Problem Set 8

published: 23 June 2022
discussion: 1 July 2022

Data Compression With And Without Deep Probabilistic Models
Prof. Robert Bamler, University of Tuebingen

Course materials available at https://robamler.github.io/teaching/compress22/

Problem 8.1: Understanding the ELBO

This problem will give you some intuition for the terms that make up the evidence lower bound (ELBO) that was introduced in the lecture. In fact, we'll introduce three equivalent formulations of the ELBO, and we'll find an interpretation of each term in each of these three formulations.

In the lecture, we introduced the ELBO as follows,

$$\mathrm{ELBO}(\phi) = \mathbb{E}_{Q_\phi(Z|\mathbf{X}=\mathbf{x})}\big[\log P(Z, \mathbf{X}=\mathbf{x}) - \log Q_\phi(Z \,|\, \mathbf{X}=\mathbf{x})\big]. \tag{1}$$

Here, $P(Z, \mathbf{X})$ is the probabilistic model of the *generative process* with latent variables Z and observed variables (i.e., the message) $\mathbf{X}$. This generative model is typically given as a product, $P(Z, \mathbf{X}) = P(Z)P(\mathbf{X}|Z)$, of a prior $P(Z)$ and a likelihood $P(\mathbf{X}|Z)$. Further, $Q_\phi(Z=z \,|\, \mathbf{X})$ is the variational distribution, which has variational parameters ϕ. Finally, the expectation in Eq. 1 is taken only over the latents Z (the message $\mathbf{X} = \mathbf{x}$ is fixed).

Let's assume for simplicity that both Z and $\mathbf{X}$ are discrete. For this case, we showed in the lecture that the ELBO is the negative expected net bit rate of the modified bits-back coding algorithm ("modified" because we use $Q_\phi(Z \,|\, \mathbf{X} = \mathbf{x})$ as a stand-in for the typically intractable true posterior $P(Z \,|\, \mathbf{X} = \mathbf{x})$). Thus,

$$\mathrm{ELBO}(\phi) = -\mathbb{E}_{Q_\phi(Z|\mathbf{X}=\mathbf{x})}\big[\tilde{R}^{(Z)}_{\mathrm{net}}(\mathbf{x})\big]. \tag{2}$$

This motivated us to *maximize* the ELBO over the variational parameters ϕ (so that we minimize the expected net bit rate). We'll show now that there are also a number of other ways in which we can interpret the maximization of the ELBO.

(a) The second term on the right-hand side of Eq. 1 is the entropy of the variational distribution: $H\big[Q_\phi(Z \,|\, \mathbf{X}=\mathbf{x})\big] \equiv -\mathbb{E}_{Q_\phi(Z|\mathbf{X}=\mathbf{x})}\big[\log Q_\phi(Z \,|\, \mathbf{X}=\mathbf{x})\big]$. Thus,

$$\mathrm{ELBO}(\phi) = \mathbb{E}_{Q_\phi(Z|\mathbf{X}=\mathbf{x})}\big[\log P(Z, \mathbf{X}=\mathbf{x})\big] + H\big[Q_\phi(Z \,|\, \mathbf{X}=\mathbf{x})\big]. \tag{3}$$

Imagine the entropy term was absent, i.e., pretend that we only maximize the first term on the right-hand side of Eq. 3. Argue (in words) why maximizing only this first term over the variational parameters ϕ would lead to a deterministic variational distribution $Q_{\phi^*}(Z \,|\, \mathbf{X}=\mathbf{x})$, i.e., a variational distribution that puts all probability mass on a single z^*. Thus, we would have $Q_{\phi^*}(Z=z^* \,|\, \mathbf{X}=\mathbf{x}) = 1$ and $Q_{\phi^*}(Z \neq z^* \,|\, \mathbf{X}=\mathbf{x}) = 0$. What is the value of z^*?

Now let's return to the full expression in Eq. 3 that includes the entropy term $H\big[Q_\phi(Z \,|\, \mathbf{X}=\mathbf{x})\big]$. Argue why this entropy term acts *against* the variational distribution becoming deterministic (*hint*: what is the entropy of such a deterministic distribution that puts all its probability mass on a single value?).

(b) Show that the ELBO from Eq. 1 can also be expressed as follows,

$$\text{ELBO}(\phi) = \mathbb{E}_{Q_\phi(Z|\mathbf{X}=\mathbf{x})}\big[\log P(\mathbf{X}=\mathbf{x} \,|\, Z)\big] - D_{\text{KL}}\big(Q_\phi(Z \,|\, \mathbf{X}=\mathbf{x}) \,\big|\big|\, \log P(Z)\big). \quad (4)$$

(*Hint*: it's easier to start with Eq. 4 and derive Eq. 1 from it rather than trying it the other way round.)

