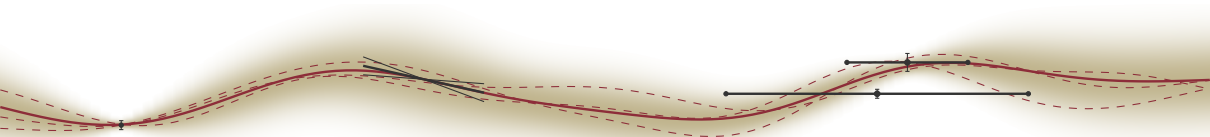
Probabilistic Machine Learning
Lecture 24
Variational Inference

Philipp Hennig

13 July 2021

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

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



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

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

- 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.

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

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

- 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.

Consider a joint distribution $p(x, z)$ with $z \in \mathbb{R}^n$

- to find a "good" but tractable approximation $q(z)$, assume that it factorizes $q(z) = \prod_i q_i(z_i)$.
- Initialize all q_i to some initial *distribution*
- Iteratively compute

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

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

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

- 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 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.

Example: The Gaussian Mixture Model

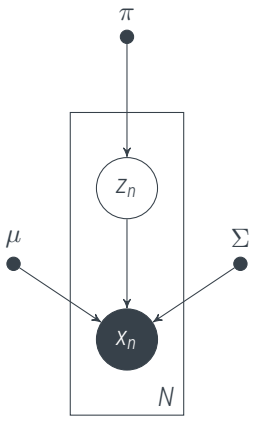
Returning to EM

Exposition after Bishop, PRML 2006, Chapter 10.2

- Remember EM for Gaussian mixtures $\theta := (\pi, \mu, \Sigma)$

$$p(x, z \mid \mu, \Sigma, \pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \cdot \mathcal{N}(x_n; \mu_k, \Sigma_k)^{z_{nk}}$$

$$= \prod_{n=1}^{N} p(z_{n:} \mid \pi) \cdot p(x_n \mid z_{n:}, \mu, \Sigma)$$



Example: The Gaussian Mixture Model

Returning to EM

Exposition after Bishop, PRML 2006, Chapter 10.2

- Remember EM for Gaussian mixtures $\theta := (\pi, \mu, \Sigma)$

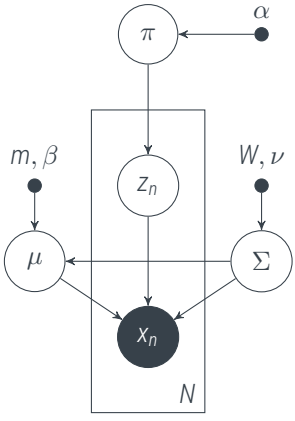
$$p(x, z \mid \mu, \Sigma, \pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \cdot \mathcal{N}(x_n; \mu_k, \Sigma_k)^{z_{nk}}$$

- For Bayesian inference, turn parameters into variables

$$p(x, z, \pi, \mu, \Sigma) = p(x, z \mid \pi, \mu, \Sigma) \cdot p(\pi) \cdot p(\mu \mid \Sigma) \cdot p(\Sigma)$$

$$p(\pi) = \mathcal{D}(\pi \mid \alpha) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_k \Gamma(\alpha_k)} \prod_k \pi_k^{\alpha_k - 1}$$

$$p(\mu \mid \Sigma) \cdot p(\Sigma) = \prod_{k=1}^{K} \mathcal{N}(\mu_k; m, \Sigma_k/\beta) \cdot \mathcal{W}(\Sigma_k^{-1}; W, \nu)$$



Constructing the Variational Approximation

Example: The Gaussian Mixture Model

[Exposition from Bishop, PRML 2006, Chapter 10.2]

- We know that the full posterior $p(z, \pi, \mu, \Sigma \mid x)$ is intractable (check the graph!)
- But let's consider an approximation $q(z, \pi, \mu, \Sigma)$ with the factorization

$$q(z, \pi, \mu, \Sigma) = q(z) \cdot q(\pi, \mu, \Sigma)$$

- from $(\star)$, we have

$$\begin{aligned}
\log q^*(z) &= \mathbb{E}_{q(\pi,\mu,\Sigma)} \left(\log p(x, z, \pi, \mu, \Sigma)\right) + \text{const.} \\
&= \mathbb{E}_{q(\pi)} \left(\log p(z \mid \pi)\right) + \mathbb{E}_{q(\mu,\Sigma)} \left(\log p(x \mid z, \mu, \Sigma)\right) + \text{const.} \\
&= \sum_{n}^{N} \sum_{k}^{K} z_{nk} \underbrace{\left(\mathbb{E}_{q(\pi)}(\log \pi_k) + \frac{1}{2} \mathbb{E}_{q(\mu,\Sigma)}(\log |\Sigma^{-1}| - (x_n - \mu_k)^\intercal \Sigma_k^{-1} (x - \mu_k)) \right)}_{=: \log \rho_{nk}} + \text{const.}
\end{aligned}$$

$$q^*(z) \propto \prod_n \prod_k \rho_{nk}^{z_{nk}} \qquad \text{define } r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nj}}, \text{ then } q^*(z) = \prod_n \prod_k r_{nk}^{z_{nk}} \text{ with } \mathbb{E}_{q(z)}[z] = r_{nk}$$

