Interim Report: Noise Contrastive Estimation with Unobserved Variables

Ben Rhodes

March 16, 2018

1 Introduction

This thesis presents a method for estimating the parameters of unnormalised probability densities with unobserved (or 'latent') variables. The method is an extension of noise-contrastive estimation, which applies only to unnormalised, fully observed models. Before introducing the method, it helps to first understand the definitions and difficulties of both unnormalised and latent variable models.

An unnormalised model, $\phi(\mathbf{u};\boldsymbol{\theta})$, is a parametrised family of functions that you can think of as scaled probability density [1] functions that do not integrate to 1 for all values of $\boldsymbol{\theta}$. That is:

$$\int_{\mathbf{u}} \phi(\mathbf{u};\boldsymbol{\theta})\, \mathrm{d}\mathbf{u} = Z(\boldsymbol{\theta}) \neq 1 \qquad\qquad \phi(\mathbf{u};\boldsymbol{\theta}) \geq 0 \tag{1}$$

where $Z(\boldsymbol{\theta})$ is called the *partition function.* The partition function is important because it allows us to normalise $\phi(\mathbf{u};\boldsymbol{\theta})$, obtaining a model $p(\mathbf{u};\boldsymbol{\theta})$ that, regardless of the value of $\boldsymbol{\theta}$, integrates to 1.

$$p(\mathbf{u};\boldsymbol{\theta}) = \frac{\phi(\mathbf{u};\boldsymbol{\theta})}{Z(\boldsymbol{\theta})}, \qquad\qquad \int_{\mathbf{u}} p(\mathbf{u};\boldsymbol{\theta})\, \mathrm{d}\mathbf{u} = 1. \tag{2}$$

Unfortunately, the partition function is rarely computable analytically, and the cost of approximating it numerically scales exponentially in general. The intractability of the partition function becomes problematic when we want to estimate good values of $\boldsymbol{\theta}$ given some data $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$, where 'good' - in practical terms - means that $p(\mathbf{x}_i;\boldsymbol{\theta})$ is large for all i (whilst still satisfying the constraint of integrating to 1). It is problematic because the standard approach of maximum likelihood estimation requires us to maximise an objective that depends on the partition function; namely, the log-likelihood:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\mathbf{x}_i;\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \phi(\mathbf{x}_i;\boldsymbol{\theta}) - n\log(Z(\boldsymbol{\theta})), \tag{3}$$

[1] unless stated otherwise, assume that all results apply equally to probability mass functions defined over discrete random variables, replacing integrals with sums as necessary.

In light of this intractability, significant work has gone into building estimators that avoid directly computing the partition function, such as pseudolikelihood, score matching, ratio matching, contrastive divergence (and its variants), and noise-contrastive estimation (!!!). It is this final method, noise-contrastive estimation, that we build on this thesis, extending its applicability to latent variable models.

A latent variable model is a family of probability densities $p(\mathbf{u}; \boldsymbol{\theta})$ for which we do not have a direct expression; instead we only have a joint $p(\mathbf{u}, \mathbf{z}; \boldsymbol{\theta})$ density, where $\mathbf{z}$ are hidden variables, not present in our data set. We thus obtain $p(\mathbf{u}; \boldsymbol{\theta})$ only after marginalisation:

$$p(\mathbf{u}; \boldsymbol{\theta}) = \int p(\mathbf{u}, \mathbf{z}; \boldsymbol{\theta}) \, \mathrm{d}\mathbf{z} \tag{4}$$

Learning the parameters $\boldsymbol{\theta}$ is problematic here for essentially the same reason as in normalised models: the presence of an intractable integral. In section (!!!) we review different strategies developed for dealing with this integral as part of the topic of approximate inference.

To see why latent variable models are important, it helps to consider two conceptually distinct (but mathematically equivalent) types of latent variables. The first type represent 'missing data'; perhaps certain pixels in an image were corrupted, generating NaNs. The second type represent compressions of the visible data, which often (but not always) capture low-dimensional causes of the visible data. Such compressions are at the core of unsupervised learning, since the ability to build parsimonious models of the hidden causes of your observations appears to be a fundamental aspect of intelligence. Thus, it is primarily the second view of latent variables that motivates the widespread use of such models.