Eq. 4 tells us that maximizing the ELBO can be interpreted as a *regularized maximum likelihood estimation*. To see this, answer the following two questions: what would be the optimal variational distribution $Q_{\phi^*}(Z \,|\, \mathbf{X}=\mathbf{x})$ if we were maximizing (i) only the first term or (ii) only the second term on the right-hand side of Eq. 4 over ϕ (no calculation required).

(c) Show that the ELBO from Eq. 1 can also be expressed as follows,

$$\text{ELBO}(\phi) = \log P(\mathbf{X}=\mathbf{x}) - D_{\text{KL}}\big(Q_\phi(Z \,|\, \mathbf{X}=\mathbf{x}) \,\big|\big|\, P(Z \,|\, \mathbf{X}=\mathbf{x})\big). \quad (5)$$

(*Hint*: it's again easier to start with Eq. 5 and derive Eq. 1 from it rather than trying it the other way round.)

Combining Eq. 5 with Eq. 2, derive an expression for how much the expected bit rate $\mathbb{E}_{Q_\phi(Z|\mathbf{X}=\mathbf{x})}\big[\tilde{R}^{(Z)}_{\text{net}}(\mathbf{x})\big]$ of the (modified) bits-back coding algorithm increases due to the fact that the algorithm replaces the true posterior $P(Z \,|\, \mathbf{X}=\mathbf{x})$ with the variational distribution $Q_\phi(Z \,|\, \mathbf{X} = \mathbf{x})$. Then explain why the process of maximizing the ELBO is called "variational *inference*", i.e., how does maximizing the right-hand side of Eq. 5 over ϕ relate to Bayesian inference?

Problem 8.2: Black-Box Variational Inference

In this problem, we discuss the actual task of maximizing the ELBO in Eq. 1.

The most efficient way to maximize the ELBO is the so-called coordinate ascent variational inference (CAVI) algorithm (see, e.g., review by Blei et al. (2017)). This algorithm can be derived by solving the equation $\nabla_\phi \,\text{ELBO}(\phi) = 0$ analytically for one coordinate ϕ_i at a time (by writing out the expectation on the right-hand side of Eq. 1 as an explicit integral over z, taking the derivative w.r.t. ϕ_i, and solving the resulting integrals analytically). While this CAVI algorithm is extremely fast (and should therefore be preferred whenever possible!), its application is limited because the resulting integrals can be solved analytically only for very special models (e.g., so-called conditional conjugate models).

Mainstream adoption of variational inference only occurred after the invention of so-called *black box variational inference (BBVI)*, which estimates expectations by sampling

instead of evaluating them analytically, thus making VI possible for (almost) arbitrary models. In this problem, you derive the main two approaches to BBVI.

(a) Let's first understand why BBVI is nontrivial: Eq. 1 expresses the ELBO as an expectation value: $\mathrm{ELBO}(\phi) = \mathbb{E}_{Q_\phi(Z|\mathbf{X}=\mathbf{x})}[\ell(\phi, Z)]$ with $\ell(\phi, Z) = \log P(Z, \mathbf{X}=\mathbf{x}) - \log Q_\phi(Z \mid \mathbf{X}=\mathbf{x})$. This *seems* similar to the typical situation in supervised learning, where the loss function is usually also expressed as some expectation value (in this case, the expectation is taken over samples from the training set). The method of choice for minimizing the loss function in supervised learning is usually the stochastic gradient descent algorithm (see below).

Why can't we just straight-forwardly apply stochastic gradient descent[1] to the maximization of the ELBO? In other words, why can't we do the following:

- draw some sample $z_\mathrm{s} \sim Q_\phi(Z \mid \mathbf{X}=\mathbf{x})$;
- evaluate the gradient $\hat{g} := \nabla_\phi \ell(\phi, z_\mathrm{s})$ w.r.t. ϕ at this sample;
- use this gradient as an estimate of $\nabla_\phi \mathrm{ELBO}(\phi)$, and update $\phi \leftarrow \phi + \rho\hat{g}$ with some small learning rate (aka step size) $\rho > 0$?

Hint: look for all places where ϕ appears in the ELBO.

In the following parts, we discuss two possible solutions to the problem from part (a).

(b) The simplest form of BBVI uses so-called reparameterization gradients (Kingma and Welling, 2014). Assume, for example, that the latent variable z is continuous and d-dimensional (i.e., $z \in \mathbb{R}^d$) and assume that the variational family is the set of all fully factorized normal distributions. Thus, Q_ϕ has the form

$$Q_\phi(Z=z \mid \mathbf{X}=\mathbf{x}) = \prod_{i=1}^{d} \mathcal{N}(z_i; \mu_i, \sigma_i^2) \tag{6}$$

where the means $\{\mu_i\}_{i=1}^d$ and standard deviations $\{\sigma_i\}_{i=1}^d$ together comprise the variational parameters ϕ over which we optimize.

