Machine Learning

Lecture 12: Variational Inference

Prof. Dr. Stephan Günnemann

Data Mining and Analytics
Technical University of Munich

28.01.2019

Reading material

Reading material

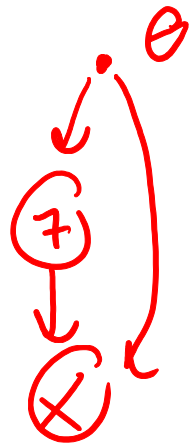
- Blei, Kucukelbir and McAuliffe -
  " Variational Inference: A Review for Statisticians " (pp. 1 - 16)
  `https://arxiv.org/abs/1601.00670`

Recall: EM algorithm

Setup

- Latent variable model with (intractable) log-likelihood

$$\log p(\mathbf{X} \mid \boldsymbol{\theta}) = \log \left(\sum_{\mathbf{Z}} p(\mathbf{Z} \mid \boldsymbol{\theta}) p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta}) \right)$$



What are we interested in?

- Maximum likelihood estimate $\boldsymbol{\theta}^*$ of the parameters $\boldsymbol{\theta}$

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \log p(\mathbf{X} \mid \boldsymbol{\theta})$$

$\max_\theta \mathbb{E}_{Z \sim p(Z \mid X, \theta^{old})} [\log p(X, Z \mid \theta)]$

- Posterior distribution of the latent variables $p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^*)$

How do we solve it?

- Alternating optimization, consisting of E & M steps.

Bayesian viewpoint

Can we do better than MLE for $\boldsymbol{\theta}$?

- We have seen many times already that maximum likelihood is prone to overfitting and has other undesirable properties.
- As good Bayesians, we could instead try to compute the full posterior

$$p(\mathbf{Z}, \boldsymbol{\theta} \mid \mathbf{X})$$

Full Bayesian approach

- We first need to place a prior $p(\boldsymbol{\theta})$ over the parameters $\boldsymbol{\theta}$.
- The full posterior is then obtained as

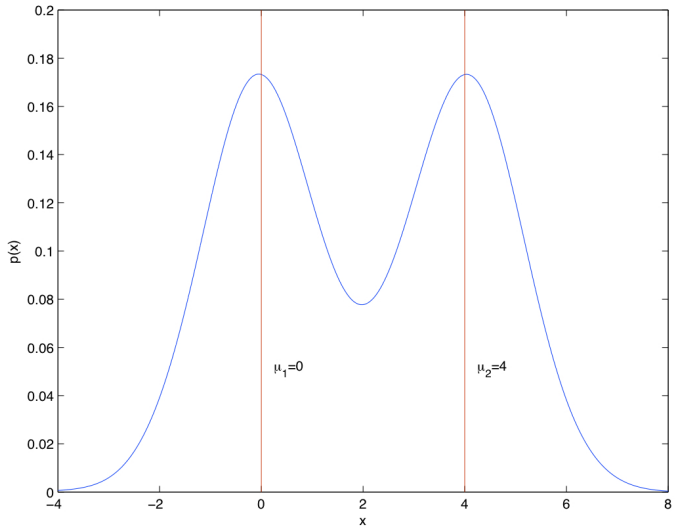
$$\begin{aligned} p(\mathbf{Z}, \boldsymbol{\theta} \mid \mathbf{X}) &= \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta})}{p(\mathbf{X})} \\ &= \frac{p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta}) p(\mathbf{Z} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int \int p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta}) p(\mathbf{Z} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} d\mathbf{Z}} \end{aligned}$$

Simplified univariate GMM (with prior on $\boldsymbol{\mu}$)

$\mu = [\mu_1, \mu_2]$

- Data is 1-dimensional: $x_i \in \mathbb{R}$, $\mu_k \in \mathbb{R}$ for all i, k.
- Cluster indicators $\mathbf{z}_i$ are chosen uniformly

$$p(\mathbf{z}_i) = \mathrm{Cat}(1/K, ..., 1/K)$$



- Each component's mean is chosen i.i.d from $\mathcal{N}(0, \sigma^2)$ with σ^2 fixed

$$p(\mu_k) = \mathcal{N}(\mu_k \mid 0, \sigma^2), \qquad p(\boldsymbol{\mu}) = \prod_{k=1}^{K} p(\mu_k)$$

$z_i = [0, 1, 0]$

- Each component has unit variance, i.e. probability of $\mathbf{x}_i$ is

$$p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu}) = \prod_{k=1}^{K} \mathcal{N}(x_i \mid \mu_k, 1)^{z_{ik}} = \mathcal{N}(x_i \mid \mathbf{z}_i^T \boldsymbol{\mu}, 1)$$

$= \mathcal{N}(x_i \mid \mu_{z_i}, 1)$

Bayesian treatment of simplified GMM

- The joint distribution is

$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\mu}) = p(\boldsymbol{\mu}) \prod_{i=1}^{N} p(\mathbf{z}_i) p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu})$$

- To obtain the posterior

$$p(\mathbf{Z}, \boldsymbol{\mu} \mid \mathbf{X}) = p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\mu}) / p(\mathbf{X})$$

  we need to compute the normalization constant

$$\begin{aligned} p(\mathbf{X}) &= \int p(\boldsymbol{\mu}) \prod_{i=1}^{N} \left[\sum_{\mathbf{z}_i} p(\mathbf{z}_i) p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu}) \right] d\boldsymbol{\mu} \\ &= \sum_{\mathbf{Z}} p(\mathbf{Z}) \int p(\boldsymbol{\mu}) \prod_{i=1}^{N} p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu}) d\boldsymbol{\mu} \end{aligned}$$

- In either case this requires $\mathcal{O}(K^N)$ operations and is intractable.

Approximate posterior inference

- We saw that posterior inference even in such an extremely simplistic model is intractable.
- Is there a way to approximate the posterior distribution that is computationally feasible?

Approximate posterior inference

Setting

- For brevity, we overload $\mathbf{z}$, and use it to denote both the parameters and latent variables [1]; $\mathbf{x}$ stands for observations as before.
- We want to find the posterior distribution

$$p(\mathbf{z} \mid \mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{x})} = \frac{p(\mathbf{x}, \mathbf{z})}{\int p(\mathbf{x}, \mathbf{z}) d\mathbf{z}}$$

where the denominator $p(\mathbf{x})$ (called evidence) is intractable.

Problem

- If we can't compute $p(\mathbf{z} \mid \mathbf{x})$, how can we approximate it?

[1] In GMM this would include cluster indicators $\mathbf{z}_i$, priors $\boldsymbol{\pi}$, means $\boldsymbol{\mu}_k$ and covariances $\boldsymbol{\Sigma}_k$.

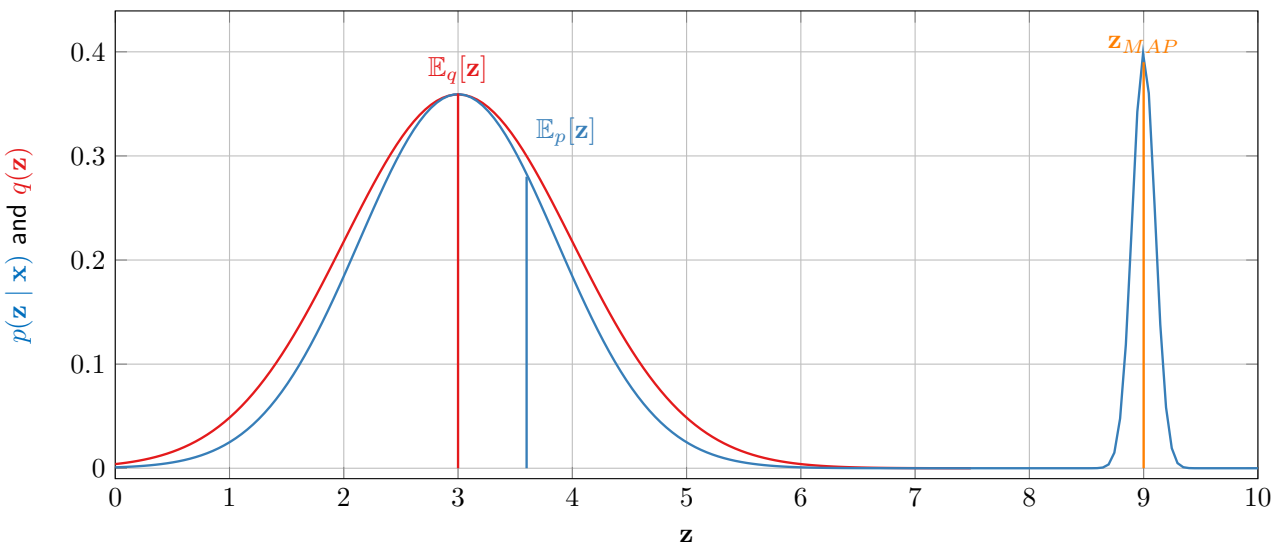
Variational inference (VI)

Problem

- We want to approximate an intractable distribution $p(\mathbf{z} \mid \mathbf{x})$.

Main idea

- Find a tractable distribution $q(\mathbf{z})$ that is "similar" to $p(\mathbf{z} \mid \mathbf{x})$.
- Use $q(\mathbf{z})$ to answer the questions about $p(\mathbf{z} \mid \mathbf{x})$ that we care about, e.g., instead of computing $\mathbb{E}_{p(\mathbf{z}|\mathbf{x})}[\mathbf{z}]$, we can compute $\mathbb{E}_{q(\mathbf{z})}[\mathbf{z}]$.



How do we find a $q(\mathbf{z})$ "similar" to $p(\mathbf{z} \mid \mathbf{x})$?

Prerequisites

- Define the search space.

  Choose the family (set) of tractable candidate distributions $\mathcal{Q}$.
  In the context of VI, we call $\mathcal{Q}$ the variational family.

- Choose the divergence D.

  Divergence $D(p \parallel q)$ measures how dissimilar distributions p and q are.

Formal problem statement

- The problem can now be formally stated as

$$q^* = \arg\min_{q \in \mathcal{Q}} D(p(\mathbf{z} \mid \mathbf{x}) \parallel q(\mathbf{z}))$$

  i.e. "Find the distribution $q^*(\mathbf{z})$ from the family of distributions $\mathcal{Q}$ that is as close as possible to $p(\mathbf{z} \mid \mathbf{x})$ in terms of divergence D."

Families of distributions

A family of distributions is a set of distributions that share some property.

One choice is a parametric family $\mathcal{Q}$, such that every distribution $q \in \mathcal{Q}$ is uniquely defined by its parameters $\boldsymbol{\nu}$.

$$\mathcal{Q} = \{q(\mathbf{z} \mid \boldsymbol{\nu}) \text{ for all valid parameters } \boldsymbol{\nu}\}$$

Examples of parametric families

- The "set of all univariate normal distributions with unit variance" is

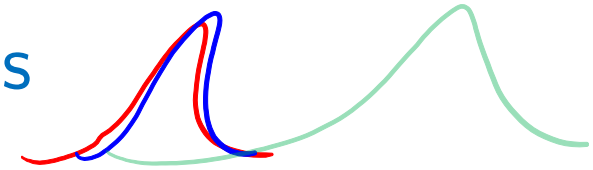
$$\mathcal{Q}_1 = \{\mathcal{N}(\mu, 1) : \mu \in \mathbb{R}\}$$

  Any choice of $\mu \in \mathbb{R}$ corresponds to a distribution $q \in \mathcal{Q}_1$.
- A more expressive family - "set of all univariate normal distributions"

$$\mathcal{Q}_2 = \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_{>0}\}$$

  Each distribution in $\mathcal{Q}_2$ is defined by a pair $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0}$.

