Variational Inference

Generative and Deep Learning (GDL)

Davide Bacciu (davide.bacciu@unipi.it)

UNIVERSITÀ DI PISA

Objectives

Introduce the basic concepts of variational learning useful for both generative models and deep learning

- Learning and inference in intractable latent variable models
- Lower bounding the maximum likelihood optimization (ELBO)
- Variational approximation: a generalized form of EM learning

Setting the stage

Problem Setup – Intractable latent variable models

Latent variables

- Unobserved RV that define a hidden generative process of observed data
- Explain complex relation between many observable variables
- E.g. hidden states in HMM

Maximum likelihood in latent variable models

$$\max_{\theta} \log P_\theta(x) = \max_{\theta} \log \int_{\boldsymbol{z}} \prod_{i=1}^{N} P_\theta(x_i|\boldsymbol{z}) P_\theta(\boldsymbol{z}) d\boldsymbol{z}$$

Short for $P(x|\theta)$

The integral generally does not admit a closed form solution and numerical integration is unfeasible, so we cannot optimize the function

The “Variational” Term

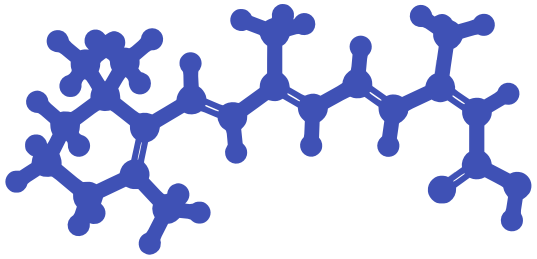
Stems from Variational Calculus

Calculus

Objective: find derivatives of functions

$$\max_{x} f(x)$$

Application: find configuration that maximizes the potential energy of a molecule

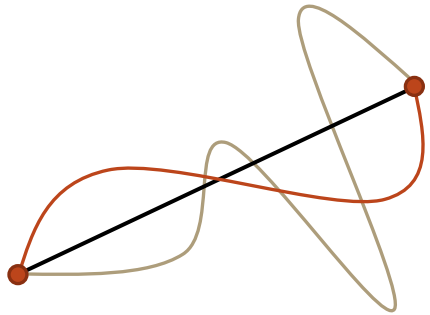


Variational Calculus

Objective: find derivatives of functionals

$$\max_{f} F(f)$$

Application: find function that connects two points with the shortest distance



We will look for “good” functions

Tractability and the EM Algorithm

- Introducing hidden variables can produce couplings between the distributions (i.e., one depending on the other) which can make their posterior intractable
- Bayesian learning introduces priors which introduce integrals in the posterior computations which are not always analytically or computationally tractable
- Posteriors are key to the EM algorithm (E-Step)

This lecture is about how we can approximate such intractable problems

- Variational view of EM (also used in variational DL)

Let us introduce some useful tools

- A measure of closeness between distributions
- A useful lower-bounding inequality