Convince yourself that, for such a variational distribution, the expectation of any function $f(z)$ can be expressed as follows,

$$\mathbb{E}_{z\sim Q_\phi(Z|\mathbf{X}=\mathbf{x})}\big[f(z)\big] = \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)}\big[f(\boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon})\big]. \tag{7}$$

Here, $\boldsymbol{\mu} \equiv (\mu_1, \ldots, \mu_d)$ and $\boldsymbol{\sigma} \equiv (\sigma_1, \ldots, \sigma_d)$ are the concatenations into vectors of the means and standard deviations, respectively. Further, $\mathcal{N}(0, I)$ denotes a d-dimensional standard normal distribution (i.e., with zero mean and unit variance in each direction), and $\odot$ denotes elementwise multiplication of two vectors.

Now use Eq. 7 to fix the problem from part (a), i.e., to come up with an unbiased estimate of $\nabla_\phi \mathrm{ELBO}(\phi)$.

[1] more precisely, stochastic gradient *ascent* since we want to *maximize*, but that's not the issue here.

(c) While the reparameterization gradient method from part (b) can be generalized to some variational distributions other than the normal distribution, it does not work on arbitrary variational distributions. In particular, reparameterization gradients don't work (without additional tricks (Jang et al., 2016; Maddison et al., 2016)) for variational distributions over *discrete* latents Z (because they would require taking derivatives w.r.t. integers).

For such variational distributions, an alternative and more general approach called score function gradient estimates (aka the "REINFORCE method") can be used (Ranganath et al., 2014). This method is actually similar to the naive approach from part (a): one first draws some random sample $z_\mathrm{s} \sim Q_\phi(Z \,|\, \mathbf{X}=\mathbf{x})$. However, in the next step, one does not simply evaluate $\nabla_\phi \, \ell(\phi, z_\mathrm{s})$ as suggested in part (a). Instead, one calculates a different gradient estimate,

$$\hat{g}(z_\mathrm{s}) := \hat{g}^{(1)}(z_\mathrm{s}) + \hat{g}^{(2)}(z_\mathrm{s}) \tag{8}$$

where

$$\begin{aligned} \hat{g}^{(1)}(z_\mathrm{s}) &:= \big(\nabla_\phi \log Q_\phi(Z = z_\mathrm{s} \,|\, \mathbf{X}=\mathbf{x})\big) \, \ell(\phi, z_\mathrm{s}); \\ \hat{g}^{(2)}(z_\mathrm{s}) &:= \nabla_\phi \ell(\phi, z_\mathrm{s}) = -\nabla_\phi \log Q_\phi(Z = z_\mathrm{s} \,|\, \mathbf{X}=\mathbf{x}). \end{aligned} \tag{9}$$

Show that $\hat{g}(z_\mathrm{s})$ is an unbiased gradient estimate of the ELBO, i.e., that

$$\mathbb{E}_{z \sim Q_\phi(Z|\mathbf{X}=\mathbf{x})}\big[\hat{g}(z)\big] = \nabla_\phi \, \mathrm{ELBO}(\phi) \tag{10}$$

Thus, $\hat{g}(z_\mathrm{s})$ can be used to optimize the ELBO with stochastic gradient descent.

Hint: write out the expectation $\mathbb{E}_{Q_\phi}[\,\cdot\,]$ in the definition of the ELBO (Eq. 1) as a weighted average over all possible values z, pull the gradient operation ∇_ϕ into the sum (or integral), and apply the product rule of differential calculus. Then compare the result to the left-hand side of Eq. 10.

(d) It turns out that the score-function gradients from Eqs. 8-9 can be simplified: we don't actually need $\hat{g}^{(2)}(z_\mathrm{s})$. Show that

$$\mathbb{E}_{z \sim Q_\phi(Z|\mathbf{X}=\mathbf{x})}\big[\hat{g}^{(2)}(z)\big] = 0. \tag{11}$$

Hint: Write out the expectation in Eq. 11 again as a weighted average, apply the chain rule of differentiation and then pull the gradient operation out of the sum (or integral) and use the fact that the probability mass function (or probability density function) $Q_\phi(Z \,|\, \mathbf{X} = \mathbf{x})$ is normalized over Z.

References

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Jang, E., Gu, S., and Poole, B. (2016). Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. In *International Conference on Learning Representations*, pages 1–9.

Maddison, C. J., Mnih, A., and Teh, Y. W. (2016). The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*.

Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR.