1}^{K} \pi_{dk}^{\alpha_{dk} - 1 + n_{dk\cdot}} \right) \cdot \left(\prod_{k=1}^{K} \frac{\Gamma(\sum_v \beta_{kv})}{\prod_v \Gamma(\beta_{kv})} \prod_{v=1}^{V} \theta_{kv}^{\beta_{kv} - 1 + n_{\cdot kv}} \right)$$

If we had Π, Θ (which we don't), then the posterior $p(C \mid \Theta, \Pi, W)$ would be easy:

$$p(C \mid \Theta, \Pi, W) = \frac{p(W, C, \Theta, \Pi)}{\sum_C p(W, C, \Theta, \Pi)} = \prod_{d=1}^{D} \prod_{i=1}^{I_d} \frac{\prod_{k=1}^{K} (\pi_{dk} \theta_{kw_{di}})^{c_{dik}}}{\sum_{k'} (\pi_{dk'} \theta_{k' w_{di}})}$$

Recap from Last Lecture

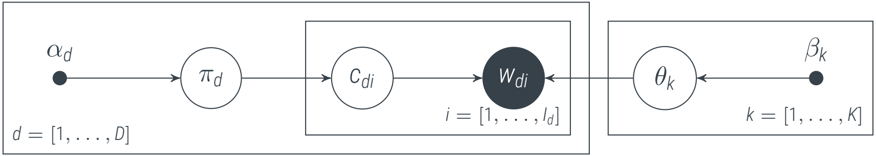
Conditional Factorization



$$p(C, \Pi, \Theta, W) = \left(\prod_{d=1}^{D} \mathcal{D}(\boldsymbol{\pi}_d; \boldsymbol{\alpha}_d)\right) \cdot \left(\prod_{d=1}^{D} \prod_{i=1}^{I_d} \left(\prod_{k=1}^{K} \pi_{dk}^{c_{dik}}\right)\right) \cdot \left(\prod_{d=1}^{D} \prod_{i=1}^{I_d} \left(\prod_{k=1}^{K} \theta_{kw_{di}}^{c_{dik}}\right)\right) \cdot \left(\prod_{k=1}^{K} \mathcal{D}(\boldsymbol{\theta}_k; \boldsymbol{\beta}_k)\right)$$

If we had Π, Θ (which we don't), then the posterior $p(C \mid \Theta, \Pi, W)$ would be easy:

$$p(C \mid \Theta, \Pi, W) = \frac{p(W, C, \Theta, \Pi)}{\sum_C p(W, C, \Theta, \Pi)} = \prod_{d=1}^{D} \prod_{i=1}^{I_d} \frac{\prod_{k=1}^{K} (\pi_{dk}\theta_{kw_{di}})^{c_{dik}}}{\sum_{k'} (\pi_{dk'}\theta_{k'w_{di}})}$$



$$p(C, \Pi, \Theta, W) = \left(\prod_{d=1}^{D} \frac{\Gamma(\sum_k \alpha_{dk})}{\prod_k \Gamma(\alpha_{dk})} \prod_{k=1}^{K} \pi_{dk}^{\alpha_{dk} - 1 + n_{dk\cdot}} \right) \cdot \left(\prod_{k=1}^{K} \frac{\Gamma(\sum_v \beta_{kv})}{\prod_v \Gamma(\beta_{kv})} \prod_{v=1}^{V} \theta_{kv}^{\beta_{kv} - 1 + n_{\cdot kv}} \right)$$

If we had C (which we don't), then the posterior $p(\Theta, \Pi \mid C, W)$ would be easy:

$$p(\Theta, \Pi \mid C, W) = \frac{p(C, W, \Pi, \Theta)}{\int p(\Theta, \Pi, C, W)\, d\Theta\, d\Pi} = \frac{\left(\prod_d \mathcal{D}(\pi_d; \alpha_d) \left(\prod_k \pi_{dk}^{n_{dk\cdot}}\right)\right) \left(\prod_k \mathcal{D}(\theta_k; \beta_k) \left(\prod_v \theta_{kv}^{n_{\cdot kv}}\right)\right)}{p(C, W)}$$

$$= \left(\prod_d \mathcal{D}(\pi_d; \alpha_{d:} + n_{d:\cdot}) \right) \left(\prod_k \mathcal{D}(\theta_k; \beta_{k:} + n_{\cdot k:}) \right)$$