Divergences between distributions



Definition

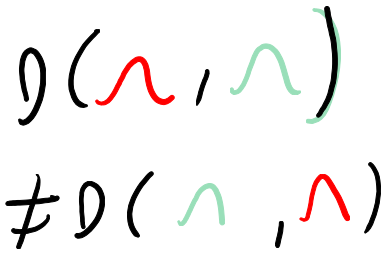
Let $\mathcal{P}$ denote the set of probability distributions over a space Ω (e.g. $\mathbb{R}^d$).

$$\mathcal{P} = \{p : p(\mathbf{x}) \text{ is a valid probability distribution over } \Omega\}$$

A divergence $D : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_{\geq 0}$ defines a measure of dissimilarity between two probability distributions $p, q \in \mathcal{P}$.

Properties

- Non-negative, $D(p \parallel q) \geq 0$ for all $p, q \in \mathcal{P}$.
- $D(p \parallel q) = 0 \iff p = q$ almost everywhere



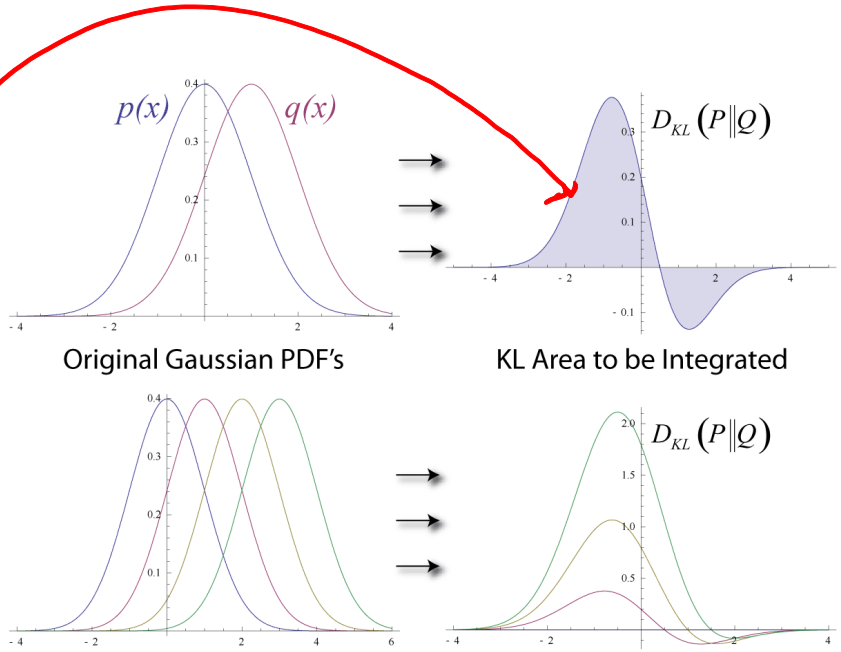
Divergence is not a distance

- Doesn't have to be symmetric, $D(p \parallel q) \neq D(q \parallel p)$ in general.
- Doesn't have to fulfill the triangle inequality.

Recall: Kullback-Leibler divergence

Kullback-Leibler divergence between two distributions $p(\mathbf{z})$ and $q(\mathbf{z})$ is defined as

$$\begin{aligned}\mathbb{KL}(p \parallel q) &= \int p(\mathbf{z}) \log \frac{p(\mathbf{z})}{q(\mathbf{z})} d\mathbf{z} \\ &= \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log p(\mathbf{z}) - \log q(\mathbf{z})\right]\end{aligned}$$



Properties

- Asymmetric, $\mathbb{KL}(p \parallel q) \neq \mathbb{KL}(q \parallel p)$ in general.
- Nonnegative, $\mathbb{KL}(p \parallel q) \geq 0$.
- $\mathbb{KL}(p \parallel q) = 0 \iff p = q$ almost everywhere.

We assume that $\mathbf{z}$ is continuous. For the discrete case we simply have to replace respective integrals by summations.

Minimizing forward KL divergence

$$\sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{x}) \cdot \log \frac{p(\mathbf{z}|\mathbf{x})}{q(\mathbf{z})}$$

First thing to try is to find q that is close to p in terms of KL divergence.

$$\begin{aligned} q^* &= \arg\min_{q \in \mathcal{Q}} \mathbb{KL}(p(\mathbf{z} \mid \mathbf{x}) \parallel q(\mathbf{z})) \\ &= \arg\min_{q \in \mathcal{Q}} \mathbb{E}_{p(\mathbf{z}|\mathbf{x})} \left[\log p(\mathbf{z} \mid \mathbf{x}) - \log q(\mathbf{z})\right] \end{aligned}$$

Computing the objective involves taking an expectation w.r.t. the intractable $p(\mathbf{z} \mid \mathbf{x})$, so we need to try something else. [1]

What about the reverse direction $\mathbb{KL}(q \parallel p)$?

[1] There exist ways to minimize $\mathbb{KL}(p \parallel q)$, such as Expectation propagation algorithm, but it's outside the scope of this lecture.

Minimizing reverse KL divergence

In the reverse direction, we are looking for the optimizer

$$
\begin{aligned}
q^* &= \arg\min_{q \in \mathcal{Q}} \mathbb{KL}(q \parallel p) \\
&= \arg\min_{q \in \mathcal{Q}} \mathbb{E}_q \left[\log q(\mathbf{z}) - \log p(\mathbf{z} \mid \mathbf{x})\right] \\
&= \arg\min_{q \in \mathcal{Q}} \mathbb{E}_{q(\mathbf{z})} \left[\log q(\mathbf{z}) - \log p(\mathbf{x}, \mathbf{z}) + \log p(\mathbf{x})\right]
\end{aligned}
$$

because $\log p(\mathbf{x})$ does not depend on $\mathbf{z}$, we can move it outside $\mathbb{E}_{q(\mathbf{z})}$

$$
= \arg\min_{q \in \mathcal{Q}} \mathbb{E}_{q(\mathbf{z})} \left[\log q(\mathbf{z}) - \log p(\mathbf{x}, \mathbf{z})\right] + \text{const}
$$

We got rid of the intractable $\log p(\mathbf{x})$, so this expression is something that we can work with!

Evidence lower bound (ELBO)

The optimization problem thus is

$$\begin{aligned} q^* &= \arg\min_{q \in \mathcal{Q}} \mathbb{KL}(q \parallel p) \\ &= \arg\max_{q \in \mathcal{Q}} \mathbb{E}_{q(\mathbf{z})} \left[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z})\right] \end{aligned}$$

We denote the optimization objective as

$$\mathcal{L}(q) := \mathbb{E}_{q(\mathbf{z})} \left[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z})\right]$$

and call it Evidence Lower BOund (ELBO).

Its name comes from the fact that (as seen in the previous slide)

$$\log p(\mathbf{x}) = \mathbb{KL}(q(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x})) + \mathcal{L}(q) \geq \mathcal{L}(q),$$

which follows from nonnegativity of $\mathbb{KL}$.

Note that the bound is tight iff $\mathbb{KL}(q(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x})) = 0$.

ELBO and entropy

We can equivalently rewrite the ELBO as following

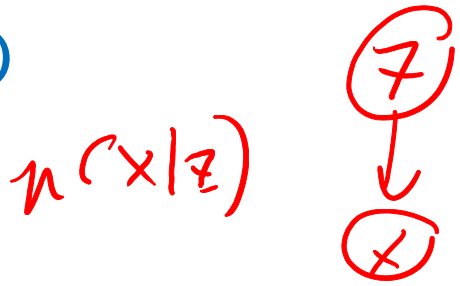
$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_{q(\mathbf{z})}\left[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z})\right] \\ &= \mathbb{E}_{q(\mathbf{z})}\left[\log p(\mathbf{x}, \mathbf{z})\right] + \mathbb{H}\left[q(\mathbf{z})\right],\end{aligned}$$

where $\mathbb{H}[q(\mathbf{z})]$ is the entropy of the distribution $q(\mathbf{z})$.

The ELBO "encourages" the variational distribution to

1. Place high probability mass on regions where the unnormalized posterior $p(\mathbf{x}, \mathbf{z})$ is high.
2. Keep entropy as large as possible, thus keeping the distribution "general" and avoiding overfitting.

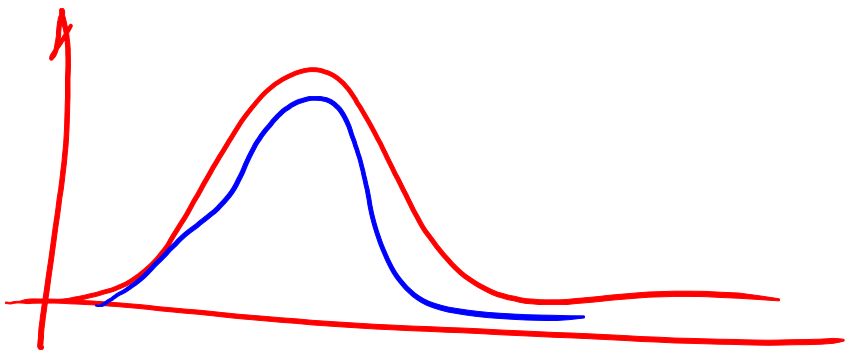
Role of ELBO



An extremely important thing to understand is that the variational distribution $q(\mathbf{z} \mid \boldsymbol{\nu})$ does not model the observed data $\mathbf{x}$.

Rather, the approximate posterior $q(\mathbf{z} \mid \boldsymbol{\nu})$ is "coupled" to the observations by the ELBO.

That is, because we are maximizing ELBO, we are bringing $q(\mathbf{z} \mid \boldsymbol{\nu})$ closer to the true posterior $p(\mathbf{z} \mid \mathbf{x})$, which in turn depends on the data.