Let us now consider the combined case of an unnormalised, latent variable model. To write down an expression for $p(\mathbf{u}; \boldsymbol{\theta})$, we combine equations 2 and 4 to get:

$$p(\mathbf{u}; \boldsymbol{\theta}) = \frac{\int \phi(\mathbf{u}, \mathbf{z}; \boldsymbol{\theta}) \, \mathrm{d}\mathbf{z}}{Z(\boldsymbol{\theta})} \tag{5}$$

where $Z(\boldsymbol{\theta})$ is as we defined it in equation 1. This expression contains two troublesome integrals, and it is this double intractability that is the fundamental technical barrier we must overcome to perform parameter estimation.

Of course, there is no use solving this problem if such models are not useful. Fortunately, this is not the case. Unnormalised, latent variable models arise naturally in the context of undirected graphical models (or 'Markov networks') which are very general tools for describing probabilistic relations between random variables. An exemplary case is the Restricted Boltzmann Machine (RBM) which is given a thorough treatment in section (!!!). The dominant approach to learning in models such as an RBM is to use a variant of contrastive divergence (as described in section (!!!background)). However, there are theoretical and empirical reasons to prefer noise-contrastive estimation to contrastive divergence (see section !!!!) in the case of fully observed models, and so we might expect the same to be true in the latent variable case. Investigating this claim both theoretically and empirically is one of principal goals of this thesis.

2 Background: Noise-contrastive estimation

NCE converts an unsupervised density estimation problem into a supervised classification problem, by training a (non-linear) logistic classifier to distinguish between the true data and samples from some reference distribution, which we refer to as noise samples.

Concretely, we first generate m samples $\{\mathbf{y}_1, \dots \mathbf{y}_m\}$ from a noise distribution $p_{\mathbf{y}}(\mathbf{y})$ and mix these samples with our data $(\mathbf{x}_1, \dots, \mathbf{x}_n)$. Our goal is now to classify whether a random sample $\mathbf{u}$ from this mixture came from the model ϕ or the noise $p_{\mathbf{y}}$. We can do this very simply by calculating posterior probabilities. If $\nu = \frac{m}{n}$, then:

$$\mathbb{P}(\mathbf{u} \sim p_{\mathbf{x}}; \boldsymbol{\theta}) = \frac{\phi(\mathbf{u}; \boldsymbol{\theta})}{\phi(\mathbf{u}; \theta) + \nu p_{\mathbf{y}}(\mathbf{u})} = \frac{1}{1 + \nu \exp\big(-h(\mathbf{u}; \boldsymbol{\theta})\big)} \tag{6}$$

$$\mathbb{P}(\mathbf{u} \sim p_{\mathbf{y}}; \boldsymbol{\theta}) = \frac{p_{\mathbf{y}}(\mathbf{u})}{\frac{1}{\nu}\phi(\mathbf{u}; \theta) + p_{\mathbf{y}}(\mathbf{u})} = \frac{1}{1 + \frac{1}{\nu} \exp\big(h(\mathbf{u}; \boldsymbol{\theta})\big)} \tag{7}$$

$$\tag{8}$$

where:

$$h(\mathbf{u}; \boldsymbol{\theta}) = \log \phi(\mathbf{u}; \boldsymbol{\theta}) - \log p_{\mathbf{y}}(\mathbf{u}) \tag{9}$$

At first, the rightmost expressions in 6 and 7 may appear unnecessary, but if we absorb the ν terms into the exponentials, then we see that the probabilities of each class (data or noise), is expressed with a sigmoid. Our classification problem can therefore be thought of as a (non-linear) logistic regression problem, with objective function:

$$J_n(\boldsymbol{\theta}) = \frac{1}{n}\left\{ \sum_{i=1}^{n} \log \mathbb{P}(\mathbf{x}_i \sim p_{\mathbf{x}}; \boldsymbol{\theta}) + \sum_{i=1}^{m} \log \big[1 - \mathbb{P}(\mathbf{y}_i \sim p_{\mathbf{x}}; \boldsymbol{\theta})\big] \right\} \tag{10}$$

$$= -\frac{1}{n}\left\{ \sum_{i=1}^{n} \log \big[1 + \nu \exp(-h(\mathbf{x}_i; \boldsymbol{\theta}))\big] + \sum_{i=1}^{m} \log \left[1 + \frac{1}{\nu} \exp(h(\mathbf{y}_i; \boldsymbol{\theta}))\right] \right\}, \tag{11}$$

which is the sample version of $J(\boldsymbol{\theta})$,

$$J(\boldsymbol{\theta}) = -\mathbb{E}_{\mathbf{x}} \log \big[1 + \nu \exp(-h(\mathbf{x}; \boldsymbol{\theta}))\big] - \nu \mathbb{E}_{\mathbf{y}} \log \big[1 + 1/\nu \exp(h(\mathbf{y}; \boldsymbol{\theta}))\big]. \tag{12}$$