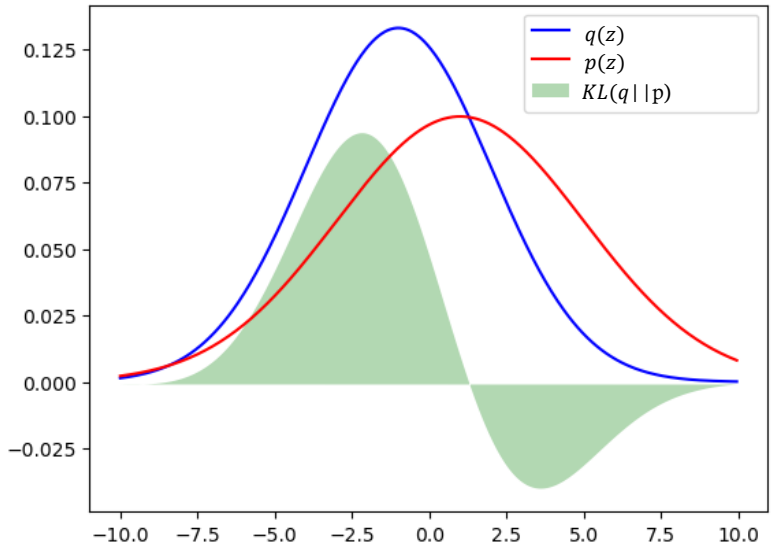
Kullback-Leibler (KL) Divergence

An information theoretic measure of closeness of two distributions p and q

$$KL(q||p) = \mathbb{E}_q\left[\log \frac{q(z)}{p(z|x)}\right] = \langle \log q(z) \rangle_q - \langle \log p(z|x) \rangle_q$$

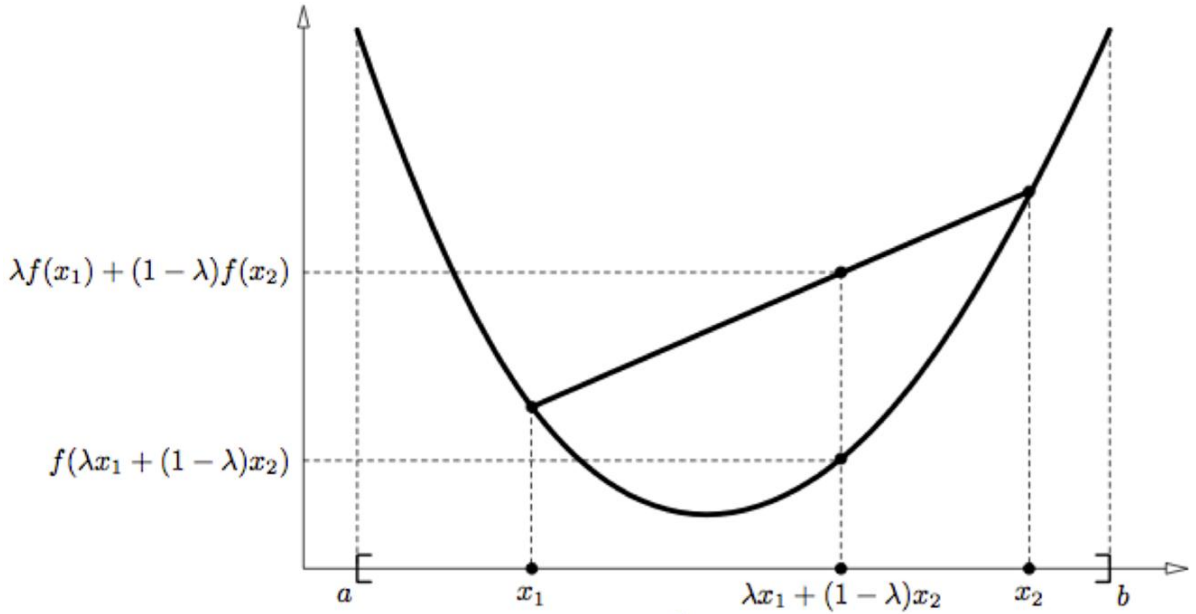
Note:

- A specialized definition for our latent variable setting
  - If q high and p high ⇒ happy
  - If q high and p low ⇒ unhappy
  - If q low ⇒ don't care (due to expectation)
- Its a divergence ⇒ it is not symmetric