The Toolbox

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Framework:

$$\int p(x_1, x_2)\, dx_2 = p(x_1) \qquad p(x_1, x_2) = p(x_1 \mid x_2)p(x_2) \qquad p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}$$

Modelling:

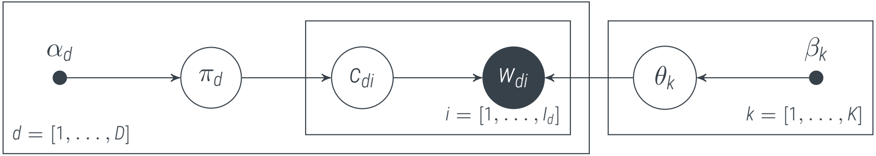
- graphical models
- Gaussian distributions
- (deep) learnt representations
- Kernels
- Markov Chains
- Exponential Families / Conjugate Priors
- Factor Graphs & Message Passing

Computation:

- Monte Carlo
- Linear algebra / Gaussian inference
- maximum likelihood / MAP
- Laplace approximations
-

Maximum Likelihood?

unfortunately, not an option



$$p(C, \Pi, \Theta, W) = \underbrace{\left(\prod_{d=1}^{D} \mathcal{D}(\boldsymbol{\pi}_d; \boldsymbol{\alpha}_d)\right)}_{p(\Pi|\boldsymbol{\alpha})} \cdot \underbrace{\left(\prod_{d=1}^{D} \prod_{i=1}^{I_d} \left(\prod_{k=1}^{K} \pi_{dk}^{c_{dik}}\right)\right)}_{p(C|\Pi)} \cdot \underbrace{\left(\prod_{d=1}^{D} \prod_{i=1}^{I_d} \left(\prod_{k=1}^{K} \theta_{kw_{di}}^{c_{dik}}\right)\right)}_{p(W|C,\Theta)} \cdot \underbrace{\left(\prod_{k=1}^{K} \mathcal{D}(\boldsymbol{\theta}_k; \boldsymbol{\beta}_k)\right)}_{p(\Theta|\boldsymbol{\beta})}$$

$$p(W \mid \Pi, \Theta) = \sum_{d,i,k} \left(\prod_{d=1}^{D} \prod_{i=1}^{I_d} \prod_{k=1}^{K} \pi_{dk} \theta_{kw_{di}}\right) \qquad \log p(W \mid \Pi, \Theta) = \log \sum (\dots) \neq \sum \log (\dots)$$

Maximizing the likelihood for Θ, Π is difficult because it does not factorize along documents or words.

"Complete Data Maximum Likelihood"

Remember that the posteriors factorize

- consider the *complete data* log likelihood

$$p(W, C \mid \Theta, \Pi) = \left(\prod_{d=1}^{D} \prod_{i=1}^{I_d} \prod_{k=1}^{K} (\pi_{dk}\theta_{kw_{di}})^{c_{dik}}\right)$$

$$\log p(W, C \mid \Theta, \Pi) = \sum_{d}^{D} \sum_{i}^{I_d} \sum_{k}^{K} c_{dik}(\log \pi_{dk} + \log \theta_{kw_{di}})$$

- maximize wrt. Π, introduce Lagrange multiplier to ensure $\sum_k \pi_{dk} = 1$

$$\frac{\partial}{\partial \pi_{e\ell}} \left(\log p(W, C \mid \Theta, \Pi)] + \lambda_e \left(\sum_{k'} \pi_{ek} - 1\right)\right) = \frac{1}{\pi_{e\ell}} \sum_{i} c_{ei\ell} + \lambda_e \stackrel{!}{=} 0$$

$$\Rightarrow \qquad \pi_{dk} = \frac{c_{d\cdot k}}{c_{d\cdot\cdot}}$$

“Complete Data Maximum Likelihood”

