On the Convergence of the ELBO to Entropy Sums

Jörg Lücke
Machine Learning Lab, Faculty VI
University of Oldenburg
26111 Oldenburg, Germany

Jan Warnken
Machine Learning Lab, Faculty VI
University of Oldenburg
26129 Oldenburg, Germany

Abstract

The variational lower bound (a.k.a. ELBO or free energy) is the central objective for many established as well as many novel algorithms for unsupervised learning. Learning algorithms change model parameters such that the variational lower bound increases. Learning usually proceeds until parameters have converged to values close to a stationary point of the learning dynamics. In this purely theoretical contribution, we show that (for a very large class of generative models) the variational lower bound is at all stationary points of learning equal to a sum of entropies. For standard machine learning models with one set of latents and one set observed variables, the sum consists of three entropies: (A) the (average) entropy of the variational distributions, (B) the negative entropy of the model's prior distribution, and (C) the (expected) negative entropy of the observable distributions. The obtained result applies under realistic conditions including: finite numbers of data points, at any stationary points (including saddle points) and for any family of (well behaved) variational distributions. The class of generative models for which we show the equality to entropy sums contains many well-known generative models. As concrete examples we discuss Sigmoid Belief Networks, probabilistic PCA and (Gaussian and non-Gaussian) mixture models. The results also apply for standard (Gaussian) variational autoencoders, which has been shown in parallel (Damm et al., 2023). The prerequisites we use to show equality to entropy sums are relatively mild. Concretely, the distributions of a given generative model have to be of the exponential family (with constant base measure), and the model has to satisfy a parameterization criterion (which is usually fulfilled). Proving the equality of the ELBO to entropy sums at stationary points (under the stated conditions) is the main contribution of this work.

1 Introduction

We consider probabilistic generative models and their optimization based on the variational lower bound. The variational bound is also known as *evidence lower bound* (ELBO; e.g., Hoffman et al., 2013) or (variational) *free energy* (Neal and Hinton, 1998). For a given set of data points, a learning algorithm based on generative models usually changes the parameters of the model until no or only negligible parameter changes are observed. Learning is then stopped and parameters will usually lie very close to a stationary point of the learning dynamics, i.e., close to a point where the derivatives of the ELBO objective w.r.t. all parameters are zero.

In this theoretical study, we investigate the ELBO objective at stationary points. As our main result, we show that for many generative models (including very common models), the ELBO becomes equal to a sum of entropies at stationary points. The result is very concise such that it can be stated here initially. We will consider standard generative models of the form:

$$\vec{z} \sim p_\Theta(\vec{z}) \quad \text{and} \quad \vec{x} \sim p_\Theta(\vec{x}\,|\,\vec{z})\,, \tag{1}$$

where $\vec{z}$ and $\vec{x}$ are latent and observed variables, respectively, and where Θ denotes the set of model parameters. The distribution $p_\Theta(\vec{z})$ will be denoted the *prior* distribution and $p_\Theta(\vec{x}\,|\,\vec{z})$ will be referred to as *noise model* or *observable distribution*. Given N data points $\vec{x}^{(n)}$, the variational lower bound is given by:

$$\mathcal{F}(\Phi,\Theta) \;=\; \frac{1}{N}\sum_n \int q_\Phi^{(n)}(\vec{z})\,\log\big(p_\Theta(\vec{x}^{(n)}\,|\,\vec{z})\big)\,\mathrm{d}\vec{z} \;-\; \frac{1}{N}\sum_n D_{\mathrm{KL}}\big[q_\Phi^{(n)}(\vec{z}), p_\Theta(\vec{z})\big]\,, \tag{2}$$

where $q_\Phi^{(n)}(\vec{z})$ denote the variational distributions used for optimization.

The main result of this work states that under relatively mild conditions for the generative model in (1) the ELBO decomposes into a sum of entropies at all stationary points of learning. Concretely, we will show that at all stationary points (i.e., at all points of vanishing derivatives of $\mathcal{F}(\Phi,\Theta)$) applies:

$$\mathcal{F}(\Phi,\Theta) \;=\; \frac{1}{N}\sum_n \mathcal{H}[q_\Phi^{(n)}(\vec{z})] \;-\; \mathcal{H}[p_\Theta(\vec{z})] \;-\; \frac{1}{N}\sum_n \mathbb{E}_{q_\Phi^{(n)}}\big\{\mathcal{H}[p_\Theta(\vec{x}\,|\,\vec{z})]\big\}\,. \tag{3}$$

Notably, the result applies for essentially any variational distribution (very weak conditions on $q_\Phi^{(n)}(\vec{z})$), at any stationary point (including saddle points), and for finitely many data points. In principle, the result can also be used for training (because for many generative models it is sufficient that only derivatives w.r.t. a subset of the parameters Θ vanish). In this contribution, we will focus on the derivation of the result itself, however, and on the conditions that have to be satisfied for a generative model in order for the result to apply. The essential condition is that the distributions $p_\Theta(\vec{z})$ and $p_\Theta(\vec{x}\,|\,\vec{z})$ are within the exponential family of distributions but we will also require some conditions for the parameterization of the models for our proof. As examples of generative models for which the result applies, we explicitly discuss sigmoid belief networks (SBNs; Neal, 1992; Saul et al., 1996), probabilistic PCA (Roweis, 1998; Tipping and Bishop, 1999) and mixture models (e.g. McLachlan and Peel, 2004). But the result applies for a broader class of models which also includes prominent models such as Factor Analysis (e.g. Everitt, 1984; Bartholomew and Knott, 1987) or variational autoencoders (VAEs; Kingma and Welling, 2014; Rezende et al., 2014). For standard (Gaussian) VAEs convergence to entropy sums was shown previously (Lücke et al., 2021), and the final version for the treatment of VAEs (Damm et al., 2023) made use of general results provided in this work. Also other hierarchical/deep models and other extensions of the basic graphical model structure (1) can be treated by using the result derived here or by using the here introduced treatment as basis for further generalizations.

The decomposition into entropies can be advantageous from the theoretical and practical perspectives. The entropy decomposition (3) can, for instance, be closed-form also in cases when the original ELBO objective is not (we will briefly discuss below). The computation of variational bounds and/or likelihoods can also become easier for other reasons, e.g., for probabilistic PCA the likelihood at stationary points can be computed from the parameter values alone (we discuss in Sec. 7).

Related work. Given an exponential family of distributions, the relation of its maximum likelihood parameter estimation to the (negative) entropy of the distribution family is well-known in mathematical statistics (e.g. Bickel and Doksum, 2015; Efron, 2023). Some of these well-established results have been used for generative models whose joint distributions take the form of an exponential family of distributions (e.g. Wainwright and Jordan, 2008). The decomposition of variational lower bounds (i.e., ELBOs) into sums of entropies has not been studied to a comparable extent. At least for idealized conditions, it is relatively straight-forward to show that at stationary points the ELBO can be written as a sum of entropies (e.g. Lücke and Henniges, 2012). At the same time, there is a long-standing interest in stationary points of variational learning that are reached under realistic conditions. For instance, likelihoods, lower bounds, and their stationary points have been studied in detail for (restricted) Boltzmann machines (e.g. Welling and Teh, 2003), elementary generative models (e.g. Frey, 1999; Tipping and Bishop, 1999) as well as for general graphical models (Freeman and Weiss, 2000; Weiss and Freeman, 2000; Yedidia et al., 2001; Opper and Saad, 2001). More recently, stationary points of deep generative models have been of considerable interest. Lucas et al. (2019), for instance, linked phenomena like mode collapse in variational autoencoders (VAEs) to the stability of stationary points of the variational lower bound; and Dai et al. (2018) analyzed a series of tractable special cases of VAEs to better understand convergence points of VAE learning. Notably both, Lucas et al. and Dai et al., emphasised the close relation of standard VAEs to principal component analysis (PCA; Roweis, 1998; Tipping and Bishop, 1999). Theoretical results on different versions of PCA (e.g. Tipping and Bishop, 1999; Xu et al., 2010; Holtzman et al., 2020, for probabilistic, sparse, and robust PCA) can consequently be of direct relevance for VAEs and visa versa.

Standard VAEs and standard generative models such as probabilistic PCA (p-PCA) or probabilistic sparse coding (e.g. Olshausen and Field, 1996) all assume Gaussian distributed observables (and standard VAEs and p-PCA also assume Gaussian latents). The previously mentioned contributions Lücke and Henniges (2012) and Damm et al. (2023) both assume Gaussianity. The relation of the ELBO to entropy sums as studied by Lücke and Henniges (2012), furthermore, required additional conditions. Concretely, equality to entropy sums was shown when the following idealized conditions apply: (A) when the limit of infinitely many data points is considered, (B) when the model distribution can exactly match the data distribution, and (C) when a global optimum is reached. Lücke and Henniges (2012) also discussed relaxations of these rather unrealistic assumptions (their Sec. 6) but those relaxations made use of specific properties of p-PCA and sparse coding like models optimized using expectation maximization. Damm et al. (2023) reported a result derived in parallel to the one presented in this work. They focus on standard VAEs, i.e., one specific (but very popular) class of generative models. For standard VAEs, their work shows equality to entropy sums also under realistic conditions (i.e., local optima, imperfect match of model assumptions and data, finite data sets etc.). Their derivations assume the standard choice of Gaussian observables and Gaussian latents. In contrast, we here address the problem of more general distributions and use elementary generative models as examples. For the example of Gaussian VAEs we refer to Damm et al. (2023) who shows that convergence to entropy sums is a property that can apply for potentially very intricate non-linear relations between latents and observables. For standard Gaussian VAEs one implication of the equality to entropy sums is that closed-form expressions to compute the ELBO can be available at stationary points also in cases when the original ELBO is not closed-form (see Damm et al., 2023).

2 The Class of Considered Generative Models

We first define the class of generative models that we consider, i.e., the models and the properties that we require to show convergence to entropy sums. We will also discuss two concrete examples and a counter-example. We keep focusing on generative models of the kind introduced above (Eqn. 1). Such models will be sufficient to communicate the concepts the proof relies on but generalizations will be discussed in Sec. 7. First, we will require a more specific notation for the parameterization of the generative model, however.

Definition A. (Generative Model)

The generative models we consider have prior distribution $p_{\vec{\Psi}}$ with parameters $\vec{\Psi}$, i.e., all parameters are arranged into one column vector with scalar entries. Analogously, the noise distribution $p_{\vec{\Theta}}$ is parameterized by $\vec{\Theta}$ such that:

$$\vec{z} \sim p_{\vec{\Psi}}(\vec{z})\,, \tag{4}$$

$$\vec{x} \sim p_{\vec{\Theta}}(\vec{x}\,|\,\vec{z})\,. \tag{5}$$

□

Arranging prior parameters and noise model parameters into column vectors will be important for further definitions and for the derivation of the main result. For our purposes, the definition of a generative model in terms of prior and noise model distributions will also be notationally more convenient than using the model's joint probability. In general, we will adopt a somewhat less abstract mathematical notation than could be applied. For instance, we will not use Lebesgue measure notation. The motivation is that the main steps of the derivation can be communicated using notations more common in the broader Machine Learning literature. Knowledge, e.g., on Lebesgue measures, Legendre transformations and duality principles are not essential for the here made definitions and the derivations of results. We will, however, briefly discuss more abstract notations in Sec. 7.

Many standard generative models are of the form as described in Definition A. Examples are those named in the introduction. For our purposes, we will require that prior and noise distribution are of the exponential family of distributions, which is the case for most standard generative models of the form of Definition A. But there are exceptions that are not (at least not directly) covered, for instance, sparse coding with mixture of Gaussians prior (Olshausen and Millman, 2000). We will denote the class of considered generative models as *exponential family generative models*.

Definition B. (EF Generative Models)

Given a generative model as given by Definition A, we say the generative model is an *exponential family model (EF model)* if there exist exponential family distributions $p_{\vec{\zeta}}(\vec{z})$ (for the latents) and $p_{\vec{\eta}}(\vec{x})$ (for the observables) such that the generative model can be reparameterized to take the following form:

$$\vec{z} \sim p_{\vec{\zeta}(\vec{\Psi})}(\vec{z}) \tag{6}$$

$$\vec{x} \sim p_{\vec{\eta}(\vec{z};\,\vec{\Theta})}(\vec{x}), \tag{7}$$

where $\vec{\zeta}(\vec{\Psi})$ maps the prior parameters $\vec{\Psi}$ to the natural parameters of distribution $p_{\vec{\zeta}}(\vec{z})$, and where $\vec{\eta}(\vec{z};\vec{\Theta})$ maps $\vec{z}$ and the noise model parameters $\vec{\Theta}$ to the natural parameters of distribution $p_{\vec{\eta}}(\vec{x})$. Furthermore, we demand that the exponential family distributions $p_{\vec{\zeta}}(\vec{z})$ and $p_{\vec{\eta}}(\vec{x})$ have a constant base measure.

□

The natural parameters of the prior can in principle also depend on $\vec{\Theta}$ because reparameterizations are conceivable that also involve part of the noise model parameters for the reparameterized prior. However, we will use the more intuitive but slightly more restricted definition above.

For our main result to apply, it will turn out that it is not sufficient for a generative model to be an EF generative model. An additional property for the mapping from standard to natural parameters is required. Relations of the property to concepts such as minimal exponential family representations (see, e.g., Wainwright and Jordan, 2008) and linear subspaces will briefly be discussed in Sec. 7 .

Definition C. (Parameterization Criterion)

Consider an EF generative model as given in Definition B with mappings to natural parameters given by $\vec{\zeta}(\vec{\Psi})$ and $\vec{\eta}(\vec{z};\vec{\Theta})$. Let us denote by $\frac{\partial\vec{\zeta}(\vec{\Psi})}{\partial\vec{\Psi}^{\mathrm{T}}}$ the Jacobian matrix of the mapping $\vec{\zeta}(\vec{\Psi})$ and by R the number of prior parameters $\vec{\Psi}$, i.e.,

$$\frac{\partial\vec{\zeta}(\vec{\Psi})}{\partial\vec{\Psi}^{\mathrm{T}}} := \left(\frac{\partial\vec{\zeta}(\vec{\Psi})}{\partial\Psi_1},\ldots,\frac{\partial\vec{\zeta}(\vec{\Psi})}{\partial\Psi_R}\right). \tag{8}$$

Similarly, we consider a Jacobian matrix of the mapping $\vec{\eta}(\vec{z};\vec{\Theta})$, i.e.,

$$\frac{\partial\vec{\eta}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}^{\mathrm{T}}} := \left(\frac{\partial\vec{\eta}(\vec{z};\vec{\Theta})}{\partial\theta_1},\ldots,\frac{\partial\vec{\eta}(\vec{z};\vec{\Theta})}{\partial\theta_S}\right), \tag{9}$$

for which we allow for the Jacobian (9) to be constructed using just a subset $\vec{\theta}$ of the parameters $\vec{\Theta}$. The dimension of this subset is referred to as S. We then say that the *EF* generative model fulfills the *parameterization criterion* if the following two properties hold:

(A) There exists a vectorial function $\vec{\alpha}(\cdot)$ from the set of parameters $\vec{\Psi}$ to $\mathbb{R}^R$ that satisfies the equation

$$\vec{\zeta}(\vec{\Psi}) = \frac{\partial\vec{\zeta}(\vec{\Psi})}{\partial\vec{\Psi}^{\mathrm{T}}}\,\vec{\alpha}(\vec{\Psi}). \tag{10}$$

(B) There exists an (S-dim) subset $\vec{\theta}$ of $\vec{\Theta}$ and a vectorial function $\vec{\beta}(\cdot)$ from the space of parameters of the observable distribution $\vec{\Theta}$ to $\mathbb{R}^S$ that satisfies the equation

$$\vec{\eta}(\vec{z};\vec{\Theta}) \;=\; \frac{\partial \vec{\eta}(\vec{z};\vec{\Theta})}{\partial \vec{\theta}^{\mathrm{T}}}\,\vec{\beta}(\vec{\Theta}). \tag{11}$$

The subset $\vec{\theta}$ has to be non-empty ($S \geq 1$) in order for the Jacobian to be defined. Furthermore, notice that $\vec{\beta}(\vec{\Theta})$ may only depend on the parameters $\vec{\Theta}$ but not on the latent variable $\vec{z}$.

□

For mixture models, $\vec{z}$ is usually replaced by c where $c \in \{1, \ldots, C\}$.

Part A of the parameterization criterion is usually trivially fulfilled but is required for completeness. For instance, the usual mappings from standard parameters to natural parameters of standard exponential family distributions (Bernoulli, Gaussain, Beta, Gamma etc.) can be shown to fulfill part A of the criterion. Part B is somewhat more difficult because of the additional dependence of the natural parameters $\vec{\eta}$ on the latent variable $\vec{z}$.

Below, let us first discuss some examples of generative models. The examples aim at illustrating the parameterization criterion.

Example 1. (Simple SBN)

Let us consider a first example, which will be a simple form of a sigmoid belief net (SBN; Saul et al., 1996).

$$z \;\sim\; p_{\vec{\Psi}}(z) = \mathrm{Bern}(z;\,\pi),\;\text{ with } 0 < \pi < 1, \tag{12}$$

$$\vec{x} \;\sim\; p_{\vec{\Theta}}(\vec{x}\,|\,z) = \mathrm{Bern}\big(x_1;\,\mathcal{S}(v\,z)\big)\;\mathrm{Bern}\big(x_2;\,\mathcal{S}(w\,z)\big),\;\text{ where } \mathcal{S}(a) = \frac{1}{1+e^{-a}}\;, \tag{13}$$

and where $\mathrm{Bern}(z;\,\pi) = \pi^z(1-\pi)^{1-z}$ is the standard parameterization of the Bernoulli distribution (π is the probability of z being one, $p(z=1)=\pi$). The same parameterization is used for the two Bernoulli distributions of the noise model (with the sigmoidal function setting the probability). The prior parameters $\vec{\Psi}$ just consists of the parameter $\pi \in (0,1)$. The noise model parameters consists of $v \in \mathbb{R}$ and $w \in \mathbb{R}$.

The generative model (12) and (13) is an EF generative model because it can be reparameterized as follows:

$$z \sim p_{\zeta(\pi)}(z) \qquad \text{where }\; \zeta(\pi) = \log\big(\frac{1}{1-\pi}\big) \tag{14}$$

$$\vec{x} \sim p_{\vec{\eta}(z;\,\vec{\Theta})}(\vec{x}) \qquad \text{where } \vec{\Theta} = \begin{pmatrix} v \\ w \end{pmatrix} \text{ and }\; \vec{\eta}(z;\,\vec{\Theta}) = \begin{pmatrix} v\,z \\ w\,z \end{pmatrix}. \tag{15}$$

$p_\zeta(z)$ is now the Bernoulli distribution in its exponential family form with natural parameters ζ. Likewise, $p_{\vec{\eta}}(\vec{x})$ is the Bernoulli distribution (for two observables) with natural parameters $\vec{\eta}$. The standard sigmoidal function (13) is the reason for the function to natural parameters, $\vec{\eta}(z;\vec{\Theta})$, to take on a particularly concise form.

For the generative model the parameterization criterion of Definition C is fulfilled: From the parameterization as EF model, it follows for part A that the Jacobian $\frac{\partial\vec{\zeta}(\vec{\Psi})}{\partial\vec{\Psi}^{\mathrm{T}}}$ of (8) is the scalar $\frac{\partial}{\partial\pi}\zeta(\pi) = \frac{1}{1-\pi} > 0$. Therefore, equation (10) is satisfied by $\alpha(\pi) = (1-\pi)\zeta(\pi)$. For the part B of the criterion, we use all parameters $\vec{\Theta}$ to construct the Jacobian, i.e. $\vec{\theta} = \vec{\Theta} = (v, w)^{\mathrm{T}}$. The Jacobian is then a (2×2)-matrix given by

$$\frac{\partial\vec{\eta}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}^{\mathrm{T}}} = \begin{pmatrix} z & 0 \\ 0 & z \end{pmatrix}. \tag{16}$$

Hence, the vectorial function $\vec{\beta}(\vec{\Theta}) = \begin{pmatrix} v \\ w \end{pmatrix}$ fulfills equation (11), which shows part B of the criterion.

□

Example 2. (Simple Factor Analysis)

Let us consider as second example a simple form of factor analysis (FA; Everitt, 1984; Bishop, 2006).

$$z \sim \quad p_{\vec{\Psi}}(z) = \mathcal{N}\big(z;\, 0, \tilde{\tau}\big), \quad \text{where } \tilde{\tau} > 0\,, \tag{17}$$

$$\vec{x} \sim \quad p_{\vec{\Theta}}(\vec{x}\,|\,z) = \mathcal{N}\big(\vec{x};\, \vec{w}z, \Sigma\big), \quad \text{where } \Sigma = \begin{pmatrix} \tilde{\sigma}_1 & 0 \\ 0 & \tilde{\sigma}_2 \end{pmatrix}, \tilde{\sigma}_1, \tilde{\sigma}_2 > 0\,, \tag{18}$$

$$\text{and where } \vec{w} \in \mathbb{R}^2 \text{ with } ||\vec{w}|| = 1\,.$$

The parameter vectors are $\vec{\Psi} = \Psi = \tilde{\tau}$ and $\vec{\Theta} = \begin{pmatrix} \tilde{\sigma}_1 \\ \tilde{\sigma}_2 \\ w_1 \\ w_2 \end{pmatrix}$, for which we again abbreviated the variances τ^2 by $\tilde{\tau}$ and σ_d^2 by $\tilde{\sigma}_d$.

Although we use standard parameters (mean and variance) for the Gaussian distributions (and not natural parameters), the parameterization of the whole model is still somewhat unusual for factor analysis because the prior also has a parameter. However, the vector $\vec{w}$ is of unit (here Euclidean) norm such that the model parameterizes the same set of distributions as a more conventionally defined FA model. The reasons for the parameterization will become clearer later on.

The model (17) and (18) can be recognized as an EF generative model (Definition A) using the following natural parameters:

$$\vec{\zeta}(\tilde{\tau}) = \begin{pmatrix} 0 \\ -\frac{1}{2\tilde{\tau}} \end{pmatrix} \quad \text{and} \quad \vec{\eta}(z;\vec{\Theta}) = \begin{pmatrix} \frac{w_1 z}{\tilde{\sigma}_1} \\ \frac{w_2 z}{\tilde{\sigma}_2} \\ -\frac{1}{2\tilde{\sigma}_1} \\ -\frac{1}{2\tilde{\sigma}_2} \end{pmatrix}. \tag{19}$$

By computing the Jacobian of the prior natural parameter mapping

$$\frac{\partial\vec{\zeta}(\vec{\Psi})}{\partial\vec{\Psi}^{\mathrm{T}}} = \frac{\partial\vec{\zeta}(\tilde{\tau})}{\partial\tilde{\tau}} = \begin{pmatrix} 0 \\ \frac{1}{2\tilde{\tau}^2} \end{pmatrix} \tag{20}$$

and choosing $\alpha(\tilde{\tau}) = -\tilde{\tau}$, we can see directly that part A of the parameterization criterion is fulfilled. For part B of the criterion, we choose as subset of $\vec{\Theta}$ the vector $\vec{\theta} = \begin{pmatrix} \tilde{\sigma}_1 \\ \tilde{\sigma}_2 \end{pmatrix}$, i.e. $\frac{\partial}{\partial\vec{\theta}^{\mathrm{T}}} = \left(\frac{\partial}{\partial\tilde{\sigma}_1}, \frac{\partial}{\partial\tilde{\sigma}_2}\right)$. The Jacobian is then the (4×2)-matrix

$$\frac{\partial\vec{\eta}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}^{\mathrm{T}}} = \begin{pmatrix} -\frac{w_1 z}{\tilde{\sigma}_1^2} & 0 \\ 0 & -\frac{w_2 z}{\tilde{\sigma}_2^2} \\ \frac{1}{2\tilde{\sigma}_1^2} & 0 \\ 0 & -\frac{w_2 z}{\tilde{\sigma}_2^2} \end{pmatrix} \tag{21}$$

and $\vec{\beta}(\vec{\Theta}) = \begin{pmatrix} -\tilde{\sigma}_1 \\ -\tilde{\sigma}_2 \end{pmatrix}$ satisfies equation (11) such that the parameterization criterion is fulfilled.
□

Example 3. (Counter-Example: Rigid SBN)

As an example of a model for which it can not be shown that the parameterization criterion is fulfilled, we go back to the simple SBN model (Example 1). However, the natural parameters of the noise model now use only one parameter $v \in \mathbb{R}$ as follows:

$$z \sim \quad p_{\vec{\Psi}}(z) = \mathrm{Bern}(z;\, \pi) \tag{22}$$

$$\vec{x} \sim \quad p_{\vec{\Theta}}(\vec{x}\,|\,z) = \mathrm{Bern}\big(x_1;\, \mathcal{S}(v\,z)\big)\ \mathrm{Bern}\big(x_2;\, \mathcal{S}((v+1)\,z)\big),\ \text{where } \mathcal{S}(a) = \frac{1}{1+e^{-a}}\ , \tag{23}$$

so $\vec{\Theta} = v$ (now a scalar) and $\vec{\eta}(z;\, \vec{\Theta}) = \begin{pmatrix} v\,z \\ (v+1)\,z \end{pmatrix}$.