Jensen Inequality

Property of linear operators on convex/concave functions



Generalizes to

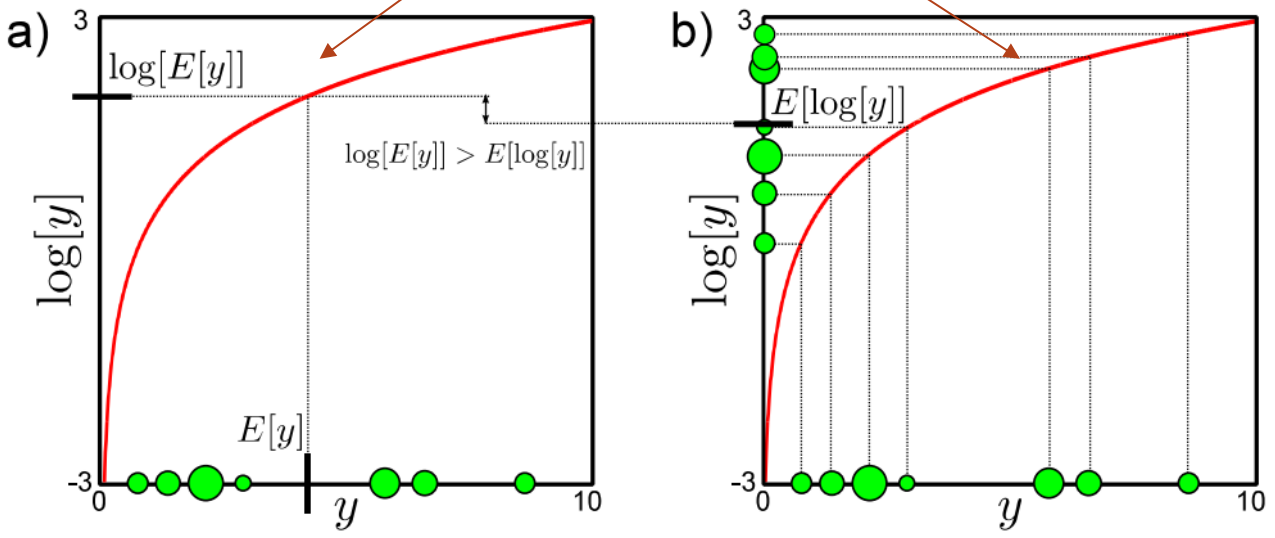
$$\frac{\sum_i a_i f(x_i)}{\sum_i a_i} \geq f\left(\frac{\sum_i a_i x_i}{\sum_i a_i}\right)$$

$$\lambda f(x_1) + (1-\lambda) f(x_2) \geq f(\lambda x_1 + (1-\lambda) x_2)$$

Jensen Inequality for our probabilistic setting

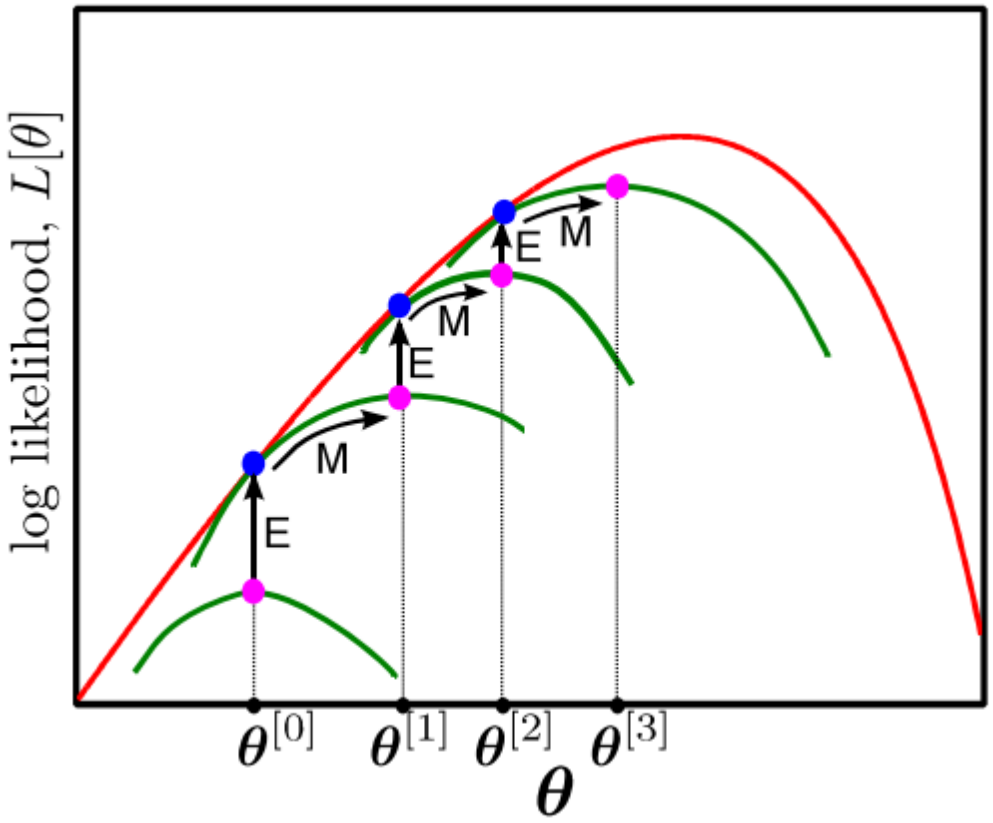
We work with a concave f (the log) and with a specific linear combination (the expectation operator $\mathbb{E}[]$)

$$f(\mathbb{E}[y]) \geq \mathbb{E}[f(y)]$$



Bounding the likelihood

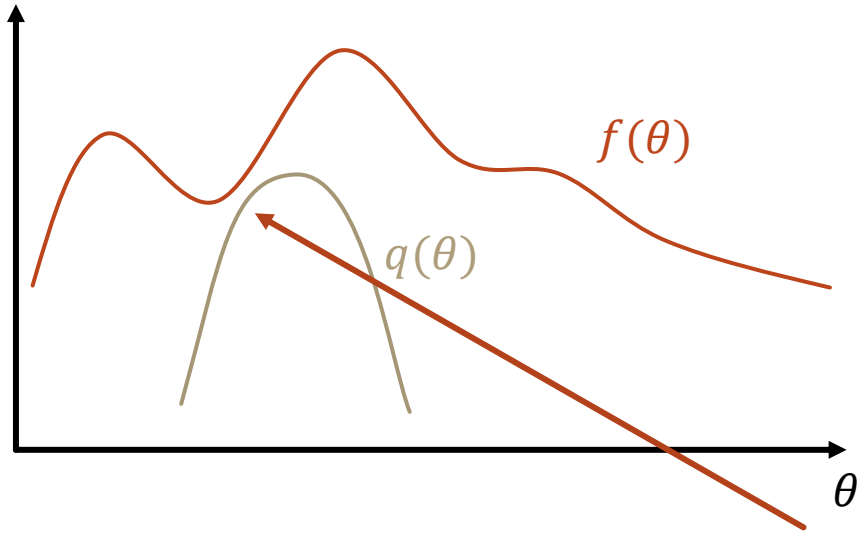
Bound maximization view of EM



- The E-Step finds a lower bound of the likelihood
- The M-Step finds a maximiser of the lower bound
- The lower bound is exact (it touches the likelihood where it is computed)

Generalizing lower bound maximization

Our goal is to compute $\max_{\theta} \log \int_{\boldsymbol{z}} \prod_{i=1}^{N} P_\theta(x_i|\boldsymbol{z}) P_\theta(\boldsymbol{z}) d\boldsymbol{z}$, in short $\max_{\theta} f(\theta)$



The lower bound need not to be is exact this time

Pick a "nice" function $q(\boldsymbol{\theta})$ that is a general lower bound of $f(\boldsymbol{\theta})$ $\forall \boldsymbol{\theta}$:

$$\forall \boldsymbol{\theta}\ f(\boldsymbol{\theta}) \geq q(\boldsymbol{\theta})$$