Remember that the posteriors factorize

- consider the *complete data* log likelihood

$$p(W, C \mid \Theta, \Pi) = \left(\prod_{d=1}^{D}\prod_{i=1}^{I_d}\prod_{k=1}^{K}(\pi_{dk}\theta_{kw_{di}})^{c_{dik}}\right)$$

$$\log p(W, C \mid \Theta, \Pi) = \sum_{d}^{D}\sum_{i}^{I_d}\sum_{k}^{K} c_{dik}(\log \pi_{dk} + \log \theta_{kw_{di}})$$

- maximize wrt. Θ, introduce Lagrange multiplier to ensure $\sum_v \theta_{dv} = 1$

$$\frac{\partial}{\partial \theta_{\ell v}}\left(\log p(W, C \mid \Theta, \Pi)] + \lambda_\ell \left(\sum_{v'} \theta_{\ell v} - 1\right)\right) = \frac{1}{\theta_{\ell v}}\sum_d \sum_i c_{ei\ell} + \lambda_v \stackrel{!}{=} 0$$

$$\Rightarrow \qquad \theta_{kv} = \frac{n_{\cdot kv}}{n_{\cdot k\cdot}}$$

(remember $n_{dkv} = \#\{i : w_{di} = v, c_{ijk} = 1\}$. Write $n_{dk:} := [n_{dk1}, \dots, n_{dkV}]$ and $n_{dk\cdot} = \sum_v n_{dkv}$)

"Complete Data Maximum Likelihood"

Remember that the posteriors factorize

- consider the *complete data* log likelihood

$$p(W, C \mid \Theta, \Pi) = \left(\prod_{d=1}^{D} \prod_{i=1}^{I_d} \prod_{k=1}^{K} (\pi_{dk} \theta_{kw_{di}})^{c_{dik}} \right)$$

$$\log p(W, C \mid \Theta, \Pi) = \sum_{d}^{D} \sum_{i}^{I_d} \sum_{k}^{K} c_{dik} (\log \pi_{dk} + \log \theta_{kw_{di}})$$

- to maximize wrt. C, simply set

$$c_{dik} = \begin{cases} 1 & \text{if } k = \arg\max_{k'} (\log \pi_{dk'} + \log \theta_{k' w_{di}}) \\ 0 & \text{else} \end{cases}$$

Note again that

$$p(C \mid \Theta, \Pi, W) = \frac{p(W, C, \Theta, \Pi)}{\sum_C p(W, C, \Theta, \Pi)} = \prod_{d=1}^{D} \prod_{i=1}^{I_d} \frac{\prod_{k=1}^{K} (\pi_{dk} \theta_{kw_{di}})^{c_{dik}}}{\sum_{k'} (\pi_{dk'} \theta_{k' w_{di}})}$$

$$= \prod_{d=1}^{D} \prod_{i=1}^{I_d} \prod_{k=1}^{K} \tilde{\gamma}_{dik}^{c_{dik}} \quad \text{where} \quad \gamma_{dik} := \pi_{dk} \theta_{kw_{di}} \quad \text{and} \quad \tilde{\gamma}_{dik} := \gamma_{dik} / \sum_{k'} \gamma_{dik'}$$

with $\tilde{\gamma}$, we can compute the *expected* (complete data) log likelihood

$$p(W, C \mid \Theta, \Pi) = \left(\prod_{d=1}^{D} \prod_{i=1}^{I_d} \prod_{k=1}^{K} (\pi_{dk} \theta_{kw_{di}})^{c_{dik}} \right)$$

$$\log p(W, C \mid \Theta, \Pi) = \sum_{d}^{D} \sum_{i}^{I_d} \sum_{k}^{K} c_{dik} (\log \pi_{dk} + \log \theta_{kw_{di}}) = \sum_{d} \sum_{k} n_{dk\cdot} \log \pi_{dk} + \sum_{k} \sum_{v} n_{\cdot kv} \log \theta_{kv}$$

