Notes on Co-ordinate Ascent Variational Inference (CAVI)

Wilker Aziz

November 3, 2017

1 ELBO

Consider a joint distribution of latent variables $\mathbf{z} = z_1^m$ and observations $\mathbf{x} = x_1^n$

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) \tag{1}$$

where $\mathbf{z}$ help govern the distribution of the data. In a Bayesian model, we draw the latent variables from a prior $p(\mathbf{z})$ and relate them to observations through an observation model—or likelihood—$p(\mathbf{x}|\mathbf{z})$.[1] Inference in Bayesian models then consists in conditioning on data and computing the posterior $p(\mathbf{z}|\mathbf{x})$.

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{z})p(\mathbf{x}|\mathbf{z})}{p(\mathbf{x})} \tag{2}$$

The posterior can be used to making decisions, predictions, exploratory research, amongst other applications. Unfortunately, $p(\mathbf{z}|\mathbf{x})$ is not typically available in closed-form due to a generally intractable marginalisation.[2]

$$p(\mathbf{x}) = \int p(\mathbf{z})p(\mathbf{x}|\mathbf{z})\mathrm{d}\mathbf{z} \tag{3}$$

Variational inference (VI) (Jordan et al., 1999) (see (Blei et al., 2017) for a review) suggests we approximate the posterior via optimisation. We start by positing a family of

[1] In a Bayesian model every unobserved aspect of our story is a latent variable. The prior distribution may itself depend on some parameter α but in Bayesian modelling we either consider those *fixed* hyperparameters (that undergo no tuning nor optimisation) or we further extend the model hierarchy by imposing a prior $p(\alpha)$ and treating α as any other latent variable.

[2] The quantity $p(\mathbf{x})$ is of central importance in Bayesian model selection—it is known as marginal likelihood or model evidence. Note that the evidence—as well as the posterior—implicitly conditions on a particular model, i.e. $p(\mathbf{x}) \triangleq p(\mathbf{x}|\mathcal{M})$, where a model $\mathcal{M}$ is a set of conditional independence assumptions and a parametric form. Unlike in frequentist models, a Bayesian model's evidence does not depend on a particular parameter value.

distributions $\mathcal{Q}$ and proceed by finding the member of that family which minimises KL divergence to the exact posterior.

$$q^* = \underset{q \in \mathcal{Q}}{\arg\min}\ \mathrm{KL}\left[q(\mathbf{Z}) || p(\mathbf{Z}|\mathbf{x})\right] \tag{4}$$

Minimising this KL is equivalent to maximising the so called *evidence lowerbound* (ELBO).

$$\underset{q \in \mathcal{Q}}{\arg\max}\ \underbrace{\mathbb{E}_{q(\mathbf{Z})}\left[\log \frac{p(\mathbf{x}, \mathbf{Z})}{q(\mathbf{Z})}\right]}_{\mathrm{ELBO}(q)} \tag{5}$$

The objective in (4) can be justified by expressing the log-evidence as an expectation wrt an arbitrary density $q(\mathbf{z})$.[3]

$$\begin{aligned}
\log p(\mathbf{x}) &= \log \int p(\mathbf{x}, \mathbf{z}) \mathrm{d}\mathbf{z} && \text{(6a)}\\
&= \log \int p(\mathbf{x}, \mathbf{z}) \frac{q(\mathbf{z})}{q(\mathbf{z})} \mathrm{d}\mathbf{z} && \text{(6b)}\\
&= \log \mathbb{E}_{q(\mathbf{Z})}\left[\frac{p(\mathbf{x}, \mathbf{Z})}{q(\mathbf{Z})}\right] && \text{(6c)}\\
&\overset{\mathrm{JI}}{\geq} \underbrace{\mathbb{E}_{q(\mathbf{Z})}\left[\log \frac{p(\mathbf{x}, \mathbf{Z})}{q(\mathbf{Z})}\right]}_{\mathrm{ELBO}(q)} && \text{(6d)}\\
&= \mathbb{E}_{q(\mathbf{Z})}\left[\log \frac{p(\mathbf{Z}|\mathbf{x}) p(\mathbf{x})}{q(\mathbf{Z})}\right] && \text{(6e)}\\
&= \underbrace{\mathbb{E}_{q(\mathbf{Z})}\left[\log \frac{p(\mathbf{Z}|\mathbf{x})}{q(\mathbf{Z})}\right]}_{-\mathrm{KL}[q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{x})]} + \log p(\mathbf{x}) && \text{(6f)}
\end{aligned}$$

We have derived a lowerbound on the log-evidence (6d) by using Jensen's inequality.[4] Also note from (6f)

$$\log p(\mathbf{x}) \geq \log p(\mathbf{x}) - \underbrace{\mathrm{KL}\left[q(\mathbf{Z}) || p(\mathbf{Z}|\mathbf{x})\right]}_{\mathrm{gap}} \tag{7}$$

that the gap is exactly $\mathrm{KL}[q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{x})]$. Minimising the gap tightens the bound and justifies the principle stated in Equation (4). However, expressed as such, that objective is of little use—due to the intractable marginalisation in (2) we cannot compute the

[3] We are constrained to picking $q(\mathbf{z})$ such that it has support everywhere where $p(\mathbf{z})$ has support.

