Solutions to Problem Set 8

discussed:
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 $\mathbf{z}^*$?

Solution: Maximizing $\mathbb{E}_{Q_\phi(\mathbf{Z})}[\log P(\mathbf{Z}, \mathbf{X}{=}\mathbf{x})]$ over ϕ amounts to searching for the distribution $Q_\phi \in \mathcal{Q}$ such that, if we draw samples $\mathbf{z} \sim Q_\phi(\mathbf{Z})$, these

samples have high $\log P(\mathbf{Z}{=}\mathbf{z}, \mathbf{X}{=}\mathbf{x})$ in expectation. Clearly, we'll get the highest expectation if all samples $\mathbf{z} \sim Q_\phi(\mathbf{Z})$ maximize $\log P(\mathbf{Z}{=}\mathbf{z}, \mathbf{X}{=}\mathbf{x})$. Thus, $\mathbb{E}_{Q_\phi(\mathbf{Z})}[\log P(\mathbf{Z}, \mathbf{X}{=}\mathbf{x})]$ is maximized by a deterministic distribution $Q_{\phi^*}(\mathbf{Z})$ as described in the problem, which puts all probability mass on $\mathbf{z}^* = \arg\max_{\mathbf{Z}} \log P(\mathbf{Z}, \mathbf{X}{=}\mathbf{x}) = \arg\max_{\mathbf{Z}} P(\mathbf{Z}, \mathbf{X}{=}\mathbf{x})$. The value $\mathbf{z}^*$ is called the "maximum a-posteriori" (MAP) since it also maximizes the posterior $P(\mathbf{Z} \,|\, \mathbf{X}{=}\mathbf{x}) = \frac{P(\mathbf{Z}, \mathbf{X}=\mathbf{x})}{P(\mathbf{X}=\mathbf{x})}$ (since the denominator $P(\mathbf{X}{=}\mathbf{x})$) is a constant. ■

(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?).

Solution: Including the entropy term $H_{Q_\phi}[\mathbf{Z}]$ in the objective function penalizes distributions $Q_\phi(\mathbf{Z})$ with low entropy. For a deterministic distribution $Q_{\phi^*}(\mathbf{Z})$ as described in the problem, we find $H_{Q_{\phi^*}}[\mathbf{Z}] = 0$, which is the smallest possible entropy of a discrete random variable. Thus, such a deterministic distribution is particularly penalized. Variational inference favors variational distributions that capture some estimate of our uncertainty of $\mathbf{Z}$ (this will become more evident in parts (b) and (c) below).

Technical remark: the fact that $H_{Q_{\phi^*}}[\mathbf{Z}] := \mathbb{E}_{Q_{\phi^*}(\mathbf{Z})}[-\log Q_{\phi^*}(\mathbf{Z})]$ is zero for a deterministic distribution can be understood most easily if one interprets the expectation $\mathbb{E}_{Q_{\phi^*}(\mathbf{Z})}[\cdots]$ as an average over many samples $\mathbf{z} \sim Q_{\phi^*}(\mathbf{Z})$. For a deterministic distribution, all these samples equal $\mathbf{z}^*$, which satisfies $Q_{\phi^*}(\mathbf{Z}{=}\mathbf{z}^*) = 1$ and thus $-\log Q_{\phi^*}(\mathbf{Z}{=}\mathbf{z}^*) = 0$. If one instead follows the measure-theoretical definition of the expectation from the lecture, i.e., $H_{Q_{\phi^*}}[\mathbf{Z}] = -\sum_{\mathbf{z}} Q_{\phi^*}(\mathbf{Z}{=}\mathbf{z}) \log Q_{\phi^*}(\mathbf{Z}{=}\mathbf{z})$, then most of the terms in the sum are expressions of the form "$0 \times \infty$". In measure theory, such expressions are considered to evaluate to zero. This is well motivated if we consider a distribution that is *almost* deterministic, but that still assigns a small probability $\epsilon > 0$ to all $\mathbf{Z} \neq \mathbf{z}^*$. Taking the limit $\epsilon \to 0$, we indeed find, using L'Hôpital's rule ($\star$): $\lim_{\epsilon \to 0} (\epsilon \log \epsilon) = \lim_{\epsilon \to 0} \frac{\log \epsilon}{1/\epsilon} \overset{(\star)}{=} \lim_{\epsilon \to 0} \frac{1/\epsilon}{-1/\epsilon^2} = -\lim_{\epsilon \to 0} \epsilon = 0$. ■

(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.)

