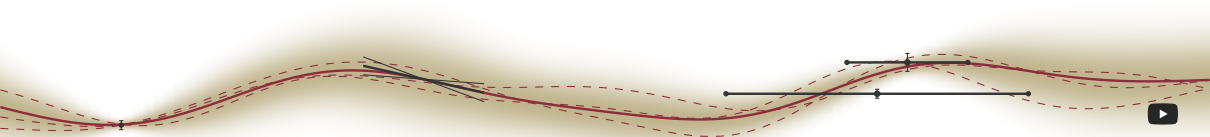
Probabilistic Machine Learning
Lecture 22
Variational Inference

Philipp Hennig

07 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

The Toolbox

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
- EM / variational approximations

Recap: The EM Algorithm — General Form

from last lecture

Setting:

- Want to find *maximum likelihood* (or MAP) estimate for a model $p(x \mid \theta)$, and happen to know a **latent** variable z that factorizes $p(x, z \mid \theta)$

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

Idea: Initialize θ_0, then iterate between

E Compute the approximation $q(z) = p(z \mid x, \theta_{\text{old}})$, which sets $D_{\text{KL}}(q \| p(z \mid x, \theta) = 0$ and maximizes the **Expectation Lower Bound**

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

M **Maximize** the bound (rather than the log likelihood itself in θ to

$$\theta_{\text{new}} = \arg\max_\theta \mathcal{L}(q, \theta)$$

Lemma

Consider the probability distribution $p(x, z)$ and an arbitrary probability distribution $q(z)$ such that $q(z) > 0$ whenever $p(z) = \sum_x p(x, z) > 0$. Then the following equality holds:

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

$$\textit{where} \quad \mathcal{L}(q) := \int q(z) \log \left(\frac{p(x, z)}{q(z)} \right) dz \quad \textit{and} \quad D_{KL}(q \| p) := - \int q(z) \log \left(\frac{p(z \mid x)}{q(z)} \right) dz.$$

- $-\mathcal{L}(q)$ is known as the **Variational Free Energy** in physics, because

$$-\mathcal{L}(q) = -\mathbb{E}_q(\log p(x, z)) - \mathbb{H}(q) \quad \text{cf.} \quad F = U - TS$$

- note that $D_{KL}(q \| p) \geq 0$ with "$=$" iff $p \equiv q$. Thus $\mathcal{L}(q) \leq p(x)$, and it is also known as the **Expectation Lower Bound (ELBO)**

Historical Side-Note

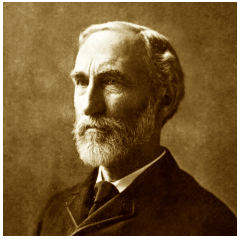
Machine Learning is the application of scientific modelling to *everything*

Hermann v. Helmholtz
1821–1894

image:L. Meder

"Energy"

$$\mathcal{F} = U - TS$$



Josia W. Gibbs
1839–1903

image:unknown

"Enthalpy"

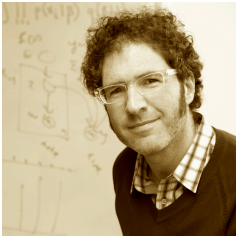
$$\mathcal{H} = U + pV$$

Ludwig Boltzmann
1844–1906

image:wikipedia

"Entropy"

$$\mathcal{G} = H - TS$$



David M. Blei

image:Columbia U

"ELBO"

$$\mathcal{L} = \mathbb{E}_q(\log p(x,z)) + \mathrm{H}(q)$$

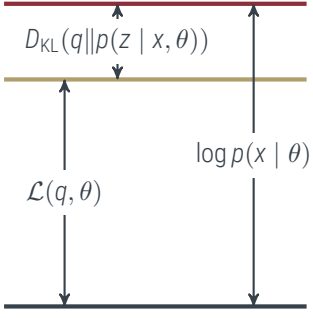
EM maximizes the ELBO / minimizes Free Energy