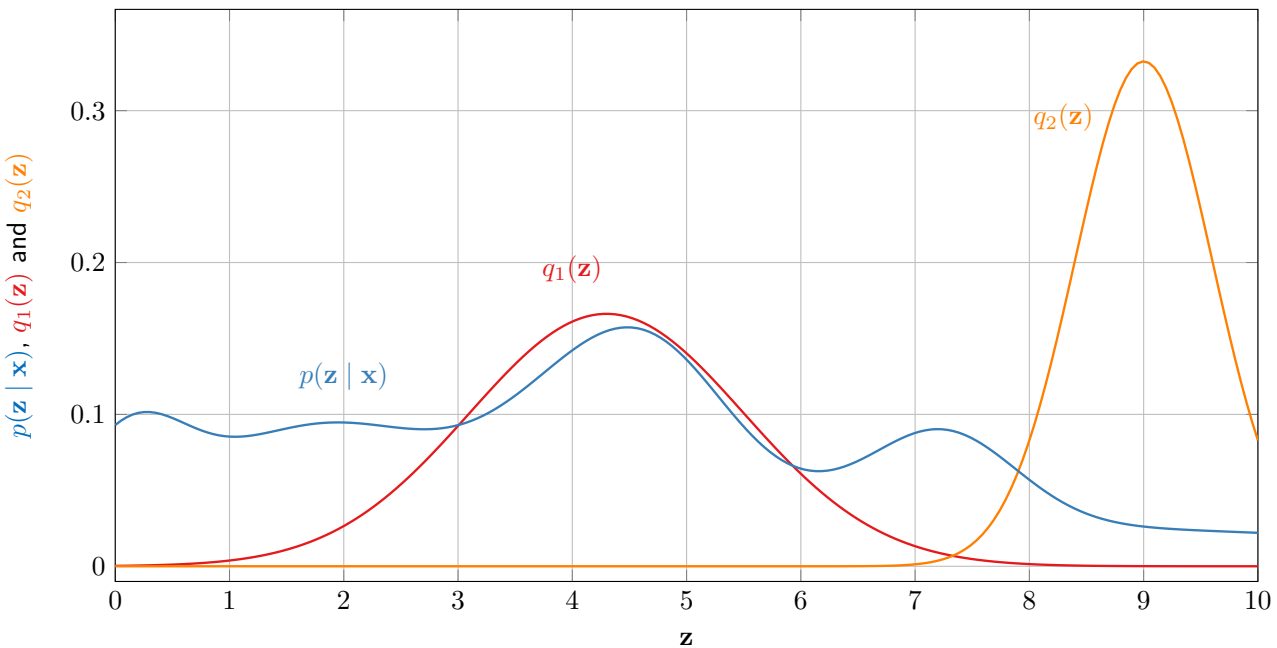
Approximating the posterior

The first variational distribution $q_1(\mathbf{z})$ is a good choice:

- The KL divergence $\mathbb{KL}(q_1(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x}))$ is low.
- The ELBO $\mathcal{L}(q_1(\mathbf{z}))$ is high.

The second variational distribution $q_2(\mathbf{z})$ is a bad choice:

- The KL divergence $\mathbb{KL}(q_2(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x}))$ is high.
- The ELBO $\mathcal{L}(q_2(\mathbf{z}))$ is low.



What we did so far

$q(z) \approx p(z|x)$

- We want to approximate the intractable posterior $p(\mathbf{z} \mid \mathbf{x})$ with a tractable variational distribution $q(\mathbf{z})$.
- Select a variational family $\mathcal{Q}$ of candidate distributions.
- Use KL-divergence to measure dissimilarity between distributions.
- Find the best approximating distribution q^* in $\mathcal{Q}$ by solving

$$\begin{aligned} q^* &= \arg\min_{q \in \mathcal{Q}} \mathbb{KL}(q(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x})) \\ &= \arg\min_{q \in \mathcal{Q}} \mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z}) - \log p(\mathbf{z} \mid \mathbf{x})] \end{aligned}$$

- This is equivalent to maximizing the ELBO

$$\begin{aligned} q^* &= \arg\max_{q \in \mathcal{Q}} \mathcal{L}(q) \\ &= \arg\max_{q \in \mathcal{Q}} \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z})] \end{aligned}$$

Inference as optimization

We want to find the "best" distribution q^* in $\mathcal{Q}$ that maximizes the ELBO

$$q^* = \arg\max_{q \in \mathcal{Q}} \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z})]$$

If we choose $\mathcal{Q}$ to be some parametric family $\mathcal{Q} = \{q(\mathbf{z} \mid \boldsymbol{\nu})\}_{\boldsymbol{\nu}}$, we simply have to find the "best" parameters $\boldsymbol{\nu}^*$ that maximize the ELBO

$$\boldsymbol{\nu}^* = \arg\max_{\boldsymbol{\nu}} \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\nu})}[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z} \mid \boldsymbol{\nu})]$$

Figure: Blei et al. - "Variational Inference: Foundations and Modern Methods"

Inference as optimization

We want to find the "best" distribution q^* in $\mathcal{Q}$ that maximizes the ELBO

$$q^* = \arg\max_{q \in \mathcal{Q}} \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z})]$$

If we choose $\mathcal{Q}$ to be some parametric family $\mathcal{Q} = \{q(\mathbf{z} \mid \boldsymbol{\nu})\}_{\boldsymbol{\nu}}$, we simply have to find the "best" parameters $\boldsymbol{\nu}^*$ that maximize the ELBO

$$\boldsymbol{\nu}^* = \arg\max_{\boldsymbol{\nu}} \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\nu})}[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z} \mid \boldsymbol{\nu})]$$

We turned inference into an optimization problem!

ELBO is just a function of $\boldsymbol{\nu}$, we can simply use our favorite optimization algorithm, such as Gradient Descent.

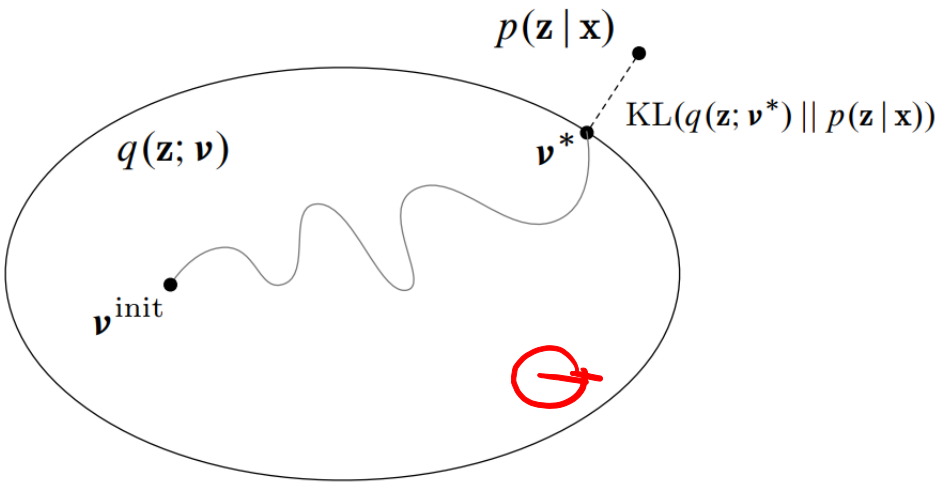
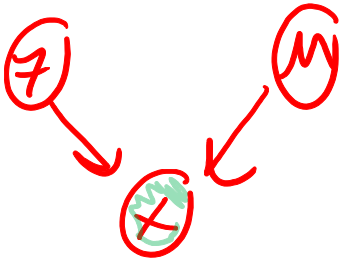
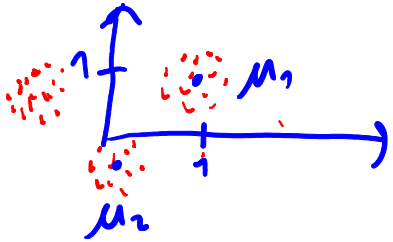


Figure: Blei et al. - "Variational Inference: Foundations and Modern Methods"

Bayesian (simplified) GMM with variational inference

Model





- Priors

1) GENERATE μ_1, μ_2, μ_3

$$p(\mathbf{z}_i) = \text{Cat}(1/K, ..., 1/K) \qquad p(\mathbf{Z}) = \prod_{i=1}^{N} p(\mathbf{z}_i)$$

2) FOR EVERY DATAPOINT i

$$p(\mu_k) = \mathcal{N}(\mu_k \mid 0, \sigma^2) \qquad p(\boldsymbol{\mu}) = \prod_{k=1}^{K} p(\mu_k)$$

a) GENERATE $z_i \sim \text{CAT}(\cdot)$

- Probability of a single sample

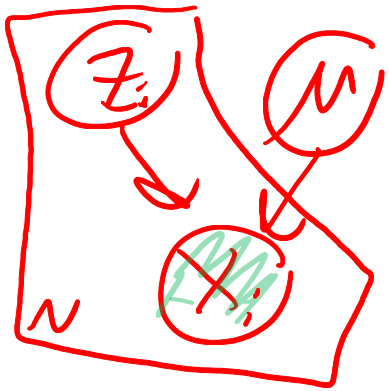
b) GENERATE $x_i \sim \mathcal{N}(\mu_{z_i}, 1)$

$$p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu}) = \mathcal{N}(x_i \mid \mathbf{z}_i^T \boldsymbol{\mu}, 1) \quad = \mathcal{N}(x_i \mid \mu_{z_i}, 1)$$

- Joint probability of the entire dataset

$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\mu}) = \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{N} p(\mathbf{z}_i) p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu})$$

Choosing the variational distribution



We are interested in the true posterior

$$p(\mathbf{Z}, \boldsymbol{\mu} \mid \mathbf{X}) = p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\mu})/p(\mathbf{X})$$

To make our life easier, we choose a $q(\mathbf{Z}, \boldsymbol{\mu})$ that fully factorizes

$$q(\mathbf{Z}, \boldsymbol{\mu} \mid \mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi}) = \prod_{k=1}^{K} q(\mu_k \mid m_k, s_k^2) \prod_{i=1}^{N} q(\mathbf{z}_i \mid \boldsymbol{\psi}_i),$$

where

$$q(\mu_k \mid m_k, s_k^2) = \mathcal{N}(\mu_k \mid m_k, s_k^2)$$

$$q(\mathbf{z}_i \mid \boldsymbol{\psi}_i) = \mathrm{Cat}(\mathbf{z}_i \mid \boldsymbol{\psi}_i)$$

with the variational parameters

$$\mathbf{m} = \{m_1, ..., m_K\},\ \mathbf{s}^2 = \{s_1^2, ..., s_K^2\},\ \boldsymbol{\Psi} = \{\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_N\}.$$

Note that it should hold $s_k^2 > 0$ for all k, and $\boldsymbol{\psi}_i \in \Delta^{K-1}$ for all i for $q(\mathbf{Z}, \boldsymbol{\mu} \mid \mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi})$ to be a valid probability distribution.

Mean field variational inference

A factorizable multivariate distribution can be written as a product

$$q(\mathbf{z} \mid \boldsymbol{\nu}) = q(z_1, ..., z_M \mid \boldsymbol{\nu}_1, ..., \boldsymbol{\nu}_M) = \prod_{j=1}^{M} q_j(z_j \mid \boldsymbol{\nu}_j)$$

Each component z_j is governed by its own distribution $q_j(\cdot \mid \boldsymbol{\nu}_j)$.

$q(z_1) \cdot q(z_2) \approx \mu(z_1, z_2 \mid x)$

What does this mean for our variational posterior?

- Factorized q doesn't capture correlations between z_j's.
- This limits the expressiveness, but simplifies optimization.
- In the context of VI, choosing a factorizable q is called mean field variational inference.

