Latent variable models

Course of Machine Learning
Master Degree in Computer Science
University of Rome "Tor Vergata"
a.a. 2024-2025

Giorgio Gambosi

Inference and maximum likelihood

The construction of a model of a probability distribution given a finite sample of data drawn from that distribution is one of the central problems in pattern recognition and machine learning.

A standard approach to the problem involves parametric models in which a specific form for the density is proposed which contains a number of parameters. Values for these parameters are then determined from an observation $\mathbf{x}$ consisting of a data vector on domain $\mathcal{X}$, which is assumed to be an instantiation of a random variable $\mathbf{x}$, distributed according to some (unknown) distribution $p(\mathbf{x};\boldsymbol{\theta})$ with parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$.[1]

We have then an **inference** problem, aiming to find a "good" way to model $\mathbf{x}$, that is finding a specific probability distribution that better represents $\mathbf{x}$. In our context, since $p(\mathbf{x};\boldsymbol{\theta})$ has been arbitrarily assumed, the task is to estimate the value $\boldsymbol{\theta}^*$ so that $p(\mathbf{x};\boldsymbol{\theta}^*)$ better represents $\mathbf{x}$, assuming that we are bound to use the distribution $p(\mathbf{x};\boldsymbol{\theta})$. A common approach here is to define $\boldsymbol{\theta}^*$ as the value of $\boldsymbol{\theta}$ such that the probability of $\mathbf{x}$ is maximum if $p(\mathbf{x};\boldsymbol{\theta}^*)$ is the assumed distribution of $\mathbf{x}$. This estimate of the parameters of the available sample is called **point estimation**.

Given a parameter value $\boldsymbol{\theta}$, the **evidence** $p(\mathbf{x};\boldsymbol{\theta})$ of $\mathbf{x}$ is the joint probability $p(\mathbf{x};\boldsymbol{\theta})$ of $\mathbf{x}$. This is also denoted as **likelihood** $L(\boldsymbol{\theta}|\mathbf{x})$ when seen as a function of $\boldsymbol{\theta}$.

The best value $\boldsymbol{\theta}^*$ of $\boldsymbol{\theta}$ to model $\mathbf{x}$ can be computed by maximizing the likelihood wrt $\boldsymbol{\theta}$, that is

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\text{argmax}}\, L(\boldsymbol{\theta}|\mathbf{x})$$

or, equivalently,

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\text{argmax}}\, \log L(\boldsymbol{\theta}|\mathbf{x}) = \underset{\boldsymbol{\theta}}{\text{argmax}}\, l(\boldsymbol{\theta}|\mathbf{x})$$

where $l(\boldsymbol{\theta}|\mathbf{x}) \triangleq \log L(\boldsymbol{\theta}|\mathbf{x})$ is the **log-likelihood** of $\mathbf{x}$.[2]

If the model is simple enough that equation[3]

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{x}) = \nabla_{\boldsymbol{\theta}} \log p(\mathbf{x}|\boldsymbol{\theta}) = \mathbf{0}$$

has a finite set of closed-form solutions $\{\boldsymbol{\theta}_0^*, \boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_r^*\}$, the values $l(\boldsymbol{\theta}_0^*|\mathbf{x}), l(\boldsymbol{\theta}_1^*|\mathbf{x}), \ldots, l(\boldsymbol{\theta}_r^*|\mathbf{x})$ can be computed and compared to find the global maximum.

[1] Observe that we may interpret $p(\mathbf{x};\boldsymbol{\theta})$ as a function $p(\mathbf{x};\boldsymbol{\theta})$ of $\mathbf{x}$ given the value of the parameter $\boldsymbol{\theta}$ or as a function $p(\mathbf{x};\boldsymbol{\theta})$ of $\boldsymbol{\theta}$ given an observed value of $\mathbf{x}$.

[2] The equality holds because both functions have the same stationary points (since $\frac{d}{dx}\log f(x) = \frac{1}{f(x)}\frac{d}{dx}f(x)$ and $f(x)$ is bounded) and $f(x_1) > f(x_2)$ iff $\log f(x_1) > \log f(x_2)$, since log is a monotone increasing function.

[3] This corresponds to the system of equations $\frac{\partial l(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_i} = 0$ for $i = 1, \ldots, n$

Even in the case that closed-form solutions cannot be computed, if the likelihood $L(\boldsymbol{\theta}|\mathbf{x})$ can be evaluated for any $\boldsymbol{\theta}$ its gradient can be also computed (either by expressing it analytically or through automatic differentiation) for any $\boldsymbol{\theta}$, and **gradient ascent** methods such as

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} + \eta \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathbf{x})\big|_{\boldsymbol{\theta}^{(i)}}$$

can be iteratively applied to converge to some stationary point.

Values for parameters are indeed usually determined from a whole set of observations, that is a data set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ consisting of n data vectors on $\mathcal{X}$. Such values are assumed to be independent instantiation of $\mathbf{x}$, distributed according to the same (unknown) distribution $p(\mathbf{x}; \boldsymbol{\theta})$.