Solution: Writing out the KL-divergence (see end of Lecture 3), we find

$$\begin{aligned}
&\mathbb{E}_{Q_\phi(\mathbf{Z})}\big[\log P(\mathbf{X}{=}\mathbf{x} \,|\, \mathbf{Z})\big] - D_{\mathrm{KL}}\big(Q_\phi(\mathbf{Z}) \,\big\|\, P(\mathbf{Z})\big) \\
&= \mathbb{E}_{Q_\phi(\mathbf{Z})}\big[\log P(\mathbf{X}{=}\mathbf{x} \,|\, \mathbf{Z}) + \log P(\mathbf{Z}) - \log Q_\phi(\mathbf{Z})\big] \\
&= \mathbb{E}_{Q_\phi(\mathbf{Z})}\big[\log P(\mathbf{Z}, \mathbf{X}{=}\mathbf{x}) - \log Q_\phi(\mathbf{Z})\big] \overset{\text{(Eq. 1)}}{=} \mathrm{ELBO}(\phi, \mathbf{x}).
\end{aligned}$$

■

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

Solution: Following an argument analogous to part (a) (i), maximizing $\mathbb{E}_{Q_\phi(\mathbf{Z})}[\log P(\mathbf{X} = \mathbf{x} \,|\, \mathbf{Z})]$ over ϕ would again lead to a deterministic variational distribution that puts all probability mass on a single value $\mathbf{z}^*$. This time, however, $\mathbf{z}^*$ would not be the MAP estimate but instead the maximum likelihood estimate (MLE), $\mathbf{z}^* = \arg\max_{\mathbf{Z}} P(\mathbf{X}{=}\mathbf{x} \,|\, \mathbf{Z})$. ■

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

Solution: Since the KL-divergence is always nonnegative (Problem 3.1 (a)), the maximum value of the negative KL-divergence (i.e., the minimal value of the KL-divergence) is achieved if $D_{\mathrm{KL}}\big(Q_\phi(\mathbf{Z}) \,\big\|\, P(\mathbf{Z})\big) = 0$, which is satisfied exactly if the variational distribution $Q_\phi(\mathbf{Z})$ matches the prior $P(\mathbf{Z})$ almost everywhere (see again Problem 3.1 (a)). ■

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?

Solution: The KL-term in Eq. 5 can be seen as a regularizer as it is data-independent and encourages $Q_\phi(\mathbf{Z})$ to stay close to the prior $P(\mathbf{Z})$, which is typically a relatively broad distribution with a relatively high entropy, i.e., it doesn't overfit to any particular data. Without this KL-term, optimization would find the MLE, i.e., the value of the latent variable that best explains the observed data, with no regard to any prior knowledge. In a machine learning setup, the MLE would therefore be prone to overfitting, i.e., matching accidental patterns in the data and failing to generalize to new data. For a well-chosen prior, the KL-term counteracts the tendency to overfit, which is referred to as regularizing. ■

(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.)

Solution: Writing out again KL-divergence and pulling the evidence $\log P(\mathbf{X}{=}\mathbf{x})$ (which does not depend on $\mathbf{z}$) into the expectation, we find

$$\begin{aligned}
&\log P(\mathbf{X}{=}\mathbf{x}) - D_{\mathrm{KL}}\big(Q_\phi(\mathbf{Z}) \,\big\|\, P(\mathbf{Z} \,|\, \mathbf{X}{=}\mathbf{x})\big) \\
&= \mathbb{E}_{Q_\phi(\mathbf{Z})}\big[\log P(\mathbf{X}{=}\mathbf{x}) + \log P(\mathbf{Z} \,|\, \mathbf{X}{=}\mathbf{x}) - \log Q_\phi(\mathbf{Z})\big] \\
&= \mathbb{E}_{Q_\phi(\mathbf{Z})}\big[\log P(\mathbf{Z}, \mathbf{X}{=}\mathbf{x}) - \log Q_\phi(\mathbf{Z})\big] \overset{\text{(Eq. 1)}}{=} \mathrm{ELBO}(\phi, \mathbf{x})
\end{aligned}$$

■

(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?