If we assume that the true data generating distribution is $p(\mathbf{u}; \hat{\boldsymbol{\theta}})$, and that this distribution is contained within our unnormalised family $\phi(\mathbf{u}; \boldsymbol{\theta})$, then optimising the above objective function guarantees that $\boldsymbol{\theta}$ converges (in probability) to $\hat{\boldsymbol{\theta}}$ in the limit of infinite data.

Why should the *normalised* data generating distribution live in an *unnormalised* family? This assumption may seem odd at first, but note that it is simple to give the model family $\phi(\mathbf{u}; \boldsymbol{\theta})$ an extra degree of freedom by multiplying it with a scaling parameter c. This new parameter is estimated in conjunction with $\boldsymbol{\theta}$, so that we automatically obtain a normalised solution. [2]

[2] note, this is *not* the same as normalising the model i.e computing the partition function, which provides the normalisation constant for *every* value of $\boldsymbol{\theta}$

It turns out that for r sufficiently large, the choice of noise distribution q is not important. However, increasing r uses more computational resource, so we still need to select an appropriate distribution. It remains an area of active research to select q based on theoretically justified properties, with the main constraints currently being that it should be cheap to sample from, non-zero wherever p is non-zero, and similar to the underlying the dataset.

In the case of latent variable models, the extended version of NCE that we present in this thesis is able to partially automate the construction of a suitable noise distribution, which is a significant advance over hand-crafted choices.

2.1 variational lower bound on NCE objective function

Describe result here, attributing it to unpublished work of Guttmann's.

This section contains an abbreviated derivation of a variational lower bound to the NCE objective function in equation 12. [3]

The problem with standard NCE objective function is the presence of the intractable integral over latent variables. It appear twice: In term 1 and term 2.

The standard method to deal with an objective function containing this integral is to lower bound it with a new objective function that does not involve any intractable terms. We can then apply gradient-based optimisation to maximise the lower bound. This is the essential premise of variational inference, and typically involves the following steps:

Re-express the intractable integral using Importance sampling with auxiliary distribution q. Call this new integral I. Apply Jensen's inequality to a concave function g(I). Parameterise the auxiliary distribution such that we can optimise the lower bound with respect to its parameters.

In particular, the Evidence-lower-bound [!!!] starts with the log-likelihood: Math of log-likelihood re-expressed with importance sampling And then applies Jensen's inequality: More math Optimising this lower bound with respect to the parameters of q is non-trivial, and we return to this matter later.

For now, let's apply the same strategy to our case, starting with term 1. We can no longer apply Jensen's to the g(I) = log(I) as we do in the ELBO. Instead, we apply it to $g(I) = -\log(1 + \nu p_y/I)$. The same choice of g won't work for lower bounding the second term. However, we don't necessarily need to. We we can simply re-express it with importance sampling, reusing the auxiliary distribution we got from the first term. The final result is: math

List the properties of model, noise and variational distribution required to use this objective function.

[3] I should emphasise that the derivation of this lower bound is *not* my work, but the unpublished work of Michael Guttmann.

3 Research Questions

- Can we successfully perform parameter estimation using the variational lower bound given in section (!!!)?
- How does this new method empirically compare to others, such as contrastive divergence, when applied across multiple models and datasets?
- What are the theoretical properties of the new estimation method? This is obviously too broad a question, and needs to be concretised. Which specific theoretical properties to investigate depend upon the empirical work, but two lines of research seem worthwhile. The first is concerned with convergence properties: is the method guaranteed to converge? What is the estimator's asymptotic variance? The second line of enquiry involves mathematical comparisons of the new method to methods such as contrastive divergence.
- Investigation of models for which we cannot easily apply contrastive divergence. Contrastive divergence requires us to be able to sample from the posterior over latent variables. The new method, however, only needs a variational approximation of the posterior. It is possible then, that we could learn the parameters of a new class of models for which learning was previously very difficult. For instance, a modified RBM with connections between latent variables, or with multiple layers of latent variables.

4 Results

This section contains two main subsections: the first outlines current results, whilst the second outlines work to be completed during the remainder of the thesis. In the first section, we cover:

- Application of the new objective function to a simple problem, namely an unnormalised mixture of gaussians.
- Preliminary results applying the new objective function to a Restricted Boltzmann Machine.

In the remaining work section, we cover the following:

- Outline of empirical work to be done during the remainder of the thesis.
- Suggestions for further theoretical work to be done during the remainder of the thesis.

4.1 Unnormalised mixture of gaussians

4.2 Restricted Boltzmann Machine

4.3 Further empirical work

4.4 Further theoretical work

5 Present conclusions and remaining work

Our goal in this preliminary version of the thesis was twofold. Firstly, to give an exposition of an extension to noise-contrastive estimation which applies to unnormalised models with latent variables. Secondly, to evaluate the performance of this new method in the context of two models: a simple unnormalised mixture of gaussians, and a Restricted Boltzmann Machine.

Both evaluations are ongoing, so it is not yet clear how the new method compares to others, such as contrastive divergence. Nevertheless, current results demonstrate that the method can be successfully used for density estimation, at least on simple problems. The remainder of the thesis will need to critically assess how the method scales with model complexity, sample size and diversity of data generating distributions.

Other topics to explore in remainder of thesis: Theoretical comparison of new optimisation method with contrastive divergence. Once I have clearer empirical results on the differences between the two methods, it might be valuable to explain these differences with theory. Theoretical results on the convergence properties of this new estimator. The original NCE paper presented such a result, although an analogous theorem might be hard to establish for this new method. (EM convergence theorem?) RBMs have tractable posterior distributions, which makes learning easier. What about models with intractable posteriors? Deep Boltzmann machines?

6 Appendix

6.1 Lower bound for the first term in the NCE objective function

We first re-write $-\log[1 + \nu \exp(-h(\mathbf{u}; \boldsymbol{\theta}))]$ by inserting the definition of h

$$-\log[1 + \nu \exp(-h(\mathbf{u}; \boldsymbol{\theta}))] = -\log\left[1 + \nu \exp\left(-\log \frac{\phi(\mathbf{u}; \boldsymbol{\theta})}{p_{\mathbf{y}}(\mathbf{u})}\right)\right] \tag{13}$$

$$= -\log\left[1 + \nu \exp\left(\log \frac{p_{\mathbf{y}}(\mathbf{u})}{\phi(\mathbf{u}; \boldsymbol{\theta})}\right)\right] \tag{14}$$

$$= -\log\left[1 + \nu \frac{p_{\mathbf{y}}(\mathbf{u})}{\phi(\mathbf{u}; \boldsymbol{\theta})}\right] \tag{15}$$

Let further

$$r(\mathbf{u}; \boldsymbol{\theta}) = \exp(h(\mathbf{u}; \boldsymbol{\theta})) = \frac{\phi(\mathbf{u}; \boldsymbol{\theta})}{p_{\mathbf{y}}(\mathbf{u})} \tag{16}$$

and

$$g(r) = -\log\left(1 + \nu \frac{1}{r}\right), \tag{17}$$

we then obtain

$$-\log[1 + \nu \exp(-h(\mathbf{u}; \boldsymbol{\theta}))] = g(r(\mathbf{u}; \boldsymbol{\theta})). \tag{18}$$

The first derivative of g with respect to r is

$$\frac{\mathrm{d}g}{\mathrm{d}r} = \frac{-1}{1 + \nu \frac{1}{r}} \frac{-\nu}{r^2} \tag{19}$$

$$= \frac{\nu}{r^2 + \nu r} \tag{20}$$

The second derivative of g with respect to r is thus

$$\frac{\mathrm{d}^2 g}{\mathrm{d}r^2} = -\nu \frac{2r + \nu}{(r^2 + \nu r)^2}, \tag{21}$$

which is always negative for $r > 0$. The function $g(r)$ is thus concave.

We now introduce an auxiliary distribution $q_u(\mathbf{z})$ over $\mathbf{z}$ and write

$$\phi(\mathbf{u}; \boldsymbol{\theta}) = \int \phi(\mathbf{u}, \mathbf{z}; \boldsymbol{\theta}) \, \mathrm{d}\mathbf{z} \tag{22}$$

$$= \int q_u(\mathbf{z}) \frac{\phi(\mathbf{u}, \mathbf{z}; \boldsymbol{\theta})}{q_u(\mathbf{z})} \, \mathrm{d}\mathbf{z} \tag{23}$$

$$= \mathbb{E}_{\mathbf{z} \sim q_u}\left[\frac{\phi(\mathbf{u}, \mathbf{z}; \boldsymbol{\theta})}{q_u(\mathbf{z})}\right] \tag{24}$$

This corresponds to estimating $\phi(\mathbf{u},\boldsymbol{\theta})$ by importance sampling. The subscript for $q_u(\mathbf{z})$ is meant to indicate that we can use different auxiliary distributions for different values of $\mathbf{u}$.

We can thus write $r(\mathbf{u};\boldsymbol{\theta})$ as

$$\begin{aligned} r(\mathbf{u};\boldsymbol{\theta}) &= \frac{1}{p_{\mathbf{y}}(\mathbf{u})}\mathbb{E}_{\mathbf{z}\sim q_u}\left[\frac{\phi(\mathbf{u},\mathbf{z};\boldsymbol{\theta})}{q_u(\mathbf{z})}\right] && (25)\\ &= \mathbb{E}_{\mathbf{z}\sim q_u}\left[\frac{\phi(\mathbf{u},\mathbf{z};\boldsymbol{\theta})}{q_u(\mathbf{z})p_{\mathbf{y}}(\mathbf{u})}\right] && (26)\end{aligned}$$

and since g is concave, we have

$$\begin{aligned} -\log[1+\nu\exp(-h(\mathbf{u};\boldsymbol{\theta}))] &= g(r(\mathbf{u};\boldsymbol{\theta})) && (27)\\ &= g\left(\mathbb{E}_{\mathbf{z}\sim q_u}\left[\frac{\phi(\mathbf{u},\mathbf{z};\boldsymbol{\theta})}{q_u(\mathbf{z})p_{\mathbf{y}}(\mathbf{u})}\right]\right) && (28)\\ &\geq \mathbb{E}_{\mathbf{z}\sim q_u} g\left(\frac{\phi(\mathbf{u},\mathbf{z};\boldsymbol{\theta})}{q_u(\mathbf{z})p_{\mathbf{y}}(\mathbf{u})}\right). && (29)\end{aligned}$$

With the definition of g, we thus obtain

$$-\log[1+\nu\exp(-h(\mathbf{u};\boldsymbol{\theta}))] \geq -\mathbb{E}_{\mathbf{z}\sim q_u}\log\left[1+\nu\frac{q_u(\mathbf{z})p_{\mathbf{y}}(\mathbf{u})}{\phi(\mathbf{u},\mathbf{z};\boldsymbol{\theta})}\right]. \qquad (30)$$

We now plug this relation into $J(\boldsymbol{\theta})$ in (12), where the values $\mathbf{u}$ are given by the values of the random vector $\mathbf{x}$. We thus write $q(\mathbf{z}|\mathbf{x})$ instead of $q_u(\mathbf{z})$. Doing so, we obtain

$$\begin{aligned} J(\boldsymbol{\theta}) &= -\mathbb{E}_{\mathbf{x}}\log\left[1+\nu\exp(-h(\mathbf{x};\boldsymbol{\theta}))\right] - \nu\mathbb{E}_{\mathbf{y}}\log\left[1+1/\nu\exp(h(\mathbf{y};\boldsymbol{\theta}))\right] && (31)\\ &\geq -\mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x})}\log\left[1+\nu\frac{q(\mathbf{z}|\mathbf{x})p_{\mathbf{y}}(\mathbf{x})}{\phi(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}\right] - \nu\mathbb{E}_{\mathbf{y}}\log\left[1+1/\nu\exp(h(\mathbf{y};\boldsymbol{\theta}))\right]. && (32)\end{aligned}$$

We see that $q(\mathbf{z}|\mathbf{x})$ can be interpreted as a noise distribution for the unobserved variables. Together, $p_{\mathbf{y}}q$ can be considered a noise distribution for both the observed and unobserved data.

