Problem Set 8

published: 14 June 2023
discussion: 21 June 2023

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

Course materials available at https://robamler.github.io/teaching/compress23/

Problem 8.1: Understanding the ELBO

In the lecture, we introduced the evidence lower bound (ELBO),

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

Here, $P(\mathbf{Z}, \mathbf{X}) = P(\mathbf{Z})P(\mathbf{X} \,|\, \mathbf{Z})$ models a *generative process* with latent variables $\mathbf{Z}$ and observed variables (i.e., the message) $\mathbf{X}$. Further, $Q_\phi(\mathbf{Z})$ is the variational distribution, which has variational parameters ϕ.

This problem will give you some intuition for the ELBO. Let's assume for simplicity that both $\mathbf{Z}$ and $\mathbf{X}$ are discrete. We showed in the lecture that the ELBO is then the negative expected net bit rate of bits-back coding with the approximate posterior $Q_\phi(\mathbf{Z})$,

$$\mathrm{ELBO}(\phi, \mathbf{x}) = -\mathbb{E}_{\mathbf{s}}\big[R^{\mathrm{net}}_\phi(\mathbf{x} \,|\, \mathbf{s})\big] \tag{2}$$

where $\mathbf{s}$ is a random bit string ("side information") from which we decode with the model $Q_\phi(\mathbf{Z})$ in bits-back coding. Eq. 2 motivated us to maximize the ELBO over the variational parameters ϕ (as this is equivalent to minimizing the expected net bit rate):

$$\phi^* := \arg\max_\phi \mathrm{ELBO}(\phi, \mathbf{x}). \tag{3}$$

You'll now show three different ways in which maximizing the ELBO is usually motivated in the (non-compression) literature.

(a) The term $\mathbb{E}_{Q_\phi(\mathbf{Z})}[-\log Q_\phi(\mathbf{Z})]$ on the right-hand side of Eq. 1 is the entropy $H_{Q_\phi}[\mathbf{Z}]$ of $\mathbf{Z}$ under the variational distribution. Thus, we can express the ELBO as follows,

$$\mathrm{ELBO}(\phi, \mathbf{x}) = \mathbb{E}_{Q_\phi(\mathbf{Z})}\big[\log P(\mathbf{Z}, \mathbf{X}{=}\mathbf{x})\big] + H_{Q_\phi}\big[\mathbf{Z}\big]. \tag{4}$$

(i) Imagine the entropy term $H_{Q_\phi}[\mathbf{Z}]$ was absent, i.e., pretend that we maximize only the first term on the right-hand side of Eq. 4. Argue (in words) that setting $\phi^* = \arg\max_\phi \mathbb{E}_{Q_\phi(\mathbf{Z})}\big[\log P(\mathbf{Z}, \mathbf{X} = \mathbf{x})\big]$ would make the resulting distribution $Q_{\phi^*}(\mathbf{Z})$ deterministic, i.e., there would be some $\mathbf{z}^*$ such that $Q_{\phi^*}(\mathbf{Z}{=}\mathbf{z}^*) = 1$ and $Q_{\phi^*}(\mathbf{Z}{\neq}\mathbf{z}^*) = 0$ (assuming that this distribution is part of the variational family). What is the value of $\mathbf{z}^*$?

(ii) Now let's return to the full expression in Eq. 4 that includes the entropy term $H_{Q_\phi}[\mathbf{Z}]$. 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,

$$\mathrm{ELBO}(\phi, \mathbf{x}) = \mathbb{E}_{Q_\phi(\mathbf{Z})}\big[\log P(\mathbf{X}{=}\mathbf{x} \,|\, \mathbf{Z})\big] - D_{\mathrm{KL}}\big(Q_\phi(\mathbf{Z}) \,\big\|\, P(\mathbf{Z})\big). \tag{5}$$

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

Eq. 5 tells us that maximizing the ELBO over ϕ can be interpreted as *regularized maximum likelihood estimation*. To see this, answer the following two questions (no calculation required): what distribution $Q_{\phi^*}(\mathbf{Z})$ would we get if we maximized

(i) only the first term on the right-hand side of Eq. 5; or

(ii) only the second term on the right-hand side of Eq. 5?

In reality, we maximize over the sum of both terms, and so $Q_{\phi^*}(\mathbf{Z})$ interpolates between (i) and (ii). Which of the two terms in Eq. 5 can be seen as a regularizer?

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

$$\mathrm{ELBO}(\phi, \mathbf{x}) = \log P(\mathbf{X}{=}\mathbf{x}) - D_{\mathrm{KL}}\big(Q_\phi(\mathbf{Z}) \,\big\|\, P(\mathbf{Z} \,|\, \mathbf{X}{=}\mathbf{x})\big). \tag{6}$$

(*Hint*: it's again easier to start with Eq. 6 and derive Eq. 1 from it rather than trying it the other way round.)

