Graphical models, inference and learning

Filipe Rodrigues

2015

1 Probabilistic graphical models

Probabilities are at the heart of modern machine learning. Probability theory provides us with a consistent framework for quantifying and manipulating uncertainty, which is caused by limitations in our ability to observe the world, our ability to model it, and possibly even because of its innate nondeterminism (Koller and Friedman, 2009). It is, therefore, essential to account for uncertainty when building models of reality. However, probabilistic models can sometimes be quite complex. Hence, it is important to have a simple and compact manner of expressing them.

Probabilistic graphical models provide an intuitive way of representing the structure of a probabilistic model, which not only gives us insights about the properties of the model, such as conditional independencies, but also helps us design new models. A probabilistic graphical model consists of *nodes*, which represent random variables, and *edges* that express probabilistic relationships between the variables. Graphical models can be either undirected or directed. In the latter, commonly known as *Bayesian networks* (Jensen, 1996), the directionality of the edges is used to convey causal relationships (Pearl, 2014). This thesis will make extensive use of directed graphs and a special type of graphs called *factor graphs*, which generalize both directed and undirected graphs. Factor graphs are useful for solving inference problems and enabling efficient computations.

1.1 Bayesian networks

Consider an arbitrary joint distribution $p(a, \mathbf{b}, \mathbf{c})$ over the random variables a, $\mathbf{b} = \{b_n\}_{n=1}^N$ and $\mathbf{c} = \{c_n\}_{n=1}^N$ that we want to model. This joint distribution can be factorized in various ways. For instance, making use of the chain rule (or product rule) of probability, it can be verified that $p(a, \mathbf{b}, \mathbf{c}) = p(a)\, p(\mathbf{b}|a)\, p(\mathbf{c}|\mathbf{b}, a)$ and $p(a, \mathbf{b}, \mathbf{c}) = p(\mathbf{c})\, p(a, \mathbf{b}|\mathbf{c})$ are both equivalently valid factorizations of $p(a, \mathbf{b}, \mathbf{c})$. By linking variables, a probabilistic graphical model specifies how a joint distribution factorizes. Furthermore, by omitting the links between certain variables, probabilistic graphical models convey a set of conditional independencies, which simplifies the factorization.

Figure 1 shows an example of a Bayesian network model representing a factorization of the joint distribution $p(a, \mathbf{b}, \mathbf{c})$. Notice that, instead of writing out the multiple nodes for $\{b_n\}_{n=1}^N$ and $\{c_n\}_{n=1}^N$ explicitly, a rectangle with the label N was used to indicate that the structure within it repeats N times. This rectangle is called a *plate*. Also, we adopted the convention of using large circles to represent

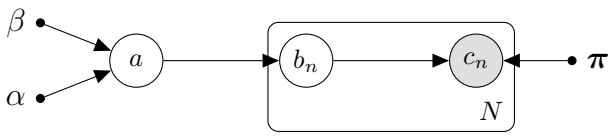

Figure 1: Example of a (directed) graphical model.

random variables (a, b_n and c_n) and small solid circles to denote deterministic parameters (α, β and $\boldsymbol{\pi}$) (Bishop, 2006). Observed variables are identified by shading their nodes. The unobserved variables, also known as *hidden* or *latent* variables, are indicated using unshaded nodes.

By reading off the dependencies expressed in the probabilistic graphical model of Figure 1, the joint distribution of the model, given the parameters α, β and $\boldsymbol{\pi}$, factorizes as

$$p(a, \mathbf{b}, \mathbf{c}|\alpha, \beta, \boldsymbol{\pi}) = p(a|\alpha, \beta) \prod_{n=1}^{N} p(b_n|a)\, p(c_n|b_n, \boldsymbol{\pi}). \tag{1}$$

Hence, rather than encoding the probability of every possible assignment to all the variables in the domain, the joint probability breaks down into a product of smaller factors, corresponding to conditional probability distributions over a much smaller space of possibilities, thus leading to a substantially more compact representation that requires significantly less parameters.

So far we have not discussed the form of the individual factors. It turns out that, for *generative models* such as the one in Figure 1, a great way to do so is through what is called the *generative process* of the model. Generative models specify how to randomly generate observable data, such as c_n in our example, typically given some latent variables, such as a and b_n. They contrast with *discriminative models* by being full probabilistic models of all the variables, whereas discriminative approaches model only the *target* variables conditional on the observed ones. A generative process is then a description of how to sample observations according to the model.