Solution: For a fixed generative model P, the perfect variational distribution would be the true posterior $P(\mathbf{Z} \,|\, \mathbf{X}{=}\mathbf{x})$, if this distribution was contained in the variational family (which is typically not the case because the true posterior is typically too complicated). Setting $Q_\phi(\mathbf{Z}) = P(\mathbf{Z} \,|\, \mathbf{X}{=}\mathbf{x})$ would set the KL-divergence on the right-hand side of Eq. 6 to zero. This would maximize the ELBO since the remaining term $\log P(\mathbf{X}{=}\mathbf{x})$ (called "evidence") is a constant for a constant generative model. ■

(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?

Solution: Bits-back coding with the true posterior has a net bit rate that is independent of $\mathbf{z}$ and given by $-\log P(\mathbf{X}{=}\mathbf{x})$. If we replace the true posterior with a variational distribution $Q_\phi(\mathbf{Z})$, then the net bit rate depends on $\mathbf{z}$. In expectation, the net bit rate is the negative ELBO (see Eq. 2), i.e., according to Eq. 6, the expected net bit rate is then $-\log P(\mathbf{X}{=}\mathbf{x}) + D_{\mathrm{KL}}(Q_\phi(\mathbf{Z}) \,\|\, P(\mathbf{Z} \,|\, \mathbf{X}{=}\mathbf{x}))$. Thus, replacing the true posterior with a variational distribution leads to an overhead of $D_{\mathrm{KL}}(Q_\phi(\mathbf{Z}) \,\|\, P(\mathbf{Z} \,|\, \mathbf{X}{=}\mathbf{x}))$. ■

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 $\ell(\phi, \mathbf{x}, \mathbf{z})$ is the expression in the expectation in Eq. 1, i.e.,

$$\ell(\phi, \mathbf{x}, \mathbf{z}) = \log P(\mathbf{Z}, \mathbf{X}{=}\mathbf{x}) - \log Q_\phi(\mathbf{Z}) \tag{8}$$

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

$$q_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\mathbf{z}) = \prod_{i=1}^{d} \mathcal{N}(z_i; \mu_i, \sigma_i^2). \tag{9}$$

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

$$\mathcal{N}(z_i; \mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(z_i - \mu_i)^2}{2\sigma_i^2}\right]. \tag{10}$$

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

$$\mathrm{ELBO}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{x}) = \mathbb{E}_{\mathbf{z}\sim Q_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\mathbf{Z})}\big[\ell(\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{x}, \mathbf{z})\big] = \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)}\big[\ell(\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{x}, \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon})\big] \tag{11}$$

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.

Solution: The claim here is that, if we draw a sample $\epsilon_i \sim \mathcal{N}(0, 1)$ from a standard normal distribution and then scale ϵ_i by σ_i and shift it by μ_i, the result is distributed from a normal distribution $\mathcal{N}(\mu_i, \sigma_i^2)$ with mean μ_i and standard

deviation σ_i. In fact, this is how sampling from a normal distribution with user-defined mean and standard deviation is typically implemented in software libraries.

I didn't expect a formal derivation here. It's more important that you understand why Eq. 11 holds. But one way to formally derive the equivalence would go back to how we defined probability density functions (PDFs) in the lecture: the statement that $q_{\boldsymbol{\mu},\boldsymbol{\sigma}}$ defined in Eqs. 9-10 is a PDF of $Q_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\mathbf{Z})$ means that, for all measurable functions f, the expectation $\mathbb{E}_{Q_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\mathbf{Z})}[f(\mathbf{Z})]$ can be expressed as $\mathbb{E}_{Q_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\mathbf{Z})}[f(\mathbf{Z})] = \int q_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\mathbf{z}) f(\mathbf{z})\,\mathrm{d}\mathbf{z}$. We now integrate by substituting $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$, i.e., $z_i = \mu_i + \sigma_i \epsilon_i$ for all $i \in \{1, \ldots, d\}$,

$$\begin{aligned}
\mathbb{E}_{Q_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\mathbf{Z})}\big[f(\mathbf{Z})\big] &= \int q_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\mathbf{z}) f(\mathbf{z})\,\mathrm{d}\mathbf{z} \\
&= \int \left| \det\left(\frac{\partial \mathbf{z}}{\partial \boldsymbol{\epsilon}}\right) \right| q_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon})\, f(\boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon})\,\mathrm{d}\boldsymbol{\epsilon} \\
&= \mathbb{E}_{\boldsymbol{\epsilon} \sim Q_0}\big[f(\boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon})\big]
\end{aligned}$$

where, in the last equality, we interpreted the integral over $\boldsymbol{\epsilon}$ as an expectation under a distribution Q_0, which has the PDF

$$q_0(\boldsymbol{\epsilon}) = \left| \det\left(\frac{\partial \mathbf{z}}{\partial \boldsymbol{\epsilon}}\right) \right| q_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}).$$