a more general view

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

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

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



EM maximizes the ELBO / minimizes Free Energy

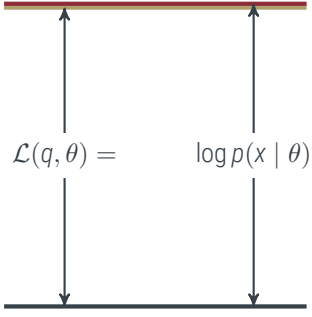
a more general view

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

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

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

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



EM maximizes the ELBO / minimizes Free Energy

a more general view

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

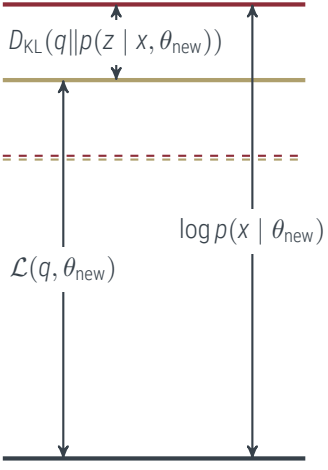
$$\mathcal{L}(q, \theta) = \sum_z q(z) \log \left(\frac{p(x, z \mid \theta)}{q(z)} \right)$$

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

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 / minimize Free Energy**

$$\begin{aligned} \theta_{\mathrm{new}} &= \arg\max_\theta \sum_z q(z) \log p(x, z \mid \theta) \\ &= \arg\max_\theta \mathcal{L}(q, \theta) + \sum_z q(z) \log q(z) \end{aligned}$$



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

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

- For EM, we minimized KL-divergence to find $q = p(z \mid x, \theta)$ (E), then maximized $\mathcal{L}(q, \theta)$ in θ.
- What if we treated the parameters θ as a *probabilistic* variable for full Bayesian inference?

$$z \leftarrow z \cup \theta$$

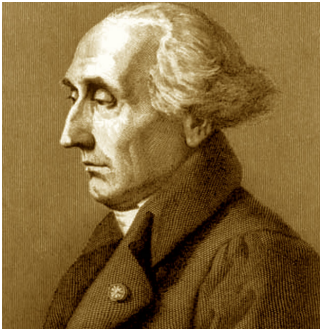
- Then we could just maximize $\mathcal{L}(q(z))$ wrt. q (not z!) to implicitly minimize $D_{\mathrm{KL}}(q \| p(z \mid x))$, because $\log p(x)$ is constant. This is an **optimization in the space of distributions** q, not (necessarily) in parameters of such distributions, and thus a very powerful notion.
- In general, this will be intractable, because the optimal choice for q is exactly $p(z \mid x)$. But maybe we can help out a bit with approximations. Amazingly, we often don't need to impose strong approximations. Sometimes we can get away with just imposing restrictions on the **factorization** of q, not its analytic form.

The Calculus of Variations

One of the big ideas they don't teach you in school

Feynman image: Nobel foundation

Leonhard Euler
1707–1783



Joseph-Louis Lagrange
1736–1813

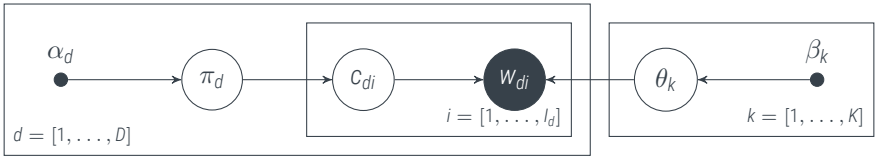
Richard P. Feynman
1918–1988

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

Can we maximize $\mathcal{L}$ w.r.t. q, *without* explicitly parametrising q?

A Motivation

what, exactly makes inference hard?