Returning to our previous example of Figure 1, a possible generative process is as follows:[1]

1. Draw $a|\alpha, \beta \sim \text{Beta}(a|\alpha, \beta)$
2. For each n
   (a) Draw $b_n|a \sim \text{Bernoulli}(b_n|a)$
   (b) Draw $c_n|b_n, \boldsymbol{\pi} \sim \text{Bernoulli}(c_n|\pi_{b_n})$

Given this generative process, we know that, for example, the variable a follows a beta distribution with parameters α and β. Similarly, the conditional probability of c_n given b_n is a Bernoulli distribution with parameter π_{b_n}. Generative processes are then an excellent way of presenting a generative model, and they complement the framework of probabilistic graphical models by conveying additional details. Also, when designing models of reality, it is often useful to think generatively and describe how the observed data came to be. Hence, we shall make extensive use of generative processes throughout this thesis for presenting models.

[1] A familiar reader might recognize this as an example of a mixture model.

1.2 Factor graphs

Directed and undirected probabilistic graphical models allow a global function of several variables to be expressed as a product of factors over subsets of those variables. These factors can be, for example, probability distributions, as we saw with Bayesian networks (see Eq. 1). Factor graphs (Kschischang et al., 2001) differ from directed and undirected graphical models by introducing additional nodes for explicitly representing the factors, which allows them to represent a wider spectrum of distributions (Koller and Friedman, 2009). Figure 2 shows an example of a factor graph over the variables a, b, c and d, where the factors are represented using small solid squares.

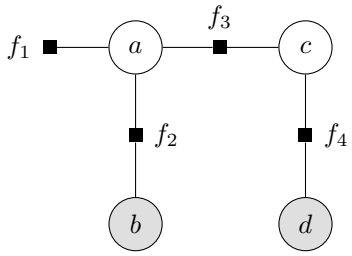

Figure 2: Example of a factor graph.

Like Bayesian networks, factor graphs encode a joint probability distribution over a set of variables. However, in factor graphs, the factors do not need to be probability distributions. For example, the factor graph in Figure 2 encodes the following factorization of the joint probability distribution over the variables a, b, c and d

$$p(a, b, c, d) = \frac{1}{Z} f_1(a)\, f_2(a, b)\, f_3(a, c)\, f_4(c, d). \tag{2}$$

Notice how the factors are now arbitrary functions of subsets of variables. Hence, a normalization constant Z is required to guarantee that the joint distribution $p(a, b, c, d)$ is properly normalized. If the factors correspond to normalized probability distributions, the normalization constant Z can be ignored.

As we shall see later, a great advantage of factor graphs is that they allow the development of efficient inference algorithms by propagating *messages* in the graph (Kschischang et al., 2001; Murphy, 2012) (see Section 2.3).

2 Bayesian inference

Having specified the probabilistic model, the next step is to perform *inference*. Inference is the procedure that allows us to answer various types of questions about the data being modeled, by computing the posterior distribution of the latent variables given the observed ones. For instance, in the example of Figure 1, we would like to compute the posterior distribution of a and $\mathbf{b}$, given the observations $\mathbf{c}$. Bayesian inference is a particular method for performing statistical inference, in which Bayes' rule is used to update the posterior distribution of a certain variable(s) as new evidence is acquired.

Bayesian inference can be exact or approximate. In this thesis we will make use of both exact inference and approximate inference procedures, namely *variational inference* (Jordan et al., 1999; Wainwright and Jordan, 2008) and *expectation propagation* (EP) (Minka, 2001).

2.1 Exact inference

Without loss of generality, let $\mathbf{z} = \{z_m\}_{m=1}^M$ denote the set of latent variables in a given model, and let $\mathbf{x} = \{x_n\}_{n=1}^N$ denote the observations. Using Bayes' rule, the posterior distribution of $\mathbf{z}$ can be computed as

$$\overbrace{p(\mathbf{z}|\mathbf{x})}^{\text{posterior}} = \frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{x})} = \frac{\overbrace{p(\mathbf{x}|\mathbf{z})}^{\text{likelihood}}\overbrace{p(\mathbf{z})}^{\text{prior}}}{\underbrace{p(\mathbf{x})}_{\text{evidence}}}. \tag{3}$$

The model evidence, or *marginal likelihood*, can be computed by making use of the sum rule of probability to give

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}|\mathbf{z})\, p(\mathbf{z}), \tag{4}$$

where the summation is replaced by integration in the case that $\mathbf{z}$ is continuous instead of discrete.