- Compute the *Expected* complete log likelihood

$$
\begin{aligned}
\mathbb{E}_{p(C|\gamma)}[\log p(W, C \mid \Theta, \Pi)] &= \sum_C^{} \sum_d^D \sum_i^{I_d} \sum_k^K \tilde{\gamma}_{dik} c_{dik} (\log \pi_{dk} + \log \theta_{kw_{di}}) \\
&= \sum_d^D \sum_i^{I_d} \sum_k^K \tilde{\gamma}_{dik} (\log \pi_{dk} + \log \theta_{kw_{di}})
\end{aligned}
$$

- maximize wrt. Π, introduce Lagrange multiplier to ensure $\sum_k \pi_{dk} = 1$

$$
\frac{\partial}{\partial \pi_{e\ell}} \left(\mathbb{E}_{p(C|\gamma)}[\log p(W, C \mid \Theta, \Pi)] + \lambda_e \left(\sum_{k'} \pi_{ek} - 1 \right) \right) = \frac{1}{\pi_{e\ell}} \sum_i \tilde{\gamma}_{ei\ell} + \lambda_e \stackrel{!}{=} 0
$$

$$
\Rightarrow \qquad \pi_{dk} = \frac{\tilde{\gamma}_{d \cdot k}}{\sum_{k'} \tilde{\gamma}_{d \cdot k'}}
$$

- Compute the *Expected* complete log likelihood

$$\begin{aligned}\mathbb{E}_{p(C|\gamma)}[\log p(W, C \mid \Theta, \Pi)] &= \sum_C \sum_d^D \sum_i^{I_d} \sum_k^K \tilde{\gamma}_{dik} c_{dik} (\log \pi_{dk} + \log \theta_{kw_{di}}) \\ &= \sum_d^D \sum_i^{I_d} \sum_k^K \tilde{\gamma}_{dik} (\log \pi_{dk} + \log \theta_{kw_{di}})\end{aligned}$$

- *Maximize* wrt. Θ, introduce Lagrance multiplier to ensure $\sum_v \theta_{kv} = 1$

$$\frac{\partial}{\partial \theta_{\ell v}} \left(\mathbb{E}_{p(C|\gamma)}[\log p(W, C \mid \Theta, \Pi)] + \lambda_\ell \left(\sum_{v'} \theta_{\ell v'} - 1 \right) \right) = \frac{1}{\theta_{\ell v}} \sum_d \sum_i \tilde{\gamma}_{di\ell} + \lambda_\ell \stackrel{!}{=} 0$$

$$\Rightarrow \qquad \theta_{kv} = \frac{\sum_{d,i} \mathbb{I}(w_{di} = v) \tilde{\gamma}_{dik}}{\sum_{v'} \sum_{d,i} \mathbb{I}(w_{di} = v') \tilde{\gamma}_{dik}}$$

The EM algorithm

Goal: maximize the likelihood $p(x \mid \theta)$ wrt. parameters θ. Identify a latent variable z such that the *complete (data) likelihood* $p(x, z \mid \theta)$ has convenient structure. Then, instead of trying to maximize

$$\log p(x \mid \theta) = \log \sum_z p(x, z \mid \theta),$$

iterate between computing the ***Expected*** complete likelihood and ***Maximizing*** it:

$$\mathbb{E}_z \log p(x, z \mid \theta) = \sum_z p(z \mid x, \theta) \log p(x, z \mid \theta),$$

Why is this a good idea?

