Probabilistic Inference and Learning
Lecture 20
Variational Inference

Philipp Hennig

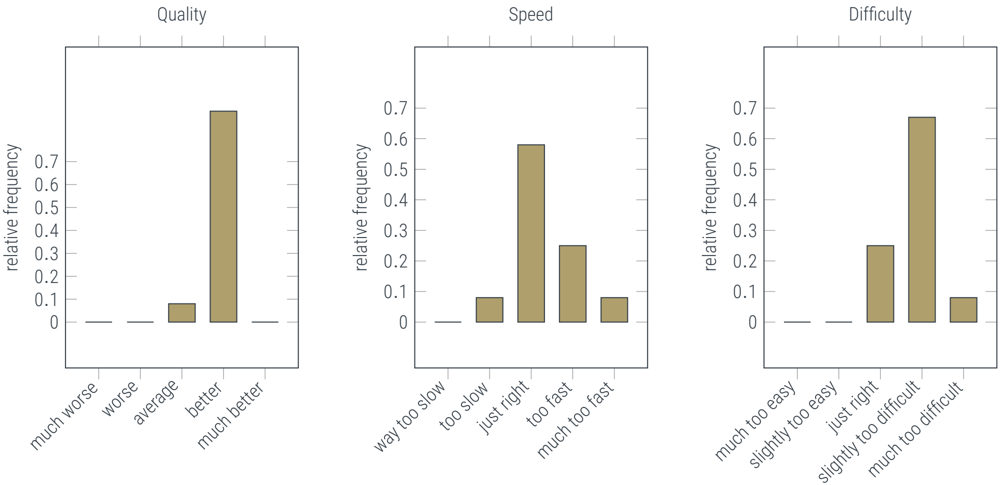
07 January 2019

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

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

Last Lecture: Debrief

Feedback dashboard



Last Lecture: Debrief

Detailed Feedback

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Things you did not like:

- only 4 Mondays left ;(
- cold lecture hall

Things you did not understand:

- When will the lecture notes be updated?

Things you enjoyed:

- recap of content before Christmas
- historical context
- speed
- recap of $\nabla_{\mu,\Sigma,\pi}$

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Recap: The EM Algorithm — General Form

from last lecture

Setting:

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

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

- 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 $D_{\text{KL}}(q \| p(z \mid x, \theta)) = 0$ and $\mathcal{L}(q, \theta_{\text{old}}) = \log p(x \mid \theta)$
2. Set θ_{new} to the **Maximize** the **Expectation Lower Bound / minimize the Variational Free Energy**

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

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

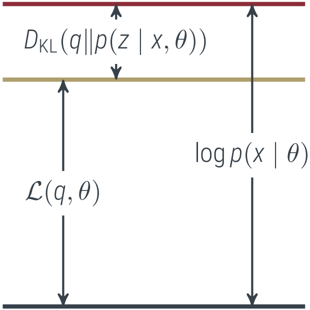
EM maximizes the ELBO / minimizes Free Energy

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

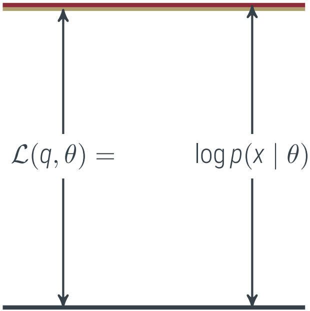
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

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

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

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

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



EM maximizes the ELBO / minimizes Free Energy

a more general view

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

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

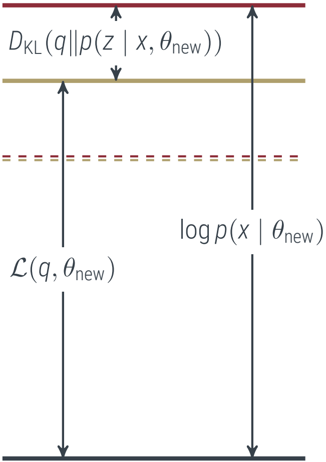
$$D_{\mathrm{KL}}(q \| p(z \mid x,\theta)) = -\sum_z q(z) \log\left(\frac{p(z \mid x,\theta)}{q(z)}\right)$$

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**

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

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



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

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

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

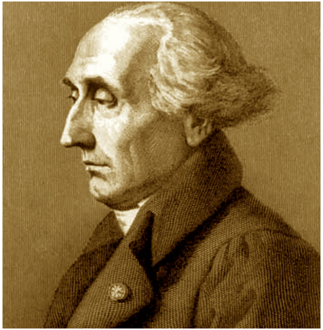
$$z \leftarrow z \cup \theta$$

- 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

Leonhard Euler
1707–1783



Joseph-Louis Lagrange
1736–1813

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

Richard P. Feynman
1918–1988 (Nobel Prize 1965)

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**

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

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

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

where $\log \tilde{p}(x, z_j) = \mathbb{E}_{q, i \neq j}[\log p(x,z)] + \text{const.}$

Mean Field Theory

Factorizing variational approximations

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

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

  and maximize wrt. q_j. Doing so *minimizes* $D_{\text{KL}}(q(z_j) \| \tilde{p}(x, z_j))$, 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).