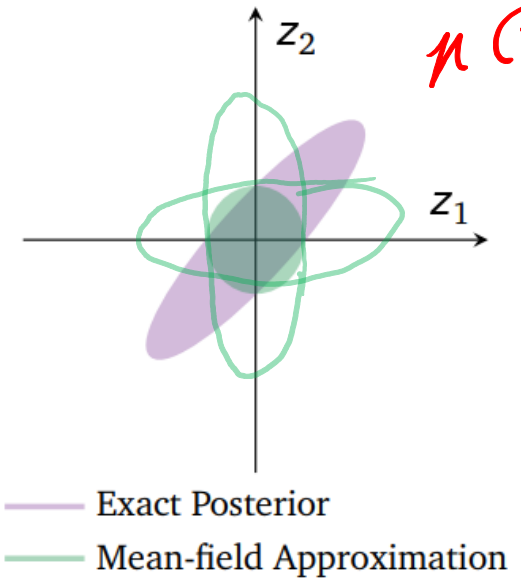


Figure from Blei, Kucukelbir and McAuliffe -" Variational Inference: A Review for Statisticians "

VI optimization problem for simple GMM

We want to find $q(\mathbf{Z}, \boldsymbol{\mu} \mid \mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi})$ that is close to $p(\mathbf{Z}, \boldsymbol{\mu} \mid \mathbf{X})$ in terms of KL-divergence. For this we need to maximize the ELBO

$$\max_{\mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi}} \mathcal{L}(\mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi}).$$

Let's denote the variational parameters as

$$\boldsymbol{\nu} = \{\mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi}\} = \{m_1, ..., m_K, s_1^2, ..., s_K^2, \boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_N\}.$$

We need to solve the following optimization problem

$$\max_{\boldsymbol{\nu}} \mathcal{L}(\boldsymbol{\nu})$$

VI optimization problem for simple GMM

We want to find $q(\mathbf{Z}, \boldsymbol{\mu} \mid \mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi})$ that is close to $p(\mathbf{Z}, \boldsymbol{\mu} \mid \mathbf{X})$ in terms of KL-divergence. For this we need to maximize the ELBO

$$\max_{\mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi}} \mathcal{L}(\mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi}).$$

Let's denote the variational parameters as

$$\boldsymbol{\nu} = \{\mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi}\} = \{m_1, ..., m_K, s_1^2, ..., s_K^2, \boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_N\}.$$

We need to solve the following optimization problem

$$\max_{\boldsymbol{\nu}} \mathcal{L}(\boldsymbol{\nu})$$

Let's do it with gradient ascent!

For this we need to compute the gradient of ELBO w.r.t. $\boldsymbol{\nu}$

$$\nabla_{\boldsymbol{\nu}} \mathcal{L}(\boldsymbol{\nu}) = \nabla_{\boldsymbol{\nu}} \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\nu})} [\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z} \mid \boldsymbol{\nu})]$$

Gradient of the ELBO

We want to compute the gradient

$$\nabla_{\boldsymbol{\nu}} \mathcal{L}(\boldsymbol{\nu}) = \nabla_{\boldsymbol{\nu}} \underbrace{\mathbb{E}_{q(\mathbf{z}|\boldsymbol{\nu})}[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z} \mid \boldsymbol{\nu})]}_{\text{we need to find a closed-form expression for this expectation}}$$

The easiest way to proceed is by finding a closed-form expression for the expectation and then computing the gradient.

Let's write down the ELBO for our problem and simplify it.

ELBO

The general expression for ELBO is

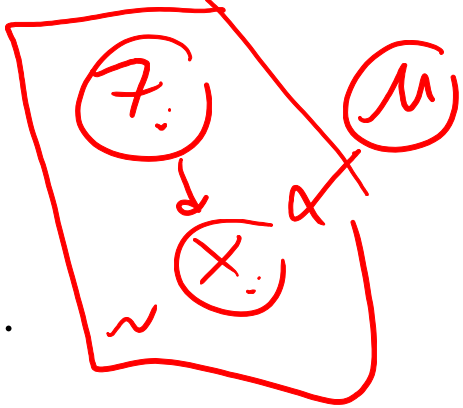
$$\mathcal{L}(\boldsymbol{\nu}) = \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\nu})}[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z} \mid \boldsymbol{\nu})].$$

For our problem this translates to

$$\mathcal{L}(\mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi}) = \mathbb{E}_q \left[\log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\mu})\right] - \mathbb{E}_q \left[\log q(\mathbf{Z}, \boldsymbol{\mu} \mid \mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi})\right]$$

ELBO

$q(z, \mu | \ldots) = q(z | \ldots) \cdot q(\mu | \ldots)$



The general expression for ELBO is

$$\mathcal{L}(\boldsymbol{\nu}) = \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\nu})}[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z} \mid \boldsymbol{\nu})].$$

For our problem this translates to

$$\mathcal{L}(\mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi}) = \mathbb{E}_q \left[\log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\mu})\right] - \mathbb{E}_q \left[\log q(\mathbf{Z}, \boldsymbol{\mu} \mid \mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi})\right]$$

$$= \mathbb{E}_q \left[\log p(\boldsymbol{\mu})\right] + \mathbb{E}_q \left[\log p(\mathbf{Z})\right] + \mathbb{E}_q \left[\log p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\mu})\right]$$
$$- \mathbb{E}_q \left[\log q(\mathbf{Z} \mid \boldsymbol{\Psi})\right] - \mathbb{E}_q \left[\log q(\boldsymbol{\mu} \mid \mathbf{m}, \mathbf{s}^2)\right]$$

ELBO

The general expression for ELBO is

$$\mathcal{L}(\boldsymbol{\nu}) = \mathbb{E}_{q(\mathbf{z}|\boldsymbol{\nu})}[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z} \mid \boldsymbol{\nu})].$$

For our problem this translates to

$$\begin{aligned}
\mathcal{L}(\mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi}) &= \mathbb{E}_q\left[\log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\mu})\right] - \mathbb{E}_q\left[\log q(\mathbf{Z}, \boldsymbol{\mu} \mid \mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi})\right] \\
&= \mathbb{E}_q\left[\log p(\boldsymbol{\mu})\right] + \mathbb{E}_q\left[\log p(\mathbf{Z})\right] + \mathbb{E}_q\left[\log p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\mu})\right] \\
&\quad - \mathbb{E}_q\left[\log q(\mathbf{Z} \mid \boldsymbol{\Psi})\right] - \mathbb{E}_q\left[\log q(\boldsymbol{\mu} \mid \mathbf{m}, \mathbf{s}^2)\right] \\
&= \sum_{k=1}^{K} \mathbb{E}_q[\log p(\mu_k)] \\
&\quad + \sum_{i=1}^{N} \left(\mathbb{E}_q[\log p(\mathbf{z}_i)] + \mathbb{E}_q[\log p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu})]\right) \\
&\quad - \sum_{i=1}^{N} \mathbb{E}_q[\log q(\mathbf{z}_i \mid \boldsymbol{\psi}_i)] - \sum_{k=1}^{K} \mathbb{E}_q[\log q(\mu_k \mid m_k, s_k^2)].
\end{aligned}$$

ELBO

Let's consider each term of ELBO one by one

$$\mathcal{L}(\mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi}) = \sum_{k=1}^{K} \mathbb{E}_q[\log p(\mu_k)] \tag{1}$$

$$+ \sum_{i=1}^{N} \mathbb{E}_q[\log p(\mathbf{z}_i)] \tag{2}$$

$$+ \sum_{i=1}^{N} \mathbb{E}_q[\log p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu})] \tag{3}$$

$$- \sum_{i=1}^{N} \mathbb{E}_q[\log q(\mathbf{z}_i \mid \boldsymbol{\psi}_i)] \tag{4}$$

$$- \sum_{k=1}^{K} \mathbb{E}_q[\log q(\mu_k \mid m_k, s_k^2)]. \tag{5}$$

(1) Simplifying $\sum_k \mathbb{E}_q[\log p(\mu_k)]$

$\mathbb{E}_q[\mu_k] = m_k$

$\mathbb{E}_q[\mu_k^2]$

The prior on μ_k is

σ^2

$$p(\mu_k) = \mathcal{N}(\mu_k \mid 0, 1) = 1/2 \exp(-\mu_k^2/2)$$

Therefore

$$\sum_{k=1}^{K} \mathbb{E}_q[\log p(\mu_k)] = \sum_{k=1}^{K} \mathbb{E}_q[\log(\exp(-\mu_k^2/2)) + \text{const.}]$$

$$= \sum_{k=1}^{K} \mathbb{E}_q[-\mu_k^2]/2 + \text{const.}$$

Since $q(\mu_k) = \mathcal{N}(\mu_k \mid m_k, s_k^2)$, we know that $\mathbb{E}_q[\mu_k^2] = m_k^2 + s_k^2$, so

$$= \boxed{-\frac{1}{2}\sum_{k=1}^{K}(m_k^2 + s_k^2)} + \text{const.}$$

$\mathbb{E}_q[\mu_k^2] = \int q(\mu_k) \cdot \mu_k^2 \, d\mu_k$

(2) Simplifying $\sum_i \mathbb{E}_q[\log p(\mathbf{z}_i)]$

ONE HOT VECTOR NOTATION

The prior on $\mathbf{z}_i$ is

$$p(\mathbf{z}_i) = \mathrm{Cat}(1/K, ..., 1/K) = \prod_{k=1}^{K} \left(\frac{1}{K}\right)^{z_{ik}}$$

Therefore

$$\begin{aligned}
\sum_{i=1}^{N} \mathbb{E}_q[\log p(\mathbf{z}_i)] &= \sum_{i=1}^{N} \mathbb{E}_q \left[\log \left(\prod_{k=1}^{K} \left(\frac{1}{K}\right)^{z_{ik}}\right)\right] \\
&= \sum_{i=1}^{N} \mathbb{E}_q \left[\sum_{k=1}^{K} z_{ik} \log \frac{1}{K}\right] \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{E}_q[z_{ik}] \cdot (-\log K) \\
&= \boxed{-\log K \cdot \sum_{i=1}^{N} \sum_{k=1}^{K} \psi_{ik}}
\end{aligned}$$

$q(z_i \mid \psi_i) = \mathrm{CAT}(z_i \mid \psi_i)$

$q(z_i = 1 \mid \psi_i) = \psi_{i1}$

$q(z_i = 2 \mid \psi_i) = \psi_{i2}$

$\vdots$

$\mathbb{E}_q[z_{ik}]$

$= \mathbb{E}_q[z_i = k]$

$= \sum_{z_i} q(z_i) \cdot [z_i = k]$

$= \sum_{k'} q(z_i = k') \cdot [z_i = k]$

$\psi_{ik} = q(z_i = k) = q(z_i = k) \cdot [z_i = k]$

(3) Simplifying $\sum_i \mathbb{E}_q[\log p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu})]$

The likelihood for a single sample is

$$p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu}) = \mathcal{N}(x_i \mid \mathbf{z}_i^T \boldsymbol{\mu}, 1) = \prod_{k=1}^{K} \left(\mathcal{N}(x_i \mid \mu_k, 1)\right)^{z_{ik}}$$