At this point, it is important to introduce a broad class of probability distributions called the *exponential family* (Duda and Hart, 1973; Bernardo and Smith, 2009). A distribution over $\mathbf{z}$ with parameters $\boldsymbol{\eta}$ is a member of the exponential family if it can be written in the form

$$p(\mathbf{z}|\boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta})}\, h(\mathbf{z})\, \exp(\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{z})), \tag{5}$$

where $\boldsymbol{\eta}$ are called the *natural parameters*, $\mathbf{u}(\mathbf{z})$ is a vector of *sufficient statistics* and $h(\mathbf{z})$ is a scaling constant, often equal to 1. The normalization constant $Z(\boldsymbol{\eta})$, also called the *partition function*, ensures that the distribution is normalized.

Many popular distributions belong to the exponential family, such as the Gaussian, exponential, beta, Dirichlet, Bernoulli, multinomial and Poisson (Bernardo and Smith, 2009). Exponential family members have many interesting properties, which make them so appealing for modelling random variables. For example, the exponential family has finite-sized sufficient statistics, which means that the data can be compressed into a fixed-sized summary without loss of information.

A particularly useful property of exponential family members is that they are closed under multiplication. This means that if we multiply together two exponential family distributions $p(\mathbf{z})$ and $p(\mathbf{z}')$, the product $p(\mathbf{z},\mathbf{z}') = p(\mathbf{z})\, p(\mathbf{z}')$ will also be in the exponential family. This property is closely related to the concept of *conjugate priors*. In general, for a given posterior distribution $p(\mathbf{z}|\mathbf{x})$, we seek a prior distribution $p(\mathbf{z})$ so that when multiplied by the likelihood $p(\mathbf{x}|\mathbf{z})$, the posterior has the same functional form as the prior. This is called a conjugate prior. For any member of the exponential family there exists a conjugate prior (Bishop, 2006; Bernardo and Smith, 2009). For example, the conjugate prior for the parameters of a multinomial

distribution is the Dirichlet distribution, while the conjugate prior for the mean of a Gaussian is another Gaussian. As we shall see, the choice of conjugate priors greatly simplifies the calculations involved in Bayesian inference. Furthermore, the fact that the posterior keeps the same functional form as the prior, allows the development of online learning algorithms, where the posterior is used as the new prior, as new observations are sequentially acquired.

2.2 Variational inference

Unfortunately, for various models of practical interest, it is infeasible to evaluate the posterior distribution exactly or to compute expectations with respect to it. There are several reasons for this. For example, it might be the case where the dimensionality of the latent space is too high to work with directly, or because the form of the posterior distribution is so complex that computing expectations is not analytically tractable, or even because some of the required integrations might not have closed-form solutions. Consider, for example, the case of the model of Figure 1. The posterior distribution over the latent variables a and $\mathbf{b}$ is given by

$$p(a, \mathbf{b}|\mathbf{c}) = \frac{p(a, \mathbf{b}, \mathbf{c})}{p(\mathbf{c})} = \frac{p(a|\alpha, \beta) \prod_{n=1}^{N} p(b_n|a)\, p(c_n|b_n, \boldsymbol{\pi})}{\int_a \sum_{\mathbf{b}} p(a|\alpha, \beta) \prod_{n=1}^{N} p(b_n|a)\, p(c_n|b_n, \boldsymbol{\pi})}. \tag{6}$$

The numerator can be easily evaluated for any combination of the latent variables, but the denominator is intractable to compute. In such cases, where computing the exact posterior distribution is infeasible, we need to resort to approximate inference algorithms, which turn the computation of posterior distributions into a tractable problem, by trading off computation time for accuracy.

We can differentiate between two major classes of approximate inference algorithms, depending on whether they rely on stochastic or deterministic approximations. Stochastic techniques for approximate inference, such as Markov chain Monte Carlo (MCMC) (Gilks, 2005), rely on sampling and have the property that given infinite computational resources they can generate exact results. For example, MCMC methods are based on Monte Carlo approximations, whose main idea is to use repeated sampling to approximate the desired distribution. MCMC methods iteratively construct a Markov chain of samples, which, at the some point, converges. At this stage, the sample draws are close to the true posterior distribution and they can be collected to approximate the required expectations. However, in practice, it is hard to determine when a chain has converged or "mixed". Furthermore, the number of samples required for the chain to mix can be very large. As a consequence, MCMC methods tend to be computationally demanding, which generally restricts their application to small-scale problems (Bishop, 2006). On the other hand, deterministic methods, such as variational inference and expectation propagation, are based on analytical approximations to the posterior distribution. Therefore, they tend to scale better to large-scale inference problems, making them better suited for the models proposed in this thesis.