We find for the Jacobi determinant

$$\det\left(\frac{\partial \mathbf{z}}{\partial \boldsymbol{\epsilon}}\right) = \det\big(\mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_d)\big) = \prod_{i=1}^{d} \sigma_i$$

and thus,

$$\begin{aligned}
q_0(\boldsymbol{\epsilon}) &= \left(\prod_{i=1}^{d} \sigma_i\right) q_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}) \\
&\overset{\text{(Eqs. 9-10)}}{=} \prod_{i=1}^{d} \frac{\sigma_i}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{\big((\mu_i + \sigma_i \epsilon_i) - \mu_i\big)^2}{2\sigma_i^2} \right] \\
&= \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{\epsilon_i^2}{2} \right] = \prod_{i=1}^{d} \mathcal{N}(\epsilon_i; 0, 1) = \mathcal{N}(\boldsymbol{\epsilon}; 0, I)
\end{aligned}$$

Thus, we've shown that $\mathbb{E}_{Q_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\mathbf{Z})}[f(\mathbf{Z})] = \mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(0,I)}[f(\boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon})]$ for all measurable functions f. Eq. 11 applies this relation to the function $f(\mathbf{z}) := \ell(\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{x}, \mathbf{z})$. ■

(b) Using Eq. 11, come up with a gradient estimate $\hat{g}(\boldsymbol{\epsilon})$ that satisfies $\mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(0,I)}\big[\hat{g}(\boldsymbol{\epsilon})\big] = \nabla_{\boldsymbol{\mu},\boldsymbol{\sigma}} \mathrm{ELBO}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{x})$ (no calculation required; your result should fit into half a line).

Solution: Since the distribution $\mathcal{N}(0, I)$ under which we evaluate the expectation on the right-hand side of Eq. 11 does not depend on the variational parameters $\boldsymbol{\mu}$

and $\boldsymbol{\sigma}$, we can pull the gradient operator "$\nabla_{\boldsymbol{\mu},\boldsymbol{\sigma}}$" into the expectation and obtain:

$$\nabla_{\boldsymbol{\mu},\boldsymbol{\sigma}} \text{ELBO}(\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{x}) = \mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(0,I)}\big[\hat{g}(\boldsymbol{\epsilon})\big] \quad \text{where} \quad \hat{g}(\boldsymbol{\epsilon}) = \nabla_{\boldsymbol{\mu},\boldsymbol{\sigma}}\, \ell(\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{x}, \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}).$$

■

(c) The code below is taken verbatim from the paper "Black-Box Stochastic Variational Inference in Five Lines of Python" (Duvenaud and Adams, 2015). Convince yourself that it implements the right-hand side of Eq. 11. 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`.

Solution: The first line in the function body extracts the variational parameters `mu` (i.e., the means $\boldsymbol{\mu}$) and `cov` (i.e., the variances $(\sigma_i^2)_{i=1}^d$) from the function argument `variational_params` (which corresponds to ϕ). Note that the second half of `variational_params` contains the *logarithm* of the variances rather than the variances themselves. This is generally a good idea because (i) it ensures that the variances are always positive (as they are obtained by taking the exponential of a real valued function) and (ii) optimizing variances in log space is numerically easier since it effectively adjusts gradient steps relative to the size of each σ_i (thus allowing SGD to converge to accurate estimates of σ_i for dimensions i where it should be small while still being able to reach large values of σ_i for other dimensions).

The second line in the function body implements the expression $\boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$ from Eq. 11, where the function `npr.randn` draws $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$. The `return` statement evaluates the part of $\ell(\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{x}, \mathbf{z})$ that contains the generative model P, which is provided as a function argument `logprob_func` (misspelled as `logprob` in the `return` statement). For the part of $\ell(\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{x}, \mathbf{z})$ that contains the variational distribution $Q_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\mathbf{Z})$, the expectation $\mathbb{E}_{Q_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\mathbf{Z})}[-\log Q_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\mathbf{Z})] = H_{Q_{\boldsymbol{\mu},\boldsymbol{\sigma}}}[\mathbf{Z}]$ is calculated analytically. The last line of the code obtains $\hat{g}(\mathbf{Z})$ via automatic differentiation.