Therefore

$$\sum_{i=1}^{N} \mathbb{E}_q[\log p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu})] = \sum_{i=1}^{N} \mathbb{E}_q \left[\log \left(\prod_{k=1}^{K} \left(\mathcal{N}(x_i \mid \mu_k, 1)\right)^{z_{ik}}\right)\right]$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{E}_q \left[z_{ik} \log \left(\mathcal{N}(x_i \mid \mu_k, 1)\right)\right]$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{E}_q \left[z_{ik}\left(-1/2(x_i - \mu_k)^2\right)\right] + \text{const.}$$

$$= -\frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{K} \psi_{ik}(x_i^2 - 2x_i m_k + m_k^2 + s_k^2) + \text{const.}$$

(4) Simplifying $-\sum_i \mathbb{E}_q[\log q(\mathbf{z}_i \mid \boldsymbol{\psi}_i)]$

We chose the variational distribution for $\mathbf{z}_i$ as

$$q(\mathbf{z}_i \mid \boldsymbol{\psi}_i) = \mathrm{Cat}(\boldsymbol{\psi}_i) = \prod_{k=1}^{K} \psi_{ik}^{z_{ik}}$$

The term we want to simplify involves the entropy of this distribution.

$$\begin{aligned}
\mathbb{H}[q(\mathbf{z}_i \mid \boldsymbol{\psi}_i)] &= -\mathbb{E}_q[\log q(\mathbf{z}_i \mid \boldsymbol{\psi}_i)] \\
&= -\sum_{k=1}^{K} \mathbb{E}_q[z_{ik}] \log \psi_{ik} \\
&= -\sum_{k=1}^{K} \psi_{ik} \log \psi_{ik}
\end{aligned}$$

Which means that

$$-\sum_{i=1}^{N} \mathbb{E}_q[\log q(\mathbf{z}_i \mid \boldsymbol{\psi}_i)] = \boxed{-\sum_{i=1}^{N}\sum_{k=1}^{K} \psi_{ik} \log \psi_{ik}}$$

(5) Simplifying $-\sum_k \mathbb{E}_q[\log q(\mu_k \mid m_k, s_k^2)]$

We chose the variational distribution for μ_k as

$$q(\mu_k \mid m_k, s_k^2) = \mathcal{N}(\mu_k \mid m_k, s_k^2) = \frac{1}{\sqrt{2\pi s_k^2}} \exp\left(-\frac{(\mu_k - m_k)^2}{2s_k^2}\right)$$

The term we want to simplify involves the entropy of this distribution.

$$\begin{aligned}
\mathbb{H}\left[q(\mu_k \mid m_k, s_k^2)\right] &= -\mathbb{E}_q\left[\log q(\mu_k \mid m_k, s_k^2)\right] \\
&= -\mathbb{E}_q\left[-\frac{1}{2}\log(2\pi s_k^2) - \frac{\mu_k^2 - 2\mu_k m_k + m_k^2}{2s_k^2}\right] \\
&= \frac{1}{2}\log(2\pi s_k^2) + 1
\end{aligned}$$

Which means that

$$-\sum_{k=1}^{K} \mathbb{E}_q[\log q(\mu_k \mid m_k, s_k^2)] = \boxed{\frac{1}{2}\sum_{k=1}^{K} \log s_k^2} + \text{const.}$$

Putting everything together

By combining parts $(1) - (5)$, we get the closed-form expression for ELBO

$$\begin{aligned}\mathcal{L}(\mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi}) = &-\frac{1}{2}\sum_{k=1}^{K}(m_k^2 + s_k^2) - \log K \cdot \sum_{i=1}^{N}\sum_{k=1}^{K}\psi_{ik} \\ &- \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{K}\psi_{ik}(x_i^2 - 2x_i m_k + m_k^2 + s_k^2) \\ &- \sum_{i=1}^{N}\sum_{k=1}^{K}\psi_{ik}\log\psi_{ik} + \frac{1}{2}\sum_{k=1}^{K}\log s_k^2\end{aligned}$$

This is just a function of the variational parameters $\boldsymbol{\nu} = \{\mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi}\}$.

We can compute the gradients of this expression (or use TensorFlow, etc).

Putting everything together

By combining parts $(1) - (5)$, we get the closed-form expression for ELBO

$$\begin{aligned}\mathcal{L}(\mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi}) = &-\frac{1}{2}\sum_{k=1}^{K}(m_k^2 + s_k^2) - \log K \cdot \sum_{i=1}^{N}\sum_{k=1}^{K}\psi_{ik} \\ &- \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{K}\psi_{ik}(x_i^2 - 2x_i m_k + m_k^2 + s_k^2) \\ &- \sum_{i=1}^{N}\sum_{k=1}^{K}\psi_{ik}\log\psi_{ik} + \frac{1}{2}\sum_{k=1}^{K}\log s_k^2\end{aligned}$$

This is just a function of the variational parameters $\boldsymbol{\nu} = \{\mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi}\}$.

We can compute the gradients of this expression (or use TensorFlow, etc).

However, there is one more thing to do before we can use gradient ascent.

Reparametrization trick

Our optimization problem is actually constrained:

We need to make sure that $\boldsymbol{\psi}_i \in \Delta^{K-1}$ and $s_k^2 > 0$ for all i and k.

We could use Lagrange multipliers or projected gradient ascent.

But there exists an easier way!

Reparametrization trick

Our optimization problem is actually constrained:

We need to make sure that $\boldsymbol{\psi}_i \in \Delta^{K-1}$ and $s_k^2 > 0$ for all i and k.

We could use Lagrange multipliers or projected gradient ascent.

But there exists an easier way!

Removing the constraints with the reparametrization trick

- We define a new variable $\boldsymbol{a}_i \in \mathbb{R}^K$ for each $\boldsymbol{\psi}_i \in \Delta^{k-1}$.
- We set $\boldsymbol{\psi}_i$ to be a function of $\boldsymbol{a}_i$

$$\boldsymbol{\psi}_i(\boldsymbol{a}_i) = \text{softmax}(\boldsymbol{a}_i)$$

- Optimization is now done over $\boldsymbol{a}_i$ (unconstrained!)
- Since $\boldsymbol{\psi}_i$ is output of a softmax, we get $\sum_{k=1}^{K} \boldsymbol{\psi}_{ik} = 1$ for free.
- If we know $\nabla_{\boldsymbol{\psi}_i} \mathcal{L}(\boldsymbol{\psi})$, we can obtain $\nabla_{\boldsymbol{a}_i} \mathcal{L}(\boldsymbol{\psi}(\boldsymbol{a}))$ by chain rule!

Reparametrization trick

Our optimization problem is actually constrained:

We need to make sure that $\boldsymbol{\psi}_i \in \Delta^{K-1}$ and $s_k^2 > 0$ for all i and k.

We could use Lagrange multipliers or projected gradient ascent.

But there exists an easier way!

Removing the constraints with the reparametrization trick

- We define a new variable $\boldsymbol{a}_i \in \mathbb{R}^K$ for each $\boldsymbol{\psi}_i \in \Delta^{k-1}$.
- We set $\boldsymbol{\psi}_i$ to be a function of $\boldsymbol{a}_i$

$$\boldsymbol{\psi}_i(\boldsymbol{a}_i) = \text{softmax}(\boldsymbol{a}_i)$$

- Optimization is now done over $\boldsymbol{a}_i$ (unconstrained!)
- Since $\boldsymbol{\psi}_i$ is output of a softmax, we get $\sum_{k=1}^{K} \boldsymbol{\psi}_{ik} = 1$ for free.
- If we know $\nabla_{\boldsymbol{\psi}_i} \mathcal{L}(\boldsymbol{\psi})$, we can obtain $\nabla_{\boldsymbol{a}_i} \mathcal{L}(\boldsymbol{\psi}(\boldsymbol{a}))$ by chain rule!

Think on your own: how to enforce the $s_k^2 > 0$ constraint with the RT?

Direct optimization of ELBO

We want to solve the optimization problem

$$\boldsymbol{\nu}^* = \arg\max_{\boldsymbol{\nu}} \mathcal{L}(\boldsymbol{\nu})$$

to find a good approximation $q(\mathbf{z} \mid \boldsymbol{\nu}^*)$ to the posterior $p(\mathbf{z} \mid \mathbf{x})$.

For this, we take the following steps

1. Write down the ELBO.
2. Simplify all the $\mathbb{E}_q$'s and get a closed-form expression for $\mathcal{L}(\boldsymbol{\nu})$.
3. Compute the gradients $\nabla_{\boldsymbol{\nu}} \mathcal{L}(\boldsymbol{\nu})$.
4. Apply reparametrization trick if necessary.
5. Do gradient ascent.

Alternatives to direct optimization

Drawbacks of direct optimization

- We had to assume that $q(\mathbf{z} \mid \boldsymbol{\nu})$ takes a specific form, e.g. $q(\mu_k)$ is normal and $q(\mathbf{z}_i)$ categorical.
- If we make wrong assumptions about q, this will lead to bad results.
- Gradient ascent requires setting hyperparameters (learning rate, momentum, tolerance, etc).

There exists a different method that

- Makes fewer assumptions.
- Has (almost) no hyperparameters.

Recall: Mean field variational inference

A factorizable multivariate distribution can be written as a product

$$q(\mathbf{z} \mid \boldsymbol{\nu}) = q(z_1, ..., z_M \mid \boldsymbol{\nu}_1, ..., \boldsymbol{\nu}_M) = \prod_{j=1}^{M} q_j(z_j \mid \boldsymbol{\nu}_j)$$

Each component z_j is governed by its own distribution $q_j(\cdot \mid \boldsymbol{\nu}_j)$.

What does this mean for our variational posterior?

- Factorized q doesn't capture correlations between z_j's.
- This limits the expressiveness, but simplifies optimization.
- This assumption is enough to derive a new algorithm for VI.

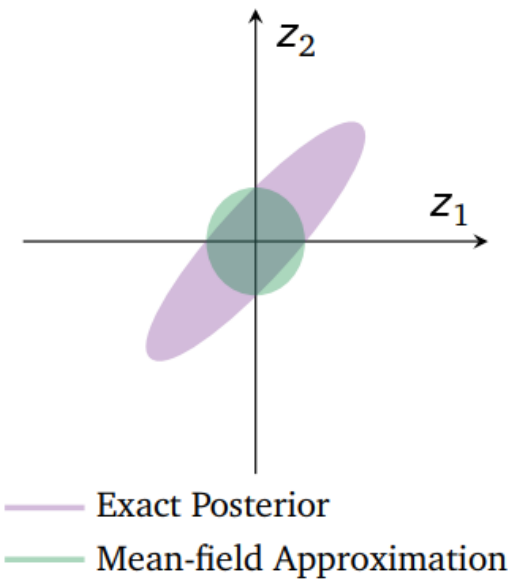


Figure from Blei, Kucukelbir and McAuliffe -" Variational Inference: A Review for Statisticians "

Coordinate ascent mean-field variational inference

So what do we gain from the factorization assumption?