Variational inference, or *variational Bayes* (Jordan et al., 1999; Wainwright and Jordan, 2008), constructs an approximation to the true posterior distribution $p(\mathbf{z}|\mathbf{x})$ by considering a family of tractable distributions $q(\mathbf{z})$. A tractable family can be obtained by relaxing some constraints in the true distribution. Then, the inference

problem is to optimize the parameters of the new distribution so that the approximation becomes as close as possible to the true posterior. This reduces inference to an optimization problem.

The closeness between the approximate posterior $q(\mathbf{z})$, known as the *variational distribution*, and the true posterior $p(\mathbf{z}|\mathbf{x})$ can be measured by the Kullback-Leibler (KL) divergence (MacKay, 2003), which is given by

$$\mathbb{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) = \int_{\mathbf{z}} q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x})}. \tag{7}$$

Notice that the KL divergence is an asymmetric measure. Hence, we could have chosen the reverse KL divergence, $\mathbb{KL}(p(\mathbf{z}|\mathbf{x})||q(\mathbf{z}))$, but that would require us to be able to take expectations with respect to $p(\mathbf{z}|\mathbf{x})$. In fact, that would lead to a different kind of approximation algorithm, called expectation propagation, which shall be discussed in Section 2.3.

Unfortunately, the KL divergence in (7) cannot be minimized directly. However, we can find a function that we can minimize, which is equal to it up to an additive constant, as follows

$$\begin{aligned}
\mathbb{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) &= \mathbb{E}_q\left[\log \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x})}\right] \\
&= \mathbb{E}_q[\log q(\mathbf{z})] - \mathbb{E}_q[\log p(\mathbf{z}|\mathbf{x})] \\
&= \mathbb{E}_q[\log q(\mathbf{z})] - \mathbb{E}_q\left[\log \frac{p(\mathbf{z},\mathbf{x})}{p(\mathbf{x})}\right] \\
&= -\underbrace{(\mathbb{E}_q[\log p(\mathbf{z},\mathbf{x})] - \mathbb{E}_q[\log q(\mathbf{z})])}_{\mathcal{L}(q)} + \underbrace{\log p(\mathbf{x})}_{\text{const.}}.
\end{aligned} \tag{8}$$

The $\log p(\mathbf{x})$ term of (8) does not depend on q and thus it can be ignored. Minimizing the KL divergence is then equivalent to maximizing $\mathcal{L}(q)$, which is called the *evidence lower bound*. The fact that $\mathcal{L}(q)$ is a lower bound on the log model evidence, $\log p(\mathbf{x})$, can be emphasized by recalling Jensen's inequality to notice that, due to the concavity of the logarithmic function, $\log \mathbb{E}[p(\mathbf{x})] \geqslant \mathbb{E}[\log p(\mathbf{x})]$. Thus, Jensen's inequality can be applied to the logarithm of the model evidence to give

$$\begin{aligned}
\log p(\mathbf{x}) &= \log \int_{\mathbf{z}} p(\mathbf{z},\mathbf{x}) \\
&= \log \int_{\mathbf{z}} \frac{q(\mathbf{z})}{q(\mathbf{z})} p(\mathbf{z},\mathbf{x}) \\
&= \log \mathbb{E}_q\left[\frac{p(\mathbf{z},\mathbf{x})}{q(\mathbf{z})}\right] \\
&\geqslant \underbrace{\mathbb{E}_q[\log p(\mathbf{z},\mathbf{x})] - \mathbb{E}_q[\log q(\mathbf{z})]}_{\mathcal{L}(q)}.
\end{aligned} \tag{9}$$

The evidence lower bound $\mathcal{L}(q)$ is tight when $q(\mathbf{z}) \approx p(\mathbf{z}|\mathbf{x})$, in which case $\mathcal{L}(q) \approx \log p(\mathbf{x})$. The goal of variational inference is then to find the parameters of the variational distribution $q(\mathbf{z})$, known as the *variational parameters*, that maximize the evidence lower bound $\mathcal{L}(q)$.

The key to make variational inference work is to find a tractable family of approximate distributions $q(\mathbf{z})$ for which the expectations in (9) can be easily computed. The most common choice for $q(\mathbf{z})$ is a fully factorized distribution, such that

$q(\mathbf{z}) = \prod_{m=1}^{M} q(z_m)$. This is called a *mean-field* approximation. In fact, mean field theory is by itself a very important topic in statistical physics (Parisi, 1988).