The EM algorithm

Goal: maximize the likelihood $p(x \mid \theta)$ wrt. parameters θ. Identify a latent variable z such that the *complete (data) likelihood* $p(x, z \mid \theta)$ has convenient structure. Then, instead of trying to maximize

$$\log p(x \mid \theta) = \log \sum_z p(x, z \mid \theta),$$

iterate between computing the ***Expected*** complete likelihood and ***Maximizing*** it:

$$\mathbb{E}_z \log p(x, z \mid \theta) = \sum_z p(z \mid x, \theta) \log p(x, z \mid \theta),$$

Why is this a good idea?

An observation: By Jensen's inequality (log is concave!)

$$\sum_z q(z) \log p(x, z \mid \theta) + \mathbb{H}(q) \leq \log \sum_z p(x, z \mid \theta)$$

An Observation

maximizing a lower bound

- We constructed an approximate distribution $q(z) = p(z \mid x, \theta)$ (in the concrete case, $q(C) = p(C \mid W, \Theta, \Pi)$ for our latent quantity.
- For *any* such approximation $q(z)$:

$$\begin{aligned}
\log p(x \mid \theta) &= \log \int p(x, z \mid \theta)\, dz \\
&= \log \int q(z) \frac{p(x, z \mid \theta)}{q(z)}\, dz \\
&\geq \int q(z) \log \frac{p(x, z \mid \theta)}{q(z)}\, dz =: \mathcal{L}(q)
\end{aligned}$$

Theorem (Jensen's inequality (Jensen,1906))

Let (Ω, A, μ) be a probability space, g be a real-valued, μ-integrable function and ϕ be a convex function on the real line. Then

$$\phi\left(\int_\Omega g\, d\mu\right) \leq \int_\Omega \phi \circ g\, d\mu.$$

An Observation

maximizing a lower bound

- We constructed an approximate distribution $q(z) = p(z \mid x, \theta)$ (in the concrete case, $q(C) = p(C \mid W, \Theta, \Pi)$ for our latent quantity.
- For *any* such approximation $q(z)$:

$$\begin{aligned}\log p(x \mid \theta) &= \log \int p(x, z \mid \theta)\, dz \\ &= \log \int q(z) \frac{p(x, z \mid \theta)}{q(z)}\, dz \\ &\geq \int q(z) \log \frac{p(x, z \mid \theta)}{q(z)}\, dz =: \mathcal{L}(q)\end{aligned}$$

- Thus, by maximizing the RHS in θ in the M-step, we increase a lower bound on the LHS (the target quantity)
- But can we be sure that this increases the LHS?
- To show that this is the case, we will now establish that the E-step makes the bound *tight* at the local θ.

An Insight

Distance of our bound

$$\begin{aligned}\mathcal{L}(q) &= \int q(z) \log \frac{p(x,z \mid \theta)}{q(z)}\, dz = \int q(z) \log \frac{p(z \mid x,\theta) \cdot p(x \mid \theta)}{q(z)}\, dz \\ &= \int q(z) \log \frac{p(z \mid x,\theta)}{q(z)}\, dz + \log p(x \mid \theta) \int q(z)\, dz \\ \text{thus } \log p(x \mid \theta) &= \mathcal{L}(q) - \int q(z) \log \frac{p(z \mid x,\theta)}{q(z)} = \mathcal{L}(q) + D_{\mathrm{KL}}(q \| p(z \mid x,\theta))\end{aligned}$$

The Kullback-Leibler divergence satisfies

- $D_{\mathrm{KL}}(q \| p) \geq 0$
- $D_{\mathrm{KL}}(q \| p) = 0 \qquad \Leftrightarrow q \equiv p$

