Problem Set 8

published: 18 June 2021
discussion: 21 June 2021

Data Compression With Deep Probabilistic Models
Prof. Robert Bamler, University of Tuebingen

Course material available at https://robamler.github.io/teaching/compress21/

Problem 8.1: Understanding the ELBO

In the lecture, we introduced the evidence lower bound, or ELBO,

$$\mathrm{ELBO}(\theta, \phi) = \mathbb{E}_{\mathbf{z} \sim Q_\phi(\mathbf{Z}|\mathbf{X}=\mathbf{x})}\big[\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})\big] \tag{1}$$

where $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{z})\, p_\theta(\mathbf{x}|\mathbf{z})$ is the joint probability *density* of the generative model P_θ, which has model parameters θ, and $q_\phi(\mathbf{z}|\mathbf{x})$ is the probability density of the variational distribution (or "approximate posterior") Q_ϕ, which has amortized variational parameters ϕ.

We arrived at the ELBO by considering the negative expected net bit rate of a modified variant of bits-back coding that uses Q_ϕ as a stand-in for the true posterior distribution. Based on the fact that the regular bits-back coding algorithm is optimal, we argued that our modification to the algorithm cannot reduce the net bit rate. This lead us to the conclusion that the ELBO is indeed a lower bound on the evidence, i.e.,

$$\mathrm{ELBO}(\theta, \phi) \le \log p_\theta(\mathbf{x}) \qquad \forall \theta, \phi. \tag{2}$$

In this problem, you will derive important equivalent formulations of the ELBO that will allow you to interpret what happens when we maximize the ELBO over θ and ϕ. In doing so, you will also prove the important Eq. 2 in a more direct way.

(a) Show by simple regrouping of the terms in Eq. 1 that

$$\mathrm{ELBO}(\theta, \phi) = \mathbb{E}_{\mathbf{z} \sim Q_\phi(\mathbf{Z}|\mathbf{X}=\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] - D_{\mathrm{KL}}\big[Q_\phi(\mathbf{Z}|\mathbf{X}=\mathbf{x}) \,\big|\big|\, P_\theta(\mathbf{Z})\big]. \tag{3}$$

What would the encoder and decoder networks learn if, instead of maximizing the ELBO, one would optimize only the first or only the second term on the right-hand side of Eq. 3, respectively?

(b) Show again by simple regrouping of the terms in Eq. 1 that

$$\mathrm{ELBO}(\theta, \phi) = \log p_\theta(\mathbf{x}) - D_{\mathrm{KL}}\big[Q_\phi(\mathbf{Z}|\mathbf{X}=\mathbf{x}) \,\big|\big|\, P_\theta(\mathbf{Z}|\mathbf{X}=\mathbf{x})\big]. \tag{4}$$

Notice that this confirms Eq. 2 since we know that the Kullback-Leibler divergence D_{KL} is always nonnegative.

What is the name of the first term on the right-hand side of Eq. 4, and why would we want to maximize it?

In many applications of variational inference, the generative model $P(\mathbf{X}, \mathbf{Z})$ is fixed (i.e., there are no free model parameters θ). In these applications, maximizing the ELBO just amounts to minimizing the KL-term on the right-hand side of Eq. 4 over the variational parameters ϕ. What does this minimization achieve, and why is the method called "variational *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 via the so-called coordinate ascent variational inference (CAVI) algorithm [Blei et al., 2017]. Roughly speaking, this algorithm can be derived by solving the equation $\nabla_{\phi_i}\text{ELBO}(\theta, \phi) = 0$ for one coordinate ϕ_i at a time, by writing out the expectation $\mathbb{E}$ on the right-hand side of Eq. 1 as an explicit integral over $\mathbf{z}$, taking the derivative, 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 admit an analytic solution 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 generalizes the method to arbitrary model architectures. In this problem, you will prove the validity of two different approaches to BBVI.

(a) Let's first understand the problem: the ELBO in Eq. 1 is an expectation of a known quantity $\log p_\theta(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x})$ under a known distribution Q_ϕ. This seems very similar to the typical situation in supervised learning, where we usually have to minimize some loss function that is the expectation of some known expression over samples from the training set.

So why couldn't we just use the same techniques that we know from supervised learning and maximize the ELBO with regular stochastic gradient descent[1]? In other words, why can't we just do the following:

- draw some sample $z_{\text{sample}} \sim Q_\phi(\mathbf{Z}|\mathbf{X} = \mathbf{x})$;
- evaluate the gradients of $\log p_\theta(\mathbf{x}, z_{\text{sample}}) - \log q_\phi(z_{\text{sample}}|\mathbf{x})$ with respect to θ and ϕ;
- use these gradients as an estimate of the gradient of $\text{ELBO}(\theta, \phi)$, and update θ and ϕ with a small gradient step.