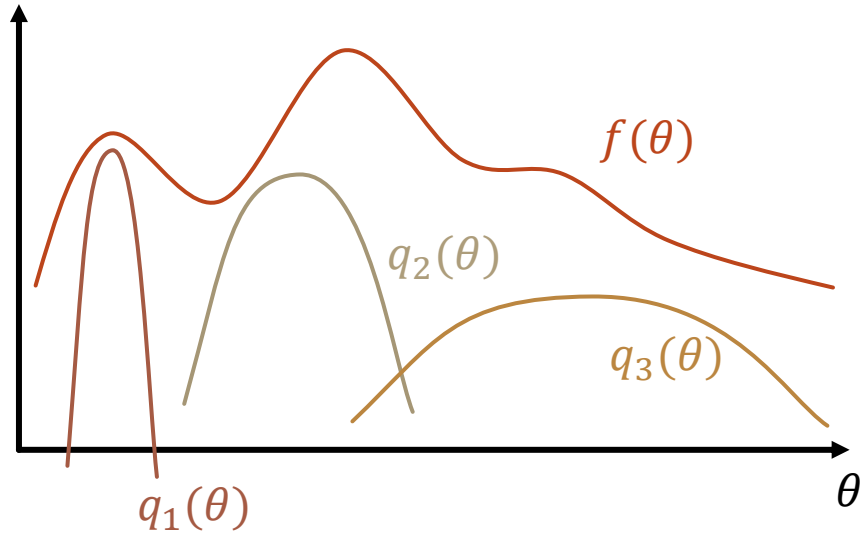
Maximizing a $q(\boldsymbol{\theta})$ gives a lower bound on the solution of the original problem!

$$\max_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \geq \max_{\boldsymbol{\theta}} q(\boldsymbol{\theta})$$

Which $q(\boldsymbol{\theta})$?

Generalizing lower bound maximization

Our goal is to compute $\max_{\theta} \log \int_{\boldsymbol{z}} \prod_{i=1}^{N} P_\theta(x_i|\boldsymbol{z}) P_\theta(\boldsymbol{z}) d\boldsymbol{z}$, in short $\max_{\theta} f(\theta)$



Instead of using a single lower bound, consider a family of lower bounds with uncountably many elements

$$\mathcal{Q} = \{q_1, q_2, q_3, \dots\}$$

We can get even closer to the optimal solution of our original problem:

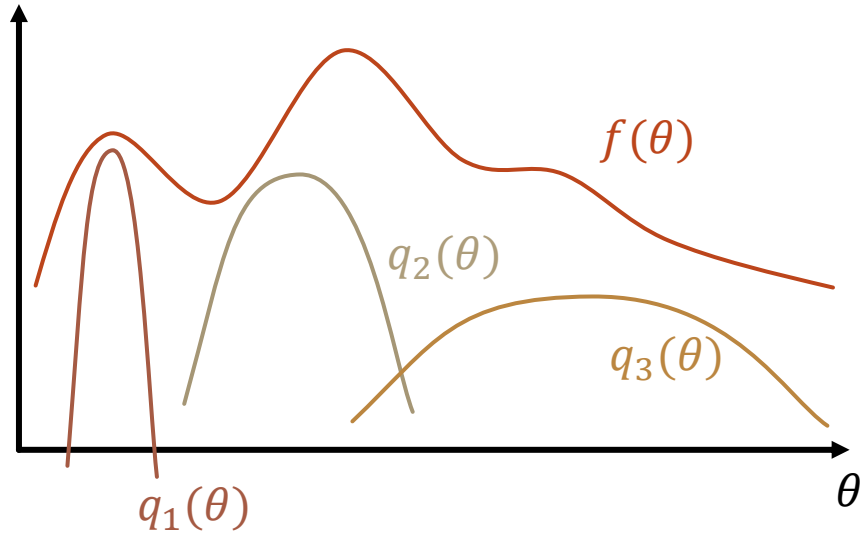
$$\max_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \geq \max_{q \in \mathcal{Q}} \max_{\boldsymbol{\theta}} q(\boldsymbol{\theta})$$

A functional!

How to choose $\mathcal{Q}$ now?