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

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

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

Factorizing Approximations

A surprisingly subtle approximation with strong implications

- in general, maximizing $\mathcal{L}(q)$ wrt. $q(z)$ is hard, because the extremum is exactly at $q(z) = p(z \mid x)$
- but let's assume that $q(z)$ **factorizes**

$$q(z) = \prod_i q_i(z_i) = \prod_i^n q_i$$

- then the bound simplifies. Let's focus on one particular variable z_j:

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_i^n q_i \left(\log p(x, z) - \sum_i \log q_i \right) dz \\
&= \int q_j \left(\int \log p(x, z) \prod_{i \neq j} q_i \, dz_i \right) dz_j - \int q_j \log q_j dz_j + \text{const.} \\
&= \int q_j \log \tilde{p}(x, z_j) \, dz_j - \int q_j \log q_j \, dz_j + \text{const.} \\
\text{where } \log \tilde{p}(x, z_j) &= \mathbb{E}_{q, i \neq j}[\log p(x, z)] + \text{const.}
\end{aligned}$$

- Iteratively compute

$$\begin{aligned}\mathcal{L}(q) &= \int q_j \log \tilde{p}(x, z_j)\, dz_j - \int q_j \log q_j\, dz_j + \text{const.} \\ &= -D_{\mathrm{KL}}(q_j(z) \| \tilde{p}(x, z_j)) + \text{const.} = -D_{\mathrm{KL}}(q_j(z) \| \tilde{p}(z_j \mid x)) + \text{const.}\end{aligned}$$

  and maximize wrt. q_j. Doing so *minimizes* $D_{\mathrm{KL}}(q(z_j) \| \tilde{p}(x, z_j))$, thus the minimum is at q_j^* with

$$\log q_j^*(z_j) = \log \tilde{p}(x, z_j) + \text{const.} = \mathbb{E}_{q, i \neq j}(\log p(x, z)) + \text{const.} \qquad (\star)$$

- note that this expression identifies a **function** q_j, not some parametric form.
- the optimization converges, because $-\mathcal{L}(q)$ can be shown to be *convex* wrt. q.

In physics, this trick is known as **mean field theory** (because an n-body problem is separated into n separate problems of individual particles who are affected by the "mean field" $\tilde{p}$ summarizing the expected effect of all other particles).

Variational Approximations

- arise from imposing *algebraic* (e.g. factorisation, not *necessarily* parametric) constraints on $q(z)$ and then maximizing the ELBO

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

- (As we will see below), it is often enough to just impose factorisation, not explicit forms for the approximation
- maximizing $\mathcal{L}$ is equivalent to minimizing

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

What kind of approximations does this yield?

Recap: Kullback-Leibler Divergence

from Lecture 14

Definition (Kullback-Leibler divergence)

Let P and Q be probability distributions over $\mathbb{X}$ with pdf's $p(x)$ and $q(x)$, respectively. The **KL-divergence from** Q *to* P is defined as

$$D_{\mathrm{KL}}(P\|Q) := \int \log\left(\frac{p(x)}{q(x)}\right) dp(x)$$

(I will often write $D_{\mathrm{KL}}(p\|q)$ instead)

Some properties:

- $D_{\mathrm{KL}}(P\|Q) \neq D_{KL}(Q\|P)$
- $D_{\mathrm{KL}}(P\|Q) \geq 0, \forall P, Q$ (**Gibbs' inequality**), and
- $D_{\mathrm{KL}}(P\|Q) = 0 \Leftrightarrow p \equiv q$ almost everywhere

Solomon Kullback (1907–1994)

Richard Leibler (1914–2003)

KL Divergence — Pictoral View

$D_{\mathrm{KL}}(q\|p)$ is *zero-enforcing*, $D_{\mathrm{KL}}(p\|q)$ is *nonzero-enforcing*

[images from Bishop, PRML, 2006, Fig. 10.3]

- $D_{\mathrm{KL}}(p\|q) = -\int p(z) \log \left(\frac{q(z)}{p(z)} \right) dz$ is **large** if $q(z) \approx 0$ where $p(z) \gg 0$
- $D_{\mathrm{KL}}(q\|p) = -\int q(z) \log \left(\frac{p(z)}{q(z)} \right) dz$ is **large** if $q(z) \gg 0$ where $p(z) \approx 0$