Warning: While the paper demonstrates how easy it is to implement reparameterization gradients, it unfortunately sets a bad example in the way how it calculates the entropy, which is both unnecessarily computationally expensive and potentially numerically unstable for high dimensional latent spaces. The expression `mvn.entropy(mu, np.diag(cov))` in the `return` statement explicitly constructs the $d \times d$ covariance matrix $\Sigma := \operatorname{diag}(\sigma_1^2, \dots, \sigma_d^2)$ and then calculates the entropy of $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$. The entropy of a multidimensional normal distribution with arbitrary covariance matrix Σ is $\frac{1}{2} \log\det\Sigma + \text{const}$, where calculating the determinant has run-time complexity $O(d^3)$, and its gradient is potentially numerically unstable. It would be much better to exploit the fact that $Q_{\boldsymbol{\mu},\boldsymbol{\sigma}}(\mathbf{Z}) = \prod_{i=1}^d \mathcal{N}(z_i, \mu_i, \sigma_i^2)$ factorizes over all i ("mean field approximation"), which implies that the entropy separates (see Problem 4.3 (d)): $H_{Q_{\boldsymbol{\mu},\boldsymbol{\sigma}}}[\mathbf{Z}] = \sum_{i=1}^d H_{\mathcal{N}(\mu_i,\sigma_i^2)}[Z_i] = \sum_{i=1}^d \log\sigma_i + \text{const}$. It can thus be calculated in $O(d)$ time and differentiated numerically safely. ■

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)),

$$\hat{g}(\mathbf{z}) := \hat{g}^{(1)}(\mathbf{z}) + \hat{g}^{(2)}(\mathbf{z}) \tag{12}$$

where

$$\begin{aligned} \hat{g}^{(1)}(\mathbf{z}) &:= \big(\nabla_\phi \log Q_\phi(\mathbf{Z}{=}\mathbf{z})\big)\, \ell(\phi, \mathbf{x}, \mathbf{z}) \\ \hat{g}^{(2)}(\mathbf{z}) &:= \nabla_\phi \ell(\phi, \mathbf{x}, \mathbf{z}) = -\nabla_\phi \log Q_\phi(\mathbf{Z}{=}\mathbf{z}) \end{aligned} \tag{13}$$

with $\ell(\phi, \mathbf{x}, \mathbf{z})$ defined in Eq. 8.

(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.

Solution: We'll do the derivation for discrete $\mathbf{Z}$. For continuous $\mathbf{Z}$, the proof is analogous, except that sums are replaced by integrals and probability mass functions are replaced by probability density functions.

For the left-hand side of Eq. 7, we obtain

$$\begin{aligned}
\mathbb{E}_{\mathbf{z}\sim Q_\phi(\mathbf{Z})}\big[\hat{g}(\mathbf{z})\big] &= \sum_{\mathbf{z}} Q_\phi(\mathbf{Z}{=}\mathbf{z})\hat{g}(\mathbf{z}) \\
&= \sum_{\mathbf{z}} Q_\phi(\mathbf{Z}{=}\mathbf{z})\Big(\big(\nabla_\phi \log Q_\phi(\mathbf{Z}{=}\mathbf{z})\big)\,\ell(\phi,\mathbf{x},\mathbf{z}) + \nabla_\phi \ell(\phi,\mathbf{x},\mathbf{z})\Big) \\
&= \sum_{\mathbf{z}} Q_\phi(\mathbf{Z}{=}\mathbf{z})\left(\frac{\nabla_\phi Q_\phi(\mathbf{Z}{=}\mathbf{z})}{Q_\phi(\mathbf{Z}{=}\mathbf{z})}\,\ell(\phi,\mathbf{x},\mathbf{z}) + \nabla_\phi \ell(\phi,\mathbf{x},\mathbf{z})\right) \\
&= \sum_{\mathbf{z}} \Big[\big(\nabla_\phi Q_\phi(\mathbf{Z}{=}\mathbf{z})\big)\ell(\phi,\mathbf{x},\mathbf{z}) + Q_\phi(\mathbf{Z}{=}\mathbf{z})\,\nabla_\phi \ell(\phi,\mathbf{x},\mathbf{z})\Big]
\end{aligned}$$