Recap: Kullback-Leibler Divergence

from Lecture 14

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

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

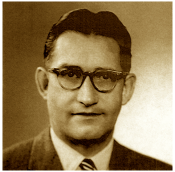
$$D_{\mathrm{KL}}(P\|Q) := \int \log\left(\frac{p(x)}{q(x)}\right) dp(x)$$

(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)

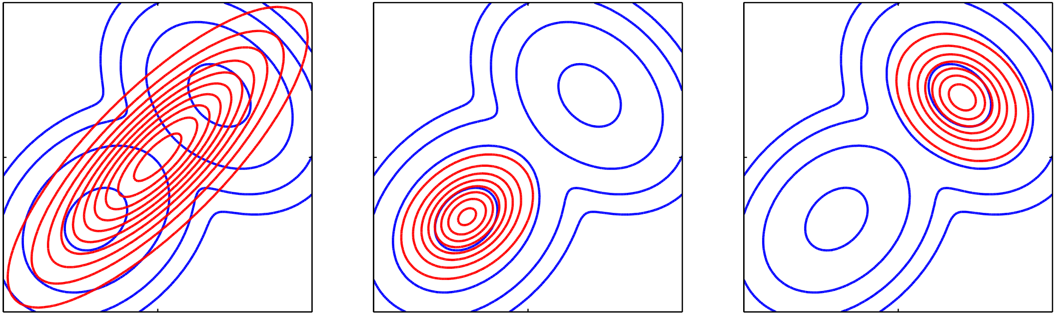


Richard Leibler (1914–2003)

KL Divergence — Pictoral View

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

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



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

The Toolbox

Five principal methods for dealing with computational complexity in probabilistic inference

1. **Maximum Likelihood (ML) / Maximum A-Posteriori (MAP)** estimation:
   To estimate θ in $p(D \mid \theta)$ or $p(\theta \mid D)$, set $\hat{\theta} = \arg\max_\theta p$.
2. **Laplace** Approximation: $p(\theta \mid D) \approx \mathcal{N}\left(\theta; \hat{\theta}, -(\nabla\nabla^\intercal \log p(\theta \mid D))^{-1}\right)$
3. **Variational Inference:**
   To approximate $p(\theta \mid D)$, impose structure on $q(\theta)$, then minimize $D_{\mathrm{KL}}(q\|p)$
4. ????
5. **Numerical Quadrature:**
   To marginalize θ, compute $\int p(f \mid \theta)\, d\theta \approx \sum_i w_i \cdot p(f \mid \theta_i)$

Disclaimer: The listed items are neither mutually exclusive nor collectively exhaustive. Some of the methods are intricately interrelated.

Back to the Gaussian Mixture Example

- The **Wishart** distribution is the conjugate prior (exponential family) to the Gaussian with unknown precision $\Xi \in \mathbb{R}^{d \times d}$ (symmetric positive definite) (it is the multivariate Version of the Gamma distribution)

$$\mathcal{W}(\Xi; W, \nu) = \frac{1}{2^{\nu d/2}|W|^{\nu/2}\Gamma_d(\nu/2)}|\Xi|^{(\nu-d-1)/2}\exp(-\operatorname{tr}(W^{-1}\Xi)/2)$$

$$\prod_i^n \mathcal{N}(x_i; \mu, \Sigma)\mathcal{W}(\Sigma^{-1}; W, \nu) \propto \mathcal{W}\left(\Sigma^{-1}; \left(W^{-1} + \sum_i (x_i - \mu)(x_i - \mu)^\intercal\right)^{-1}, \nu + n\right)$$

- The **Normal-inverse-Wishart** is the conjugate prior to the Gaussian with unknown mean and precision (updated parameters left out for brevity)

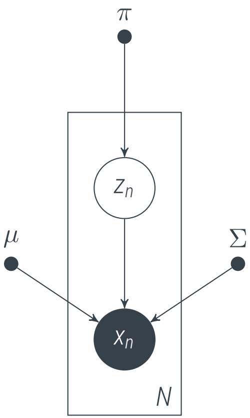
$$\prod_i^n \mathcal{N}(x_i; \mu, \Sigma) \cdot \mathcal{N}(\mu, \mu_0, \gamma_0^{-1}\Sigma)\mathcal{W}(\Sigma^{-1}; W, \nu) \propto \mathcal{N}(\mu, \mu_n, \gamma_n^{-1}\Sigma)\mathcal{W}(\Sigma^{-1}; W_n, \nu_n)$$

