# Solutions to Problem Set 8

*discussed:*
*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 $\phi$. 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 $\phi$ (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 $\phi$ 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 $H\big[Q_\phi(Z \mid \mathbf{X}=\mathbf{x})\big]$. 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?).

**Solution:** The first term on the right-hand side of Eq. 3 is an expectation of a fixed function ($\log P(Z, \mathbf{X}=\mathbf{x})$) under the variational distribution, over whose parameters we optimize. We can understand the expectation as a weighted average, where the weights have to add up to one. Clearly, if we want to make such a weighted average as large as possible, we have to put all available weight on the largest term (more technical argument: taking a weighted average is (trivially) a convex operation). Thus, the first term on the right-hand side of Eq. 3 is maximized by a deterministic variational distribution $\mathbb{E}_{Q_{\phi^*}}$ that puts all weight on $z^* := \arg\max_z \log P(Z=z, \mathbf{X}=\mathbf{x})$, provided that this distribution is part of the variational family. Thus, maximizing only the first term on the right-hand side of Eq. 3 would result in a simple maximum a-posteriori (MAP) estimate. (There's an edge case where the maximum of $P(Z, \mathbf{X}=\mathbf{x})$ is degenerate; in this case, we may divide the weight in $Q_{\phi^*}$ arbitrarily between the degenerate maxima, so there's an infinite number of solutions for $Q_{\phi^*}$, and some but not all of these are fully deterministic probability distributions.)

If we now consider again the maximization of the full expression in Eq. 3 then we can see that the entropy term $H\big[Q_\phi(Z \mid \mathbf{X}=\mathbf{x})\big]$ punishes such deterministic variational distributions: the entropy of a deterministic (discrete) distribution is zero, which is the smallest possible value given that the entropy of *any* (discrete) distribution is nonnegative. Therefore, the entropy term provides an incentive for the maximization of the ELBO to go beyond a simple MAP estimate. As we'll show in part (c) below, the two contributions to the ELBO (expected log joint and entropy) conspire together so that the variational distribution that maximizes the ELBO is in some specific sense the one that's closest to the true posterior. ■

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

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

(*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 $Q_{\phi^*}(Z \mid \mathbf{X}=\mathbf{x})$ if we were maximizing (i) only the first term or (ii) only the second term on the right-hand side of Eq. 4 over $\phi$ (no calculation required).

**Solution:** The equivalence between Eqs. 1 and 4 follows directly from inserting

the definition of the KL-divergence into Eq. 4,

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

By an argument analogous to part (a), maximizing only the first term on the right-hand side of Eq. 4 would lead to a deterministic variational distribution that—this time—puts all its probability mass on the maximum likelihood solution $z^* = \arg\max_z \log P(\mathbf{X} = \mathbf{x}\,|\,Z = z)$. By contrast, maximizing only the second term on the right-hand side of Eq. 4 would correspond to minimizing (because of the minus sign) the KL-divergence from the prior to the variational distribution, which takes its minimal value (zero) if the variational distribution is equal to the prior. We can interpret maximizing both terms as a regularized maximum-likelihood estimation: the expected log likelihood term tries to fit the variational distribution to the data $\mathbf{x}$ while the KL-term prevents the method from being overly confident. ■

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

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

(*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 $\mathbb{E}_{Q_\phi(Z|\mathbf{X}=\mathbf{x})}\big[\tilde{R}^{(Z)}_{\mathrm{net}}(\mathbf{x})\big]$ 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 $\phi$ relate to Bayesian inference?

**Solution:** The equivalence between Eqs. 1 and 5 follows directly from inserting the definition of the KL-divergence into Eq. 5,

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

where we used the fact that $\log P(\mathbf{X}=\mathbf{x})$ is a constant (as opposed to a function of $Z$) and can thus be pulled into the expectation.