Using a mean-field approximation corresponds to assuming that the latent variables $\{z_i\}_{i=1}^{M}$ are independent of each other. Hence, the expectations in (9) become sums of simpler expectations. For example, the term $\mathbb{E}_q[\log q(\mathbf{z})]$ becomes $\mathbb{E}_q[\log q(\mathbf{z})] = \sum_{m=1}^{M} \mathbb{E}_q[\log q(z_m)]$. The evidence lower bound, $\mathcal{L}(q)$, can then be optimized by using a coordinate ascent algorithm that iteratively optimizes the variational parameters of the approximate posterior distribution of each latent variable $q(z_m)$ in turn, holding the others fixed, until a convergence criterium is met. This ensures convergence to a local maximum of $\mathcal{L}(q)$.

2.3 Expectation propagation

Expectation propagation (EP) (Minka, 2001) is another deterministic method for approximate inference. It differs from variational inference by considering the reverse KL divergence $\mathbb{KL}(p||q)$ instead of $\mathbb{KL}(q||p)$. This gives the approximation different properties.

Consider an arbitrary probabilistic graphical model encoding a joint probability distribution over observations $\mathbf{x} = \{x_n\}_{n=1}^{N}$ and latent variables $\mathbf{z} = \{z_m\}_{m=1}^{M}$, so that it factorizes as a product of factors $f_i(\mathbf{z})$

$$p(\mathbf{z}, \mathbf{x}) = \prod_i f_i(\mathbf{z}), \tag{10}$$

where we omitted the dependence of the factors on the observations for the ease of exposition and to keep the presentation coherent with the literature (Minka, 2001; Bishop, 2006; Murphy, 2012). The posterior distribution of the latent variables is then given by

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{z}, \mathbf{x})}{p(\mathbf{x})} = \frac{1}{p(\mathbf{x})} \prod_i f_i(\mathbf{z}). \tag{11}$$

The model evidence $p(\mathbf{x})$ is obtained by marginalizing over the latent variables, i.e. $p(\mathbf{x}) = \int_{\mathbf{z}} \prod_i f_i(\mathbf{z})$, where the integral is replaced by a summation in the case that $\mathbf{z}$ is discrete. However, without loss of generality, we shall assume for the rest of this section that $\mathbf{z}$ is continuous.

In expectation propagation, we consider an approximation to the posterior distribution of the form

$$q(\mathbf{z}) = \frac{1}{Z_{\text{EP}}} \prod_i \widetilde{f}_i(\mathbf{z}), \tag{12}$$

where the normalization constant Z_{EP} is required to ensure that the distribution integrates to unity. Just as with variational inference, the approximate posterior $q(\mathbf{z})$ needs to be restricted in some way, in order for the required computations to be tractable. In particular, we shall assume that the approximate factors $\widetilde{f}_i(\mathbf{z})$ belong to the exponential family, so that the product of all the factors is also in the exponential family and thus can be described by a finite set of sufficient statistics.

As previously mentioned, expectation propagation considers the reverse KL, $\mathbb{KL}(p||q)$. However, minimizing the global KL divergence between the true posterior

and the approximation, $\mathbb{KL}(p(\mathbf{z}|\mathbf{x})||q(\mathbf{z}))$, is generally intractable. Alternatively, one could consider minimizing the local KL divergences between the different individual factors, $\mathbb{KL}(f_i(\mathbf{z})||\widetilde{f}_i(\mathbf{z}))$, but that would give no guarantees that the product of all the factors $\prod_i \widetilde{f}_i(\mathbf{z})$ would be a good approximation to $\prod_i f_i(\mathbf{z})$, and actually, in practice, it leads to poor approximations (Bishop, 2006). EP uses a tractable compromise between these two alternatives, where the approximation is made by optimizing each factor in turn in the context of all the remaining factors.