Generalizing lower bound maximization

Approximately solve $\max_{\theta} f(\theta)$ for some intractable f



Algorithm hint

1. Construct a lower bound family
$$\mathcal{Q} = \{q_1, q_2, q_3, \dots\}$$
2. Solve the optimization problem
$$\max_{q \in \mathcal{Q}, \boldsymbol{\theta}} q(\boldsymbol{\theta})$$

Pick $\mathcal{Q}$ to be a tractable family of distributions!
(e.g. Gaussians with fixed variance)

Bounding Log-Likelihood with Jensen

The log-likelihood for a model with a single hidden variable Z and parameters θ (assume single sample for simplicity) is

$$\log P(x|\theta) = \log \int_z P(x,z|\theta)dz = \log \int_z \frac{Q(z|\phi)}{Q(z|\phi)} P(x,z|\theta)dz$$

which holds for $Q(z|\phi) \neq 0$ with parameters ϕ

Given the definition of expectation this rewrites as (Jensen)

$$\log P(x|\theta) = \log \mathbb{E}_Q \left[\frac{P(x,z)}{Q(z)}\right] \geq \mathbb{E}_Q \left[\log \left(\frac{P(x,z)}{Q(z)}\right)\right]$$

The Evidence Lower BOund (ELBO)

$$= \underbrace{\mathbb{E}_Q[\log P(x,z)]}_{\text{Expectation of Joint Distribution}} - \underbrace{\mathbb{E}_Q[\log Q(z)]}_{\text{Entropy}} = \mathcal{L}(x,\theta,\phi)$$

How Good is this Lower Bound?

$$\log P(x|\theta) - \mathcal{L}(x, \theta, \phi) = ?$$

Inserting the definition of $\mathcal{L}(x, \theta, \phi)$

$$\log P(x|\theta) - \int_z Q(z) \log \frac{P(x, z)}{Q(z)} dz$$

Introducing $Q(z)$ by marginalization ($\int_z Q(z) = 1$)

$$\int_z Q(z) \log P(x) dz - \int_z Q(z) \log \frac{P(x, z)}{Q(z)} dz =$$

$$\int_z Q(z) \log \frac{P(x)Q(z)}{P(x, z)} \, dz$$

How Good is this Lower Bound?

$$\log P(x|\theta) - \mathcal{L}(x, \theta, \phi) = ?$$

Inserting the definition of $\mathcal{L}(x, \theta, \phi)$

$$\log P(x|\theta) - \int_z Q(z) \log \frac{P(x, z)}{Q(z)} dz$$

Introducing $Q(z)$ by marginalization ($\int_z Q(z) = 1$)

$$\int_z Q(z) \log P(x) dz - \int_z Q(z) \log \frac{P(x, z)}{Q(z)} dz =$$

$$\mathbb{E}_Q \left[\log \frac{Q(z)}{P(z|x)} \right] = KL\big(Q(z|\phi) || P(z|x, \theta)\big)$$

Lower Bound and Variational Inference

$$\log P(x|\theta) = \mathcal{L}(x, \theta, \phi) + KL\big(Q(z|\phi)||P(z|x, \theta)\big)$$

- Since $KL(\quad)$ is nonnegative $\mathcal{L}(x, \theta, \phi)$ is effectively a lower bound to $\log P(x|\theta)$ for any distribution $Q(z|\phi)$
  - Different choices of $Q(z|\phi)$ lead to different lower bounds
- How tight the bound is depends on how close $Q(z|\phi)$ is to the true posterior $P(z|x, \theta)$ in terms of KL divergence.
  - When they coincide $\Rightarrow \mathcal{L}(\boldsymbol{\theta}, q) = \log p_{\boldsymbol{\theta}}(\boldsymbol{x})$

Variational Inference

Find the parameters θ and the distribution $Q(z|\phi)$ that maximize the lower bound