Part A of the parameterization criterion remains satisfied. For part B of the criterion, $\vec{\theta} = \vec{\Theta} = v$ is the only non-empty subset of $\vec{\Theta}$. The corresponding Jacobian is the (2×1)-matrix

$$\frac{\partial\vec{\eta}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}^{\mathrm{T}}} = \begin{pmatrix} z \\ z \end{pmatrix}. \tag{24}$$

Consequently, there is no function $\beta(v)$ satisfying equation (11) in Part B of the parameterization criterionn since $v \neq 1 + v$ for all $x \in \mathbb{R}$.
□

The counter-example shows that the parameterization criterion is not necessarily fulfilled for a generative model. Other counter-examples would be simple FA (Example 2) for which one sigma is fixed (e.g. $\tilde{\sigma}_2$). Also generative models with the Poisson distribution would be a counter-example but for another reason (because the Poisson distribution does not have a constant base measure). In general, however, one can observe the parameterization criterion to be satisfied for many (maybe most) conventional generative models. We will later discuss concrete models in more detail. But first we show that equality to entropy sums applies for all EF generative models that satisfy the parameterization criterion.

3 Equality to Entropy Sums at Stationary Points

Let us consider the optimization of a generative model of Definition A given a set of N data points $\vec{x}^{(n)}$. When considering the ELBO objective (2) we, as usual, assume the variational parameters Φ to represent a set of parameters different from $\vec{\Psi}$ and $\vec{\Theta}$. The variational distributions $q_{\Phi}^{(n)}$ are usually defined to approximate posterior distributions of the generative model well but here we will not require any conditions on the distributions (and we will not use a vector notation for Φ). Given distributions $q_{\Phi}^{(n)}(\vec{z})$, the ELBO is given by:

$$\begin{aligned}\mathcal{F}(\Phi,\vec{\Psi},\vec{\Theta}) &= \tfrac{1}{N}\sum_n \int q_{\Phi}^{(n)}(\vec{z})\log\big(p_{\vec{\Theta}}(\vec{x}^{(n)}\,|\,\vec{z})\,p_{\vec{\Psi}}(\vec{z})\big)\,\mathrm{d}\vec{z} - \tfrac{1}{N}\sum_n \int q_{\Phi}^{(n)}(\vec{z})\log\big(q_{\Phi}^{(n)}(\vec{z})\big)\mathrm{d}\vec{z}\\ &= \tfrac{1}{N}\sum_n \int q_{\Phi}^{(n)}(\vec{z})\log\big(p_{\vec{\Theta}}(\vec{x}^{(n)}\,|\,\vec{z})\big)\,\mathrm{d}\vec{z} - \tfrac{1}{N}\sum_n D_{\mathrm{KL}}\big[q_{\Phi}^{(n)}(\vec{z}), p_{\vec{\Psi}}(\vec{z})\big]\,, \qquad (25)\end{aligned}$$

where the first and second line represent the most common standard forms. The only unusual part of our notation is the use of two separate symbols for prior and noise model parameters ($\vec{\Psi}$ and $\vec{\Theta}$) and a notation that denotes them as vectors. For our purposes, and for generative models as defined in Definition A, we first decompose the ELBO into three summands as follows:

$$\mathcal{F}(\Phi,\vec{\Psi},\vec{\Theta}) = \mathcal{F}_1(\Phi) - \mathcal{F}_2(\Phi,\vec{\Psi}) - \mathcal{F}_3(\Phi,\vec{\Theta}), \quad \text{where} \qquad (26)$$

$$\mathcal{F}_1(\Phi) \;= -\tfrac{1}{N}\sum_n \int q_{\Phi}^{(n)}(\vec{z})\log\big(q_{\Phi}^{(n)}(\vec{z})\big)\mathrm{d}\vec{z} \qquad (27)$$

$$\mathcal{F}_2(\Phi,\vec{\Psi}) \;= -\tfrac{1}{N}\sum_n \int q_{\Phi}^{(n)}(\vec{z})\log\big(p_{\vec{\Psi}}(\vec{z})\big)\,\mathrm{d}\vec{z} \qquad (28)$$

$$\mathcal{F}_3(\Phi,\vec{\Theta}) \;= -\tfrac{1}{N}\sum_n \int q_{\Phi}^{(n)}(\vec{z})\log\big(p_{\vec{\Theta}}(\vec{x}^{(n)}\,|\,\vec{z})\big)\mathrm{d}\vec{z} \qquad (29)$$

To finalize our notational preliminaries, we will use the *aggregated posterior* (e.g. Makhzani et al., 2015; Tomczak and Welling, 2018; Aneja et al., 2021) as an abbreviation, i.e., we define (noting that the entity is an *approximation* to the posterior average):

$$\overline{q}_{\Phi}(\vec{z}) \;=\; \frac{1}{N}\sum_{n=1}^{N} q_{\Phi}^{(n)}(\vec{z})\,. \qquad (30)$$

For the proof of our main result, we will require the parameterization criterion given in Definition C. As a preparation, we will show a property that applies for the (transposed) Jacobians $\frac{\partial\vec{\zeta}^{\mathrm{T}}(\vec{\Psi})}{\partial\vec{\Psi}}$ and $\frac{\partial\vec{\eta}^{\mathrm{T}}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}}$ if the parameterization criterion in Definition C is fulfilled.

Lemma 1.

Consider an EF generative model as given in Definition B and let the dimensionalities of the natural parameter vectors $\vec{\zeta}$ and $\vec{\eta}$ be K and L, respectively. Following our notation of the Jacobians in Definition C, we denote the transposed Jacobian of the mapping $\vec{\zeta}(\vec{\Psi})$ by

$$\frac{\partial \vec{\zeta}^{\mathrm{T}}(\vec{\Psi})}{\partial \vec{\Psi}} := \Big(\frac{\partial}{\partial \vec{\Psi}} \zeta_1(\vec{\Psi}), \ldots, \frac{\partial}{\partial \vec{\Psi}} \zeta_K(\vec{\Psi}) \Big) \tag{31}$$

and the transposed Jacobian of the mapping $\vec{\eta}(\vec{z};\vec{\Theta})$ w.r.t. a subset $\vec{\theta}$ of $\vec{\Theta}$ by

$$\frac{\partial \vec{\eta}^{\mathrm{T}}(\vec{z};\vec{\Theta})}{\partial \vec{\theta}} := \Big(\frac{\partial}{\partial \vec{\theta}} \eta_1(\vec{z};\vec{\Theta}), \ldots, \frac{\partial}{\partial \vec{\theta}} \eta_L(\vec{z};\vec{\Theta}) \Big). \tag{32}$$

Then, under the assumption that the parameterization criterion (Definition C) is fulfilled, the following two statements also apply:

(A) For any vectorial function $\vec{f}(\Phi, \vec{\Psi})$ from the sets of parameters Φ and $\vec{\Psi}$ to the (K-dim) space of natural parameters of the prior it holds:

$$\frac{\partial \vec{\zeta}^{\mathrm{T}}(\vec{\Psi})}{\partial \vec{\Psi}} \vec{f}(\Phi, \vec{\Psi}) \;=\; \vec{0} \;\;\Rightarrow\;\; \vec{\zeta}^{\mathrm{T}}(\vec{\Psi}) \vec{f}(\Phi, \vec{\Psi}) \;=\; 0. \tag{33}$$

(B) For any variational parameter Φ there is a non-empty subset $\vec{\theta}$ of $\vec{\Theta}$ such that the following applies for all vectorial functions $\vec{g}_1(\vec{z};\vec{\Theta}), \ldots, \vec{g}_N(\vec{z};\vec{\Theta})$ from the parameter set $\vec{\Theta}$ and from the latent space $\Omega_{\vec{z}}$ to the (L-dim) space of natural parameters of the noise model:

$$\sum_{n=1}^{N} \int q_{\Phi}^{(n)}(\vec{z}) \frac{\partial \vec{\eta}^{\mathrm{T}}(z;\vec{\Theta})}{\partial \vec{\theta}} \vec{g}_n(\vec{z};\vec{\Theta}) \,\mathrm{d}\vec{z} = \vec{0} \tag{34}$$

$$\Rightarrow \;\; \sum_{n=1}^{N} \int q_{\Phi}^{(n)}(\vec{z}) \vec{\eta}^{\mathrm{T}}(\vec{z};\vec{\Theta}) \vec{g}_n(\vec{z};\vec{\Theta}) \,\mathrm{d}\vec{z} = 0. \tag{35}$$

In the case when $\vec{z}$ is a discrete latent variable, the integrals in (34) and (35) become a sum over all discrete states of $\vec{z}$.

Proof

To verify the first part of the lemma, we choose a function $\vec{\alpha}(\vec{\Psi})$ satisfying Eqn. 10. This is possible since we assume that the parameterization criterion (Definition C) is fulfilled. Moreover consider an arbitrary vectorial function $\vec{f}(\Phi, \vec{\Psi})$ that satiesfies

$$\frac{\partial \vec{\zeta}^{\mathrm{T}}(\vec{\Psi})}{\partial \vec{\Psi}} \vec{f}(\Phi, \vec{\Psi}) = \vec{0}. \tag{36}$$

Then we can conclude:

$$\vec{\zeta}^{\mathrm{T}}(\vec{\Psi}) \vec{f}(\Phi, \vec{\Psi}) = \vec{\alpha}^{\mathrm{T}}(\vec{\Psi}) \frac{\partial \vec{\zeta}^{\mathrm{T}}(\vec{\Psi})}{\partial \vec{\Psi}} \vec{f}(\Phi, \vec{\Psi}) = 0. \tag{37}$$

Consequently, the first part of the lemma is verified. For the proof of the second part, let Φ be any variational parameter. We consider a subset $\vec{\theta}$ of $\vec{\Theta}$ and a function $\vec{\beta}(\vec{\Theta})$ satisfying Eqn. 11. Then all vectorial functions $\vec{g}_1(\vec{z};\vec{\Theta}),\ldots,\vec{g}_N(\vec{z};\vec{\Theta})$ fulfilling the equation

$$\sum_{n=1}^{N}\int q_{\Phi}^{(n)}(\vec{z})\frac{\partial\vec{\eta}^{\mathrm{T}}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}}\vec{g}_n(\vec{z};\vec{\Theta})\,\mathrm{d}\vec{z}=\vec{0} \tag{38}$$

also fulfill the equation

$$\sum_{n=1}^{N}\int q_{\Phi}^{(n)}(\vec{z})\vec{\eta}^{\mathrm{T}}(\vec{z};\vec{\Theta})\vec{g}_n(\vec{z};\vec{\Theta})\,\mathrm{d}\vec{z}=\vec{\beta}^{\mathrm{T}}(\vec{\Theta})\sum_{n=1}^{N}\int q_{\Phi}^{(n)}(\vec{z})\frac{\partial\vec{\eta}^{\mathrm{T}}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}}\vec{g}_n(\vec{z};\vec{\Theta})\,\mathrm{d}\vec{z}=0, \tag{39}$$

which proofs the second part of the lemma. Furthermore, it is easy to see that the proof also works in the discrete case if the integrals are replace by sums.

□

Because we will only use Lemma 1 and not directly the parameterization criterion (Definition C) to show convergence of the ELBO tho entropy sums, the question may arise whether the two statements in Lemma 1 are actually equivalent to the parameterization criterion given in Definition C. If there is no equivalence, the parameterization criterion is only sufficient for the statements of Lemma 1 but not necessary. In this case, Lemma 1 would describe a larger set of models for which we can show convergence of the ELBO to entropy sums than the parameterization criterion (Definition C), and we would lose some models by using the parameterization criterion. However, given mild assumptions, the property in Lemma 1 can indeed be shown be equivalent to the parameterization criterion (Definition C). We will provide the proof in the appendix. Therefore, Definition C does, in practice, not represent a condition more restrictive than the more intricate properties represented by (33), (34) and (35).

We can now proceed by stating and proving our main result.

Theorem 1. (Equality to Entropy Sums)

If $p_{\vec{\Psi}}(\vec{z})$ and $p_{\vec{\Theta}}(\vec{x}\,|\,\vec{z})$ is an exponential family generative model (Definition B) that fulfills the parameterization criterion of Definition C then at all stationary points of the ELBO (25) it applies that

$$\mathcal{F}(\Phi,\vec{\Psi},\vec{\Theta})\;=\;\frac{1}{N}\sum_{n=1}^{N}\mathcal{H}[q_{\Phi}^{(n)}(\vec{z})]\;\;-\;\mathcal{H}[\,p_{\vec{\Psi}}(\vec{z})]\;\;-\;\mathbb{E}_{\,\bar{q}_{\Phi}}\big\{\mathcal{H}[\,p_{\vec{\Theta}}(\vec{x}\,|\,\vec{z})]\big\}\,. \tag{40}$$

Proof

The first summand of the ELBO in Eqn. 26 is already given in the form of an (average) entropy ($\mathcal{F}_1(\Phi)=\frac{1}{N}\sum_n\mathcal{H}[q_{\Phi}^{(n)}(\vec{z})]$). To prove the claim of Theorem 1, we will at stationary points show equality to entropies for the second and the third summand of Eqn. 26, separately.

$\mathcal{F}_2(\Phi, \vec{\Psi})$ at stationary points

We have assumed that the generative model is an EF generative model. Therefore we can rewrite $\mathcal{F}_2(\Phi, \vec{\Psi})$ of Eqn. 28 using the reparameterization of $p_{\vec{\Psi}}(\vec{z})$ in terms of the exponential family distribution $p_{\vec{\zeta}(\vec{\Psi})}(\vec{z})$:

$$\begin{aligned} \mathcal{F}_2(\Phi, \vec{\Psi}) &= -\tfrac{1}{N} \textstyle\sum_n \int q_\Phi^{(n)}(\vec{z}) \log\big(p_{\vec{\Psi}}(\vec{z})\big)\, \mathrm{d}\vec{z} \\ &= \tfrac{1}{N} \textstyle\sum_n \int q_\Phi^{(n)}(\vec{z}) \Big(-\log\big(p_{\vec{\zeta}(\vec{\Psi})}(\vec{z})\big)\Big)\, \mathrm{d}\vec{z}\,. \end{aligned} \tag{41}$$

As $p_{\vec{\zeta}}(\vec{z})$ is an exponential family distribution, its negative logarithm can be written as:

$$-\log\big(p_{\vec{\zeta}(\vec{\Psi})}(\vec{z})\big) = A\big(\vec{\zeta}(\vec{\Psi})\big) - \vec{\zeta}^{\mathrm{T}}_{(\vec{\Psi})} \vec{T}(\vec{z}) - \log\big(h(\vec{z})\big)\,, \tag{42}$$

where A is the partition function, $\vec{T}$ the sufficient statistics and h is the base measure. We will later require the derivative of the (negative) log probability w.r.t. $\vec{\Psi}$. The derivative is given by:

$$\frac{\partial}{\partial\vec{\Psi}}\Big(-\log\big(p_{\vec{\zeta}(\vec{\Psi})}(\vec{z})\big)\Big) = \frac{\partial\vec{\zeta}^{\mathrm{T}}(\vec{\Psi})}{\partial\vec{\Psi}} \Big(\vec{A}'\big(\vec{\zeta}(\vec{\Psi})\big) - \vec{T}(\vec{z})\Big) \tag{43}$$

$$\text{where}\quad \vec{A}'(\vec{\zeta}(\vec{\Psi})) := \frac{\partial}{\partial\vec{\zeta}} A(\vec{\zeta})\Big|_{\vec{\zeta}=\vec{\zeta}(\vec{\Psi})} \tag{44}$$

and where $\frac{\partial\vec{\zeta}^{\mathrm{T}}(\vec{\Psi})}{\partial\vec{\Psi}}$ is the transposed Jacobian matrix defined in (31).