(i) Assume, for now, that the variational family $\mathcal{Q} = \{Q_\phi\}_\phi$ contains *all* probability distributions over $\mathbf{Z}$, and that the generative model P is fixed (we'll discuss how to learn the generative model next week). What would be the optimal variational distribution $Q_{\phi^*}(\mathbf{Z})$ that maximizes the right-hand side of Eq. 6? Maximizing the ELBO is called "variational inference" because it is related to Bayesian inference. Can you explain what the relation is?

(ii) In practice, the variational family $\mathcal{Q}$ is only a subset of all probability distributions over $\mathbf{Z}$. Since we maximize the ELBO only over variational distributions from $\mathcal{Q}$, the resulting optimal variational distribution $Q_{\phi^*}(\mathbf{Z})$ will typically be somewhat different from what you found in subpart (i) above. This mismatch will lead to an overhead in the expected net bit rate when we use $Q_{\phi^*}(\mathbf{Z})$ for bits-back coding (see Eq. 2). Which term in Eq. 6 expresses this overhead?

Black-Box Variational Inference (BBVI)

Problems 8.2 and 8.3 below discuss two methods for maximizing the ELBO (Eq. 1) numerically. 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)). While this algorithm is extremely fast (and should therefore be preferred whenever possible!), its application is limited to generative models and variational families where relevant parts of the expectation $\mathrm{ELBO}(\phi, \mathbf{x}) = \mathbb{E}_{Q_\phi(\mathbf{Z})}\big[\log P(\mathbf{Z}, \mathbf{X}{=}\mathbf{x}) - \log Q_\phi(\mathbf{Z})\big]$ can be evaluated analytically, and where one can then analytically solve equations of the form $\nabla_{\phi_i} \mathrm{ELBO}(\phi, \mathbf{x}) = 0$. This would essentially forbid the use of neural networks.

Mainstream adoption of variational inference only occurred after the invention of so-called *black box variational inference* (BBVI), which replaces analytic calculations of integrals over $\mathbf{z}$ by numerical estimates based on samples $\mathbf{z} \sim Q_\phi(\mathbf{Z})$, and analytic solutions of the equation $\nabla_{\phi_i} \mathrm{ELBO}(\phi, \mathbf{x}) = 0$ by stochastic gradient descent (SGD). As discussed in the lecture, SGD requires an *unbiased gradient estimate* $\hat{g}(\mathbf{z})$ that we can calculate from one (or more) samples $\mathbf{z} \sim Q_\phi(\mathbf{Z})$ and that satisfies

$$\mathbb{E}_{\mathbf{z}\sim Q_\phi(\mathbf{Z})}\big[\hat{g}(\mathbf{z})\big] = \nabla_\phi \mathrm{ELBO}(\phi, \mathbf{x}) = \nabla_\phi\Big(\mathbb{E}_{\mathbf{z}\sim Q_\phi(\mathbf{Z})}\big[\ell(\phi, \mathbf{x}, \mathbf{z})\big]\Big) \tag{7}$$

where $\ell(\phi, \mathbf{x}, \mathbf{z}) = \log P(\mathbf{Z}, \mathbf{X}{=}\mathbf{x}) - \log Q_\phi(\mathbf{Z})$ (see Eq. 1). We saw in the lecture that obtaining an unbiased gradient estimate for the ELBO is nontrivial since the distribution $Q_\phi(\mathbf{Z})$ from which we draw $\mathbf{z}$ in Eq. 7 itself depends on ϕ, and so the gradient estimate has to take into account that changing ϕ also changes which samples $\mathbf{z}$ we obtain. In Problems 8.2 and 8.3 below, you'll derive two solutions to this issue.

Problem 8.2: BBVI With Reparameterization Gradients

Assume, for example, that $\mathbf{z} \in \mathbb{R}^d$ lives in some *continuous* space of dimension d, and that the variational family is the set of all fully factorized normal distributions, i.e., the variational distribution $Q_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\mathbf{Z})$ has a probability density function

$$q_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\mathbf{z}) = \prod_{i=1}^{d} \mathcal{N}(\mathbf{z}_i; \mu_i, \sigma_i^2). \tag{8}$$

Here, the means $\boldsymbol{\mu} \equiv (\mu_i)_{i=1}^d$ and standard deviations $\boldsymbol{\sigma} \equiv (\sigma_i)_{i=1}^d$ together comprise the variational parameters ϕ over which we maximize the ELBO, and $\mathcal{N}$ denotes the so-called normal distribution, which has the density function

$$\mathcal{N}(\mathbf{z}_i; \mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(\mathbf{z}_i - \mu_i)^2}{2\sigma_i^2} \right]. \tag{9}$$

(a) Convince yourself that, for such a variational distribution, the ELBO can be expressed as follows,

$$\text{ELBO}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{x}) = \mathbb{E}_{\mathbf{z}\sim Q_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\mathbf{Z})}\left[\ell(\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{x}, \mathbf{z})\right] = \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)}\left[\ell(\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{x}, \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon})\right] \tag{10}$$

where $\mathcal{N}(0, I)$ is the d-dimensional standard normal distribution (i.e., with mean 0 and standard deviation 1 in each direction), and $\odot$ is elementwise multiplication.

