Latent Variables, EM, Variational Inference
10708, Fall 2020
Pradeep Ravikumar

1 Introduction

So far we have considered the case where all of the variables are fully observed. But in many cases, we might be interested in models where some of the variables are never observed. This then begs the question: should we work with models where some of the variables can never be observed? How would we then know that the model is correct? This is the camp of Logical Positivism, a philosophical viewpoint that posits that what you cannot observe, should not be used for *scientific* reasoning. However a lot of science involves categorization, and the exact categories (e.g. group of patients with a certain syndrome, group of animals into a distinct species) could be viewed as unobserved variables.

1.1 Example: Mixture Models

The simplest latent variable PGM is where we have a scalar latent variable Z which is the parent of an observed random vector X. When Z is discrete, taking values in a finite set $\{1, \ldots, k\}$, the marginal distribution over X is called a **mixture model**. Let us consider the popular case of a Gaussian mixture model. This can be cast as an exponential family distribution, as follows. Let Z be a multinomial variable taking values in $\{1, \ldots, k\}$, so that is an exponential family distribution with sufficient statistics $\{\mathbb{I}[Z = j]\}_{j\in[k]}$, and corresponding canonical parameters $\{\alpha_j\}_{j\in[k]}$:

$$P(z) = \exp(\sum_{j=1}^{k} \alpha_j \mathbb{I}[z = j]).$$

Suppose $X \in \mathbb{R}$ when conditioned on $Z = j$ is Gaussian with sufficient statistics x, x^2, and corresponding canonical parameters $\{\gamma_j, \gamma_j'\}$:

$$P(x|Z = j) \propto \exp(\gamma_j x + \gamma_j' x^2),$$

so that:

$$P(x|z) \propto \exp\left(\sum_{j\in[k]} \gamma_j x \mathbb{I}[z = j] + \sum_{j\in[k]} \gamma_j' x^2 \mathbb{I}[z = j]\right).$$

We can thus write the entire joint distribution $P(x, z)$ as:

$$P(x, z) \propto \exp\left(\sum_{j=1}^{k} \alpha_j \mathbb{I}[z = j] + \sum_{j\in[k]} \gamma_j x \mathbb{I}[z = j] + \sum_{j\in[k]} \gamma_j' x^2 \mathbb{I}[z = j]\right),$$

so that the joint has sufficient statistics $\{\mathbb{I}(z = j), \mathbb{I}[z = j]x, \mathbb{I}[z = j]x^2\}_{j \in [k]}$.

This can also be extended to the case where we have a vector of latent variables $Z = (Z_1, \ldots, Z_m)$ and a vector of observed variables $X = (X_1, \ldots, X_p)$. A common setting is where $m = p$, and where Z follows a discrete PGM:

$$p(z) \propto \exp\left\{\sum_{s \in V} \alpha_s(z_s) + \sum_{(s,t) \in E(G)} \alpha_{st}(z_s, z_t)\right\},$$

and where each X_s given Z_s is independent of rest of variables, so that we can write:

$$p(x|z) = \prod_{s \in [p]} p(x_s|z_s),$$

thus yielding the joint:

$$p(x, z) \propto \exp\left\{\sum_{s \in V} \alpha_s(z_s) + \sum_{(s,t) \in E(G)} \alpha_{st}(z_s, z_t)\right\} \prod_{s \in [p]} p(x_s|z_s).$$

Thus, here the PGM dependencies are all in the layer of latent variables, each of which then specify a corresponding observation random variable. Such latent PGMs are used in computer vision; in denoising a noisy image for instance, Z_s could denote the denoised unobserved pixel value, while X_s denotes the noisy pixel value. When the latent PGM graph structure is a chain graph, then this is the so-called Hidden Markov Model.

1.2 Example: Factor Analysis

When Z is continuous, and for the specific case where both $P(Z)$and $P(X|Z)$ are Gaussian, the marginal distribution over X is called a **factor analysis model**.

1.3 Example: Latent Dirichlet Allocation

The mixture models above comprise a two-level hierarchical model: a layer of latent variables, and then a layer of observed varaibles. One could have further levels of latent variables that are hierarchically specified. A popular instance of this is **Latent Dirichlet Allocation**, also known as a **topic model**.