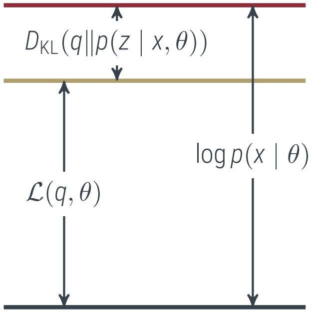
EM maximizes the ELBO / minimizes Free Energy

a more general view

$$\log p(x \mid \theta) = \mathcal{L}(q, \theta) + D_{\mathrm{KL}}(q \| p(z \mid x, \theta))$$

$$\mathcal{L}(q, \theta) = \int q(z) \log \left(\frac{p(x, z \mid \theta)}{q(z)} \right) dz$$

$$D_{\mathrm{KL}}(q \| p(z \mid x, \theta)) = - \int q(z) \log \left(\frac{p(z \mid x, \theta)}{q(z)} \right) dz$$



EM maximizes the ELBO / minimizes Free Energy

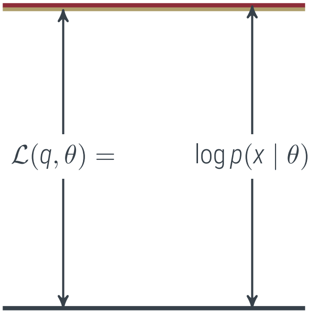
a more general view

$$\log p(x \mid \theta) = \mathcal{L}(q, \theta) + D_{\mathrm{KL}}(q \| p(z \mid x, \theta))$$

$$\mathcal{L}(q, \theta) = \int q(z) \log \left(\frac{p(x, z \mid \theta)}{q(z)} \right) dz$$

$$D_{\mathrm{KL}}(q \| p(z \mid x, \theta)) = - \int q(z) \log \left(\frac{p(z \mid x, \theta)}{q(z)} \right) dz$$

E -step: $q(z) = p(z \mid x, \theta_{\mathrm{old}})$, thus $D_{\mathrm{KL}}(q \| p(z \mid x, \theta_i)) = 0$



$$\log p(x \mid \theta) = \mathcal{L}(q, \theta) + D_{\mathrm{KL}}(q \| p(z \mid x, \theta))$$

$$\mathcal{L}(q, \theta) = \int q(z) \log \left(\frac{p(x, z \mid \theta)}{q(z)} \right) dz$$

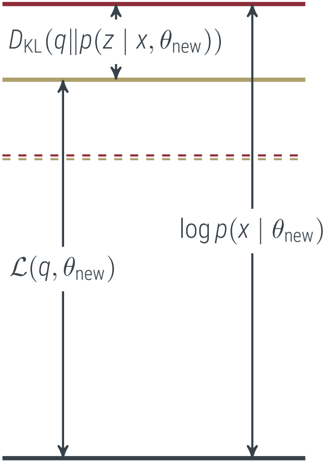
$$D_{\mathrm{KL}}(q \| p(z \mid x, \theta)) = - \int q(z) \log \left(\frac{p(z \mid x, \theta)}{q(z)} \right) dz$$

E -step: $q(z) = p(z \mid x, \theta_{\mathrm{old}})$, thus $D_{\mathrm{KL}}(q \| p(z \mid x, \theta_i)) = 0$

M -step: **Maximize ELBO**

$$\theta_{\mathrm{new}} = \arg\max_{\theta} \int q(z) \log p(x, z \mid \theta)\, dz$$

$$= \arg\max_{\theta} \mathcal{L}(q, \theta) + \int q(z) \log q(z)\, dz$$



EM Algorithm — General Form

for further generalization

Setting:

- Want to find *maximum likelihood* (or MAP) estimate for a model involving a **latent** variable

$$\theta_* = \arg\max_\theta \left[\log p(x \mid \theta)\right] = \arg\max_\theta \left[\log\left(\int p(x, z \mid \theta)\, dz\right)\right]$$

- Assume that the summation inside the log makes analytic optimization intractable
- but that optimization would be analytic if z was known (i.e. if there were only one term in the sum)