The hypothesis of independence between the data points in $\mathbf{X}$ implies then that

$$L(\boldsymbol{\theta}|\mathbf{X}) \stackrel{\Delta}{=} p(\mathbf{X}; \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i; \boldsymbol{\theta})$$

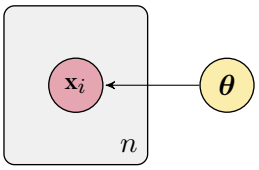

Figure 1: The dataset consists of a set of n independent realizations of the random variable $\mathbf{x}$ distributed according to a given parametric distribution $p(\mathbf{x}; \boldsymbol{\theta})$ with parameter $\boldsymbol{\theta}$.

The best value $\boldsymbol{\theta}^*$ of $\boldsymbol{\theta}$ to model $\mathbf{X}$ can then be computed by maximizing either the likelihood

$$L(\boldsymbol{\theta}|\mathbf{X}) = \prod_{i=1}^{n} p(\mathbf{x}_i; \boldsymbol{\theta})$$

or the log-likelihood

$$l(\boldsymbol{\theta}|\mathbf{X}) = \sum_{i=1}^{n} \log p(\mathbf{x}_i; \boldsymbol{\theta})$$

of the whole dataset wrt $\boldsymbol{\theta}$, and the same considerations made above for the case of a single data item (which can actually be seen as a dataset of size 1) apply, by referring to $L(\boldsymbol{\theta}|\mathbf{X})$ instead of $L(\boldsymbol{\theta}|\mathbf{x})$

Inferring $\boldsymbol{\theta}^*$ only provides a limited information about $\mathbf{X}$[4], since it just amounts to identify how to best represent the available data within a chosen probabilistic framework for $\mathbf{x}$. In unsupervised learning, however, we are interested to identifying patterns and structures in a set of observations $\mathbf{X}$, such as for example clusters of similar items or hidden dependence among features.

In order to deal with these tasks, more sophisticated models can be exploited, including additional random variables which provide the information we wish to extract from data.

The goal of such models is to express the distribution $p(\mathbf{x})$ of the variables $x_1, \ldots, x_d$ in terms of a (usually smaller) vector of additional variables $\mathbf{z} = (z_1, \ldots, z_q)$ on domain $\mathcal{Z}$. This is achieved by first decomposing the joint distribution $p(\mathbf{x}, \mathbf{z})$ into the product of the marginal distribution $p(\mathbf{z})$ of the additional variables and the conditional distribution $p(\mathbf{x}|\mathbf{z})$ of the data variables given the additional ones.

[4]In the following we shall refer to a set $\mathbf{X}$ of observations, which in a particualr case will be a single observation $\mathbf{x}$.

Latent variable models

This more general, albeit complex, approach is based on referring to a probabilistic model defined on a set of random variables larger than the set of the ones observed in the dataset.

More in detail, we refer to a probabilistic model $p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = \boldsymbol{\phi} \cup \boldsymbol{\psi}$ defined on:

- an **observed** random variable $\mathbf{x}$, whose instantiations can be observed
- a **latent** random variable $\mathbf{z}$, whose instantiations cannot be observed.

In the general case, both the domain $\mathcal{X}$ of $\mathbf{x}$ and the domain $\mathcal{Z}$ of $\mathbf{z}$ are continuous. Later, we shall also deal with the simpler case when $\mathcal{Z}$ is discrete, where integrals shall be substituted by sums.

Since $p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = p(\mathbf{x}|\mathbf{z}; \boldsymbol{\theta})p(\mathbf{z}; \boldsymbol{\theta})$ we assume that $\boldsymbol{\theta}$ can be partitioned into two disjoint sets of parameters $\boldsymbol{\phi}, \boldsymbol{\psi}$, where $\boldsymbol{\phi}$ affects the distribution $p(\mathbf{z}; \boldsymbol{\phi})$ of the latent variable, while $\boldsymbol{\psi}$ affects the conditional distribution $p(\mathbf{x}|\mathbf{z}; \boldsymbol{\psi})$. Hence, we may write

$$p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = p(\mathbf{x}|\mathbf{z}; \boldsymbol{\psi})p(\mathbf{z}; \boldsymbol{\phi})$$

Observe that even if we assume that, by hypothesis, the joint distribution $p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$[5] is manageable (in the sense that its value can be easily computed for any $\mathbf{x}, \mathbf{z}, \boldsymbol{\theta}$ and also that its gradient wrt $\boldsymbol{\theta}$ is easy to derive) computing the evidence $p(\mathbf{x}|\boldsymbol{\theta})$ may be hard or even unfeasible, since

$$p(\mathbf{x}; \boldsymbol{\theta}) = \int_{\mathcal{Z}} p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) d\mathbf{z} = \int_{\mathcal{Z}} p(\mathbf{x}|\mathbf{z}; \boldsymbol{\psi})p(\mathbf{z}; \boldsymbol{\phi}) d\mathbf{z}$$

and the usually multidimensional integral can be impossible to derive analytically and unfeasible to compute numerically since any approximation may require considering a very large number of points.