Constructing the Variational Approximation

using $q(z, \pi, \mu, \Sigma) = q(z) \cdot q(\pi, \mu, \Sigma)$

[Exposition from Bishop, PRML 2006, Chapter 10.2]

- Define some convenient notation:

$$N_k := \sum_{n=1}^{N} r_{nk} \qquad \bar{x}_k := \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} x_n \qquad S_k := \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (x_n - \bar{x}_k)(x_n - \bar{x}_k)^\intercal$$

- from $(\star)$, we have

$$\begin{aligned}
\log q^*(\pi, \mu, \Sigma) &= \mathbb{E}_{q(z)} \left(\log p(x, z, \pi, \mu, \Sigma) \right) + \text{const.} \\
&= \mathbb{E}_{q(z)} \left(\log p(\pi) + \sum_k \log p(\mu_k, \Sigma_k) + \log p(z \mid \pi) + \sum_n \log p(x_n \mid z, \mu, \Sigma) \right) \\
&= \log p(\pi) + \sum_{k=1}^{K} \log p(\mu_k, \Sigma_k) + \mathbb{E}_{q(z)} (\log p(z \mid \pi)) \\
&\quad + \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}(z_{nk}) \log \mathcal{N}(x_n; \mu_k, \Sigma_k) + \text{const.}
\end{aligned}$$

Constructing the Variational Approximation

using $q(z, \pi, \mu, \Sigma) = q(z) \cdot q(\pi, \mu, \Sigma)$

[Exposition from Bishop, PRML 2006, Chapter 10.2]

$$\log q^*(\pi, \mu, \Sigma) = \log p(\pi) + \sum_{k=1}^{K} \log p(\mu_k, \Sigma_k) + \mathbb{E}_{q(z)}(\log p(z \mid \pi))$$
$$+ \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}(z_{nk}) \log \mathcal{N}(x_n; \mu_k, \Sigma_k) + \text{const.}$$

- The bound exposes an **induced factorization** into $q(\pi, \mu, \Sigma) = q(\pi) \cdot \prod_{k=1}^{K} q(\mu_k, \Sigma_k)$

$$\text{where } \log q(\pi) = \log p(\pi) + \mathbb{E}_{q(z)}(\log p(z \mid \pi)) + \text{const.}$$
$$= (\alpha - 1) \sum_k \log \pi_k + \sum_k \sum_n r_{nk} \log \pi_k + \text{const.}$$
$$q(\pi) = \mathcal{D}(\pi, \alpha_k := \alpha + N_k) \quad \text{with } N_k = \sum_n r_{nk}$$

Constructing the Variational Approximation

using $q(z, \pi, \mu, \Sigma) = q(z) \cdot q(\pi, \mu, \Sigma)$

[Exposition from Bishop, PRML 2006, Chapter 10.2]

$$\log q^*(\pi, \mu, \Sigma) = \log p(\pi) + \sum_{k=1}^{K} \log p(\mu_k, \Sigma_k) + \mathbb{E}_{q(z)}(\log p(z \mid \pi))$$
$$+ \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}(z_{nk}) \log \mathcal{N}(x_n; \mu_k, \Sigma_k) + \text{const.}$$

- The bound exposes an **induced factorization** into $q(\pi, \mu, \Sigma) = q(\pi) \cdot \prod_{k=1}^{K} q(\mu_k, \Sigma_k)$

where (leaving out some tedious algebra) $q^*(\mu_k, \Sigma_k) = \mathcal{N}(\mu_k; m_k, \Sigma_k/\beta_k)\mathcal{W}(\Sigma_k^{-1}; W_k, \nu_k)$

$$\text{with } \beta_k := \beta + N_k \qquad m_k := \frac{1}{\beta_k}(\beta m + N_k \bar{x}_k) \qquad \nu_k := \nu + N_k$$

$$W_k^{-1} := W^{-1} + N_k S_k + \frac{\beta N_k}{\beta + N_k}(\bar{x}_k - m)(\bar{x}_k - m)^\intercal$$