We will show that the summand $\mathcal{F}_2(\Phi, \vec{\Psi})$ will at stationary points be equal to the prior entropy. Therefore, we first rewrite $-\log\big(p_{\vec{\zeta}(\vec{\Psi})}(\vec{z})\big)$ using the prior entropy. By using the definition of the entropy

$$\mathcal{H}[p_{\vec{\zeta}(\vec{\Psi})}(\vec{z})] = \int p_{\vec{\zeta}(\vec{\Psi})}(\vec{z})\Big(-\log\big(p_{\vec{\zeta}(\vec{\Psi})}(\vec{z})\big)\Big)\mathrm{d}\vec{z} \tag{45}$$

and (42) we obtain:

$$\mathcal{H}[p_{\vec{\Psi}}(\vec{z})] = \mathcal{H}[p_{\vec{\zeta}(\vec{\Psi})}(\vec{z})] = A\big(\vec{\zeta}(\vec{\Psi})\big) - \vec{\zeta}^{\mathrm{T}}_{(\vec{\Psi})}\vec{A}'\big(\vec{\zeta}(\vec{\Psi})\big) - B(\vec{\zeta}(\vec{\Psi})) \tag{46}$$

$$\Rightarrow\ A\big(\vec{\zeta}(\vec{\Psi})\big) = \mathcal{H}[p_{\vec{\zeta}(\vec{\Psi})}(\vec{z})] + \vec{\zeta}^{\mathrm{T}}_{(\vec{\Psi})}\vec{A}'\big(\vec{\zeta}(\vec{\Psi})\big) + B(\vec{\zeta}(\vec{\Psi})) \tag{47}$$

$$\text{where}\quad B(\vec{\zeta}) := \int p_{\vec{\zeta}}(\vec{z}) \log\big(h(\vec{z})\big)\,\mathrm{d}\vec{z}\,. \tag{48}$$

By combining (42) and (47), the negative logarithm of the prior probability becomes:

$$\begin{aligned} -\log\big(p_{\vec{\zeta}(\vec{\Psi})}(\vec{z})\big) &= \mathcal{H}[p_{\vec{\zeta}(\vec{\Psi})}(\vec{z})] + \vec{\zeta}^{\mathrm{T}}_{(\vec{\Psi})}\big(\vec{A}'(\vec{\zeta}(\vec{\Psi})) - \vec{T}(\vec{z})\big) + B(\vec{\zeta}(\vec{\Psi})) - \log\big(h(\vec{z})\big) \\ &= \mathcal{H}[p_{\vec{\zeta}(\vec{\Psi})}(\vec{z})] + \vec{\zeta}^{\mathrm{T}}_{(\vec{\Psi})}\big(\vec{A}'(\vec{\zeta}(\vec{\Psi})) - \vec{T}(\vec{z})\big) \end{aligned} \tag{49}$$

where $\vec{A}'(\vec{\zeta})$ is defined as in Eqn. 44, and where the last two terms have cancelled because we assumed a constant base measure $h(\vec{z})$. Inserting the negative logarithm (49) back into (41), we therefore obtain for the second summand of the ELBO:

$$\begin{aligned}
\mathcal{F}_2(\Phi,\vec{\Psi}) &= \frac{1}{N}\sum_n \int q_\Phi^{(n)}(\vec{z})\Big(\mathcal{H}[p_{\vec{\zeta}(\vec{\Psi})}(\vec{z})] + \vec{\zeta}^{\mathrm{T}}_{(\vec{\Psi})}\big(\vec{A}'(\vec{\zeta}_{(\vec{\Psi})}) - \vec{T}(\vec{z})\big)\Big)\,\mathrm{d}\vec{z} \\
&= \mathcal{H}[p_{\vec{\zeta}(\vec{\Psi})}(\vec{z})] + \vec{\zeta}^{\mathrm{T}}_{(\vec{\Psi})}\frac{1}{N}\sum_n \int q_\Phi^{(n)}(\vec{z})\big(\vec{A}'(\vec{\zeta}_{(\vec{\Psi})}) - \vec{T}(\vec{z})\big)\mathrm{d}\vec{z} \\
&= \mathcal{H}[p_{\vec{\zeta}(\vec{\Psi})}(\vec{z})] + \vec{\zeta}^{\mathrm{T}}_{(\vec{\Psi})}\big(\vec{A}'(\vec{\zeta}_{(\vec{\Psi})}) - \mathbb{E}_{\bar{q}_\Phi}\{\vec{T}(\vec{z})\}\big)\,, \quad \text{where} \;\; \bar{q}_\Phi(\vec{z}) := \frac{1}{N}\sum_n q_\Phi^{(n)}(\vec{z})\,, \quad (50) \\
&= \mathcal{H}[p_{\vec{\zeta}(\vec{\Psi})}(\vec{z})] + \vec{\zeta}^{\mathrm{T}}_{(\vec{\Psi})}\;\vec{f}(\Phi,\vec{\Psi})\,, \quad (51)
\end{aligned}$$

$$\text{where} \;\; \vec{f}(\Phi,\vec{\Psi}) := \vec{A}'(\vec{\zeta}(\vec{\Psi})) - \mathbb{E}_{\bar{q}_\Phi}\{\vec{T}(\vec{z})\}\,. \quad (52)$$

Our aim will now be to show that the second term of (51) vanishes at stationary points of $\mathcal{F}(\Phi,\vec{\Psi},\vec{\Theta})$. For all stationary points of $\mathcal{F}(\Phi,\vec{\Psi},\vec{\Theta})$ applies that the derivatives w.r.t. all parameters vanish. We have assumed the parameters Φ, $\vec{\Psi}$ and $\vec{\Theta}$ to be separate sets of parameters. $\mathcal{F}_2(\Phi,\vec{\Psi})$ is, therefore, the only summand of $\mathcal{F}(\Phi,\vec{\Psi},\vec{\Theta})$ which depends on $\vec{\Psi}$. By using (41) and (43) we obtain at stationary points:

$$\begin{aligned}
\vec{0} = \frac{\partial}{\partial\vec{\Psi}}\mathcal{F}_2(\Phi,\vec{\Psi}) = &\quad \frac{1}{N}\sum_n \int q_\Phi^{(n)}(\vec{z})\;\frac{\partial}{\partial\vec{\Psi}}\Big(-\log\big(p_{\vec{\zeta}(\vec{\Psi})}(\vec{z})\big)\Big)\;\mathrm{d}\vec{z} \quad (53) \\
= &\quad \frac{1}{N}\sum_n \int q_\Phi^{(n)}(\vec{z})\;\frac{\partial\vec{\zeta}^{\mathrm{T}}(\vec{\Psi})}{\partial\vec{\Psi}}\Big(\big(\vec{A}'(\vec{\zeta}(\vec{\Psi})) - \vec{T}(\vec{z})\big)\Big)\,\mathrm{d}\vec{z} \quad (54) \\
= &\quad \frac{\partial\vec{\zeta}^{\mathrm{T}}(\vec{\Psi})}{\partial\vec{\Psi}}\;\big(\vec{A}'(\vec{\zeta}_{(\vec{\Psi})}) - \mathbb{E}_{\bar{q}_\Phi}\{\vec{T}(\vec{z})\}\big)\,, \quad (55)
\end{aligned}$$

where $\bar{q}_\Phi$ is defined in (50). Hence, at all stationary points (w.r.t. parameters $\vec{\Psi}$) the following concise condition holds:

$$\frac{\partial\vec{\zeta}^{\mathrm{T}}(\vec{\Psi})}{\partial\vec{\Psi}}\;\vec{f}(\Phi,\vec{\Psi}) = \vec{0}\,, \quad (56)$$

where $\vec{f}(\Phi,\vec{\Psi})$ is the same function as introduced in (52). We can now apply Lemma 1 as we assumed our generative model to fulfill the criterion given in Definition C. Part A of the lemma is formulated for arbitrary functions $\vec{f}$ with arbitrary parameters Φ and arbitrary values for $\vec{\Theta}$. So if the criterion is satisfied, (33) also applies for the $\vec{f}$ as defined in (52) with Φ being the variational parameters. Using Lemma 1 we thus conclude from (56) that at stationary points applies:

$$\vec{\zeta}^{\mathrm{T}}_{(\vec{\Psi})}\;\vec{f}(\Phi,\vec{\Psi}) = 0\,. \quad (57)$$