The same considerations apply also for what regards computing the inverse conditional distribution of $\mathbf{z}$ given $\mathbf{x}$. In fact, since

$$p(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}) = \frac{p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})}{p(\mathbf{x}; \boldsymbol{\theta})}$$

this requires computing the evidence $p(\mathbf{x}; \boldsymbol{\theta})$.

We make the following assumption:

Hypothesis 1 For all $\mathbf{x}, \mathbf{z}, \boldsymbol{\theta}$ it is feasible to compute the joint probability $p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$. Moreover, it is also feasible to compute local maxima wrt $\boldsymbol{\theta}$ of such distribution, or of its logarithm, either because the equation $\nabla_{\boldsymbol{\theta}} p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = \mathbf{0}$ is solvable or because such gradient can be evaluated for all $\boldsymbol{\theta}$, making it possible to apply numerical methods like gradient descent.

This clearly makes it possible to compute a (local) maximum $\boldsymbol{\theta}^*$ of such probability (that is of the likelihood and the log likelihood)

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \log p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, l(\boldsymbol{\theta}|\mathbf{x}, \mathbf{z})$$

Moreover, we shall consider the following additional hypothesis, dealing with the simple case when it holds and the more complex and more general case when it does not hold.

Hypothesis 2 It is feasible, for all $\mathbf{x}, \boldsymbol{\theta}$, to compute the marginal probability

$$p(\mathbf{x}; \boldsymbol{\theta}) = \int_{\mathcal{Z}} p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) d\mathbf{z}$$

[5] Or equivalently both the prior distribution $p(\mathbf{z}; \boldsymbol{\phi})$ of the latent variable and the conditional distribution $p(\mathbf{x}|\mathbf{z}; \boldsymbol{\psi})$ of the observed variable given the latent one

for all $\mathbf{x}$. As a consequence, the gradient $\nabla_{\boldsymbol{\theta}} p(\mathbf{x};\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}}$ can be evaluated, either by deriving it in closed form or by applying automatic differentiation. This makes finding (local) maxima of the likelihood feasible, at least by means of gradient descent methods.

Equivalently, in this case, it is also feasible to compute the inverse conditional probability $p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}) = \dfrac{p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{p(\mathbf{x};\boldsymbol{\theta})}$. Hence, the values of both parameters θ_i and latent variables $\mathbf{z}_i$ can be inferred: the first ones through MLE, while the other ones by exploiting the knowledge of $p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})$.

Generation in latent variable models

Once a latent variable model $p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})$ is given, new elements in $\mathcal{X}$ can be generated by first sampling the latent variable distribution $p(\mathbf{z})$, thus obtaining a value $\mathbf{z}$ and then sampling the conditional distribution $p(\mathbf{x}|\mathbf{z})$, as shown in Figure 2[6]. A different generation policy is when, starting from a value $\mathbf{x} \in \mathcal{X}$, we wish to generate a new value $\bar{\mathbf{x}} \in \mathcal{X}$ which is related (similar, in many practical case) to $\mathbf{x}$. This can be performed by first sampling the inverse conditional distribution $p(\mathbf{z}|\mathbf{x})$, obtaining a latent variable value $\mathbf{z}$ dependent on $\mathbf{x}$, and then sampling the conditional distribution $p(\mathbf{x}|\mathbf{z})$, thus obtaining $\bar{\mathbf{x}}$.