Variational Approximations

- arise from imposing *algebraic* (e.g. factorisation, not *necessarily* parametric) constraints on $q(z)$ and then maximizing the ELBO

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

- (As we will see below), it is often enough to just impose factorisation, not explicit forms for the approximation
- maximizing $\mathcal{L}$ is equivalent to minimizing

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

- Variational Approximations are *zero-enforcing*, and thus often "tighter" or "narrower" than the true posterior. But they are not local, like Laplace approximations

Why do we even use exponential families if we can build abstract variational bounds?

Exponential Family Approximations

The connection to EM is not accidental

- What has happened here? Why the connection to EM?
- Consider an **exponential family** joint distribution

$$p(x, z \mid \eta) = \prod_{n=1}^{N} \exp\left(\eta^\intercal \phi(x_n, z_n) - \log Z(\eta)\right)$$

$$\text{with conjugate prior } p(\eta \mid \nu, v) = \exp\left(\eta^\intercal v - \nu \log Z(\eta) - \log F(\nu, v)\right)$$

- and assume $q(z, \eta) = q(z) \cdot q(\eta)$. Then q is in the same exponential family, with

$$\log q^*(z) = \mathbb{E}_{q(\eta)}(\log p(x, z \mid \eta)) + \text{const.} = \sum_{n=1}^{N} \mathbb{E}_{q(\eta)}(\eta)^\intercal \phi(x_n, z_n)$$

$$q^*(z) = \prod_{n=1} \exp\left(\mathbb{E}(\eta)^\intercal \phi(x_n, z_n) - \log Z(\mathbb{E}(\eta))\right) \quad \text{(note induced factorization)}$$

Exponential Family Approximations

The connection to EM is not accidental

- What has happened here? Why the connection to EM?
- Consider an **exponential family** joint distribution

$$p(x, z \mid \eta) = \prod_{n=1}^{N} \exp\left(\eta^\intercal \phi(x_n, z_n) - \log Z(\eta)\right)$$

$$\text{with conjugate prior } p(\eta \mid \nu, v) = \exp\left(\eta^\intercal v - \nu \log Z(\eta) - \log F(\nu, v)\right)$$

- and assume $q(z, \eta) = q(z) \cdot q(\eta)$. Then q is in the same exponential family, with

$$\begin{aligned}\log q^*(\eta) &= \log p(\eta \mid \nu, v) + \mathbb{E}_z(\log p(x, z \mid \eta)) + \text{const.} \\ &= -\nu \log Z(\eta) + \eta^\intercal v + \sum_{n=1}^{N} -\log Z(\eta) + \eta^\intercal \mathbb{E}_z(\phi(x_n, z_n)) + \text{const.}\end{aligned}$$

$$q^*(\eta) = \exp\left(\eta^\intercal \left(v + \sum_{n=1}^{N} \mathbb{E}_z(\phi(x_n, z_n))\right) - (\nu + N) \log Z(\eta) - \text{const.}\right)$$

Variational Inference

- is a general framework to construct approximating **probability distributions** $q(z)$ to non-analytic posterior distributions $p(z \mid x)$ by minimizing the **functional**

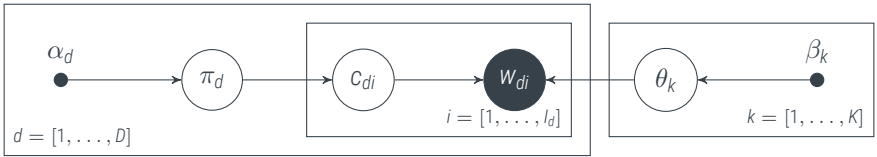
$$q^* = \arg\min_{q \in \mathcal{Q}} D_{KL}(q(z) \| p(z \mid x)) = \arg\max_{q \in \mathcal{Q}} \mathcal{L}(q)$$

- the beauty is that we get to *choose* q, so one can nearly always find a tractable approximation.
- If we impose the *mean field approximation* $q(z) = \prod_i q(z_i)$, get

$$\log q_j^*(z_j) = \mathbb{E}_{q, i \neq j}(\log p(x, z)) + \text{const}..$$

- for Exponential Family p things are particularly simple: we only need the expectation under q of the sufficient statistics.