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`

Constructing the Variational Approximation

closing the loop

[Exposition from Bishop, PRML 2006, Chapter 10.2]

- Recall from above:

$$\log q^*(z) = \sum_n^N \sum_k^K z_{nk} \underbrace{\left(\mathbb{E}_{q(\pi)}(\log \pi_k) + \frac{1}{2} \mathbb{E}_{q(\mu,\Sigma)}(\log |\Sigma^{-1}| - (x_n - \mu_k)^\intercal \Sigma_k^{-1}(x - \mu_k)) \right)}_{=:\log \rho_{nk}} + \text{const.}$$

- now we can evaluate ρ_{nk}, using tabulated identities

$$\log \tilde{\pi}_k := \mathbb{E}_{\mathcal{D}(\pi;\alpha_k)}(\log \pi_k) = \digamma(\alpha_k) - \digamma\left(\sum_k \alpha_k\right)$$

and for the Wishart:

$$\log \tilde{|\Sigma^{-1}|}_k := \mathbb{E}_{\mathcal{W}(\Sigma_k^{-1};W_k,\nu_k)}(\log |\Sigma_k^{-1}|) = \sum_{d=1}^D \digamma\left(\nu_k + \frac{1-d}{2}\right) + D \log 2 + \log |W_k|$$

$$\mathbb{E}_{\mathcal{N}(\mu_k;m_k,\Sigma_k/\beta_k)\mathcal{W}(\Sigma_k^{-1};W_k,\nu_k)}((x_n - \mu_k)^\intercal \Sigma^{-1}(x_n - \mu_k)) = D\beta_k^{-1} + \nu_k (x_n - m_k)^\intercal W_k (x_n - m_k)$$

Constructing the Variational Approximation

connection to EM

[Exposition from Bishop, PRML 2006, Chapter 10.2]

- this yields the update equation

$$\mathbb{E}_q(z_{nk}) = r_{nk} \propto \tilde{\pi}_k |\tilde{\Sigma^{-1}}|^{1/2} \exp\left(-\frac{D}{2\beta_k} - \frac{\nu_k}{2}(x_n - m_k)^\intercal W_k (x_n - m_k)\right)$$

  compare this with the EM-update

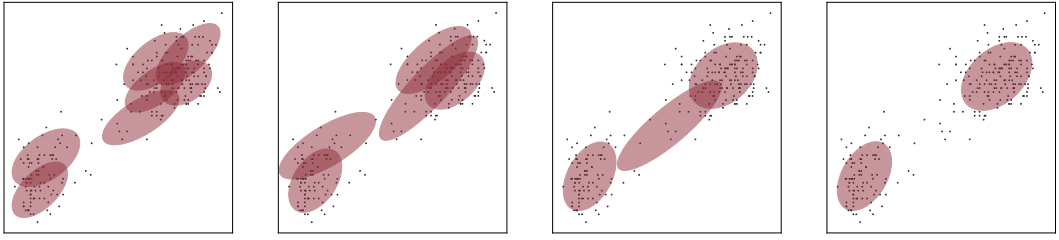
$$r_{nk} \propto \pi_k |\Sigma^{-1}|^{1/2} \exp\left(-\frac{1}{2}(x_n - \mu_k)^\intercal \Sigma_k^{-1}(x_n - \mu_k)\right)$$

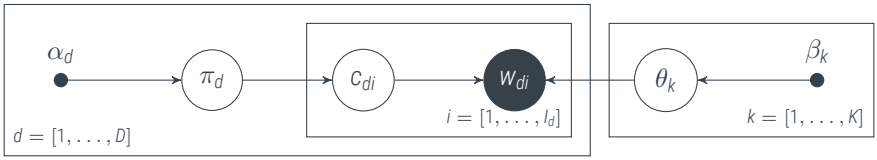
- Here, variational Inference is the Bayesian version of EM: Instead of maximizing the likelihood for $\theta = (\mu, \Sigma, \pi)$, we maximize a variational bound.
- One advantage of this is that the posterior can actually "decide" to ignore components:

Example

The Old Faithful Dataset, using $\alpha = 10^{-3}$

[from Bishop, PRML 2006, Fig. 10.6 / Ann-Kathrin Schalkamp]





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 $n_{dk\cdot} = \sum_v n_{dkv}$, 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

Mean-Field Theory: Putting Lectures 18–20 to use

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**

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

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

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**

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

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

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

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

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 $\sum_k \tilde{\alpha}_{dk} = \text{const.}$)

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

The ELBO

useful for monitoring / bug-fixing

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

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

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 (iterative maximum likelihood)
- variational inference / mean field