The KL-divergence on the right-hand side of Eq. 5 is zero if the variational distribution equals the true posterior, $Q_\phi(Z\,|\,\mathbf{X} = \mathbf{x}) = P(Z\,|\,\mathbf{X} = \mathbf{x})$. Thus, by combining Eq. 5 with Eq. 2, we find that the difference between the net bit rate $\tilde{R}^{(Z)}_{\mathrm{net}}(\mathbf{x})$ of the modified bits-back algorithm and the bit rate $R_{\mathrm{net}}(\mathbf{x})$ of the exact

bits-back algorithm is, in expectation, the KL-diveregence from the true posterior to the variational distribution,

$$\mathbb{E}_{Q_\phi(Z|\mathbf{X}=\mathbf{x})}\big[\tilde{R}^{(Z)}_{\text{net}}(\mathbf{x})\big] - R_{\text{net}}(\mathbf{x}) = D_{\text{KL}}\big(Q_\phi(Z\,|\,\mathbf{X}=\mathbf{x})\,\big|\big|\,P(Z\,|\,\mathbf{X}=\mathbf{x})\big) \geq 0.$$

Maximizing the ELBO in Eq. 5 not only minimizes the expected bit rate of the modified bits-back algorithm. It can also be understood as searching for the variational distribution that minimizes the KL-divergence from the true posterior distribution (second term on the right-hand side of Eq. 5). In this sense, the resulting optimal variational distribution $Q_{\phi^*}(Z\,|\,\mathbf{X}=\mathbf{x})$ with $\phi^* := \arg\max_\phi \text{ELBO}(\phi)$ can be regarded as an approximation to the true posterior. ■

## 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 $\phi_i$ 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. $\phi_i$, 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: $\text{ELBO}(\phi) = \mathbb{E}_{Q_\phi(Z|\mathbf{X}=\mathbf{x})}[\ell(\phi, Z)]$ with $\ell(\phi, Z) = \log P(Z, \mathbf{X}=\mathbf{x}) - \log Q_\phi(Z\,|\,\mathbf{X}=\mathbf{x})$. 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 $z_\text{s} \sim Q_\phi(Z\,|\,\mathbf{X}=\mathbf{x})$;
- evaluate the gradient $\hat{g} := \nabla_\phi\, \ell(\phi, z_\text{s})$ w.r.t. $\phi$ at this sample;

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

- 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 $\phi$ appears in the ELBO.

**Solution:** The gradient step $\phi \leftarrow \phi + \rho\hat{g}$ in stochastic gradient descent has to be constructed from an *unbiased* gradient estimate $\hat{g}$, i.e., an estimate that satisfies $\mathbb{E}_{Q_\phi(Z|\mathbf{X}=\mathbf{x})}[\hat{g}] = \nabla_\phi \mathrm{ELBO}(\phi)$. However, the above estimate does not satisfy this requirement because it only takes the gradient of the term inside the expectation in Eq. 1. This neglects the fact that the distribution $Q_\phi(Z \mid \mathbf{X}=\mathbf{x})$ over which the expectation is taken depends on $\phi$ itself. This dependency also contributes to the gradient:

$$\begin{aligned}
\nabla_\phi \mathrm{ELBO}(\phi) &= \nabla_\phi \Big( \mathbb{E}_{Q_\phi(Z|\mathbf{X}=\mathbf{x})} \big[ \ell(\phi, Z) \big] \Big) \\
&= \nabla_\phi \Big( \sum_z Q_\phi(Z=z \mid \mathbf{X}=\mathbf{x})\, \ell(\phi, z) \Big) \\
&= \sum_z \big[ \nabla_\phi Q_\phi(Z=z \mid \mathbf{X}=\mathbf{x}) \big] \ell(\phi, z) + \mathbb{E}_{Q_\phi(Z|\mathbf{X}=\mathbf{x})}[\hat{g}].
\end{aligned}$$

Thus, $\mathbb{E}_{Q_\phi(Z|\mathbf{X}=\mathbf{x})}[\hat{g}] \neq \nabla_\phi \mathrm{ELBO}(\phi)$ in general, i.e., $\hat{g}$ is *not* an unbiased gradient estimate. ■

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, $Q_\phi$ 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 $\phi$ 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 $\nabla_\phi \mathrm{ELBO}(\phi)$.

**Solution:** The equivalence between Eqs. 6 and 7 simply follows from the fact that a normal distribution with mean $\mu_i$ and standard deviation $\sigma_i$ can be obtained from stretching a standard normal distribution by $\sigma_i$ and shifting it by $\mu_i$ (in fact, this is the most typical way how one draws samples from a normal distribution in practice: use a library function to draw a sample from a standard normal distribution, then scale and shift it).