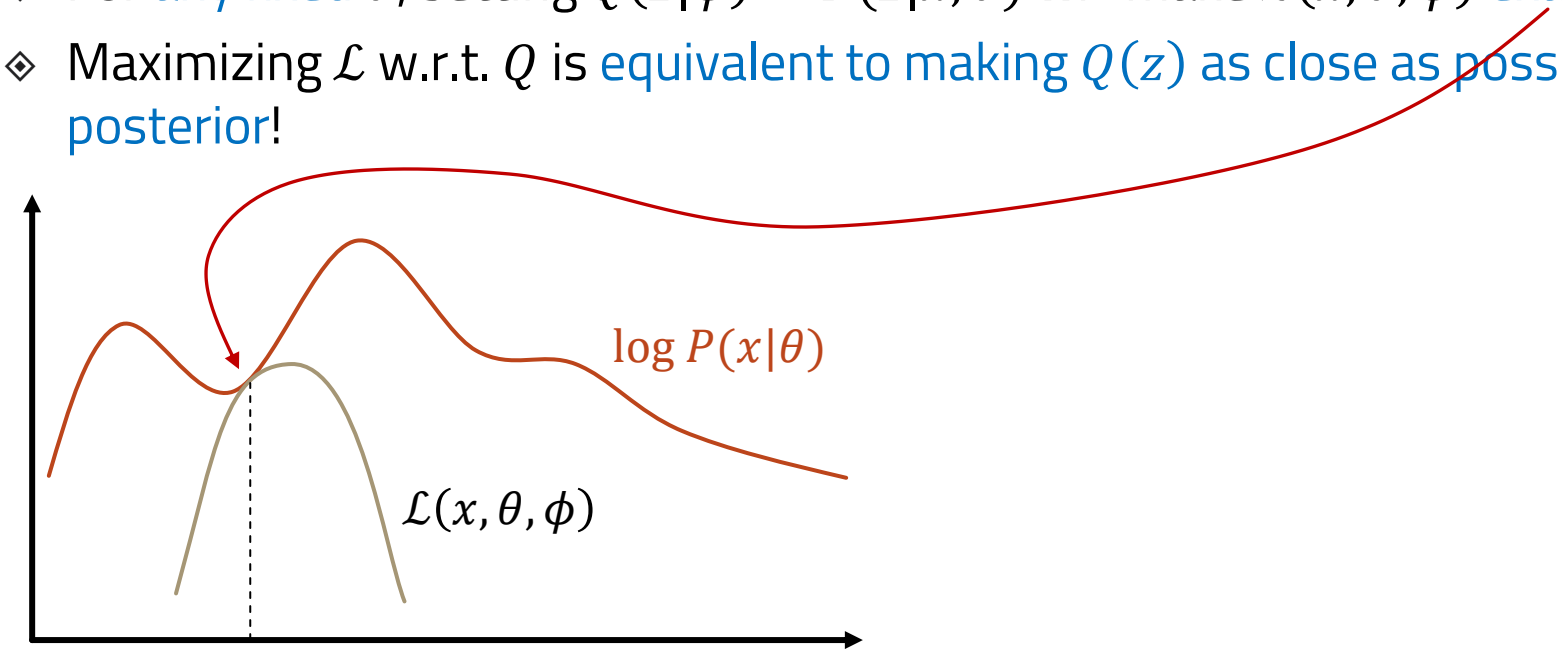
$$\max_{\boldsymbol{\theta}, Q} \mathcal{L}(x, \theta, Q)$$

An alternative optimization problem

- We can rewrite the variational bound as

$$\mathcal{L}(x, \theta, \phi) = \log P(x|\theta) - KL\big(Q(z|\phi)||P(z|x, \theta)\big)$$

- For any fixed θ, setting $Q(z|\phi) = P(z|x, \theta)$ will make $\mathcal{L}(x, \theta, \phi)$ exactly equal to $\log P(x|\theta)$
- Maximizing $\mathcal{L}$ w.r.t. Q is equivalent to making $Q(z)$ as close as possible (in KL terms) to the posterior!