Variational Inference is an extremely flexible and powerful approximation method. Its downside is that constructing the bound and update equations can be tedious. For a quick test, variational inference is often not a good idea. But for a deployed product, it can be the most powerful tool in the box.



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

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

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

A Variational Bound

Find the *best simple* way to describe a *complex* thing

reminder: $n_{dkv} = \#\{i : w_{di} = v, c_{ijk} = 1\}$. $n_{dk\cdot} = \sum_v n_{dkv}$, etc.

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

- The posterior $p(\Pi, \Theta, C \mid W)$ is intractable. We want an approximation q that *factorises*

$$q(\Pi, \Theta, C) = q(C) \cdot q(\Pi, \Theta)$$

- To find the *best* such approximation – the one that *minimizes* $D_{\mathrm{KL}}(q \| p(\Pi, \Theta, C \mid W))$, we *maximize* the **ELBO** (minimize variational free energy)

$$\mathcal{L}(q) = \int q(C, \Theta, \Pi) \log\left(\frac{p(C, \Pi, \Theta, W)}{q(C, \Theta, \Pi)}\right) dC\, d\Theta\, d\Pi$$

Constructing the Bound

reminder: $n_{dkv} = \#\{i : w_{di} = v, c_{dik} = 1\}$. $n_{dk\cdot} = \sum_v n_{dkv}$, etc.

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

Recall from above: To maximize the ELBO of a factorized approximation, compute the **mean field**

$$\log q^*(z_i) = \mathbb{E}_{z_j, j\neq i}(\log p(x, z)) + \text{const.}$$

$$\log q^*(C) = \mathbb{E}_{q(\Pi,\Theta)}\left(\sum_{d,i,k} \log(\pi_{dk}\theta_{kw_{di}})\right) + \text{const.} \quad = \sum_{d,i}\sum_{k=1}^{K} c_{dik} \underbrace{\left(\mathbb{E}_{q(\Pi,\Theta)}(\log \pi_{dk}\theta_{dw_{di}})\right)}_{=:\log \gamma_{dik}} + \text{const.}$$

Thus, $q(C) = \prod_{d,i} q(\boldsymbol{c}_{di})$ with $q(\boldsymbol{c}_{di}) = \prod_k \tilde{\gamma}_{dik}^{c_{dik}}$, where $\tilde{\gamma}_{dik} = \gamma_{dik} / \sum_k \gamma_{dik}$
(Note: Thus, $\mathbb{E}_q(c_{dik}) = \tilde{\gamma}_{dik}$)

Constructing the Bound

reminder: $n_{dkv} = \#\{i : w_{di} = v, c_{dik} = 1\}$. $n_{dk\cdot} = \sum_v n_{dkv}$, etc.

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

Recall from above: To maximize the ELBO of a factorized approximation, compute the **mean field**

$$\log q^*(z_i) = \mathbb{E}_{z_{j\neq i}}(\log p(x, z)) + \text{const.}$$

$$\log q^*(\Pi, \Theta) = \mathbb{E}_{\prod_{d,i} q(\mathbf{c}_{di:})}\left(\sum_{d,k}(\alpha_{dk} - 1 + n_{dk\cdot}) \log \pi_{dk} + \sum_{k,v}(\beta_{kv} - 1 + n_{\cdot kv}) \log \theta_{kv}\right) + \text{const.}$$