The second term in (32),

$$-\nu\mathbb{E}_{\mathbf{y}}\log\left[1+1/\nu\exp(h(\mathbf{y};\boldsymbol{\theta}))\right]$$

still features the term $h(\mathbf{y};\boldsymbol{\theta})$ that contains an intractable integral,

$$h(\mathbf{y};\boldsymbol{\theta}) = \log\frac{\int\phi(\mathbf{y},\mathbf{z};\boldsymbol{\theta})\,\mathrm{d}\mathbf{z}}{p_{\mathbf{y}}(\mathbf{y})} \qquad (33)$$

We can estimate $\exp(h(\mathbf{y};\boldsymbol{\theta})) = r(\mathbf{y};\boldsymbol{\theta})$ using important sampling as in (26). While we here use the same q_u as in (26), it would be possible to work with a different auxiliary distribution q_u. We then obtain an objective function that can be computed by taking sample averages,

$$\mathcal{J}_1(\boldsymbol{\theta}, q) = -\mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x})} \log \left[1 + \nu \frac{q(\mathbf{z}|\mathbf{x})p_{\mathbf{y}}(\mathbf{x})}{\phi(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})}\right] - \nu\mathbb{E}_{\mathbf{y}} \log \left[1 + \frac{1}{\nu}\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{y})} \left[\frac{\phi(\mathbf{y}, \mathbf{z}; \boldsymbol{\theta})}{q(\mathbf{z}|\mathbf{y})p_{\mathbf{y}}(\mathbf{y})}\right]\right]. \tag{34}$$

We have that $J(\boldsymbol{\theta}) \geq \mathcal{J}_1(\boldsymbol{\theta}, q)$ for all q. We maximise $J(\boldsymbol{\theta})$ by iterating between

- maximisation of $\mathcal{J}_1(\boldsymbol{\theta}, q)$ with respect to $\boldsymbol{\theta}$, and
- tightening of the bound by maximising $\mathcal{J}_1(\boldsymbol{\theta}, q)$ with respect to the variational distribution q.

The second step would be done by choosing a certain parametric family of conditional distributions q and optimising its parameters.

This approach is like using importance sampling to compute the intractable integral. The advantage of maximising the bound is that it helps us to choose a good auxiliary distribution (importance distribution) q. But this had to be validated experimentally by comparison to a standard importance sampling approach.

6.2 Lower bound for the second term in the NCE objective function

Alternatively, we can also derive a lower bound for the second term in (32) that can also be computed as a sample average.

We first expand the term $\log\left[1 + 1/\nu \exp(h(\mathbf{u}; \boldsymbol{\theta}))\right]$ as before

$$-\log[1 + 1/\nu \exp(h(\mathbf{u}; \boldsymbol{\theta}))] = -\log\left[1 + 1/\nu \exp\left(\log\frac{\phi(\mathbf{u}; \boldsymbol{\theta})}{p_{\mathbf{y}}(\mathbf{u})}\right)\right] \tag{35}$$

$$= -\log\left[1 + \frac{1}{\nu}\frac{\phi(\mathbf{u}; \boldsymbol{\theta})}{p_{\mathbf{y}}(\mathbf{u})}\right] \tag{36}$$

With $g_2(r) = -\log(1 + r/\nu)$ and r defined as before in Equation (16) as

$$r(\mathbf{u}; \boldsymbol{\theta}) = \frac{\phi(\mathbf{u}; \boldsymbol{\theta})}{p_{\mathbf{y}}(\mathbf{u})} \tag{37}$$

we have

$$-\log[1 + 1/\nu \exp(h(\mathbf{u}; \boldsymbol{\theta}))] = g_2(r(\mathbf{u}; \boldsymbol{\theta})) \tag{38}$$

Taking second derivatives shows that $g_2(r)$ is a convex function,

$$\frac{\mathrm{d}g_2}{\mathrm{d}r} = -\frac{1/\nu}{1 + r/\nu} \tag{39}$$

$$\frac{\mathrm{d}^2 g_2}{\mathrm{d}r^2} = \frac{1/\nu^2}{(1 + r/\nu)^2} > 0 \quad \text{for } r \geq 0. \tag{40}$$