Two flavours of variational inference

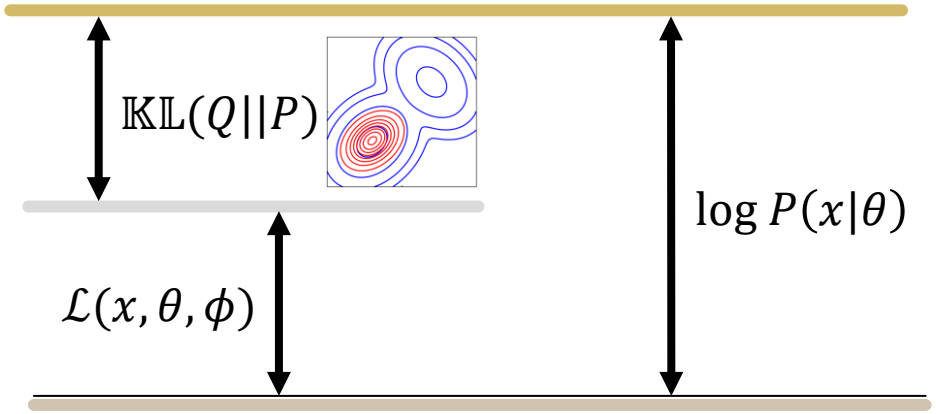
EM maximizing the bound

E-step: compute

$$\boldsymbol{\phi}^{(k)} = \arg\max_{\boldsymbol{\phi}} \mathcal{L}\left(\boldsymbol{x}, \boldsymbol{\theta}^{(k)}, \boldsymbol{\phi}\right)$$

M-step: compute

$$\boldsymbol{\theta}^{(k)} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}\left(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\phi}^{(k)}\right)$$



Minimizing KL

E-step: compute

$$\boldsymbol{\phi}^{(k)} = \arg\min_{\boldsymbol{\phi}} KL\left(Q(z|\phi)||P\left(\boldsymbol{z}|\boldsymbol{x}, \theta^{(k)}\right)\right)$$

M-step: compute

$$\boldsymbol{\theta}^{(k)} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}\left(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\phi}^{(k)}\right)$$

Closed form computation of the KL is available only for very specific distributions!

Can be addressed through sampling-based approximation

Wrap-up

Evidence Lower Bound (ELBO)

We can assume the existence of a probability $Q(z|\phi)$ which allows to bound the likelihood $P(x|\theta)$ from below using $\mathcal{L}(x, \theta, \phi)$

The term $\mathcal{L}(x, \theta, \phi)$ is called **variational bound** or **evidence lower bound (ELBO)**

The optimal bound is obtained for $KL\big(Q(z|\phi)||P(z|x,\theta)\big) = 0$, that is if we choose $\boldsymbol{Q(z|\phi) = P(z|x,\theta)}$

Minimizing KL is equivalent to maximize the ELBO ⇒ change a sampling problem with an optimization problem

Variational View of Expectation Maximization

EM Learning Reformulated

Maximum likelihood learning with hidden variables can be approached by maximization of the ELBO

$$\max_{\theta,\phi} \sum_{n=1}^{N} \mathcal{L}(x_n, \theta, \phi)$$

where θ are the model parameters and ϕ serve in $Q(z|\phi)$

- If $P(z|x,\theta)$ is tractable ⇒ use it as $Q(z|\phi)$ (optimal ELBO)
- O.w. choose $Q(z|\phi)$ as a tractable family of distributions
  - find ϕ that minimize $KL(Q(z|\phi)||P(z|x,\theta))$, or
  - find ϕ that maximize $\mathcal{L}(\cdot,\phi)$

Take home messages

- Variational inference solves a double optimization problem
  - Over the family of lower bounds
  - Over the parameters of the generative model
- EM can be seen as an alternating version of this optimization problem
  - E-step: optimize with respect to variational distribution Q
  - M-step: optimize with respect to the generative model P
  - The E-step can also by solved as a KL-divergence optimization
- Maximization/minimization problems can be solved by gradient ascent/descent
- Exact EM can be recovered in the variational framework by having Q as the exact posterior
  - The bound becomes an equality

Next Lecture

Latent topic models

- Hidden variable models where latents have a specialized semantics
- A class of generative models for which variational or approximated methods are needed
- Example model: Latent Dirichlet Allocation