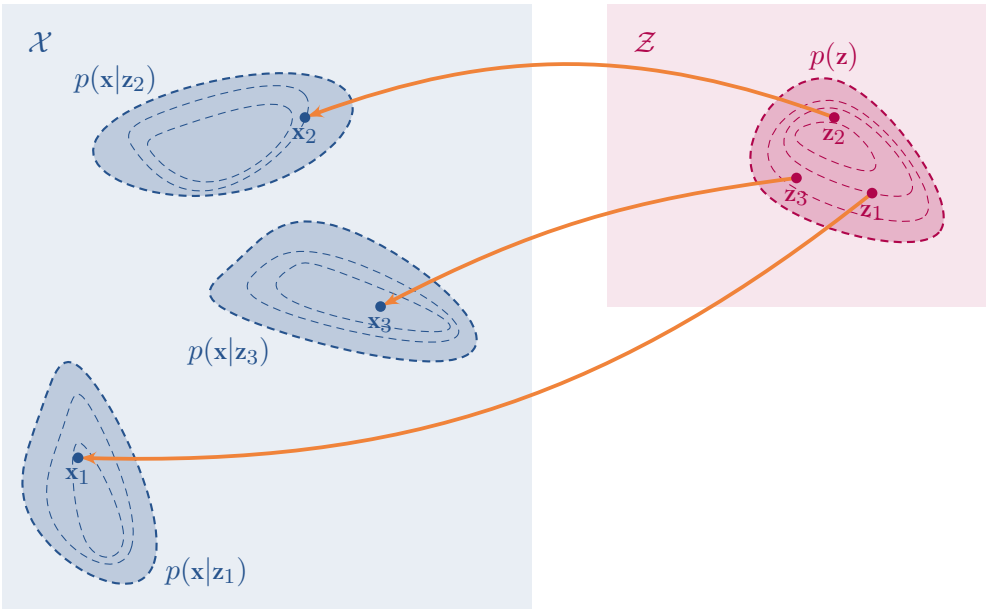

Figure 2: Generation of new items in a latent variable model

Inference in latent variable models

In terms of inference, we consider two random variables:

1. An **observed random variable** $\mathbf{x}$, whose instantiation $\mathbf{x}$ is available.
2. An additional **latent variable** $\mathbf{z}$ whose instantiation $\mathbf{z}$ is not observable.

and define the **complete dataset** $(\mathbf{x},\mathbf{z})$, including both instantiations. We remark that only $\mathbf{x}$ is actually available.

In general, we may also refer to a dataset $\mathbf{X}$ of n observations $\mathbf{x}_1,\ldots,\mathbf{x}_n$, which are cosidered as instantiations of $\mathbf{x}$ under the hypothesis that each of them has been independently generated according to the previous procedure.

[6]We shall not consider here the problem of how to tackle the task of sampling points from any given distribution.

Thus, the existence of a set $\boldsymbol{Z}$ of n latent instantiations $\mathbf{z}_1, \ldots, \mathbf{z}_n$ of $\mathbf{z}$ is assumed, where $\mathbf{x}_i$ was sampled from $p(\mathbf{x}|\mathbf{z}_i; \boldsymbol{\psi})$ and $\mathbf{z}_i$ was sampled from $p(\mathbf{z}; \boldsymbol{\phi})$.

It is easy to show that, under such hypothesis, the dataset distributions $p(\mathbf{X})$, $p(\boldsymbol{Z})$, $p(\mathbf{X}, \boldsymbol{Z})$, $p(\mathbf{X}|\boldsymbol{Z})$ and $p(\boldsymbol{Z}|\mathbf{X})$ are immediately related to the corresponding distributions for $\mathbf{x}$ and $\mathbf{z}$. In particular:

$$
\begin{aligned}
p(\boldsymbol{Z}) &= \prod_{i=1}^{n} p(\mathbf{z}_i) \\
p(\mathbf{X}|\boldsymbol{Z}) &= \prod_{i=1}^{n} p(\mathbf{x}_i|\mathbf{z}_i) \\
p(\mathbf{X}, \boldsymbol{Z}) &= \prod_{i=1}^{n} p(\mathbf{x}_i, \mathbf{z}_i) \\
p(\mathbf{X}) &= \prod_{i=1}^{n} p(\mathbf{x}_i) \\
p(\boldsymbol{Z}|\mathbf{X}) &= \prod_{i=1}^{n} p(\mathbf{z}_i|\mathbf{x}_i)
\end{aligned}
$$

As done before, we shall refer in the following to the case of a dataset $\mathbf{X}$, $\boldsymbol{Z}$ of n pairs $\{(\mathbf{x}_1, \mathbf{z}_1), \ldots, (\mathbf{x}_n, \mathbf{z}_n)\}$, where the case of a single instantion $(\mathbf{x}, \mathbf{z})$ of the random variables $\mathbf{x}, \mathbf{z}$ is just the case $n = 1$.

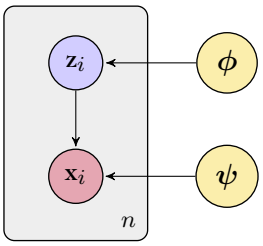

Figure 3: The dataset is composed of a set of n independent realizations of a pair of random variables $\mathbf{x}, \mathbf{z}$, with the latent variable $\mathbf{z}$ distributed according to a given parametric distribution $p(\mathbf{z}; \boldsymbol{\phi})$ with parameter $\boldsymbol{\phi}$, while $\mathbf{x}$ is dependent on $\mathbf{z}$ and distributed according to a given conditional parametric distribution $p(\mathbf{x}|\mathbf{z}; \boldsymbol{\psi})$ with parameter $\boldsymbol{\psi}$.