Using (51) we consequently obtain at all stationary points for the summand $\mathcal{F}_2(\Phi,\vec{\Psi})$:

$$\mathcal{F}_2(\Phi,\vec{\Psi}) = \mathcal{H}[p_{\vec{\zeta}(\vec{\Psi})}(\vec{z})] + \vec{\zeta}^{\mathrm{T}}_{(\vec{\Psi})}\;\vec{f}(\Phi,\vec{\Psi}) = \mathcal{H}[p_{\vec{\zeta}(\vec{\Psi})}(\vec{z})] = \mathcal{H}[p_{\vec{\Psi}}(\vec{z})]\,. \quad (58)$$

$\mathcal{F}_3(\Phi,\vec{\Theta})$ at stationary points

The proof part for $\mathcal{F}_3(\Phi,\vec{\Theta})$ will be analog to that of $\mathcal{F}_2(\Phi,\vec{\Psi})$ above but slightly more intricate because the noise distribution is conditional on $\vec{z}$. We have assumed that the generative model is an EF generative model. Therefore we can rewrite $\mathcal{F}_3(\Phi,\vec{\Theta})$ of Eqn. 29 using the reparameterization of $p_{\vec{\Theta}}(\vec{x}\,|\,\vec{z})$ in terms of the exponential family distribution $p_{\vec{\eta}(\vec{z};\,\vec{\Theta})}(\vec{x})$:

$$\begin{aligned}\mathcal{F}_3(\Phi,\vec{\Theta}) &= -\tfrac{1}{N}\sum_n \int q_\Phi^{(n)}(\vec{z})\log\big(p_{\vec{\Theta}}(\vec{x}^{(n)}\,|\,\vec{z})\big)\,\mathrm{d}\vec{z}\\ &= \tfrac{1}{N}\sum_n \int q_\Phi^{(n)}(\vec{z})\Big(-\log\big(p_{\vec{\eta}(\vec{z};\,\vec{\Theta})}(\vec{x}^{(n)})\big)\Big)\,\mathrm{d}\vec{z}\,. \end{aligned}\tag{59}$$

As $p_{\vec{\eta}}(\vec{x})$ is an exponential family distribution, its negative logarithm is given by:

$$-\log\big(p_{\vec{\eta}(\vec{z};\,\vec{\Theta})}(\vec{x})\big) = A\big(\vec{\eta}(\vec{z};\,\vec{\Theta})\big) - \vec{\eta}^{\mathrm{T}}_{(\vec{z};\,\vec{\Theta})}\vec{T}(\vec{x}) - \log\big(h(\vec{x})\big)\,. \tag{60}$$

Note that we will use the same symbols for sufficient statistics $\vec{T}(\vec{x})$, partition function $A(\vec{\eta})$ and base measure $h(\vec{x})$ as for the prior in Eqn. 42. As prior and noise model distribution are (in general) different members of the exponential family, these entities (in general) differ, of course. But as they can be distinguished from context, we avoid the introduction of further symbols.

We will later require the derivative of the (negative) log probability w.r.t. $\vec{\theta}$, i.e., w.r.t. a subset of the parameters $\vec{\Theta}$. The derivative is given by:

$$\frac{\partial}{\partial\vec{\theta}}\Big(-\log\big(p_{\vec{\eta}(\vec{z};\,\vec{\Theta})}(\vec{x})\big)\Big) = \frac{\partial\vec{\eta}^{\mathrm{T}}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}}\Big(\big(\vec{A}'(\vec{\eta}(\vec{z};\vec{\Theta})) - \vec{T}(\vec{x})\big)\Big) \tag{61}$$

$$\text{where}\;\; \vec{A}'(\vec{\eta}(\vec{z};\vec{\Theta})\,) := \frac{\partial}{\partial\vec{\eta}}A(\vec{\eta}\,)\Big|_{\vec{\eta}=\vec{\eta}(\vec{z};\vec{\Theta})} \tag{62}$$

and where $\frac{\partial\vec{\eta}^{\mathrm{T}}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}}$ is the transposed Jacobian matrix defined in (32).

We will show that the summand $\mathcal{F}_3(\Phi,\vec{\Theta})$ will at stationary points be equal to the noise model entropy. Therefore, we first rewrite $-\log\big(p_{\vec{\eta}(\vec{z};\vec{\Theta})}(\vec{x})\big)$ using the noise model entropy. By again using the definition of the entropy

$$\mathcal{H}[\,p_{\vec{\eta}(\vec{z};\vec{\Theta})}(\vec{x})] = \int p_{\vec{\eta}(\vec{z};\vec{\Theta})}(\vec{x})\Big(-\log\big(p_{\vec{\eta}(\vec{z};\vec{\Theta})}(\vec{x})\big)\Big)\mathrm{d}\vec{x} \tag{63}$$

and (60) we obtain:

$$\mathcal{H}[\,p_{\vec{\Theta}}(\vec{x}\,|\,\vec{z})] = \mathcal{H}[\,p_{\vec{\eta}(\vec{z};\vec{\Theta})}(\vec{x})] = A\big(\vec{\eta}(\vec{z};\vec{\Theta})\big) - \vec{\eta}^{\mathrm{T}}_{(\vec{z};\,\vec{\Theta})}\vec{A}'\big(\vec{\eta}(\vec{z};\vec{\Theta})\big) - B(\vec{\eta}(\vec{z};\vec{\Theta})) \tag{64}$$

$$\Rightarrow\;\; A\big(\vec{\eta}(\vec{z};\vec{\Theta})\big) = \mathcal{H}[\,p_{\vec{\eta}(\vec{z};\vec{\Theta})}(\vec{x})] + \vec{\eta}^{\mathrm{T}}_{(\vec{z};\,\vec{\Theta})}\vec{A}'\big(\vec{\eta}(\vec{z};\vec{\Theta})\big) + B(\vec{\eta}(\vec{z};\vec{\Theta}))\,, \tag{65}$$

$$\text{where we abbreviated}\;\; B(\vec{\eta}\,) := \int p_{\vec{\eta}}(\vec{x})\,\log\big(h(\vec{x})\big)\,\mathrm{d}\vec{x}\,. \tag{66}$$