Let us now see in more detail how the posterior approximation of EP is done. Suppose we want to refine the factor approximation $\widetilde{f}_j(\mathbf{z})$, and let $p^{\backslash j}(\mathbf{z})$ and $q^{\backslash j}(\mathbf{z})$ be the product of all the other factors (exact or approximate) that do not involve j, i.e. $p^{\backslash j}(\mathbf{z}) \triangleq \prod_{i \neq j} f_i(\mathbf{z})$ and $q^{\backslash j}(\mathbf{z}) \triangleq \prod_{i \neq j} \widetilde{f}_i(\mathbf{z})$. This defines the *context* of a factor. Ideally, in order to optimize a given factor $\widetilde{f}_j(\mathbf{z})$, we would like to minimize the KL divergence $\mathbb{KL}(p(\mathbf{z}|\mathbf{x})||q(\mathbf{z}))$, which can be written as

$$\mathbb{KL}\left(\frac{1}{p(\mathbf{x})} f_j(\mathbf{z})\, p^{\backslash j}(\mathbf{z}) \middle|\middle| \frac{1}{Z_{\text{EP}}} \widetilde{f}_j(\mathbf{z})\, q^{\backslash j}(\mathbf{z})\right), \tag{13}$$

but, as previously mentioned, this is intractable to compute. We can make this tractable by assuming that the approximations we already made, $q^{\backslash j}(\mathbf{z})$, are a good approximation for the rest of the distribution, i.e. $q^{\backslash j}(\mathbf{z}) \approx p^{\backslash j}(\mathbf{z})$. This corresponds to making the approximation of the factor $\widetilde{f}_j(\mathbf{z})$ in the context of all the other factors, which ensures that the approximation is most accurate in the regions of high posterior probability as defined by the remaining factors (Minka, 2001). Of course the closer the context approximation $q^{\backslash j}(\mathbf{z})$ is to the true context $p^{\backslash j}(\mathbf{z})$, the better the approximation for the factor $\widetilde{f}_j(\mathbf{z})$ will be. EP starts by initializing the factors $\widetilde{f}_i(\mathbf{z})$ and then iteratively refines each of these factors one at the time, much like the coordinate ascent algorithm used in variational inference iteratively optimizes the evidence lower bound with respect to one of the variational parameters.

Let $q(\mathbf{z})$ be the current posterior approximation and let $\widetilde{f}_j(\mathbf{z})$ be the factor we wish to refine. The context $q^{\backslash j}(\mathbf{z})$, also known as the *cavity* distribution, can be obtained either by explicitly multiplying all the other factors except $\widetilde{f}_j(\mathbf{z})$ or, more conveniently, by dividing the current posterior approximation $q(\mathbf{z})$ by $\widetilde{f}_j(\mathbf{z})$

$$q^{\backslash j}(\mathbf{z}) = \frac{q(\mathbf{z})}{\widetilde{f}_j(\mathbf{z})}. \tag{14}$$

Notice that $q^{\backslash j}(\mathbf{z})$ corresponds to an unnormalized distribution, so that it requires its own normalization constant Z_j in order to be properly normalized. We then wish to estimate the new approximate distribution $q^{\text{new}}(\mathbf{z})$ that minimizes the KL divergence

$$\mathbb{KL}\left(\frac{1}{Z_j} f_j(\mathbf{z})\, q^{\backslash j}(\mathbf{z}) \middle|\middle| q^{\text{new}}(\mathbf{z})\right). \tag{15}$$

It turns out that, as long as $q^{\text{new}}(\mathbf{z})$ is in the exponential family, this KL divergence can be minimized by setting the expected sufficient statistics of $q^{\text{new}}(\mathbf{z})$ to the corresponding moments of $Z_j^{-1} f_j(\mathbf{z})\, q^{\backslash j}(\mathbf{z})$ (Koller and Friedman, 2009; Murphy, 2012), where the normalization constant is given by $Z_j = \int_{\mathbf{z}} f_j(\mathbf{z})\, q^{\backslash j}(\mathbf{z})$. The revised factor

can then be computed as

$$\widetilde{f}_j(\mathbf{z}) = Z_j \, \frac{q^{\text{new}}(\mathbf{z})}{q^{\setminus j}(\mathbf{z})}. \tag{16}$$

In many situations, it is useful to interpret the expectation propagation algorithm as message-passing in a factor graph. This perspective can be obtained by viewing the approximation $\widetilde{f}_j(\mathbf{z})$ as the message that factor j sends to the rest of the network, and the context $q^{\setminus j}(\mathbf{z})$ as the collection of messages that factor j receives. The algorithm then alternates between computing expected sufficient statistics and propagating these in the graph, hence the name "expectation propagation".