- The resulting joint optimization problem is still non-convex.

$$\max_{q_1,\ldots,q_M} \mathcal{L}(q_1, \ldots, q_M)$$

- However, each subproblem

$$q_j^*(z_j) = \arg\max_{q_j} \mathcal{L}(q_j)$$

has a closed form solution

$$q_j^*(z_j) \propto \exp\left(\mathbb{E}_{q_{-j}}\left[\log p(\mathbf{x}, \mathbf{z})\right]\right)$$

where

$$\mathbb{E}_{q_{-j}}\left[\log p(\mathbf{x}, \mathbf{z})\right] = \int \log p(\mathbf{x}, \mathbf{z}) \prod_{i \neq j} q_i(z_i) dz_i.$$

Proof: Optimal solution for q_j

The ELBO is given as

$$\mathcal{L}(q) = \mathbb{E}_q[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z})]$$

Only keep the terms with q_j. We use iterated expectation for the first term, and mean-field (factorization) assumption for the second.

$$= \mathbb{E}_{q_j}\left[\mathbb{E}_{q_{-j}}[\log p(\mathbf{x}, \mathbf{z})]\right] - \mathbb{E}_{q_j}[\log q_j(z_j)] - \underbrace{\mathbb{E}_{q_{-j}}\left[\sum_{i \neq j} \log q_i(z_i)\right]}_{\text{const. in } q_j}$$

This is equal to KL divergence between $q_j(z_j)$ and $\exp\left(\mathbb{E}_{q_{-j}}[\log p(\mathbf{x}, \mathbf{z})]\right)$

$$= -\mathbb{KL}\left(q_j(z_j) \parallel \exp\left(\mathbb{E}_{q_{-j}}[\log p(\mathbf{x}, \mathbf{z})]\right)\right) + \text{const.}$$

The KL divergence is minimized when the two distribution are equal

$$q_j^*(z_j) \propto \exp\left(\mathbb{E}_{q_{-j}}[\log p(\mathbf{x}, \mathbf{z})]\right)$$

which is exactly the optimal solution from the previous slide.

Coordinate ascent mean-field variational inference

Coordinate ascent variational inference (CAVI)

- Using this result, we can derive a simple iterative algorithm, where we sequentially update the factors q_j until ELBO converges.

$$q_j^*(z_j) \propto \exp\left(\mathbb{E}_{q_{-j}}\left[\log p(\mathbf{z}, \mathbf{x})\right]\right)$$

- The CAVI algorithm is guaranteed to converge, because ELBO increases in each iteration, and is also bounded above by $\log p(\mathbf{x})$.

$\mathbb{E}_q[q(z)] = \mathbb{E}_q\left[\prod_i q_i(z_i)\right]$

$= \prod_i \mathbb{E}_q[q_i(z_i)]$

$= \prod_i \mathbb{E}_{q_i}[q_i(z_i)]$

$KL(q,p)$

$= \sum q \cdot \log \frac{q}{p}$

$= \mathbb{E}_q\left[\log \frac{q}{p}\right]$

$= \mathbb{E}_q[\log q] - \mathbb{E}_q[\log p]$

Coordinate ascent mean-field variational inference

Coordinate ascent variational inference (CAVI)

- Using this result, we can derive a simple iterative algorithm, where we sequentially update the factors q_j until ELBO converges.

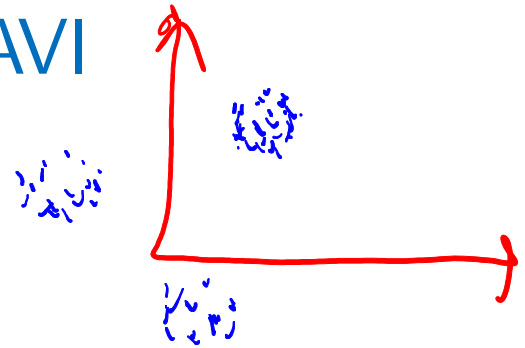
$$q_j^*(z_j) \propto \exp\left(\mathbb{E}_{q_{-j}}\left[\log p(\mathbf{z}, \mathbf{x})\right]\right)$$

- The CAVI algorithm is guaranteed to converge, because ELBO increases in each iteration, and is also bounded above by $\log p(\mathbf{x})$.

Free-form optimization

- The coolest thing is that we don't make any assumptions about the form of q (other than mean-field factorization).
- We don't have to think what distribution to choose for q - CAVI tells us what the optimal one is (almost) for free!

Again: Bayesian (simplified) GMM with CAVI



Model

- Priors

$$p(\mathbf{z}_i) = \text{Cat}(1/K, ..., 1/K) \qquad p(\mathbf{Z}) = \prod_{i=1}^{N} p(\mathbf{z}_i)$$

$q(\mu) \approx p(\mu \mid x)$

$$p(\mu_k) = \mathcal{N}(\mu_k \mid 0, \sigma^2) \qquad p(\boldsymbol{\mu}) = \prod_{k=1}^{K} p(\mu_k)$$

- Probability of a single sample

$$p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu}) = \mathcal{N}(x_i \mid \mathbf{z}_i^T \boldsymbol{\mu}, 1)$$

- Joint probability of the entire dataset

$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\mu}) = \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{N} p(\mathbf{z}_i) p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu})$$

Choosing the variational distribution

We are interested in the true posterior

$$p(\mathbf{Z}, \boldsymbol{\mu} \mid \mathbf{X}) = p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\mu})/p(\mathbf{X})$$

Using the mean-field approach, we choose the variational posterior as

$$q(\mathbf{Z}, \boldsymbol{\mu} \mid \mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi}) = \prod_{k=1}^{K} q(\mu_k \mid m_k, s_k^2) \prod_{i=1}^{N} q(\mathbf{z}_i \mid \boldsymbol{\psi}_i)$$

where

$$q(\mu_k \mid m_k, s_k^2) = \mathcal{N}(\mu_k \mid m_k, s_k^2)$$
$$q(\mathbf{z}_i \mid \boldsymbol{\psi}_i) = \mathrm{Cat}(\mathbf{z}_i \mid \boldsymbol{\psi}_i)$$

with the variational parameters

$$\mathbf{m} = \{m_1, ..., m_K\},\ \mathbf{s}^2 = \{s_1^2, ..., s_K^2\},\ \boldsymbol{\Psi} = \{\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_N\}.$$

ELBO

We want to find $q(\mathbf{Z}, \boldsymbol{\mu} \mid \mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi}) \approx p(\mathbf{Z}, \boldsymbol{\mu} \mid \mathbf{X})$, for this we maximize

$$\mathcal{L}(\mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi}) = \mathbb{E}_q \left[\log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\mu})\right] - \mathbb{E}_q \left[\log q(\mathbf{Z}, \boldsymbol{\mu} \mid \mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi})\right]$$

$$= \mathbb{E}_q \left[\log p(\boldsymbol{\mu})\right] + \mathbb{E}_q \left[\log p(\mathbf{Z})\right] + \mathbb{E}_q \left[\log p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\mu})\right]$$
$$- \mathbb{E}_q \left[\log q(\mathbf{Z} \mid \boldsymbol{\Psi})\right] - \mathbb{E}_q \left[\log q(\boldsymbol{\mu} \mid \mathbf{m}, \mathbf{s}^2)\right]$$

$$= \sum_{k=1}^{K} \mathbb{E}_q[\log p(\mu_k)]$$
$$+ \sum_{i=1}^{N} \left(\mathbb{E}_q[\log p(\mathbf{z}_i)] + \mathbb{E}_q[\log p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu})]\right)$$
$$- \sum_{i=1}^{N} \mathbb{E}_q[\log q(\mathbf{z}_i \mid \boldsymbol{\psi}_i)] - \sum_{k=1}^{K} \mathbb{E}_q[\log q(\mu_k \mid m_k, s_k^2)].$$

Update for $q(\mathbf{z}_i)$

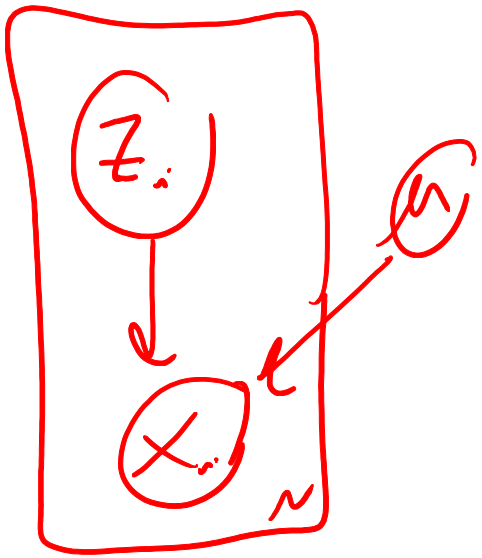
The optimal mean-field update is

$$q^*(\mathbf{z}_i \mid \boldsymbol{\psi}_i) \propto \exp\left(\mathbb{E}_{q(\boldsymbol{\mu})}\left[\mathbb{E}_{q(\mathbf{z}_1,\ldots,\mathbf{z}_{i-1},\mathbf{z}_{i+1},\ldots)}\left[\log p(\mathbf{X},\mathbf{Z},\boldsymbol{\mu})\right]\right]\right)$$

Terms dependent on $\mathbf{z}_j$ for $j \neq i$ can be absorbed into the $\propto$ sign

$$\propto \exp\left(\log p(\mathbf{z}_i) + \mathbb{E}_{q(\boldsymbol{\mu})}\left[\log p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu})\right]\right)$$

Notice, that we don't take expectation w.r.t. $q(\mathbf{z}_i)$.



Update for $q(\mathbf{z}_i)$

The optimal mean-field update is

$$q^*(\mathbf{z}_i \mid \boldsymbol{\psi}_i) \propto \exp\left(\mathbb{E}_{q(\boldsymbol{\mu})}\left[\mathbb{E}_{q(\mathbf{z}_1,\ldots,\mathbf{z}_{i-1},\mathbf{z}_{i+1},\ldots)}\left[\log p(\mathbf{X},\mathbf{Z},\boldsymbol{\mu})\right]\right]\right)$$

Terms dependent on $\mathbf{z}_j$ for $j \neq i$ can be absorbed into the $\propto$ sign

$$\propto \exp\left(\log p(\mathbf{z}_i) + \mathbb{E}_{q(\boldsymbol{\mu})}\left[\log p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu})\right]\right)$$

Notice, that we don't take expectation w.r.t. $q(\mathbf{z}_i)$.

The uniform prior

$$\log p(\mathbf{z}_i) = -\log K$$

can be absorbed into the $\propto$ sign.

What about the expected log-likelihood $\mathbb{E}_{q(\boldsymbol{\mu})}\left[\log p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu})\right]$?