A topic model is a distribution over documents. In the so-called bag of words model, we could model each document distribution simply as a product distribution over individual word distributions. A topic model modifies this in a number of ways that encourage sharing among the different words. First, each word W in a document is a mixture of multinomial distributions, over a latent "topic" variable Z. Suppose there are k mixture components, so that $Z \in [k]$, and m words in the vocabulary, so that $W \in [m]$, then as earlier, we have:

$$P(w|z) \propto \exp\left\{ \sum_{i\in[k],j\in[m]} \gamma_{ij}\mathbb{I}[z=i]\mathbb{I}[w=j]) \right\}.$$

The topic latent variable $Z \in [k]$ is in turn drawn from a multinomial distribution, with multinomial parameters $(u_1, \ldots, u_k)$, so that:

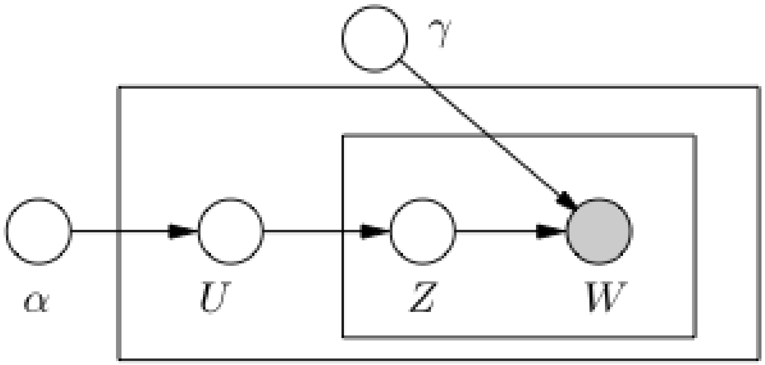
$$p(z|u) \propto \prod_{i\in[k]} u_j^{\mathbb{I}[z=j]} = \exp\{\sum_{i\in[k]} \mathbb{I}[z=j]\log u_j\}.$$

And finally, the topic multinomial parameters are drawn from a Dirichlet distribution:

$$p(u) \propto \exp\{\sum_{i\in[k]} \alpha_i \log u_i\}.$$

Thus, the overall joint is also an exponential family, but only the W variables are observed, the Z, U are never observed. The overall generative process is: draw one topic probability vector U per document, and then for each word, draw its topic latent variable Z given U, and then draw the corresponding word W given Z.

The figure below uses the "plate notation" that indicates that each graph snippet is appropriately replicated. In the above, the $Z \to W$ snippet is replicated for each word, and the U pointing to all the word snippets in a document, is replicated for each document.



2 Learning Latent Variable PGMs: Expectation Maximization

In contrast to learning fully observed PGMs, learning latent variable models is intractable even for simpler graphs. To see this, let us consider the general set up, with a latent random vector Z and an observed random vector X. Suppose the joint probability distribution is $p(x, z; \theta^*)$ for some unknown θ^*. In contrast to the fully observed case, suppose we only observe X and not Z, i.e. we are given $\{x^{(i)}\}_{i=1}^n$. If Z were also observed, the *complete log-likelihood* would be given as:

$$\ell(\theta) = \widehat{E}_{x,z} \log p(z, x; \theta),$$

where $\widehat{E}_x f(x) = \frac{1}{n}\sum_{i=1}^n f(x^{(i)})$ is the empirical expectation wrt the observed samples. Maximizing this complete log-likelihood (MLE) is a well-studied problem that even reduces to a convex optimization problem when p_θ is an exponential family distribution with canonical parameterization. But when we only observe X, we can then maximize the observed log-likelihood (also sometimes called the incomplete log-likelihood):

$$\ell_o(\theta) = \widehat{E}_x \log \sum_z p(z, x; \theta).$$

It can be seen that even when $\log p(z, x; \theta)$ is concave, it is not necessarily (and will indeed not typically) be the case that $\log \sum_z p(z, x; \theta)$ is concave. As a simple example: $\log \exp(a\theta) = a\theta$ is concave (linear), but $\log(\exp(a\theta) + \exp(b\theta))$ is no longer concave.

We can however obtain a variational characterization of the observed log-likelihood:

$$\begin{aligned}
\log p(x; \theta) &= \log \sum_z p(z, x; \theta) \\
&= \log \sum_z q(z|x) \frac{p(z, x; \theta)}{q(z|x)} \\
&\geq \sum_z q(z|x) \log \frac{p(z, x; \theta)}{q(z|x)} \\
&= \underbrace{E_{z \sim q(\cdot|x)} \log p(z, x; \theta) + H(q(z|x))}_{\mathcal{L}(x; q, \theta)}
\end{aligned}$$

where the sole inequality follows from Jensen's inequality. It can be shown that if we optimize over all possible distributions q, we get an equality, so that:

$$\log p(x; \theta) = \sup_q \mathcal{L}(x; q, \theta).$$

Note that this was the precise line of argument we used with mean-field variational inference. In both cases, we reduced a summation/integral problem over some variable to a "variational" optimization problem over distribution over the variable.