$$= \sum_{d=1}^{D}\sum_{k=1}^{K}(\alpha_{dk} - 1 + \mathbb{E}_{q(C)}(n_{dk\cdot})) \log \pi_{dk} + \sum_{k=1}^{K}\sum_{v=1}^{V}(\beta_{kv} - 1 + \mathbb{E}_{q(C)}(n_{\cdot kv})) \log \theta_{kv} + \text{const.}$$

$$q^*(\Pi, \Theta) = \prod_{d=1}^{D} \mathcal{D}\left(\boldsymbol{\pi}_d; \tilde{\alpha}_{d:} := [\alpha_{d:} + \tilde{\gamma}_{d\cdot:}]\right) \cdot \prod_{k=1}^{K} \mathcal{D}\left(\boldsymbol{\theta}_k; \tilde{\beta}_{kv} := [\beta_{kv} + \sum_{d}^{D}\sum_{i=1}^{l_d} \tilde{\gamma}_{di:}\mathbb{I}(w_{di} = v)]_{v=1,\ldots,V}\right).$$

Properties of the Dirichlet

Some tabulated identities, required for the concrete algorithm

$$p(x \mid \alpha) = \mathcal{D}(x; \alpha) = \frac{\Gamma(\hat{\alpha})}{\prod_d \Gamma(\alpha_d)} \prod_d x^{\alpha_d - 1} = \frac{1}{B(\alpha)} \prod_d x^{\alpha_d - 1} \qquad \hat{\alpha} := \sum_d \alpha_d$$

- $\mathbb{E}_p(x_d) = \frac{\alpha_d}{\hat{\alpha}}$
- $\mathrm{var}_p(x_d) = \frac{\alpha_d(\hat{\alpha} - \alpha_d)}{\hat{\alpha}^2(\hat{\alpha}+1)}$
- $\mathrm{cov}(x_d, x_i) = -\frac{\alpha_d \alpha_i}{\hat{\alpha}^2(\hat{\alpha}+1)}$
- $\mathrm{mode}(x_d) = \frac{\alpha_d - 1}{\hat{\alpha} - D}$
- $\mathbb{E}_p(\log x_d) = \digamma(\alpha_d) - \digamma(\hat{\alpha})$
- $\mathbb{H}(p) = -\int p(x) \log p(x)\, dx = -\sum_d (\alpha_d - 1)(\digamma(\alpha_d) - \digamma(\hat{\alpha})) + \log B(\alpha)$

Where $\digamma(z) = \frac{d}{dz} \log \Gamma(z)$ (the "digamma-function").

`scipy.special.digamma(z)` `https://dlmf.nist.gov/5`

The Variational Approximation

closing the loop

$$q(\boldsymbol{\pi}_d) = \mathcal{D}\left(\boldsymbol{\pi}_d; \tilde{\alpha}_{dk} := \left[\alpha_{dk} + \sum_{i=1}^{l_d} \tilde{\gamma}_{dik}\right]_{k=1,\dots,K}\right) \qquad \forall d = 1,\dots,D$$

$$q(\boldsymbol{\theta}_k) = \mathcal{D}\left(\boldsymbol{\theta}_k; \tilde{\beta}_{kv} := \left[\beta_{kv} + \sum_{d}^{D}\sum_{i=1}^{l_d} \tilde{\gamma}_{dik}\mathbb{I}(w_{di} = v)\right]_{v=1,\dots,V}\right) \qquad \forall k = 1,\dots,K$$

$$q(\mathbf{c}_{di}) = \prod_k \tilde{\gamma}_{dik}^{c_{dik}}, \qquad \forall d \;\; i = 1,\dots,l_d$$

where $\tilde{\gamma}_{dik} = \gamma_{dik} / \sum_k \gamma_{dik}$ and (note that $\sum_k \tilde{\alpha}_{dk} = \text{const.}$)

$$\begin{aligned}\gamma_{dik} &= \exp\left(\mathbb{E}_{q(\pi_{dk})}(\log \pi_{dk}) + \mathbb{E}_{q(\theta_{di})}(\log \theta_{kw_{di}})\right) \\ &= \exp\left(F(\tilde{\alpha}_{jk}) + F(\tilde{\beta}_{kw_{di}}) - F\left(\sum_v \tilde{\beta}_{kv}\right)\right)\end{aligned}$$