Update for $q(\mathbf{z}_i)$

Recalling that $\mathbf{z}_i$ is an indicator vector, we can write

$$\mathbb{E}_{q(\boldsymbol{\mu})}\left[\log p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu})\right] = \sum_{k=1}^{K} z_{ik} \mathbb{E}_{q(\boldsymbol{\mu})}\left[\log p(x_i \mid \mu_k)\right]$$

Update for $q(\mathbf{z}_i)$

Recalling that $\mathbf{z}_i$ is an indicator vector, we can write

$$\begin{aligned}\mathbb{E}_{q(\boldsymbol{\mu})}\left[\log p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu})\right] &= \sum_{k=1}^{K} z_{ik} \mathbb{E}_{q(\boldsymbol{\mu})}\left[\log p(x_i \mid \mu_k)\right] \\ &= \sum_{k=1}^{K} z_{ik} \mathbb{E}_{q(\boldsymbol{\mu})}\left[-(x_i - \mu_k)^2/2\right] + \text{const.}\end{aligned}$$

Update for $q(\mathbf{z}_i)$

Recalling that $\mathbf{z}_i$ is an indicator vector, we can write

$$\begin{aligned}
\mathbb{E}_{q(\boldsymbol{\mu})}\left[\log p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu})\right] &= \sum_{k=1}^{K} z_{ik} \mathbb{E}_{q(\boldsymbol{\mu})}\left[\log p(x_i \mid \mu_k)\right] \\
&= \sum_{k=1}^{K} z_{ik} \mathbb{E}_{q(\boldsymbol{\mu})}\left[-(x_i - \mu_k)^2/2\right] + \text{const.} \\
&= \sum_{k=1}^{K} z_{ik} \left(\mathbb{E}_{q(\boldsymbol{\mu})}\left[\mu_k\right] x_i - \mathbb{E}_{q(\boldsymbol{\mu})}\left[\mu_k^2\right]/2\right) + \text{const.}
\end{aligned}$$

Update for $q(\mathbf{z}_i)$

Recalling that $\mathbf{z}_i$ is an indicator vector, we can write

$$\mathbb{E}_{q(\boldsymbol{\mu})}\left[\log p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu})\right] = \sum_{k=1}^{K} z_{ik} \mathbb{E}_{q(\boldsymbol{\mu})}\left[\log p(x_i \mid \mu_k)\right]$$

$$= \sum_{k=1}^{K} z_{ik} \mathbb{E}_{q(\boldsymbol{\mu})}\left[-(x_i - \mu_k)^2/2\right] + \text{const.}$$

$$= \sum_{k=1}^{K} z_{ik} \left(\mathbb{E}_{q(\boldsymbol{\mu})}\left[\mu_k\right] x_i - \mathbb{E}_{q(\boldsymbol{\mu})}\left[\mu_k^2\right]/2\right) + \text{const.}$$

We already assumed, that $\mu_k \sim \mathcal{N}(m_k, s_k^2)$, so we know that

$$\mathbb{E}_{q(\boldsymbol{\mu})}\left[\mu_k\right] = m_k \quad \text{and} \quad \mathbb{E}_{q(\boldsymbol{\mu})}\left[\mu_k^2\right] = m_k^2 + s_k^2$$

and hence have the optimal update for $\boldsymbol{\psi}_i$

$$\psi_{ik}^* \propto \exp(m_k x_i - (m_k^2 + s_k^2)/2)$$

We get rid of $\propto$ by computing each ψ_{ik}, and then dividing by $\sum_k \psi_{ik}$.

Updates for $q(\mu_k)$

$\mathbb{E}_{q \cdot \mu_k}[\dots]$

Again, using the formula for optimal mean-field update we get

$$q^*(\mu_k \mid m_k, s_k^2) \propto \exp\left(\log p(\mu_k) + \sum_{i=1}^{N} \mathbb{E}_{q(\mathbf{Z})}\left[\log p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu})\right]\right)$$

To avoid clutter, we work with the unnormalized logarithm

$$\begin{aligned}\log q^*(\mu_k) &= \log p(\mu_k) + \sum_{i=1}^{N} \mathbb{E}_{q(\mathbf{Z})}\left[\log p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu})\right] + \text{const.} \\ &= \log p(\mu_k) + \sum_{i=1}^{N} \mathbb{E}_{q(\mathbf{Z})}\left[z_{ik} \log p(x_i \mid \mu_k)\right] + \text{const.}\end{aligned}$$

$\prod_k p(x_i \mid \mu_k)^{z_{ik}}$

Updates for $q(\mu_k)$

Again, using the formula for optimal mean-field update we get

$$q^*(\mu_k \mid m_k, s_k^2) \propto \exp\left(\log p(\mu_k) + \sum_{i=1}^{N} \mathbb{E}_{q(\mathbf{Z})}\left[\log p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu})\right]\right)$$

To avoid clutter, we work with the unnormalized logarithm

$$\begin{aligned}
\log q^*(\mu_k) &= \log p(\mu_k) + \sum_{i=1}^{N} \mathbb{E}_{q(\mathbf{Z})}\left[\log p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu})\right] + \text{const.} \\
&= \log p(\mu_k) + \sum_{i=1}^{N} \mathbb{E}_{q(\mathbf{Z})}\left[z_{ik} \log p(x_i \mid \mu_k)\right] + \text{const.} \\
&= -\mu_k^2/2\sigma^2 + \sum_{i=1}^{N} \cancel{\phi_{ik}} \log p(x_i \mid \mu_k) + \text{const.} \qquad \mathbb{E}_{q(\mathbf{z})}[z_{ik}] \\
&= -\mu_k^2/2\sigma^2 + \sum_{i=1}^{N} \cancel{\phi_{ik}}(-(x_i - \mu_k)^2/2) + \text{const.} \qquad \mathbb{E}_{q(\mathbf{z})}[z_{ik}]
\end{aligned}$$

Updates for $q(\mu_k)$

$$\begin{aligned}
\log q^*(\mu_k) &= \log p(\mu_k) + \sum_{i=1}^{N} \mathbb{E}_{q(\mathbf{Z})} \left[\log p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu})\right] + \text{const.} \\
&= -\mu_k^2/2\sigma^2 + \sum_{i=1}^{N} (\psi_{ik} x_i \mu_k - \psi_{ik} \mu_k^2/2) + \text{const.} \\
&= \left(\sum_{i=1}^{N} \psi_{ik} x_i\right) \mu_k - \left(1/2\sigma^2 + \sum_{i=1}^{N} \psi_{ik}/2\right) \mu_k^2 + \text{const.}
\end{aligned}$$

Updates for $q(\mu_k)$

$$\psi_{ik} = \mathbb{E}_q[z_{ik}]$$

$$\log q^*(\mu_k) = \log p(\mu_k) + \sum_{i=1}^{N} \mathbb{E}_{q(\mathbf{Z})} \left[\log p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu})\right] + \text{const.}$$

$$= -\mu_k^2/2\sigma^2 + \sum_{i=1}^{N} (\psi_{ik} x_i \mu_k - \psi_{ik}\mu_k^2/2) + \text{const.}$$

$$= \left(\sum_{i=1}^{N} \psi_{ik} x_i\right) \mu_k - \left(1/2\sigma^2 + \sum_{i=1}^{N} \psi_{ik}/2\right) \mu_k^2 + \text{const.}$$

η_1

We see that $q(\mu_k)$ is a Gaussian by completing the square and rewriting the log-density as

$$= -\frac{1}{2s_k^2}(\mu_k - m_k)^2$$

η_1

$$-\frac{\eta_1}{2\eta_2} = m = f(\eta_1, \eta_2)$$

$$-\frac{1}{2\eta_2} = s^2 = f'(\eta_1, \eta_2)$$

where

$$m_k = \frac{\sum_{i=1}^{N} \psi_{ik} x_i}{1/\sigma^2 + \sum_{i=1}^{N} \psi_{ik}} \quad \text{and} \quad s_k^2 = \frac{1}{1/\sigma^2 + \sum_{i=1}^{N} \psi_{ik}}$$

Free form optimization in action

We assumed in the very beginning that $q(\mu_k)$ was Gaussian. So it came as no surprise that we had to complete the square to get the normalized distribution.

However, even if we didn't make this assumption, the expression

$$q^*(\mu_k) \propto \exp\left(\left(\sum_{i=1}^{N} \psi_{ik} x_i\right) \mu_k - \left(1/2\sigma^2 + \sum_{i=1}^{N} \psi_{ik}\right) \mu_k^2\right)$$

tells us, that the optimal $q^*(\mu_k)$ is an exponential family distribution with sufficient statistics $\{\mu_k, \mu_k^2\}$.

The result is somewhat less exciting for $q(\mathbf{z}_i)$, because any discrete distribution on a finite support can be represented as categorical.

Free form optimization in action

We assumed in the very beginning that $q(\mu_k)$ was Gaussian. So it came as no surprise that we had to complete the square to get the normalized distribution.

However, even if we didn't make this assumption, the expression

$$q^*(\mu_k) \propto \exp\left(\left(\sum_{i=1}^{N} \psi_{ik} x_i\right) \mu_k - \left(1/2\sigma^2 + \sum_{i=1}^{N} \psi_{ik}\right) \mu_k^2\right)$$

tells us, that the optimal $q^*(\mu_k)$ is an exponential family distribution with sufficient statistics $\{\mu_k, \mu_k^2\}$.

There exists only one such distribution, and it's... Gaussian!

Even if we didn't make any assumptions about q in the beginning except factorization, we would still get the same result!

The result is somewhat less exciting for $q(\mathbf{z}_i)$, because any discrete distribution on a finite support can be represented as categorical.

CAVI algorithm for simplified GMM

Algorithm 1: CAVI for a Gaussian mixture model

Input: Data $x_{1:N}$, number of components K, prior variance of component means σ^2

Output: Variational densities $q(\mu_k \mid m_k, s_k^2)$ (Gaussian) and $q(\mathbf{z}_i \mid \boldsymbol{\psi}_i)$ (K-categorical)

Initialize: Variational parameters $\mathbf{m} = m_{1:K}$, $\mathbf{s}^2 = s^2_{1:K}$, and $\boldsymbol{\Psi} = \boldsymbol{\psi}_{1:N}$

while *the ELBO has not converged* **do**

  for $i \in \{1, \ldots, N\}$ **do**

    Set $\psi_{ik} \propto \exp(m_k x_i - (m_k^2 + s_k^2)/2)$

  end

  for $k \in \{1, \ldots, K\}$ **do**

$$\text{Set } m_k \longleftarrow \frac{\sum_i \psi_{ik} x_i}{1/\sigma^2 + \sum_i \psi_{ik}}$$

