Probabilistic Machine Learning
Lecture 23
Free Energy

Philipp Hennig

12 July 2021

Eberhard Karls Universität Tübingen

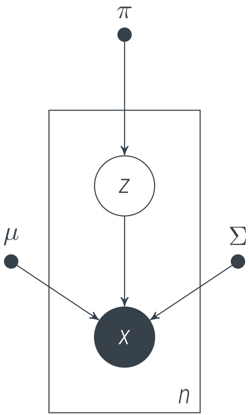
Faculty of Science
Department of Computer Science
Chair for the Methods of Machine Learning

EM for Gaussian Mixtures

re-written in generic form

- Want to maximize, as function of $\theta := (\pi_j, \mu_j, \Sigma_j)_{j=1,\dots,k}$

$$\log p(x \mid \pi, \mu, \Sigma) = \sum_i \log \left(\sum_j \pi_j \mathcal{N}(x_i; \mu_j, \Sigma_j) \right)$$



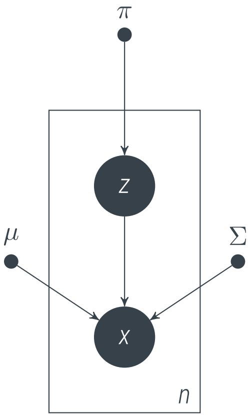
- Want to maximize, as function of $\theta := (\pi_j, \mu_j, \Sigma_j)_{j=1,\dots,k}$

$$\log p(x \mid \pi, \mu, \Sigma) = \sum_i \log \left(\sum_j \pi_j \mathcal{N}(x_i; \mu_j, \Sigma_j) \right)$$

- Instead, maximizing the "complete data" likelihood is easier:

$$\log p(x, z \mid \pi, \mu, \Sigma) = \log \prod_i^n \prod_j^k \pi_j^{z_{ij}} \mathcal{N}(x_i; \mu_j, \Sigma_j)^{z_{ij}}$$

$$= \sum_i \sum_j z_{nk} \underbrace{\left(\log \pi_j + \log \mathcal{N}(x_i; \mu_j, \Sigma_j)\right)}_{\text{easy to optimize (exponential families!)}}$$



EM for Gaussian Mixtures

re-written in generic form

1. Compute $p(z \mid x, \theta)$:

$$p(z_{ij} = 1 \mid x_i, \pi, \mu, \Sigma) = \frac{p(z_{ij} = 1)p(x_i \mid z_{ij} = 1)}{\sum_{j'}^{k} p(z_{ij'} = 1)p(x_i \mid z_{ij'} = 1)} = \frac{\pi_j \mathcal{N}(x_i; \mu_j, \Sigma_j)}{\sum_{j'} \pi_j \mathcal{N}(x_i; \mu_j, \Sigma_j)} =: r_{ij}$$

2. Maximize

$$\mathbb{E}_{p(z|x,\theta)} (\log p(x, z \mid \theta)) = \sum_i \sum_j r_{ij} \left(\log \pi_j + \log \mathcal{N}(x_i; \mu_j, \Sigma_j)\right)$$

(see earlier slides on how to solve this, much easier problem)

Generic EM Algorithm

Maximize **expected** log likelihoods

Setting:

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

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

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

Idea: Initialize θ_0, then iterate between

1. Compute $p(z \mid x, \theta_{\text{old}})$
2. Set θ_{new} to the **Maximum** of the **Expectation** of the *complete-data* log likelihood:

$$\theta_{\text{new}} = \arg\max_\theta \sum_z p(z \mid x, \theta_{\text{old}}) \log p(\underbrace{x, z}_{!} \mid \theta) = \arg\max_\theta \mathbb{E}_{p(z \mid x, \theta_{\text{old}})} [\log p(x, z \mid \theta)]$$

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

The EM algorithm

Instead of trying to maximize

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

instead maximize

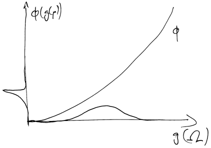
$$\mathbb{E}_z \log p(x, z \mid \theta) = \sum_z p(z \mid x, \theta) \log p(x, z \mid \theta),$$

then re-compute $p(z \mid x, \theta)$, and repeat.