For the right-hand side of Eq. 7, we obtain (using the product rule),

$$\begin{aligned}
\nabla_\phi\Big(\mathbb{E}_{\mathbf{z}\sim Q_\phi(\mathbf{Z})}\big[\ell(\phi,\mathbf{x},\mathbf{z})\big]\Big) &= \nabla_\phi\Big(\sum_{\mathbf{z}} Q_\phi(\mathbf{Z}{=}\mathbf{z})\,\ell(\phi,\mathbf{x},\mathbf{z})\Big) \\
&= \sum_{\mathbf{z}} \nabla_\phi\big[Q_\phi(\mathbf{Z}{=}\mathbf{z})\,\ell(\phi,\mathbf{x},\mathbf{z})\big] \\
&= \sum_{\mathbf{z}} \Big[\big(\nabla_\phi Q_\phi(\mathbf{Z}{=}\mathbf{z})\big)\ell(\phi,\mathbf{x},\mathbf{z}) + Q_\phi(\mathbf{Z}{=}\mathbf{z})\,\nabla_\phi \ell(\phi,\mathbf{x},\mathbf{z})\Big].
\end{aligned}$$

Thus, left-hand side and right-hand side match, and $\hat{g}(\mathbf{z})$ is an unbiased gradient estimate. ■

(b) It turns out that Eq. 12 can be simplified: we don't need $\hat{g}^{(2)}(\mathbf{z})$. Show that

$$\mathbb{E}_{\mathbf{z}\sim Q_\phi(\mathbf{Z})}\big[\hat{g}^{(2)}(\mathbf{z})\big] = 0. \tag{14}$$

Hint: Insert $\hat{g}^{(2)}(\mathbf{z}) = -\nabla_\phi \log Q_\phi(\mathbf{Z}{=}\mathbf{z})$ into Eq. 14 and write out the expectation again as a weighted average. Then use again the derivative of the logarithm, simplify, pull "∇_ϕ" out of the sum, and use the fact that $Q_\phi(\mathbf{Z})$ is normalized.

Solution: We'll do the derivation again for discrete $\mathbf{Z}$.

$$\begin{aligned}
\mathbb{E}_{\mathbf{z}\sim Q_\phi(\mathbf{Z})}\big[\hat{g}^{(2)}(\mathbf{z})\big] &= -\sum_{\mathbf{z}} Q_\phi(\mathbf{Z}{=}\mathbf{z})\,\nabla_\phi \log Q_\phi(\mathbf{Z}{=}\mathbf{z}) \\
&= -\sum_{\mathbf{z}} Q_\phi(\mathbf{Z}{=}\mathbf{z})\,\frac{\nabla_\phi\, Q_\phi(\mathbf{Z}{=}\mathbf{z})}{Q_\phi(\mathbf{Z}{=}\mathbf{z})} \\
&= -\sum_{\mathbf{z}} \nabla_\phi\, Q_\phi(\mathbf{Z}{=}\mathbf{z}) = -\nabla_\phi\Big(\sum_{\mathbf{z}} Q_\phi(\mathbf{Z}{=}\mathbf{z})\Big) = -\nabla_\phi(1) = 0.
\end{aligned}$$

Note: terms like $\hat{g}^{(2)}(\mathbf{z})$ that vanish in expectation can still be useful as so-called *control variates* that reduce gradient noise (Ranganath et al., 2014), thus speeding up convergence of SGD. For any unbiased gradient estimate $\hat{g}(\mathbf{z})$, we can obtain a whole family of unbiased gradient estimates $\hat{g}_\alpha(\mathbf{z}) := \hat{g}(\mathbf{z}) + \alpha f(\mathbf{z})$ where $\alpha \in \mathbb{R}$ and $f(\mathbf{z})$ satisfies $\mathbb{E}_{Q_\phi(\mathbf{z})}[f(\mathbf{Z})] = 0$. We can then choose α by minimizing the gradient variance $\mathbb{E}_{Q_\phi(\mathbf{z})}\big[(\hat{g}_\alpha(\mathbf{Z}) - g)^2\big]$. Here, g is the true gradient, which one does not need to know as the minimization is equivalent to minimizing $\mathbb{E}_{Q_\phi(\mathbf{z})}\big[\hat{g}_\alpha(\mathbf{Z})^2\big]$. ■

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.