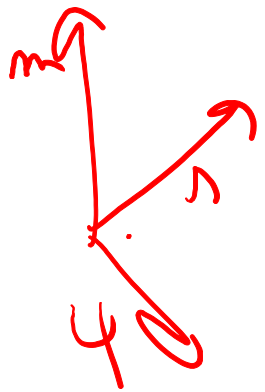
$$\text{Set } s_k^2 \longleftarrow \frac{1}{1/\sigma^2 + \sum_i \psi_{ik}}$$

  end

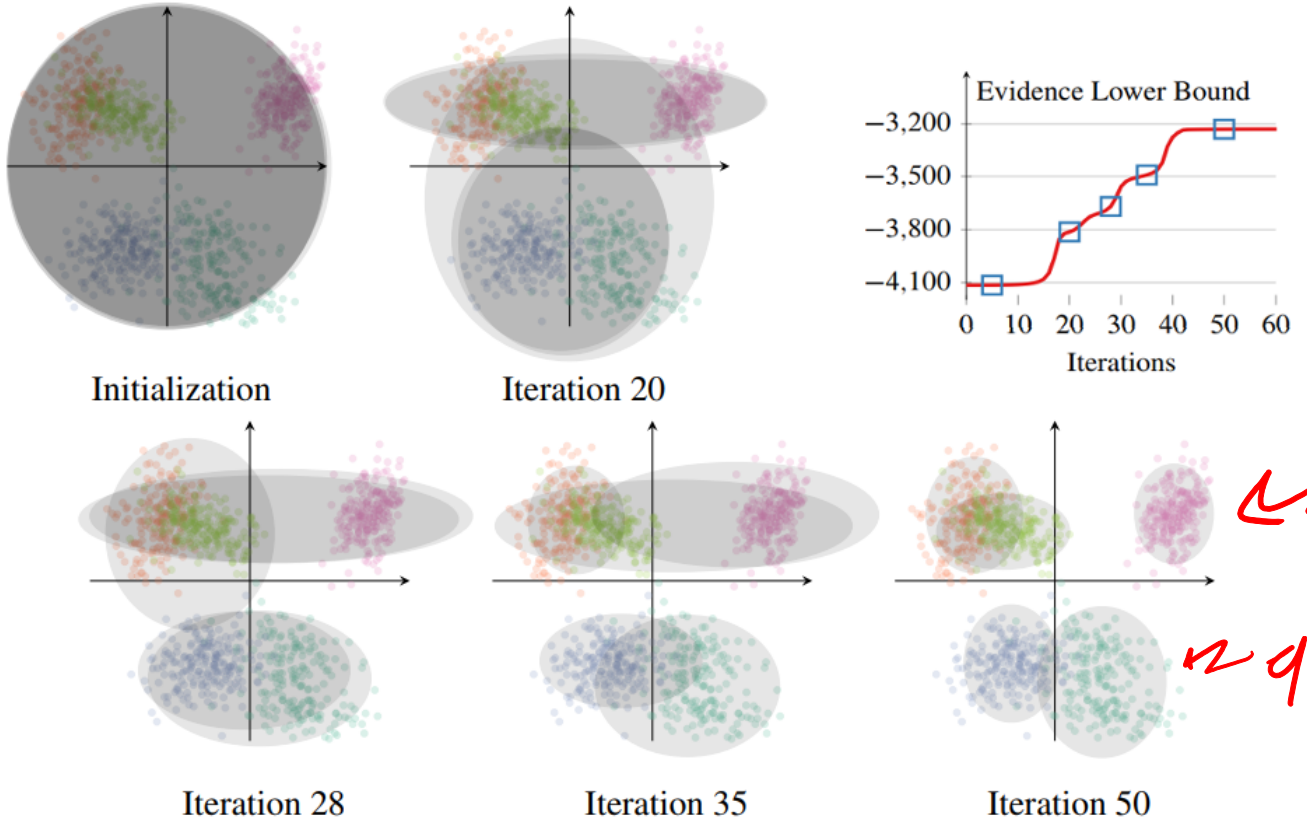
  Compute ELBO $\mathcal{L}(\mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi})$

end

return $q(\mathbf{Z}, \boldsymbol{\mu} \mid \mathbf{m}, \mathbf{s}^2, \boldsymbol{\Psi})$



CAVI algorithm for GMM in action



Notice, that the ellipses in the plots visualize $q(\mu_k)$ for each component.

While we assume the data generating process $p(x_i \mid \mathbf{z}_i, \boldsymbol{\mu})$ to have identity covariance, the shape of the clusters is captured by the distribution $q(\boldsymbol{\mu})$.

Figure from Blei, Kucukelbir and McAuliffe -" Variational Inference: A Review for Statisticians"

What happens if true posterior is in the variational family?

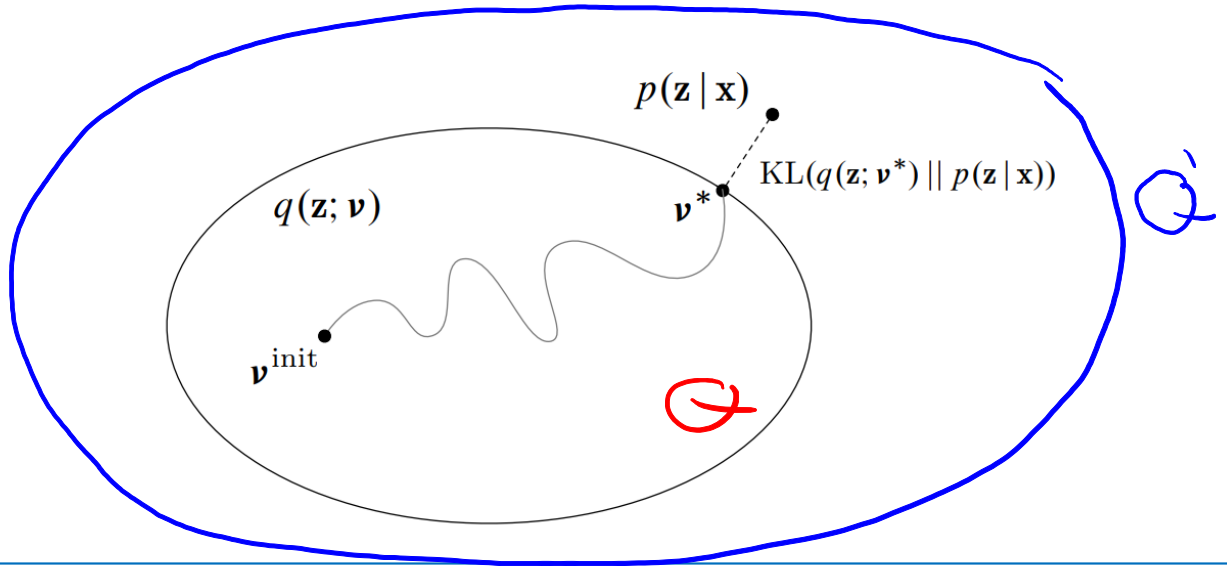
In variational inference we solve the optimization problem

$$q^* = \arg\max_{q \in \mathcal{Q}} \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z})]$$
$$= \arg\min_{q \in \mathcal{Q}} \mathbb{KL}(q(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x}))$$

What happens if the true posterior $p(\mathbf{z} \mid \mathbf{x})$ is contained in $\mathcal{Q}$?

Obviously, KL-divergence is minimized if we set

$$q(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x})$$



Dirac delta as $q(\mathbf{z})$

What if we don't want to go full Bayesian for all variables?

Let's choose Dirac delta as a variational distribution.

The Dirac delta function $\delta(z - \hat{z})$ is defined as

$$\delta(z - \hat{z}) = \begin{cases} +\infty & \text{if } z = \hat{z} \\ 0 & \text{else} \end{cases} \quad \text{and} \quad \int_{-\infty}^{\infty} \delta(z - \hat{z}) dz = 1$$

Alternatively, you can think of it as an infinitely narrow Gaussian

$$\delta(z - \hat{z}) = \lim_{\sigma^2 \to 0} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z - \hat{z})^2}{2\sigma^2}\right)$$

The Dirac delta has the property

$$\mathbb{E}_{\delta}[f] = \int_{-\infty}^{\infty} f(z)\delta(z - \hat{z})dz = f(\hat{z})$$

Dirac delta as $q(\mathbf{z})$

What happens if we let $q(\mathbf{z}) = \delta(\mathbf{z} - \hat{\mathbf{z}})$? Let's write down the ELBO

$$\mathcal{L}(q) = \mathbb{E}_{q(\mathbf{z})}\left[\log p(\mathbf{x}, \mathbf{z})\right] + \underbrace{\mathbb{H}\left[q(\mathbf{z})\right]}_{=-\infty}$$

the entropy tends to $-\infty$, but let's only consider the other term

$$= \int \log p(\mathbf{x}, \mathbf{z}) \delta(\mathbf{z} - \hat{\mathbf{z}}) d\mathbf{z} = \log p(\mathbf{x}, \hat{\mathbf{z}})$$

Which means that maximizing ELBO with a Dirac delta $q(\mathbf{z})$ ...

$$\arg\max_{\hat{\mathbf{z}}} \mathcal{L}(q) = \arg\max_{\hat{\mathbf{z}}} \log p(\mathbf{x}, \hat{\mathbf{z}})$$

$$= \arg\max_{\hat{\mathbf{z}}} [\log p(\hat{\mathbf{z}} \mid \mathbf{x}) + \underbrace{\log p(\mathbf{x})}_{\text{const.}}]$$

... is equivalent to MAP inference $\Rightarrow$ we get a point estimate $\hat{\mathbf{z}}$ of $\mathbf{z}$!

EM-algorithm as variational inference

E-step

Compute the posterior distribution (the responsibilities) for $\boldsymbol{\theta}^{old}$

$$\gamma(\mathbf{Z}) = p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{old}) \tag{6}$$

M-step

Find $\boldsymbol{\theta}^{new}$ by maximizing (while keeping $\gamma(\mathbf{Z})$ fixed).

$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}'} \mathbb{E}_{\mathbf{Z} \sim \gamma(\mathbf{Z})}[\log p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}')] \tag{7}$$

- $\gamma(\mathbf{Z})$ is just a (variational) distribution over $\mathbf{Z}$.

  By the Equation 6 we minimize $\mathbb{KL}(\gamma(\mathbf{Z}) \parallel p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{old}))$.

  This means that we maximize ELBO w.r.t. $q(\mathbf{Z})$ in the E-step!

- Having a point estimate $\boldsymbol{\theta}^{new} \iff$ using $q(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^{new})$.

  We maximize ELBO w.r.t. $q(\boldsymbol{\theta})$ in the M-step!

Variational Inference: Summary

Why do we use VI?

- To approximate an intractable posterior distribution $p(\mathbf{z} \mid \mathbf{x})$.

How does it work?

- By searching for an approximate posterior $q(\mathbf{z} \mid \boldsymbol{\nu}) \approx p(\mathbf{z} \mid \mathbf{x})$.
- Finding the best variational parameters $\boldsymbol{\nu}^*$ by optimizing ELBO

$$\boldsymbol{\nu}^* = \arg\min_{\boldsymbol{\nu}} \mathbb{KL}(q(\mathbf{z} \mid \boldsymbol{\nu}) \parallel p(\mathbf{z} \mid \mathbf{x})) = \arg\max_{\boldsymbol{\nu}} \mathcal{L}(\boldsymbol{\nu}).$$

- Option 1: Direct optimization of ELBO with gradient ascent.
- Option 2: Coordinate ascent variational inference.

Variational Inference: Summary

$$\mathbb{E}_{q(z,\theta)}[\log p(x,z,\theta) - \log q(\dots)]$$

$$= \mathbb{E}_{q(z)\cdot q(\theta)}[\dots]$$

Pros and cons

+ Much faster than the alternatives (e.g., MCMC).
+ Lots of powerful extensions (+ stochastic opt., +Deep Learning).
– Often requires tedious mathematical derivations.
– Does not converge to the true distribution in the limit.