$\Gamma_d(x) = \pi^{d(d-1)/4} \prod_{j=1}^d \Gamma(x + (1-j)/2)$

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

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

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



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

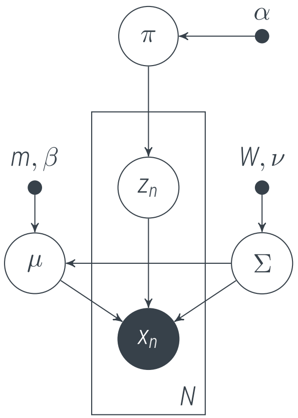
$$p(x, z \mid \mu, \Sigma, \pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \cdot \mathcal{N}(x_n; \mu_k, \Sigma_k)^{z_{nk}}$$

- For Bayesian inference, turn parameters into variables

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

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

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



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

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

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

- note that q^* factorizes over n, even though we did not impose this! An **induced factorization**

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

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

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 $q(\pi, \mu, \Sigma) = q(\pi) \cdot \prod_{k=1}^{K} q(\mu_k, \Sigma_k)$

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

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

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

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

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- 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 ($\psi(x) := \frac{d}{dx} \log \Gamma(x)$)

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

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

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

- 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

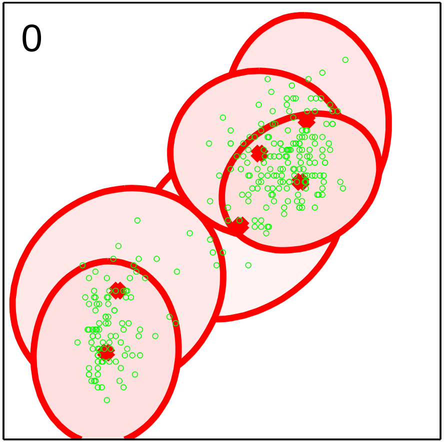
$$r_{nk} \propto \pi_k |\Sigma^{-1}|^{1/2} \exp\left(-\frac{1}{2}(x_n - \mu_k)^\intercal \Sigma_k^{-1} (x_n - \mu_k)\right)$$

- 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, because the Dirichlet prior can favor sparse π (for maximum likelihood, it is always favorable to maximize the number of components as that allows putting a lot of mass on a small number of (or single) data-point(s)):

Example

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

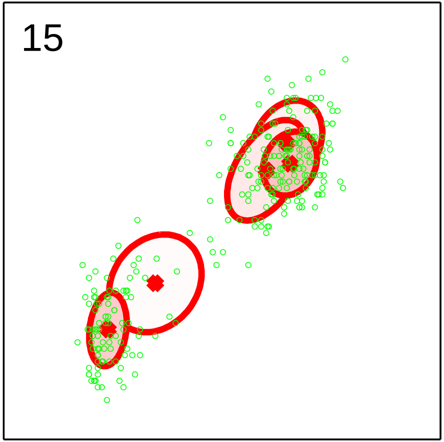
[from Bishop, PRML 2006, Fig. 10.6]



Example

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

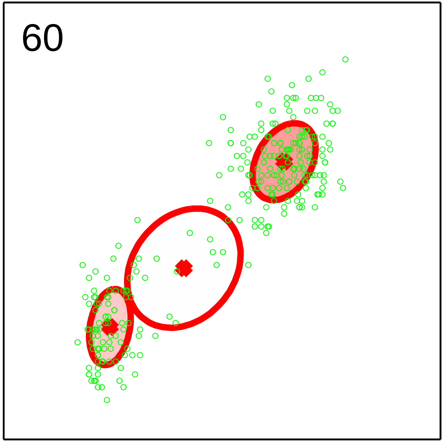
[from Bishop, PRML 2006, Fig. 10.6]



Example

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

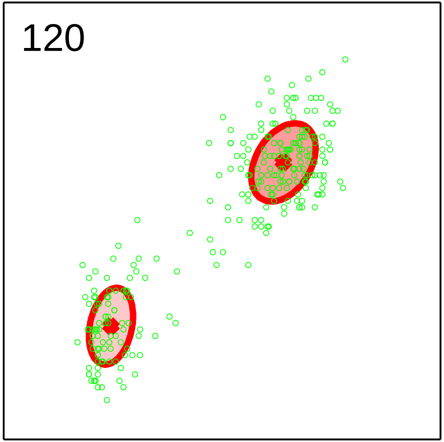
[from Bishop, PRML 2006, Fig. 10.6]



Example

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

[from Bishop, PRML 2006, Fig. 10.6]



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

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

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

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

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

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

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

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

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

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

$$\log q^*(\eta) = \log p(\eta \mid \nu, v) + \mathbb{E}_z(\log p(x, z \mid \eta)) + \text{const.}$$

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

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

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.