Thus, the task of maximizing observed log-likelihood can be rewritten as:

$$\max_{\theta} \widehat{E}_x \max_{q} \mathcal{L}(x; q, \theta).$$

We can solve this via alternating maximization:

- **E-step:** We optimize over q (why we call it the "E step" will be clear in the sequel):

$$q^{(t+1)}(z|x) = \arg\max_{q(\cdot|x)} \mathcal{L}(x; q, \theta^{(t)}).$$

- **M-step:** We optimize over θ:

$$\theta^{(t+1)} = \arg\max_{\theta} \widehat{E}_x \mathcal{L}(x; q^{(t+1)}, \theta).$$

Let us first consider the E-step.

$$\begin{aligned}
\mathcal{L}(x; q, \theta) &= \sum_z q(z|x) \log \frac{p(z, x; \theta)}{q(z|x)} \\
&= \sum_z q(z|x) \log \frac{p(z|x; \theta)}{q(z|x)} + \sum_z q(z|x) \log p(x; \theta) \\
&= -\mathrm{KL}(q(z|x), p(z|x; \theta)) + \log p(x; \theta),
\end{aligned}$$

which can be seen to be maximized over q at: $q = p(z|x; \theta)$.

Thus, the E-step involves in main computing:

$$q^{(t+1)}(z|x) = p(z|x; \theta^{(t)}).$$

Let us now consider the M-step. We have that:

$$\arg\max_{\theta} \mathbb{E}_x \mathcal{L}(x; q, \theta) = \arg\max_{\theta} \mathbb{E}_x E_{z \sim q(\cdot|x)} \log p(z, x; \theta),$$

since the other term $H(q)$ in $\mathcal{L}(x; q, \theta)$ does not depend on θ.

Thus, the M-step involves:

$$\theta^{(t+1)} \in \arg\max_{\theta} \mathbb{E}_x E_{z \sim q^{(t)}(\cdot|x)} \log p(z, x; \theta).$$

Thus, though the "E-step" entailed computing the conditional distribution $p(z|x; \theta^{(t)})$, we only care about a specific functional of this distribution, namely the expectation of the complete log-likelihood $\log p(z, x; \theta)$ with respect to this conditional distribution. We now see why this is called the E-step: since we could co-opt computing this expectation as part of the E-step. The M-step then involves simply optimizing the complete log-likelihood, where we substitute the expected value of the latent (wrt the earlier conditional distribution) instead of actual observations. But from an optimization standpoint, this inherits all the nice properties of MLE, such as being a convex estimation in the case of exponential families, among others.

So the EM algorithm is essentially an alternating maximization algorithm for a variational characterization of the observed log-likelihood. So why is it an approximate algorithm when the variational characterization is exact? This is because alternating maximization is not guaranteed to solve for the global optimum, even when each of the alternating maximizations wrt q and θ are solving a convex estimation problem; since it is not jointly convex in both.

3 EM and Exponential Families

Let us now revisit the above variational development of observed log-likelihood maximization in the context of exponential families, where we can provide a more compact development. Suppose the joint of the latent Z and the observable X is specified by an exponential family:

$$p_\theta(z, x) = \exp\left\{\langle \theta, \phi(z, x)\rangle - A(\theta)\right\}.$$

The observed log-likelihood of a sample x is then given as:

$$\begin{aligned}\ell(\theta; x) &= \log \int_{z \in \mathcal{Z}} \exp\left\{\langle \theta, \phi(z, x)\rangle - A(\theta)\right\} dz \\ &= \left[\underbrace{\log \int_{z \in \mathcal{Z}} \exp\left\{\langle \theta, \phi(z, x)\rangle dz\right\}}_{A_x(\theta)}\right] - A(\theta).\end{aligned}$$

The notation $A_x(\theta)$ (note that it depends on the sample x) is suggestive. Consider the exponential family distribution over z with sufficient statistics $\phi(z, x)$ for some fixed x. This is precisely the conditional distribution:

$$p_\theta(z|x) = \exp\{\langle \theta, \phi(z, x)\rangle - A_x(\theta)\},$$

where $A_x(\theta) = \log \int_{z \in \mathcal{Z}} \exp\{\langle \theta, \phi(z, x)\rangle\} dz$ is the log-partition function, and matching the expression we had earlier. So we have that the observed log-likelihood is given as:

$$\ell(\theta; x) = A_x(\theta) - A(\theta),$$

which is a difference of two convex functions, and hence not in general concave (or convex). Note that the complete log-likelihood on the other hand is given as:

$$\langle \theta, \phi(z,x) \rangle - A(\theta),$$

which is indeed concave. This thus starkly illustrates the difference between the simpler task of optimizing the complete log-likelihood, and the far more fraught task of optimizing the incomplete log-likelihood.