We now use Fenchel's inequality to find a function that lower-bounds (minorises) $g_2(r)$. Let $g_2^*(v)$ be the convex conjugate of $g_2(r)$,

$$g_2^*(s) = \sup_{r \geq 0} \; rs - g_2(r). \tag{41}$$

Fenchel's inequality then gives

$$g_2(r) \geq rs - g_2^*(s) \tag{42}$$

for all r and s. The interpretation of Fenchel's inequality is as follows: variable s is the slope of a tangent to $g_2(r)$ and $-g_2^*(s)$ is the correspond intercept. The convex conjugate $g_2^*(s)$ thus takes the slope as arguments and outputs the corresponding (negative) intercept of the tangent to $g_2(r)$, see e.g. Bishop Section 10.5 and Figure 10.11.

The convex conjugate $g_2^*(s)$ is obtained by solving the optimisation problem

$$\sup_{r \geq 0} \; \left\{ rs - g_2(r) \right\},$$

that is

$$\sup_{r \geq 0} \; \left\{ rs + \log(1 + r/\nu) \right\} \tag{43}$$

Taking the derivative of $rs + \log(1 + r/\nu)$ with respect to r gives

$$\frac{\mathrm{d}}{\mathrm{d}r} rs + \log(1 + r/\nu) = s + \frac{1/\nu}{1 + r/\nu} \tag{44}$$

The second derivative is negative for all r so that the maximising r satisfies

$$s = -\frac{1}{\nu + r} \iff -\frac{1}{s} = \nu + r \iff r = -\nu - \frac{1}{s}. \tag{45}$$

Since both ν and r are positive, the convex conjugate is defined on $s < 0$. The convex conjugate is

$$\begin{align} g_2^*(s) &= \sup_{r \geq 0} \; rs + \log(1 + r/\nu) \tag{46} \\ &= (-\nu - 1/s)s + \log\left[1 + \frac{1}{\nu}(-\nu - 1/s)\right] \tag{47} \\ &= -1 - \nu s + \log\left[-\frac{1}{\nu s}\right] \tag{48} \end{align}$$

The inequality is $g_2(r) \geq rs - g_2^*(s)$ thus becomes

$$g_2(r) \geq rs + 1 + \nu s - \log\left[-\frac{1}{\nu s}\right] \tag{49}$$

for all $r \geq 0$ and $s < 0$. Figure 1 shows an example for $\nu = 1$ and different values of s. The linear minorising functions touch $g_2(r)$ at $r = -\nu - 1/s$, derived above. In other

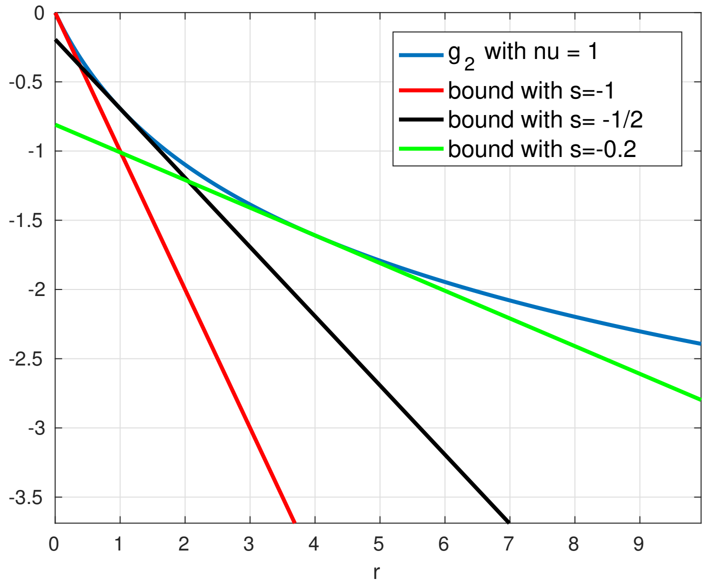

Figure 1: Minorising linear functions for the convex $g_2(r)$. For any r there is a corresponding tangent with slope s and intercept $-g_2^*(s)$.

words, for each r, there is a tangent to $g_2(r)$ with slope $s = -1/(\nu + r)$ and intercept $g_s^*(s)$.