Idea: Initialize θ_0, then iterate between

1. Compute $q(z) = p(z \mid x, \theta_{\text{old}})$, thereby setting $D_{\text{KL}}(q \| p(z \mid x, \theta)) = 0$
2. Set θ_{new} to the **Maximize** the **Expectation Lower Bound**

$$\theta_{\text{new}} = \arg\max_\theta \mathcal{L}(q, \theta) = \arg\max_\theta \int q(z) \log\left(\frac{p(x, z \mid \theta)}{q(z)}\right) dz$$

3. Check for convergence of either the log likelihood, or θ.

Some Observations

for future reference

- If $p(x, z \mid \theta)$ is an **exponential family** with θ as the natural parameters, then

$$\begin{aligned} p(x,z) &= \exp(\phi(x,z)^\intercal \theta - \log Z(\theta)) \\ \mathcal{L}(q(z),\theta) &= \mathbb{E}_{q(z)}(\phi(x,z)^\intercal\theta - \log Z(\theta)) = \mathbb{E}_{q(z)}[\phi(x,z)]^\intercal\theta - \log Z(\theta) \\ \nabla_\theta \mathcal{L}(q(z),\theta) = 0 \quad &\Rightarrow \quad \nabla_\theta \log Z(\theta) = \mathbb{E}_{p(x,z)}[\phi(x,z)] = \mathbb{E}_{q(z)}[\phi(x,z)] \end{aligned}$$

  and optimization may be analytic (example above).
- it is straightforward to extend EM to maximize a **posterior** instead of a likelihood (just add a log prior for θ)
- When we set $q(z) = p(z \mid x, \theta_{\text{old}})$, we set D_{KL} to its **minimum** $D_{\text{KL}}(q \| p(z \mid x, \theta)) = 0$, thus

$$\begin{aligned} \nabla_\theta \log p(x \mid \theta_{\text{old}}) &= \nabla_\theta \mathcal{L}(q, \theta_{\text{old}}) + \nabla_\theta D_{\text{KL}}(q \| p(z \mid x, \theta_{\text{old}})) \\ &= \nabla_\theta \mathcal{L}(q, \theta_{\text{old}}) \end{aligned}$$

  So we could also use an optimizer based on this gradient to **numerically** optimize $\mathcal{L}$.
  This is known as **generalized EM**

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

The EM algorithm:

- to find *maximum likelihood* (or MAP) estimate for a model involving a **latent** variable

$$\theta_* = \arg\max_\theta \left[\log p(x \mid \theta)\right] = \arg\max_\theta \left[\log \left(\int p(x, z \mid \theta)\, dz\right)\right]$$

- Initialize θ_0, then iterate between

E Compute $p(z \mid x, \theta_{\text{old}})$, thereby setting $D_{\text{KL}}(q \| p(z \mid x, \theta) = 0$

M Set θ_{new} to the **Maximize** the **Expectation Lower Bound**

$$\theta_{\text{new}} = \arg\max_\theta \mathcal{L}(q, \theta) = \arg\max_\theta \int q(z) \log \left(\frac{p(x, z \mid \theta)}{q(z)}\right) dz$$

- Check for convergence of either the log likelihood, or θ.

Next time: Can we make Π, Θ part of q?

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

The EM algorithm:

- to find *maximum likelihood* (or MAP) estimate for a model involving a **latent** variable

$$\theta_* = \arg\max_\theta \left[\log p(x \mid \theta)\right] = \arg\max_\theta \left[\log \left(\int p(x, z \mid \theta)\, dz \right)\right]$$

- Initialize θ_0, then iterate between

E Compute $p(z \mid x, \theta_{\text{old}})$, thereby setting $D_{\text{KL}}(q \| p(z \mid x, \theta) = 0$

M Set θ_{new} to the **Maximize** the **Expectation Lower Bound / minimize the Variational Free Energy**

$$\theta_{\text{new}} = \arg\max_\theta \mathcal{L}(q, \theta) = \arg\max_\theta \int q(z) \log \left(\frac{p(x, z \mid \theta)}{q(z)} \right) dz$$

- Check for convergence of either the log likelihood, or θ.

Next time: Can we make Π, Θ part of q?