By considering the reverse KL divergence, the approximations produced by EP have rather different properties than those produced by variational inference. Namely, while the former are "moment matching", the latter are "mode seeking". This is particularly important when the posterior is highly multimodal. Multimodality can be caused by non-identifiability in the latent space or by complex nonlinear dependencies (Bishop, 2006). When a multimodal distribution is approximated by a unimodal one using the KL divergence $\mathbb{KL}(q||p)$, the resulting approximation will fit one of the modes. Conversely, if we use the reverse KL divergence, $\mathbb{KL}(p||q)$, the approximation obtained would average across all the modes. Hence, depending on the practical application at hand, one approach is preferable over the other.

3 Parameter estimation

A probabilistic model usually consists of variables, relationships between variables, and parameters. Parameters differ from latent variables by being single-valued instead of having a probability distribution over a range of possible values associated. Section 2 described exact and approximate methods for inferring the posterior distribution of the latent variables given the observed ones. In this section, we will give an overview of common approaches to find point-estimates for the parameters of a model, that will be useful for the models proposed in this thesis.

3.1 Maximum likelihood and MAP

Let $\mathbf{x} = \{x_n\}_{n=1}^N$ be a set of observed variables and $\boldsymbol{\theta}$ denote the set of model parameters. The most widely known method for determining the values of $\boldsymbol{\theta}$ is *maximum-likelihood* estimation (MLE). As the name suggests, it consists of setting the parameters $\boldsymbol{\theta}$ to the values that maximize the likelihood of the observations. For both computational and numerical stability reasons, it is convenient to maximize the logarithm of likelihood. The maximum-likelihood estimator is then given by

$$\boldsymbol{\theta}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} \big(\log p(\mathbf{x}|\boldsymbol{\theta}) \big). \tag{17}$$

This maximization problem can be easily solved by taking derivatives of $\log p(\mathbf{x}|\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ and equating them to zero in order to obtain a solution.

In many situations, we want to incorporate prior knowledge regarding the parameters $\boldsymbol{\theta}$. This can be done by defining a prior distribution over $\boldsymbol{\theta}$, $p(\boldsymbol{\theta})$. This can be useful, for instance, for regularization purposes. Suppose $\boldsymbol{\theta}$ corresponds to a vector of weights. We can prevent these to be arbitrarily large, by assigning $\boldsymbol{\theta}$

a Gaussian prior with a small variance. In fact, as it turns out, this corresponds to a popular type of regularization called ℓ_2-regularization (see Ng (2004) for an insightful discussion on different types of regularization).

Given a prior distribution over the parameters, $p(\boldsymbol{\theta})$, the posterior distribution can be obtained by making use of Bayes' theorem

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathbf{x})}. \tag{18}$$

Since $p(\mathbf{x})$ is constant w.r.t. $\boldsymbol{\theta}$, we can find a point-estimate for the parameters by maximizing the logarithm of numerator

$$\boldsymbol{\theta}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} \big(\log p(\mathbf{x}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \big). \tag{19}$$

This is called a *maximum-a-posteriori* (MAP) estimate. Notice that, contrarily to Bayesian inference, where the full posterior distribution is computed, MAP estimation determines a single point-estimate for the parameters $\boldsymbol{\theta}$, which corresponds to the mode of the posterior distribution.

3.2 Expectation maximization

Many models of practical interest often combine observed variables with latent ones. Let $\mathbf{z} = \{z_m\}_{m=1}^M$ denote the set of latent variables in the model. Without loss of generality, we shall assume that $\mathbf{z}$ is discrete. However, the discussion would still apply if $\mathbf{z}$ was continuous, simply by replacing the summations over $\mathbf{z}$ by integrals.

Ideally, for models with unobserved variables, we would like to find the parameters $\boldsymbol{\theta}$ that maximize the (log) marginal likelihood of the data (or model evidence) given by

$$\log p(\mathbf{x}|\boldsymbol{\theta}) = \log \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}). \tag{20}$$

Unfortunately, this is generally intractable to maximize directly because of the summation that appears inside the logarithm and prevents it from acting directly on the joint distribution, which would allow us to exploit the factorization of the latter to re-write $\log p(\mathbf{x}|\boldsymbol{\theta})$ as a sum of logarithms of simpler and more tractable terms. Furthermore, this optimization problem is not convex, which makes it even harder to solve.

On the other hand, if the latent variables $\mathbf{z}$ were observed, then we could simply find the parameters $\boldsymbol{\theta}$ that maximize the *complete-data* log likelihood, $\log p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})$. Since the latent variables are not observed, we cannot maximize $\log p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})$ directly. However, we can instead maximize its expected value under a current estimate, $q(\mathbf{z})$, of the posterior distribution, $p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$, which is given by

$$\mathbb{E}_q[\log p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})] = \sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}). \tag{21}$$

Using the newly estimated parameters, we can then revise our estimate, $q(\mathbf{z})$, of the posterior distribution of the latent variables, $p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$. Iterating between these two steps gives rise to the *expectation-maximization* (EM) algorithm (Dempster et al., 1977).