An Observation

maximizing a lower bound

- We constructed an approximate distribution $q(z) = p(z \mid x, \theta)$ for our latent quantity
- For *any* such approximation $q(z)$ (if $q(z) > 0$ wherever $p(x, z \mid \theta) > 0$):



$$\log p(x \mid \theta) = \log \int p(x, z \mid \theta)\, dz \qquad = \log \int q(z) \frac{p(x, z \mid \theta)}{q(z)}\, dz$$

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

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

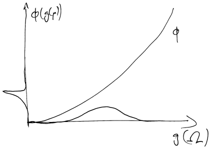
Let (Ω, A, μ) be a probability space, g be a real-valued, μ-integrable function and ϕ be a convex function on the real line. Then

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

An Observation

maximizing a lower bound

- We constructed an approximate distribution $q(z) = p(z \mid x, \theta)$ for our latent quantity
- For *any* such approximation $q(z)$ (if $q(z) > 0$ wherever $p(x, z \mid \theta) > 0$):



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

  - 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

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

$$= \int q(z) \log \frac{p(z \mid x, \theta)}{q(z)}\, dz + \log p(x \mid \theta) \int q(z)\, dz$$

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

The Kullback-Leibler divergence satisfies

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

An Insight

Distance of our bound

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

$$= \int q(z) \log \frac{p(z \mid x, \theta)}{q(z)}\, dz + \log p(x \mid \theta) \int q(z)\, dz$$

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

The Kullback-Leibler divergence satisfies

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

EM maximizes the ELBO / minimizes Free Energy

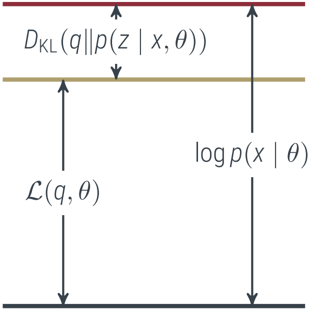
a more general view

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

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

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

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



EM maximizes the ELBO / minimizes Free Energy

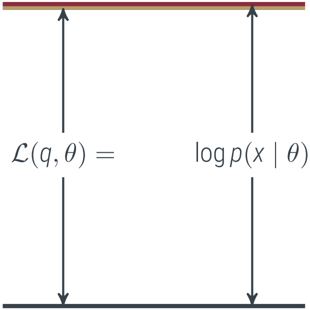
a more general view

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

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

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

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



EM maximizes the ELBO / minimizes Free Energy

a more general view

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

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

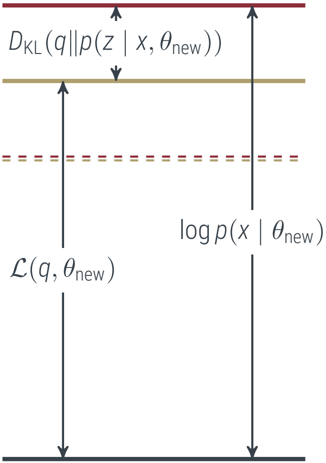
$$D_{\mathrm{KL}}(q \| p(z \mid x, \theta)) = - \int q(z) \log \left(\frac{p(z \mid x,\theta)}{q(z)} \right) dz$$

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

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

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

- 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 **Evidence 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: Gaussian Mixture Models).

Some Observations

for future reference

- It is straightforward to extend EM to maximize a **posterior** instead of a likelihood. Just add a log prior for θ:

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 **Evidence Lower Bound**

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

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

This maximizes

$$\begin{aligned} \log p(x \mid \theta) + \log p(\theta) &\leq \mathcal{L}(q, \theta) + \log p(\theta) \\ &\triangleq \log p(\theta \mid x) \end{aligned}$$

Some Observations

for future reference

- 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**

The EM algorithm:

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

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

- 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**

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

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

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
-

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

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

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

- 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.

A Historical Connection

free energy / expectation lower bounds

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

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

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

Free Energy

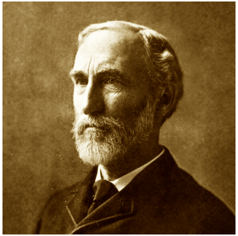
Machine Learning is the application of scientific modelling to *everything*

Hermann v. Helmholtz
1821–1894

image:L. Meder

"Energy"

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

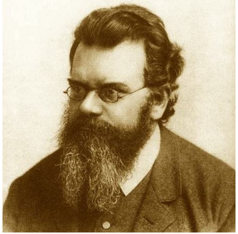


Josia W. Gibbs
1839–1903

image:unknown

"Enthalpy"

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

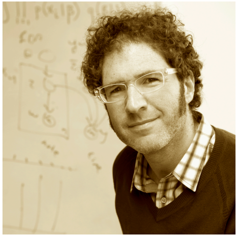


Ludwig Boltzmann
1844–1906

image:wikipedia

"Entropy"

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



David M. Blei

image:Denise Applewhite

"Evidence"

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

Variational Inference

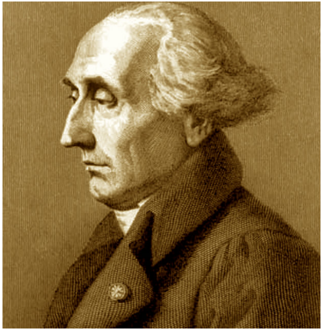
- is a general framework to construct approximating **probability distributions** $q(z)$ to non-analytic posterior distributions $p(z \mid x)$ by minimizing the **functional**

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

The 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**

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

- 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

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

  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.