Let us now revisit the conditional distribution $p_\theta(z|x)$. For fixed x, this is an exponential family, so we can consider the corresponding set of mean parameters:

$$\mathcal{M}_x = \{\mu \in \mathbb{R}^d \,|\, \exists p \text{ s.t. } \mu = \mathbb{E}_{z\sim p}[\phi(z,x)]\}.$$

We can thus consider the variational characterization of $A_x(\theta)$ (where we substitute integrating over z in the log-normalization by an optimization problem):

$$A_x(\theta) = \sup_{\mu_x \in \mathcal{M}_x} \{\langle \theta, \mu_x \rangle - A_x^*(\mu_x)\},$$

where $A_x^*(\mu_x)$ is the conjugate dual:

$$A_x^*(\mu_x) = \sup_\theta \{\langle \theta, \mu_x \rangle - A(\theta)\}.$$

We thus have the following variational characterization of the incomplete log-likelihood:

$$\begin{aligned}\ell(\theta; x) &= A_x(\theta) - A(\theta)\\ &= \sup_{\mu_x \in \mathcal{M}_x} \underbrace{\{\langle \theta, \mu_x \rangle - A_x^*(\mu_x) - A(\theta)\}}_{\mathcal{L}(\mu_x,\theta)}.\end{aligned}$$

Thus, the task of maximizing the observed log-likelihood can be written as:

$$\max_\theta \widehat{E}_x \sup_{\mu_x \in \mathcal{M}_x} \{\langle \theta, \mu_x \rangle - A_x^*(\mu_x) - A(\theta)\}.$$

We can solve this via alternating maximization, which is essentially the EM algorithm.

- **E-step:** We first optimize over the conditional distribution mean-parameters:

$$\mu_x^{(t+1)} = \arg \max_{\mu_x \in \mathcal{M}_x} \mathcal{L}(\mu_x, \theta^{(t)}).$$

- **M-step:** We then optimize over the corresponding parameters θ:

$$\theta^{(t+1)} = \arg\max_\theta \mathbb{E}_x \mathcal{L}(\mu_x^{(t+1)}, \theta).$$

Let us consider the E-step:

$$\arg \sup_{\mu_x \in \mathcal{M}_x} \{\langle \theta^{(t)}, \mu_x \rangle - A_x^*(\mu_x)\},$$

which can be seen to involve solving the forward map:

$$\mu_x^{(t+1)} = \nabla A_x(\theta^{(t)}) = \mathbb{E}_{z \sim p_\theta(\cdot|x)}[\phi(z, x)],$$

so that the "expectation" in the expectation or E step is much more obvious here than in the general case.

In the M-step, we simply solve for the complete log-likelihood:

$$\arg \sup_\theta \{\langle \theta, \widehat{E}_x \mu_x^{(t+1)} \rangle - A(\theta)\},$$

which be seen to involve the backward map:

$$\theta^{(t+1)} \in (\nabla A)^{-1} \left(\widehat{E}_x \mu_x^{(t+1)} \right),$$

based on the empirical mean (wrt the observations x) of the sufficient statistics (that are in turn the conditional mean over the latent z conditioned on x, computed in the E-step).

3.1 Variational EM

Let us now consider the complexity of the E-step:

$$\mu_x^{(t+1)} = \arg \max_{\mu_x \in \mathcal{M}_x} \{\langle \theta, \mu_x \rangle - A_x^*(\mu_x)\}.$$

This involves performing exact inference in the corresponding (conditional) exponential family. When this is intractable, we might consider using approximate inference techniques, via variational approximations of the marginal polytope $\mathcal{M}_x$, and the negative entropy $A_x^*(\mu_x)$. For instance, we could use (tree-reweighted) sum-product to obtain approximations to the marginals $\mu_x^{(t+1)}$. When performing such alternating maximization, it is in general safer to use surrogate lower bounds. When maximizing a lower bound, we have a guarantee on the lower-bound argmax: its actual objective value is at least its lower bound objective value. It is thus preferable to use mean-field approximate inference in the context of EM.