In the reparameterization in Eq. 7, the distribution $\mathcal{N}(0, I)$ over which the expectation on the right-hand side is taken no longer depends on the variational parameters $\phi$. Thus, we can take the gradient inside the expectation:

$$\nabla_\phi \,\mathrm{ELBO}(\phi) = \nabla_\phi \,\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)}\big[\ell(\phi, \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon})\big] = \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)}\big[\nabla_\phi\, \ell(\phi, \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon})\big]$$

where one mustn't forget that $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the components of the variational parameters $\phi$, i.e., the function $\ell$ inside the expectation depends on $\phi$ in both of its arguments (in practice, this is usually automatically taken into account by an automatic differentiation framework). Thus, one can maximize the ELBO with stochastic gradient descent by drawing random samples $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$, calculating gradient estimates $\hat{g}_{\text{reparam.}} := \nabla_\phi\, \ell(\phi, \boldsymbol{\mu}+\boldsymbol{\sigma}\odot\boldsymbol{\epsilon})$, and then updating the variational parameters with the gradient step $\phi \leftarrow \phi + \rho\hat{g}_{\text{reparam.}}$. ■

(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 $\nabla_\phi$ 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.

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

$$
\begin{aligned}
&\nabla_\phi \,\mathrm{ELBO}(\phi) = \\
&= \nabla_\phi \big( \mathbb{E}_{z\sim Q_\phi(Z|\mathbf{X}=\mathbf{x})} \big[ \ell(\phi, z) \big] \big) \\
&= \nabla_\phi \Big( \sum_z Q_\phi(Z=z \,|\, \mathbf{X}=\mathbf{x})\, \ell(\phi, z) \Big) \\
&= \sum_z \nabla_\phi \Big( Q_\phi(Z=z \,|\, \mathbf{X}=\mathbf{x})\, \ell(\phi, z) \Big) \\
&= \sum_z \big( \nabla_\phi Q_\phi(Z=z \,|\, \mathbf{X}=\mathbf{x}) \big) \ell(\phi, z) + \sum_z Q_\phi(Z=z \,|\, \mathbf{X}=\mathbf{x}) \big( \underbrace{\nabla_\phi \ell(\phi, z)}_{=\hat{g}^{(2)}(z)} \big) \\
&= \sum_z Q_\phi(Z=z \,|\, \mathbf{X}=\mathbf{x})\, \frac{\nabla_\phi Q_\phi(Z=z \,|\, \mathbf{X}=\mathbf{x})}{Q_\phi(Z=z \,|\, \mathbf{X}=\mathbf{x})}\, \ell(\phi, z) + \mathbb{E}_{z\sim Q_\phi(Z|\mathbf{X}=\mathbf{x})} \big[ \hat{g}^{(2)}(z) \big] \\
&= \sum_z Q_\phi(Z=z \,|\, \mathbf{X}=\mathbf{x})\, \underbrace{\big( \nabla_\phi \log Q_\phi(Z=z \,|\, \mathbf{X}=\mathbf{x}) \big)\, \ell(\phi, z)}_{=\hat{g}^{(1)}(z)} + \mathbb{E}_{z\sim Q_\phi(Z|\mathbf{X}=\mathbf{x})} \big[ \hat{g}^{(2)}(z) \big] \\
&= \mathbb{E}_{z\sim Q_\phi(Z|\mathbf{X}=\mathbf{x})} \big[ \hat{g}^{(1)}(z) + \hat{g}^{(2)}(z) \big] = \mathbb{E}_{z\sim Q_\phi(Z|\mathbf{X}=\mathbf{x})} \big[ \hat{g}(z) \big].
\end{aligned}
$$

■

(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 $Z$.

**Solution:** We'll do the derivation again for discrete $Z$.

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

*Note:* Such contributions to a gradient estimate whose expectation value is zero may still be useful because they may (if constructed well) reduce the variance of the gradient estimate (aka "gradient noise"), which speeds up convergence of stochastic gradient optimization because it allows using larger learning rates. Terms with this property are called "control variates", and there is still a lot of ongoing research about finding good control variates (e.g., in automatic ways). ■

# 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.