(b) Using Eq. 10, come up with a gradient estimate $\hat{g}(\boldsymbol{\epsilon})$ that satisfies $\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)}\left[\hat{g}(\boldsymbol{\epsilon})\right] = \nabla_{\boldsymbol{\mu},\boldsymbol{\sigma}} \text{ELBO}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{x})$ (no calculation required; your result should fit into half a line).

(c) The code below is taken verbatim from (Duvenaud and Adams, 2015). Convince yourself that it implements the right-hand side of Eq. 10. Identify each function argument and local variable with a mathematical symbol on this problem set.

```
def lower_bound(variational_params, logprob_func, D, num_samples):
    # variational_params: the mean and covariance of approximate posterior.
    # logprob_func:       the unnormalized log-probability of the model.
    # D:                  the number of parameters in the model.
    # num_samples:        the number of Monte Carlo samples to use.

    # Unpack mean and covariance of diagonal Gaussian.
    mu, cov = variational_params[:D], np.exp(variational_params[D:])

    # Sample from multivariate normal using the reparameterization trick.
    samples = npr.randn(num_samples, D) * np.sqrt(cov) + mu

    # Lower bound is the exact entropy plus a Monte Carlo estimate of energy.
    return mvn.entropy(mu, np.diag(cov)) + np.mean(logprob(samples))

# Get gradient with respect to variational params using autograd.
gradient_func = grad(lower_bound)
```

Note: the function `npr.randn` draws samples from the standard normal $\mathcal{N}(0, I)$, and `mvn.entropy` calculates $H_{Q_\phi}[\mathbf{Z}]$ analytically to reduce gradient variance. There seems to be a typo in the `return` statement: `logprob` should be `logprob_func`.

Problem 8.3: BBVI With Score Function Gradients

While the reparameterization gradients from Problem 8.2 can be generalized beyond a normal distribution, they don't exist for all variational distributions, in particular not for discrete $\mathbf{z}$ (unless an approximation is used (Jang et al., 2016; Maddison et al., 2016)). For such variational distributions, one can use the more general score function gradient estimates (aka the "REINFORCE method"; (Ranganath et al., 2014)),

$$\hat{g}(\mathbf{z}) := \hat{g}^{(1)}(\mathbf{z}) + \hat{g}^{(2)}(\mathbf{z}) \tag{11}$$

where

$$\begin{aligned} \hat{g}^{(1)}(\mathbf{z}) &:= \big(\nabla_\phi \log Q_\phi(\mathbf{Z}{=}\mathbf{z})\big)\, \ell(\phi, \mathbf{x}, \mathbf{z}); \\ \hat{g}^{(2)}(\mathbf{z}) &:= \nabla_\phi \ell(\phi, \mathbf{x}, \mathbf{z}) = -\nabla_\phi \log Q_\phi(\mathbf{Z}{=}\mathbf{z}). \end{aligned} \tag{12}$$

(a) Show that $\hat{g}(\mathbf{z})$ is an unbiased gradient estimate of the ELBO, i.e., it satisfies Eq. 7.

Hint: write out the expectations on both sides of Eq. 7 as a weighted average: $\mathbb{E}_{\mathbf{z}\sim Q_\phi(\mathbf{Z})}[\cdots] = \sum_{\mathbf{z}} Q_\phi(\mathbf{Z}{=}\mathbf{z})\,[\cdots]$. On the left-hand side, use $\nabla_\xi \log f(\xi) = \frac{1}{\xi} \nabla_\xi f(\xi)$ for $\hat{g}^{(1)}(\mathbf{z})$; on the right-hand side, pull "∇_ϕ" into the sum and use the product rule of differential calculus. Then compare both sides term by term.

(b) It turns out that Eq. 11 can be simplified: we don't need $\hat{g}^{(2)}(\mathbf{z})$. Show that

$$\mathbb{E}_{\mathbf{z}\sim Q_\phi(\mathbf{Z})}\big[\hat{g}^{(2)}(\mathbf{z})\big] = 0. \tag{13}$$

Hint: Insert $\hat{g}^{(2)}(\mathbf{z}) = -\nabla_\phi \log Q_\phi(\mathbf{Z}{=}\mathbf{z})$ into Eq. 13 and write out the expectation again as a weighted average. Then use again the derivative of the logarithm, pull "∇_ϕ" out of the sum, and use the fact that $Q_\phi(\mathbf{Z})$ is normalized.

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.

Duvenaud, D. and Adams, R. P. (2015). Black-box stochastic variational inference in five lines of python. In *NIPS Workshop on Black-box Learning and Inference*.

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

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.