We thus obtain

$$-\log[1 + 1/\nu \exp(h(\mathbf{u};\boldsymbol{\theta}))] = g_2(r(\mathbf{u};\boldsymbol{\theta})) \tag{50}$$

$$\geq r(\mathbf{u};\boldsymbol{\theta})s + 1 + \nu s - \log\left[-\frac{1}{\nu s}\right] \tag{51}$$

and inserting (26) for $r(\mathbf{u};\boldsymbol{\theta})$, we have

$$-\log[1 + 1/\nu \exp(h(\mathbf{u};\boldsymbol{\theta}))] \geq \mathbb{E}_{\mathbf{z}\sim q_u}\left[\frac{\phi(\mathbf{u},\mathbf{z};\boldsymbol{\theta})}{q_u(\mathbf{z})p_{\mathbf{y}}(\mathbf{u})}\right] s + 1 + \nu s - \log\left[-\frac{1}{\nu s}\right]. \tag{52}$$

While we here use the same q_u as for lower bound of the first term, it would be possible to work with a different distribution q_u. Moreover, we should let the value of s depend on r, or $\mathbf{u}$, because for any given value of r, certain s yield a tighter bound.

Inserting the above lower bound with $s = s(\mathbf{y})$ into (32) gives

$$
\begin{aligned}
J(\boldsymbol{\theta}) \geq & -\mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x})} \log\left[1+\nu\frac{q(\mathbf{z}|\mathbf{x})p_{\mathbf{y}}(\mathbf{x})}{\phi(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}\right] - \nu\mathbb{E}_{\mathbf{y}}\log\left[1+1/\nu\exp(h(\mathbf{y};\boldsymbol{\theta}))\right]. \quad (53)\\
\geq & -\mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x})} \log\left[1+\nu\frac{q(\mathbf{z}|\mathbf{x})p_{\mathbf{y}}(\mathbf{x})}{\phi(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}\right]\\
& + \nu\mathbb{E}_{\mathbf{y}}\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{y})}\left[\frac{\phi(\mathbf{y},\mathbf{z};\boldsymbol{\theta})}{q(\mathbf{z}|\mathbf{y})p_{\mathbf{y}}(\mathbf{y})}\right]s(\mathbf{y}) + \nu + \nu^2 s(\mathbf{y}) - \nu\log\left[-\frac{1}{\nu s(\mathbf{y})}\right] \quad (54)
\end{aligned}
$$

We thus have $J(\boldsymbol{\theta}) \geq J_1(\boldsymbol{\theta}, q) \geq J_2(\boldsymbol{\theta}, q)$, with

$$
\begin{aligned}
J_2(\boldsymbol{\theta}, q) = & -\mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x})} \log\left[1+\nu\frac{q(\mathbf{z}|\mathbf{x})p_{\mathbf{y}}(\mathbf{x})}{\phi(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}\right]\\
& + \nu\mathbb{E}_{\mathbf{y}}\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{y})}\left[\frac{\phi(\mathbf{y},\mathbf{z};\boldsymbol{\theta})}{q(\mathbf{z}|\mathbf{y})p_{\mathbf{y}}(\mathbf{y})}\right]s(\mathbf{y}) + \nu + \nu^2 s(\mathbf{y}) - \nu\log\left[-\frac{1}{\nu s(\mathbf{y})}\right] \quad (55)
\end{aligned}
$$

We can interpret $s(\mathbf{y})$ as the number of noise samples $\mathbf{z} \mid \mathbf{y}$ that needed to be sampled for each of the sampled $\mathbf{y}$. One could use a neural network to represent $s(\mathbf{y})$ and learn its parameters by making the bound tighter.

In this second approach, we would learn the parameters $\boldsymbol{\theta}$ by iterating:

- maximisation of $\mathcal{J}_2(\boldsymbol{\theta}, q)$ with respect to $\boldsymbol{\theta}$,
- tightening of the bound by maximising $\mathcal{J}_2(\boldsymbol{\theta}, q)$ with respect to the variational distribution q and with respect to the parameters of the function $s(\mathbf{y})$

Remark: The first approach seems simpler. I don't quite know whether J_2 has any advantages over J_1. Optimisation may be easier. At least it seems more amenable to stochastic gradient ascent.