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

We need

$$\begin{aligned}
\mathcal{L}(q, W) &= \mathbb{E}_q(\log p(W, C, \Theta, \Pi)) + \mathbb{H}(q) \\
&= \int q(C, \Theta, \Pi) \log p(W, C, \Theta, \Pi)\, dC\, d\Theta\, d\Pi - \int q(C, \Theta, \Pi) \log q(C, \Theta, \Pi)\, dC\, d\Theta\, d\Pi \\
&= \int q(C, \Theta, \Pi) \log p(W, C, \Theta, \Pi)\, dC\, d\Theta\, d\Pi + \sum_k \mathbb{H}(\mathcal{D}(\theta_k \;\tilde{\beta}_k)) + \sum_d \mathbb{H}(\mathcal{D}(\pi_d \;\tilde{\alpha}_d)) + \sum_{di} \mathbb{H}(\tilde{\gamma}_{di})
\end{aligned}$$

The entropies can be computed from the tabulated values. For the expectation, we use $\mathbb{E}_{q(C)}(n_{dkv}) = \sum_i \gamma_{dik} \mathbb{I}(w_{di} = v)$ and use $\mathbb{E}_{\mathcal{D}(\pi_d; \tilde{\alpha})}(\log \pi_d) = F(\tilde{\alpha_d}) - F(\hat{\tilde{\alpha}})$ from above.

Dirty secret: In practice, the ELBO itself isn't strictly necessary.

Building the Algorithm

updating and evaluating the bound

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

procedure LDA(W, α, β)
  $\tilde{\gamma}_{dik} \leftarrow$ DIRICHLET_RAND(α) // initialize
  $\mathcal{L} \leftarrow -\infty$
  while $\mathcal{L}$ not converged **do**
    for $d = 1, \ldots, D; k = 1, \ldots, K$ **do**
      $\tilde{\alpha}_{dk} \leftarrow \alpha_{dk} + \sum_i \tilde{\gamma}_{dik}$ // update document-topics distributions
    end for
    for $k = 1, \ldots, K; v = 1, \ldots, V$ **do**
      $\tilde{\beta}_{kv} \leftarrow \beta_{kv} + \sum_{d,i} \tilde{\gamma}_{dik} \mathbb{I}(w_{di} = v)$ // update topic-word distributions
    end for
    for $d = 1, \ldots, D; k = 1, \ldots, K; i = 1, \ldots, l_d$ **do**
      $\tilde{\gamma}_{dik} \leftarrow \exp(F(\tilde{\alpha}_{dk}) + F(\tilde{\beta}_{kw_{di}}) - F(\sum_v \tilde{\beta}_{kv}))$ // update word-topic assignments
      $\tilde{\gamma}_{dik} \leftarrow \tilde{\gamma}_{dik} / \tilde{\gamma}_{di\cdot}$
    end for
    $\mathcal{L} \leftarrow$ BOUND($\tilde{\gamma}, w, \tilde{\alpha}, \tilde{\beta}$) // update bound
  end while
end procedure

$$\mathcal{L}(q) = \int q(C, \Theta, \Pi) \log \left(\frac{p(C, \Pi, \Theta, W)}{q(C, \Theta, \Pi)} \right) dC\, d\Theta\, d\Pi$$

Variational Inference is a powerful mathematical tool to construct efficient approximations to intractable *probability distributions* (not just point estimates, but entire distributions). Often, just imposing factorization is enough to make things tractable. The downside of variational inference is that constructing the bound can take significant ELBOw grease. However, the resulting algorithms are often highly efficient compared to tools that require less derivation work, like Monte Carlo.

"Derive your variational bound in the time it takes for your Monte Carlo sampler to converge."

The Toolbox

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
- EM / variational approximations