[4] In (6d) we used Jensen's inequality to push the log (a concave function) through the expectation (a convex combination) getting a lowerbound on the log-evidence.

exact posterior neither access it at any given point. Maximising the ELBO is far more convenient—note that (5) does not require the intractable normaliser—and as it turns out is equivalent to minimising (4).

$$
\begin{aligned}
q^* &= \arg\min_{q \in \mathcal{Q}} \ \mathrm{KL}\left[q(\mathbf{Z}) || p(\mathbf{Z}|\mathbf{x})\right] && \text{(8a)}\\
&= \arg\min_{q \in \mathcal{Q}} \mathbb{E}_{q(\mathbf{Z})}\left[\log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{x})}\right] && \text{(8b)}\\
&= \arg\min_{q \in \mathcal{Q}} \ -\mathbb{E}_{q(\mathbf{Z})}\left[\log \frac{p(\mathbf{Z}|\mathbf{x})}{q(\mathbf{Z})}\right] && \text{(8c)}\\
&= \arg\max_{q \in \mathcal{Q}} \mathbb{E}_{q(\mathbf{Z})}\left[\log \frac{p(\mathbf{Z})p(\mathbf{x}|\mathbf{Z})}{q(\mathbf{Z})p(\mathbf{x})}\right] && \text{(8d)}\\
&= \arg\max_{q \in \mathcal{Q}} \mathbb{E}_{q(\mathbf{Z})}\left[\log \frac{p(\mathbf{Z})p(\mathbf{x}|\mathbf{Z})}{q(\mathbf{Z})}\right] - \underbrace{\log p(\mathbf{x})}_{\text{constant}} && \text{(8e)}\\
&= \arg\max_{q \in \mathcal{Q}} \ \underbrace{\mathbb{E}_{q(\mathbf{Z})}\left[\log \frac{p(\mathbf{x},\mathbf{Z})}{q(\mathbf{Z})}\right]}_{\text{ELBO}} && \text{(8f)}
\end{aligned}
$$

2 Mean field and CAVI

The mean field variation family assumes that latent variables are mutually independent and each is governed by a distinct factor.

$$q(\mathbf{z}) = \prod_{i=1}^{m} q_i(z_i) \tag{9}$$

In co-ordinate ascent we optimise one factor, e.g. $q_j(z_j)$, while holding the others

$$q_{\neg j}(\mathbf{z}_{\neg j}) \triangleq \prod_{i \neq j} q_i(z_i) \tag{10}$$

fixed—we use $\mathbf{z}_{\neg j}$ to indicate all but the jth variable and similarly $q_{\neg j}$ to indicate all but the jth factor. Let us then focus on the ELBO's dependency on the jth factor

$$\mathrm{ELBO}(q_j) = \mathbb{E}_j\left[\mathbb{E}_{\neg j}\left[\log p(\mathbf{x}, \mathbf{Z}_{\neg j}, Z_j)\right]\right] \underbrace{-\mathbb{E}_j\left[\log q_j(Z_j)\right]}_{\mathbb{H}[q_j]} - \underbrace{\mathbb{E}_{\neg j}\left[\log q_{\neg j}(\mathbf{Z}_{\neg j})\right]}_{\text{constant wrt } q_j} \tag{11}$$

where we use $\mathbb{E}_j[\cdot]$ to denote an expectation wrt the jth factor and $\mathbb{E}_{\neg j}[\cdot]$ to denote an expectation with respect to all but the jth factor.[5] Now define the log-density

$$\log \pi(\mathbf{x}, z_j) = \mathbb{E}_{\neg j}\left[\log p(\mathbf{x}, \mathbf{Z}_{\neg j}, Z_j)\right] - C \tag{12}$$

[5] Equation (11) can be written like that because the mean field family is fully factorised, thus the original expectation can be written as iterated expectations and the order of iteration is irrelevant.

where C is a log-normalising constant that makes π a proper density. In other words,

$$\pi(\mathbf{x}, z_j) \propto \exp\left\{\mathbb{E}_{\neg j}\left[\log p(\mathbf{x}, \mathbf{Z}_{\neg j}, Z_j)\right]\right\} \tag{13}$$

where solving the expectation frees the dependency on all but the jth latent variable. Now note that substituting (12) into (11)

$$\begin{aligned} \mathrm{ELBO}(q_j) &= \mathbb{E}_j\left[\log \pi(\mathbf{x}, Z_j)\right] - \mathbb{E}_j\left[\log q_j(Z_j)\right] && \text{(14a)} \\ &= -\,\mathrm{KL}\left[q_j(Z_j)||\pi(\mathbf{x}, Z_j)\right] && \text{(14b)} \end{aligned}$$

reveals that the jth component of the ELBO is maximised when the KL in (14b) is minimised—which happens when $q_j(Z_j) = \pi(\mathbf{x}, Z_j)$. Thus the optimum co-ordinate update is

$$q_j^*(z_j) \propto \exp\left\{\mathbb{E}_{\neg j}\left[\log p(\mathbf{x}, \mathbf{Z}_{\neg j}, Z_j)\right]\right\} \tag{15}$$

or equivalently

$$q_j^*(z_j) \propto \exp\left\{\mathbb{E}_{\neg j}\left[\log p(Z_j|\mathbf{Z}_{\neg j}, \mathbf{x})\right]\right\} \tag{16}$$

since $\log p(\mathbf{x})$ is constant wrt q_j.

References

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, (just-accepted).

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