Since we assume that for each $i = 1, \ldots, n$, $\mathbf{x}_i$ is observed while $\mathbf{z}_i$ is unknown, the knowledge of the value of the latent variable is only probabilistic: it may be modeled by the distribution $p(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta})$ as

$$
p(\mathbf{z}_i|\mathbf{x}_i; \boldsymbol{\theta}) = \frac{p(\mathbf{x}_i|\mathbf{z}_i; \boldsymbol{\psi}) p(\mathbf{z}_i; \boldsymbol{\phi})}{p(\mathbf{x}_i; \boldsymbol{\theta})} = \frac{p(\mathbf{x}_i|\mathbf{z}_i; \boldsymbol{\psi}) p(\mathbf{z}_i; \boldsymbol{\phi})}{\int_{\mathcal{Z}} p(\mathbf{x}_i, \mathbf{z}; \boldsymbol{\theta}) d\mathbf{z}}
$$

Assume now that, as it happens in general, Hypothesis 2 does not hold, hence the evidence $p(\mathbf{x}_i|\boldsymbol{\theta})$ is unfeasible to compute, thus making unfeasible also the computation of $p(\mathbf{z}_i|\mathbf{x}_i; \boldsymbol{\theta})$.

According to the maximum likelihood approach, we are interested to maximizing the observed dataset likelihood or its log-likelihood

$$
l(\boldsymbol{\theta}|\mathbf{X}) = \log p(\mathbf{X}; \boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\mathbf{x}_i; \boldsymbol{\theta})
$$

since $\mathbf{X}$ is the dataset actually available.

The latent variable model we are using implies instead that the likelihood of the complete dataset should be maximized

$$l(\boldsymbol{\theta}|\mathrm{X}, \boldsymbol{Z}) = \log p(\mathrm{X}, \boldsymbol{Z}; \boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\theta})$$

where $\boldsymbol{Z}$ is not available.

The Evidence Lower Bound

Let us first consider the case of a single random variable $\mathbf{x}$ (modeling a single observation), and consider any probability distribution $q(\mathbf{z})$. Then, for any $\boldsymbol{\theta}$

$$\begin{aligned}
\log p(\mathbf{x}; \boldsymbol{\theta}) &= \log p(\mathbf{x}; \boldsymbol{\theta}) \int_{\mathcal{Z}} q(\mathbf{z}) d\mathbf{z} \\
&= \int_{\mathcal{Z}} q(\mathbf{z}) \log p(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{z} \\
&= \int_{\mathcal{Z}} q(\mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta})} d\mathbf{z} \\
&= \int_{\mathcal{Z}} q(\mathbf{z}) \left(\log p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) - \log p(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta})\right) d\mathbf{z} \\
&= \int_{\mathcal{Z}} q(\mathbf{z}) \left(\log \frac{p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})}{q(\mathbf{z})} - \log \frac{p(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta})}{q(\mathbf{z})}\right) d\mathbf{z} \\
&= \int_{\mathcal{Z}} q(\mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})}{q(\mathbf{z})} d\mathbf{z} - \int_{\mathcal{Z}} q(\mathbf{z}) \log \frac{p(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta})}{q(\mathbf{z})} d\mathbf{z} \\
&= \mathcal{L}(q, \mathbf{x}, \boldsymbol{\theta}) + \mathcal{K}(q, \mathbf{x}, \boldsymbol{\theta})
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{L}(q, \mathbf{x}, \boldsymbol{\theta}) &\triangleq -D_{KL}\left(q(\mathbf{z})||p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})\right) \\
\mathcal{K}(q, \mathbf{x}, \boldsymbol{\theta}) &\triangleq D_{KL}\left(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta})\right)
\end{aligned}$$

and

$$D_{KL}\left(p_1(\mathbf{z})||p_2(\mathbf{z})\right) \triangleq \mathop{\mathbb{E}}_{p_1(\mathbf{z})}\left[\log \frac{p_1(\mathbf{z})}{p_2(\mathbf{z})}\right] = \int_{\mathcal{Z}} p_1(\mathbf{z}) \log \frac{p_1(\mathbf{z})}{p_2(\mathbf{z})} d\mathbf{z}$$

is the Kullback-Leibler divergence[7] between $p_1(\mathbf{z})$ and $p_2(\mathbf{z})$, which measures the difference between the two distributions.

In our case,

- $\mathcal{L}(q, \mathbf{x}, \boldsymbol{\theta})$ is the negative of the KL divergence between $q(\mathbf{z})$ and the joint distribution $p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$
- $\mathcal{K}(q, \mathbf{x}, \boldsymbol{\theta})$ is the KL divergence between $q(\mathbf{z})$ and the real posterior distribution $p(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta})$

As we may see, for what concerns $\mathcal{L}(q, \mathbf{x}, \boldsymbol{\theta})$ and $\mathcal{K}(q, \mathbf{x}, \boldsymbol{\theta})$,

[7] We remark that this is the definition of KL divergence in the continuous case. If $\mathbf{z}$ is a discrete variable the integral is substituted by a sum on $\mathcal{Z}$, as usual.

- they are not functions of $\mathbf{z}$, which is marginalized by the integrations
- they are functions of $\mathbf{x}$; however, usually we consider it fixed, since it is the assumed observation
- they are functions of $\boldsymbol{\theta}$
- they are functionals of the distribution q on $\mathbf{z}$; we remark however that, even if both of them depend on q, their sum is independent from that distribution.

The following basic properties of KL divergence will be useful

$$D_{KL}\left(p_1(\mathbf{z})||p_2(\mathbf{z})\right) \geq 0$$
$$D_{KL}\left(p_1(\mathbf{z})||p_2(\mathbf{z})\right) = 0 \quad \text{iff} \quad p_1(\mathbf{z}) = p_2(\mathbf{z}) \forall \mathbf{z}$$

as a consequence, it results that for all $\mathbf{x}$,

$$\mathcal{L}(q, \mathbf{x}, \boldsymbol{\theta}) \leq \log p(\mathbf{x}; \boldsymbol{\theta}) \qquad \text{whatever the distribution } q$$

This is due to the first property above, since that implies that

$$\mathcal{K}(q, \mathbf{x}, \boldsymbol{\theta}) = D_{KL}\left(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta})\right) \geq 0$$

and, then

$$\mathcal{L}(q, \mathbf{x}, \boldsymbol{\theta}) = \log p(\mathbf{x}; \boldsymbol{\theta}) - \mathcal{K}(q, \mathbf{x}, \boldsymbol{\theta}) \leq \log p(\mathbf{x}; \boldsymbol{\theta})$$

Thus, for all observations $\mathbf{x}$ and parameter values $\boldsymbol{\theta}$, $\mathcal{L}(q, \mathbf{x}, \boldsymbol{\theta})$ is indeed a lower bound for $\log p(\mathbf{x}; \boldsymbol{\theta})$, regardless of the distribution $q(\mathbf{z})$.

$\mathcal{L}(q, \mathbf{x}, \boldsymbol{\theta})$ is called the **evidence lower bound** (ELBO) for $\boldsymbol{\theta}$ and $\mathbf{x}$.[8]

Observe also that

$$\mathcal{L}(q, \mathbf{x}, \boldsymbol{\theta}) = \mathop{\mathbb{E}}_{q(\mathbf{z})}[\log p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})] - \mathop{\mathbb{E}}_{q(\mathbf{z})}[q(\mathbf{z})]$$

which shows that, given the distribution q, the ELBO differs by a constant (the entropy of $q(\mathbf{z})$), from the expectation of the log joint distribution $\log p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})$ wrt $\mathbf{z}$, assumed distributed according to $q(\mathbf{z})$.

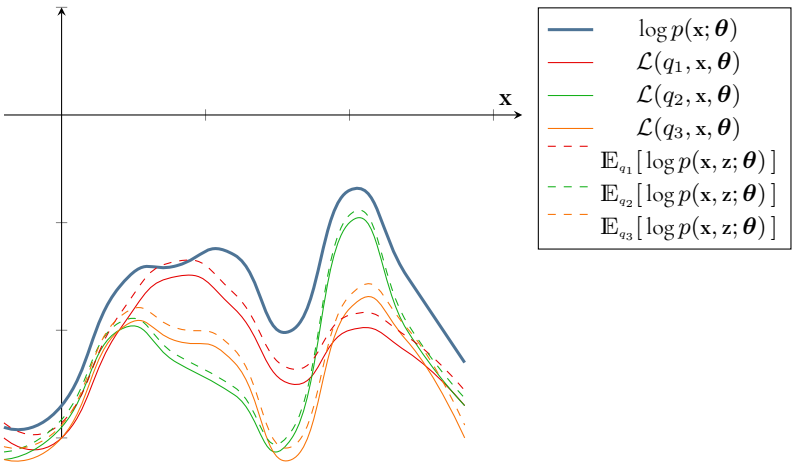

Figure 4: Each distribution q_i provides a different ELBO to the log-likelihood, given the value $\boldsymbol{\theta}$ of the parameter.

[8] This is in particular true in the case when $\mathbf{x}$ has a given value $\mathbf{x}$.

As shown in Figure 4, different distributions q_1, q_2, q_3 provide different lower bounds $\mathcal{L}(q_1, \mathbf{x}, \boldsymbol{\theta})$, $\mathcal{L}(q_2, \mathbf{x}, \boldsymbol{\theta})$, $\mathcal{L}(q_3, \mathbf{x}, \boldsymbol{\theta})$ to $\log p(\mathbf{x}; \boldsymbol{\theta})$.

A general way to deal with the iterative maximization of $\log p(\mathbf{x}; \boldsymbol{\theta})$ wrt $\boldsymbol{\theta}$ can be introduced by exploiting the ELBO as follows:

1. Given the current value $\boldsymbol{\theta}^{(k)}$, find a distribution $q^{(k)}$ which provides the best possible ELBO $\mathcal{L}(q^{(k)}, \mathbf{x}, \boldsymbol{\theta})$ to $\log p(\mathbf{x}; \boldsymbol{\theta}^{(k)})$, that is such that $\mathcal{L}(q^{(k)}, \mathbf{x}, \boldsymbol{\theta}^{(k)})$ is maximum, or equivalently $\mathcal{K}(q^{(k)}, \mathbf{x}, \boldsymbol{\theta}^{(k)})$ is minimum (Figure 5).

2. Consider the resulting ELBO

$$\mathcal{L}(q^{(k)}, \mathbf{x}, \boldsymbol{\theta}) = \mathop{\mathbb{E}}_{q^{(k)}(\mathbf{z})} [\log p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})] - \mathop{\mathbb{E}}_{q^{(k)}(\mathbf{z})} \left[q^{(k)}(\mathbf{z}) \right]$$

   as a function of $\boldsymbol{\theta}$ and maximize it wrt such parameter. Here,

   - the first term $\mathbb{E}_{q^{(k)}(\mathbf{z})}[\log p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})]$ is the expectation of the log-likelihood $\log p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}^{(k)})$ of $\mathbf{x}, \mathbf{z}$ assuming a distribution $q^{(k)}(\mathbf{z})$ of $\mathbf{z}$
   - the second term $-\mathbb{E}_{q^{(k)}(\mathbf{z})}\left[\log q^{(k)}(\mathbf{z})\right]$ is the entropy $H_{q^{(k)}(\mathbf{z})}(\mathbf{z})$ of $\mathbf{z} \sim q^{(k)}(\mathbf{z})$.

   Since the second term is a constant wrt $\boldsymbol{\theta}$, this is equivalent to the maximization of the expectation

$$Q(q^{(k)}, \mathbf{x}, \boldsymbol{\theta}) \triangleq \mathop{\mathbb{E}}_{q^{(k)}(\mathbf{z})} [\log p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})]$$

   with respect to $\boldsymbol{\theta}$.

In summary,

1. compute

$$q^{(k)} = \mathop{\text{argmax}}_{q} \mathcal{L}(q, \mathbf{x}, \boldsymbol{\theta}^{(k)})$$

2. compute

$$\boldsymbol{\theta}^{(k+1)} = \mathop{\text{argmax}}_{\boldsymbol{\theta}} \mathcal{L}(q^{(k)}, \mathbf{x}, \boldsymbol{\theta}) = \mathop{\text{argmax}}_{\boldsymbol{\theta}} Q(q^{(k)}, \mathbf{x}, \boldsymbol{\theta})$$

By construction, we have that

$$\log p(\mathbf{x}; \boldsymbol{\theta}) \geq \mathcal{L}(q_1, \mathbf{x}, \boldsymbol{\theta})$$
$$\log p(\mathbf{x}; \boldsymbol{\theta}') \geq \mathcal{L}(q_1, \mathbf{x}, \boldsymbol{\theta}') \geq \mathcal{L}(q_1, \mathbf{x}, \boldsymbol{\theta})$$

This however does not guarantee, in general, that

$$\log p(\mathbf{x}; \boldsymbol{\theta}') \geq \log p(\mathbf{x}; \boldsymbol{\theta})$$

that is the monotone increase of the log-likelihood (Figure 7).

Observe that the gap between the ELBO $\mathcal{L}(q, \mathbf{x}, \boldsymbol{\theta})$ and the log-likelihood $\log p(\mathbf{x}; \boldsymbol{\theta})$ is equal to 0 if

$$D_{KL}\left(q(\mathbf{z}) || p(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta})\right) = 0$$

which is true if $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta})$ for all $\mathbf{z}$, that is if q is the real conditional distribution $p(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta})$.

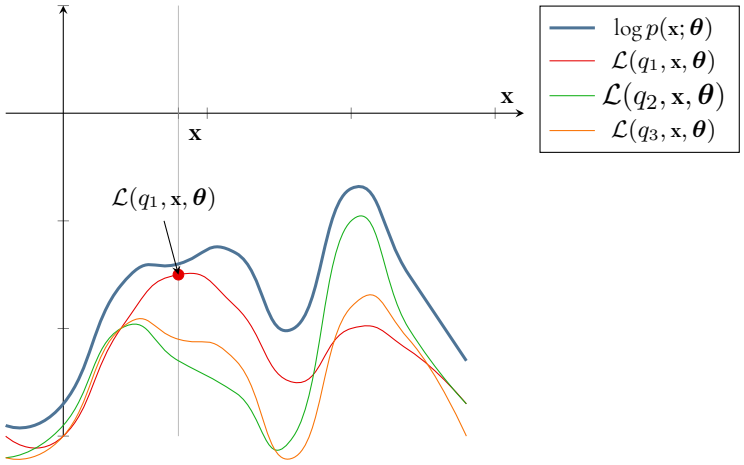

Figure 5: Different distributions provide different lower bounds to the log-likelihood, given the value $\boldsymbol{\theta}$ of the parameter.

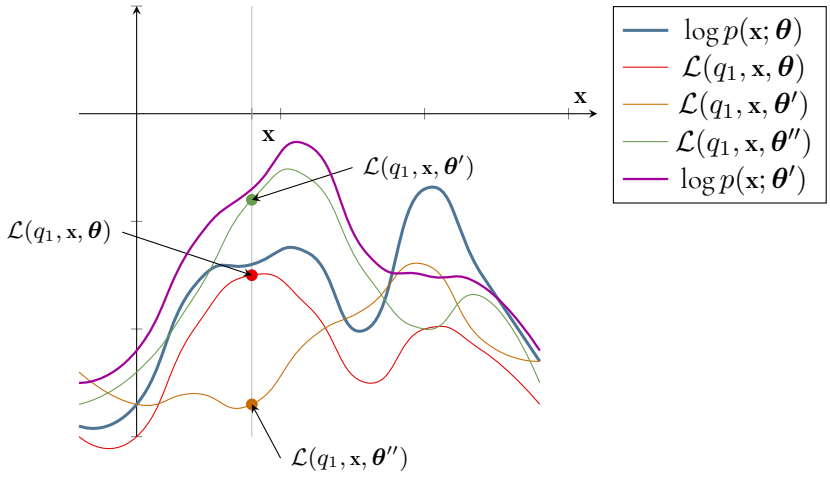

Figure 6: Different parameter values provide ELBO's to different log-likelihoods, given the same distribution q_1.

In the general case when we are dealing with a set $\mathbf{X}$ of observations, we have that

$$\begin{aligned}
\log p(\mathbf{X};\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log p(\mathbf{x}_i;\boldsymbol{\theta}) \\
&= \sum_{i=1}^{n} \mathop{\mathbb{E}}_{z\sim q_i}\left[\log \frac{p(\mathbf{x}_i,\mathbf{z};\boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{x}_i;\boldsymbol{\theta})}\right] \\
&= \sum_{i=1}^{n} \mathcal{L}(q_i,\mathbf{x}_i,\boldsymbol{\theta}) + \sum_{i=1}^{n} \mathcal{K}(q_i,\mathbf{x}_i,\boldsymbol{\theta}) \\
&\triangleq \mathcal{L}(Q,\mathbf{X},\boldsymbol{\theta}) + \mathcal{K}(Q,\mathbf{X},\boldsymbol{\theta})
\end{aligned}$$

where $Q = (q_1,\ldots,q_n)$.

Hence, the ELBO $\mathcal{L}(Q,\mathbf{X},\boldsymbol{\theta})$ on the dataset log-likelihood $\log p(\mathbf{X};\boldsymbol{\theta})$ is the sum of the ELBO's $\mathcal{L}(q_i,\mathbf{x}_i,\boldsymbol{\theta})$ on the log-likelihood $\log p(\mathbf{x}_i;\boldsymbol{\theta})$ for each observed item. The log-likelihood is also equal to the log-likelihood iff

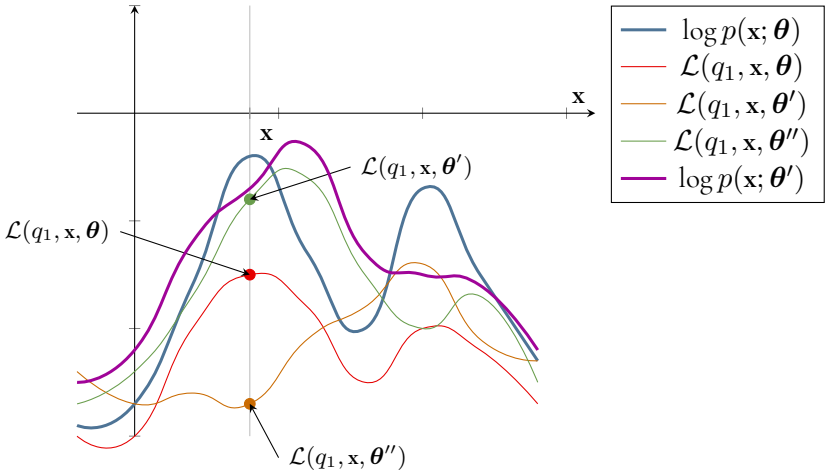

Figure 7: An example of step with not increasing log-likelihood.

$\mathcal{K}(Q, \mathbf{X}, \boldsymbol{\theta})$, that is if $q_i(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}_i)$ for $i = 1, \ldots, n$.

In this framework, the above approach requires:

1. at the first step, to identify for each $\mathbf{x}_i \in \mathbf{X}$ the best distribution $q_i(\mathbf{z})$, that is the one such that $\mathcal{L}(q_i, \mathbf{x}_i, \boldsymbol{\theta})$ is maximum ($\mathcal{K}(q_i, \mathbf{x}_i, \boldsymbol{\theta})$ is minimum)

2. at the second step, to find the value $\boldsymbol{\theta}'$ which maximizes

$$\mathcal{L}(Q, \mathbf{X}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \mathcal{L}(q_i, \mathbf{x}_i, \boldsymbol{\theta})$$