Let us now consider the complexity of the M-step:

$$\arg \sup_\theta \{\langle \theta, \widehat{E}_x \mu_x^{(t+1)} \rangle - A(\theta)\},$$

This is the learning task in the fully observed exponential family. While in general this is tractable, this could be intractable as is, and might in turn require variational approximate inference, as discussed in the lecture on learning fully observed exponential families. A

Using variational approximations to solve the E or M step in EM (which is itself a variational characterization of the observed log-likelihood) is referred to as variational EM.

4 Bayesian Models, Variational Bayes

One of the most common instances of latent variable models is in a Bayesian context, where the parameters themselves are assumed to be random, and unobserved. Let us revisit the exponential family distribution:

$$p(x|\theta) = \exp(\langle \theta, \phi(x) \rangle - A(\theta)),$$

where we now assume that θ is a random variable so that the above specifies the *conditional* distribution of X given θ. An analytically convenient prior distribution for the parameters θ is such that it has the same algebraic form as a function of θ as the likelihood above:

$$p(\theta) = \exp(\langle \beta, \theta \rangle + \alpha A(\theta) - B(\alpha, \beta)).$$

In that case, the posterior $p(\theta|x)$ can be seen to be of the same family as the prior. To see this, we first compute the marginal $p(x)$:

$$\begin{aligned} p(x) &= \int_{\theta \in \Theta} p(x|\theta) p(\theta) d\theta \\ &= \int_{\theta \in \Theta} \exp(\langle \theta, \phi(x) \rangle - A(\theta) + \langle \beta, \theta \rangle + \alpha A(\theta) - B(\alpha, \beta)) d\theta \\ &= \int_{\theta \in \Theta} \exp(\langle \theta + \beta, \phi(x) \rangle + (\alpha - 1) A(\theta) - B(\alpha, \beta)) d\theta \\ &= \left[\int_{\theta \in \Theta} \exp(\langle \theta + \beta, \phi(x) \rangle + (\alpha - 1) A(\theta)) d\theta \right] \exp(-B(\alpha, \beta)) \\ &= \exp(B(\alpha - 1, \beta + \theta) - B(\alpha, \beta)), \end{aligned}$$

so that:

$$p(\theta|x) = \exp(\langle \theta + \beta, \phi(x) \rangle + (\alpha - 1) A(\theta) - B(\alpha - 1, \beta + \theta)).$$

But for general prior distributions $p(\theta)$, this posterior in general will not have such a convenient form:

$$p(\theta|x) = \frac{\exp(\langle \theta, \phi(x) \rangle - A(\theta))}{\int_{\theta \in \Theta} \exp(\langle \theta, \phi(x) \rangle - A(\theta)) p(\theta) d\theta}.$$

The key difficulty is the integral specifying the denominator for the marginal $p(x)$. As we have seen before, we could use a variational characterization of this integral instead to simplify this expression.

$$\begin{aligned}
\log p(x) &= \log \int_{\theta \in \Theta} p(x|\theta)p(\theta)d\theta \\
&= \log \int_{\theta \in \Theta} q(\theta|x)\frac{p(x|\theta)p(\theta)}{q(\theta|x)}d\theta \\
&\geq \int_{\theta \in \Theta} q(\theta|x)\log \frac{p(x|\theta)p(\theta)}{q(\theta|x)}d\theta \\
&= \underbrace{E_{\theta \sim q(\cdot|x)} \log p(\theta, x) + H(q(\theta|x))}_{\mathcal{L}(x;q)}
\end{aligned}$$

where the sole inequality follows from Jensen's inequality. It can be shown that if we optimize over all possible distributions q, we get an equality, so that:

$$\log p(x) = \sup_q \mathcal{L}(x; q).$$

We can then write the posterior as:

$$\log p(\theta|x) = \sup_q \{\log p(x|\theta) + \log p(\theta) - \mathcal{L}(x; q)\}.$$

It is common to consider simpler parameteric class of distributions $\{q_\gamma(\theta|x)\}$, in which case we obtain a lower bound on the posterior:

$$\log p(\theta|x) \geq \sup_\gamma \{\log p(x|\theta) + \log p(\theta) - \mathcal{L}(x; q_\gamma)\}.$$

This variational characterization and approximation of the posterior is often referred to as Variational Bayes.