Hint: Focus on the optimization over ϕ and look at all places where ϕ appears in the ELBO.

In the following parts, we will consider two possible solutions to the problem from part (a). We will limit the discussion to the differentiation with respect to ϕ, since differentiation with respect to θ does not pose a problem.

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

(b) The simplest form of BBVI uses so-called reparameterization gradients [Kingma and Welling, 2014]. Assume, for example, that the variational distribution is a normal distribution,

$$q_\phi(\mathbf{z}|\mathbf{x}) = \prod_{i=1}^{K} \mathcal{N}\big(z_i; \mu_i(\mathbf{x}, \phi), \sigma_i(\mathbf{x}, \phi)^2\big) \tag{5}$$

where the means $\mu_i(\mathbf{x}, \phi)$ and standard deviations $\sigma_i(\mathbf{x}, \phi)$ together comprise the output $g_\phi(\mathbf{x})$ of the encoder network.

Convince yourself that, for such a variational distribution, the expectation of any function $t(\mathbf{z})$ can be written as follows,

$$\mathbb{E}_{\mathbf{z} \sim Q_\phi(\mathbf{Z}|\mathbf{X}=\mathbf{x})}\big[t(\mathbf{z})\big] = \mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(0,I)}\Big[t\big(\boldsymbol{\mu}(\mathbf{x}, \phi) + \boldsymbol{\sigma}(\mathbf{x}, \phi) \odot \boldsymbol{\epsilon}\big)\Big]. \tag{6}$$

Here, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$ and $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_k)$ are the concatenations into vectors of the means and standard deviations, respectively, $\mathcal{N}(0, I)$ is a k-dimensional standard Normal distribution (i.e., with zero mean and unit variance), and $\odot$ denotes elementwise multiplication.

Now use Eq. 6 to fix the problem from part (a), i.e., to come up with an *unbiased* estimate of $\nabla_\phi \operatorname{ELBO}(\theta, \phi)$.

(c) While the approach 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 not on variational distributions over discrete $\mathbf{z}$. For such variational distributions, an alternative and more general approach called score function gradient estimates (or the "REINFORCE method") can be used [Ranganath et al., 2014].

Similar to the approach in part (a), one first draws some sample $z_{\text{sample}} \sim Q_\phi(\mathbf{Z}|\mathbf{X} = \mathbf{x})$. However, in the next step, one does *not* simply evaluate $\nabla_\phi\big(\log p_\theta(\mathbf{x}, z_{\text{sample}}) - \log q_\phi(z_{\text{sample}}|\mathbf{x})\big)$. Instead, one evaluates

$$\hat{g} := \hat{g}^{(1)} + \hat{g}^{(2)} \tag{7}$$

where

$$\begin{aligned} \hat{g}^{(1)} &:= \Big(\nabla_\phi \log q_\phi(z_{\text{sample}}|\mathbf{x})\Big)\Big(\log p_\theta(\mathbf{x}, z_{\text{sample}}) - \log q_\phi(z_{\text{sample}}|\mathbf{x})\Big); \\ \hat{g}^{(2)} &:= -\nabla_\phi \log q_\phi(z_{\text{sample}}|\mathbf{x}). \end{aligned} \tag{8}$$

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

$$\mathbb{E}_{z_{\text{sample}} \sim Q_\phi(\mathbf{Z}|\mathbf{X}=\mathbf{x})}\big[\hat{g}\big] = \nabla_\phi \operatorname{ELBO}(\theta, \phi). \tag{9}$$

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

Hint: write out the expectation $\mathbb{E}$ in the definition of the ELBO (Eq. 1) as an integral, pull the gradient operation ∇_ϕ into the integral, and apply the product rule of differential calculus. Then compare the result to the left-hand side of Eq. 9.

(d) It turns out that the score-function gradients from Eqs. 7-8 can be simplified: we don't even need $\hat{g}^{(2)}$. Show that

$$\mathbb{E}_{z_{\text{sample}} \sim Q_\phi(\mathbf{Z}|\mathbf{X}=\mathbf{x})}\left[\hat{g}^{(2)}\right] = 0. \tag{10}$$

Hint: Write out the expectation again as an integral, apply the chain rule of differentiation and then pull the gradient operation out of the integral and use the fact that the density q_ϕ is normalized.

Don't forget to provide anonymous feedback to this problem set in the corresponding poll on moodle.

References

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

[Kingma and Welling, 2014] Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. In *International Conference on Learning Representations*.

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