The EM algorithm is then a general method for estimating the parameters in a probabilistic model in the presence of latent variables. It consists of two steps: the E-step and the M-step. In the E-step, the posterior distribution of the latent variables $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{\text{old}})$ is estimated given some "old" estimate of the parameter values $\boldsymbol{\theta}^{\text{old}}$. In the M-step, we find the "new" parameters $\boldsymbol{\theta}^{\text{new}}$ that maximize

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \left(\sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) \right). \tag{22}$$

The EM algorithm iterates between these two steps until a convergence criterion is satisfied. At each iteration, the algorithm guarantees that the log likelihood of the observed data, $\log p(\mathbf{x}|\boldsymbol{\theta})$, increases. In order to verify that, let us recall Eq. 8 and re-write it as

$$\log p(\mathbf{x}|\boldsymbol{\theta}) = \mathbb{E}_q[\log p(\mathbf{z}, \mathbf{x}|\boldsymbol{\theta})] + \mathcal{H}(q) + \mathbb{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})), \tag{23}$$

where we made the model parameters $\boldsymbol{\theta}$ explicit, and defined the entropy of q, $\mathcal{H}(q) \triangleq -\mathbb{E}_q[\log q(\mathbf{z})]$. Since the KL divergence is always non-negative, we have that

$$\log p(\mathbf{x}|\boldsymbol{\theta}) \geqslant \mathbb{E}_q[\log p(\mathbf{z}, \mathbf{x}|\boldsymbol{\theta})] + \mathcal{H}(q). \tag{24}$$

Hence, the right-hand side of (24) is a lower-bound on the log marginal likelihood $\log p(\mathbf{x}|\boldsymbol{\theta})$. This bound is tight when the KL divergence term vanishes from (23). The KL divergence $\mathbb{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}))$ is zero when $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$. Hence, the E-step of the EM algorithm makes this bound tight. When this bound is tight, we have that $\log p(\mathbf{x}|\boldsymbol{\theta})$ is equal to $\mathbb{E}_q[\log p(\mathbf{z}, \mathbf{x}|\boldsymbol{\theta})]$ up to an additive constant, $\mathcal{H}(q)$, which does not depend on the model parameters $\boldsymbol{\theta}$ (see Eq. 24). The expected complete-data log likelihood, $\mathbb{E}_q[\log p(\mathbf{z}, \mathbf{x}|\boldsymbol{\theta})]$, can then be used as a proxy for optimizing $\boldsymbol{\theta}$, which corresponds to the M-step of the EM algorithm.

This view of EM as optimizing a lower bound on the (log) marginal likelihood of the data highlights its close relation with variational inference. In fact, when the exact posterior over the latent variables, $p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$, is intractable to compute, variational inference can be used in the E-step to approximate it. This procedure is commonly known as *variational Bayes* EM (VBEM) (Bernardo et al., 2003; Murphy, 2012).

References

Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (2003). The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian statistics*, 7:453–464.

Bernardo, J. and Smith, A. (2009). *Bayesian theory*, volume 405. John Wiley & Sons.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B*, pages 1–38.

Duda, R. and Hart, P. (1973). *Pattern classification and scene analysis*, volume 3. Wiley.

Gilks, W. (2005). *Markov chain Monte Carlo*. Wiley Online Library.

Jensen, F. (1996). *An introduction to Bayesian networks*, volume 210. UCL Press London.

Jordan, M. I., Ghahramani, Z., Jaakkola, T., and Saul, L. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Kschischang, F., Frey, B., and Loeliger, H. (2001). Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519.

MacKay, D. (2003). *Information Theory, Inference & Learning Algorithms*. Cambridge University Press.

Minka, T. (2001). Expectation Propagation for approximate Bayesian inference. In *Proceedings of the Conference in Uncertainty in Artificial Intelligence*, pages 362–369.

Murphy, K. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

Ng, A. (2004). Feature selection, l1 vs. l2 regularization, and rotational invariance. In *Proceedings of the International Conference on Machine Learning*, page 78. ACM.

Parisi, G. (1988). Statistical field theory. *Frontiers in Physics, Addison-Wesley*.

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Wainwright, M. and Jordan, M. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305.