By combining (60) and (65), the negative logarithm of the noise distribution becomes:

$$\begin{aligned}-\log\big(p_{\vec{\eta}(\vec{z};\vec{\Theta})}(\vec{x})\big) &= \mathcal{H}[\,p_{\vec{\eta}(\vec{z};\vec{\Theta})}(\vec{x})] + \vec{\eta}^{\mathrm{T}}_{(\vec{z};\,\vec{\Theta})}\big(\vec{A}'(\vec{\eta}(\vec{z};\vec{\Theta})) - \vec{T}(\vec{x})\big) + B(\vec{\eta}(\vec{z};\vec{\Theta})) - \log\big(h(\vec{x})\big)\\ &= \mathcal{H}[\,p_{\vec{\eta}(\vec{z};\vec{\Theta})}(\vec{x})] + \vec{\eta}^{\mathrm{T}}_{(\vec{z};\,\vec{\Theta})}\big(\vec{A}'(\vec{\eta}(\vec{z};\vec{\Theta})) - \vec{T}(\vec{x})\big)\end{aligned}\tag{67}$$

where $\vec{A}'\big(\vec{\eta}(\vec{z};\,\vec{\Theta})\big)$ is defined as in Eqn. 62, and where the last two terms cancel again because we assumed a constant base measure $h(\vec{x})$ also for the noise model distribution. Inserting the negative logarithm (67) back into (59), we therefore obtain for the third summand of the ELBO:

$$\begin{aligned}
&\mathcal{F}_3(\Phi,\vec{\Theta}) && (68)\\
&= \frac{1}{N}\sum_n \int q_\Phi^{(n)}(\vec{z})\Big(\mathcal{H}[\,p_{\vec{\eta}(\vec{z};\vec{\Theta})}(\vec{x})] \,+\, \vec{\eta}^{\mathrm{T}}_{(\vec{z};\,\vec{\Theta})}\big(\vec{A}'(\vec{\eta}(\vec{z};\vec{\Theta}))\,-\,\vec{T}(\vec{x}^{(n)})\big)\Big)\,\mathrm{d}\vec{z}\\
&= \frac{1}{N}\sum_n \mathbb{E}_{q_\Phi^{(n)}}\big\{\mathcal{H}[\,p_{\vec{\eta}(\vec{z};\vec{\Theta})}(\vec{x})]\big\} \,+\, \frac{1}{N}\sum_n \int q_\Phi^{(n)}(\vec{z})\;\vec{\eta}^{\mathrm{T}}_{(\vec{z};\,\vec{\Theta})}\big(\vec{A}'(\vec{\eta}(\vec{z};\vec{\Theta}))\,-\,\vec{T}(\vec{x}^{(n)})\big)\mathrm{d}\vec{z}\\
&= \frac{1}{N}\sum_n \mathbb{E}_{q_\Phi^{(n)}}\big\{\mathcal{H}[\,p_{\vec{\eta}(\vec{z};\vec{\Theta})}(\vec{x})]\big\} \,+\, \frac{1}{N}\sum_n \int q_\Phi^{(n)}(\vec{z})\;\vec{\eta}^{\mathrm{T}}_{(\vec{z};\,\vec{\Theta})}\;\vec{g}_n(\vec{z};\vec{\Theta})\mathrm{d}\vec{z} && (69)\\
&\quad\text{where }\;\vec{g}_n(\vec{z};\vec{\Theta}) := \vec{A}'\big(\vec{\eta}(\vec{z};\,\vec{\Theta})\big)\,-\,\vec{T}(\vec{x}^{(n)})\,. && (70)
\end{aligned}$$

In analogy to the derivation for $\mathcal{F}_2(\Phi,\vec{\Psi})$, our aim will now be to show that the second term of (69) vanishes at stationary points of $\mathcal{F}(\Phi,\vec{\Psi},\vec{\Theta})$. For all stationary points of $\mathcal{F}(\Phi,\vec{\Psi},\vec{\Theta})$ applies that the derivatives w.r.t. all parameters vanish. We have assumed the parameters Φ, $\vec{\Psi}$ and $\vec{\Theta}$ to be separate sets of parameters. $\mathcal{F}_3(\Phi,\vec{\Theta})$ is therefore the only summand of $\mathcal{F}(\Phi,\vec{\Psi},\vec{\Theta})$ which depends on $\vec{\Theta}$. At stationary points we can consequently conclude for any subset $\vec{\theta}$ of $\vec{\Theta}$ that $\frac{\partial}{\partial\vec{\theta}}\mathcal{F}_3(\Phi,\vec{\Theta})=0$. By using (59) and (61) we obtain at any stationary point and for any subset $\vec{\theta}$:

$$\begin{aligned}
0 = \tfrac{\partial}{\partial\vec{\theta}}\mathcal{F}_3(\Phi,\vec{\Theta}) &= \frac{1}{N}\sum_n \int q_\Phi^{(n)}(\vec{z})\;\frac{\partial}{\partial\vec{\theta}}\Big(-\log\big(p_{\vec{\eta}(\vec{z};\vec{\Theta})}(\vec{x}^{(n)})\big)\Big)\;\mathrm{d}\vec{z} && (71)\\
&= \frac{1}{N}\sum_n \int q_\Phi^{(n)}(\vec{z})\;\frac{\partial\vec{\eta}^{\mathrm{T}}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}}\Big(\vec{A}'(\vec{\eta}(\vec{z};\vec{\Theta}))\,-\,\vec{T}(\vec{x}^{(n)})\Big)\;\mathrm{d}\vec{z}. && (72)
\end{aligned}$$

In expression (72) we recognize the functions $\vec{g}_n(\vec{z};\vec{\Theta})$ as defined in Eqn. 70. Therefore, expression (72) means that at all stationary points and for all subsets $\vec{\theta}$ of $\vec{\Theta}$ the following holds:

$$\frac{1}{N}\sum_n \int q_\Phi^{(n)}(\vec{z})\;\frac{\partial\vec{\eta}^{\mathrm{T}}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}}\;\vec{g}_n(\vec{z};\vec{\Theta})\,\mathrm{d}\vec{z} = 0\,. \qquad (73)$$

As (73) applies for all subsets $\vec{\theta}$, it also applies for the specific subset $\vec{\theta}$ of $\vec{\Theta}$ that exists according to the parameterization criterion (Definition C, part B) or Lemma 1 (part B). Using Part B of Lemma 1, we can consequently conclude from (73) that at all stationary points applies:

$$\frac{1}{N}\sum_n \int q_\Phi^{(n)}(\vec{z})\;\vec{\eta}^{\mathrm{T}}_{(\vec{z};\vec{\Theta})}\;\vec{g}_n(\vec{z};\vec{\Theta})\,\mathrm{d}\vec{z} \;=\; 0\,. \qquad (74)$$

By going back to (69) and by inserting (74), we consequently obtain at all stationary points for the summand $\mathcal{F}_3(\Phi,\vec{\Theta})$:

$$\begin{aligned}
\mathcal{F}_3(\Phi,\vec{\Theta}) &= \frac{1}{N}\sum_n \mathbb{E}_{q_\Phi^{(n)}}\big\{\mathcal{H}[\,p_{\vec{\eta}(\vec{z};\vec{\Theta})}(\vec{x})]\big\} \,+\, \frac{1}{N}\sum_n \int q_\Phi^{(n)}(\vec{z})\;\vec{\eta}^{\mathrm{T}}_{(\vec{z};\,\vec{\Theta})}\;\vec{g}_n(\vec{z};\vec{\Theta})\;\mathrm{d}\vec{z}\\
&= \frac{1}{N}\sum_n \mathbb{E}_{q_\Phi^{(n)}}\big\{\mathcal{H}[\,p_{\vec{\eta}(\vec{z};\vec{\Theta})}(\vec{x})]\big\} = \frac{1}{N}\sum_n \mathbb{E}_{q_\Phi^{(n)}}\big\{\mathcal{H}[\,p_{\vec{\Theta}}(\vec{x}\,|\,\vec{z})]\big\}\,. && (75)
\end{aligned}$$

Conclusion

We rewrote the ELBO to consist of three terms, $\mathcal{F}(\Phi,\vec{\Psi},\vec{\Theta}) = \mathcal{F}_1(\Phi) - \mathcal{F}_2(\Phi,\vec{\Psi}) - \mathcal{F}_3(\Phi,\vec{\Theta})$, with $\mathcal{F}_1(\Phi)$, $\mathcal{F}_2(\Phi,\vec{\Psi})$ and $\mathcal{F}_3(\Phi,\vec{\Theta})$ as defined by (27), (28) and (29), respectively. The first term is by definition an average entropy, $\mathcal{F}_1(\Phi) = \frac{1}{N}\sum_{n=1}^{N}\mathcal{H}[q_{\Phi}^{(n)}(\vec{z})]$. We have shown (see Eqn. 58) that $\mathcal{F}_2(\Phi,\vec{\Psi})$ becomes equal to the entropy of the prior distribution at stationary points. Furthermore, we have shown (see Eqn. 75) that $\mathcal{F}_3(\Phi,\vec{\Theta})$ becomes equal to an expected entropy at stationary points. By inserting the results (58) and (75), we have shown that at all stationary points the following applies for the ELBO (26):

$$\mathcal{F}(\Phi,\vec{\Psi},\vec{\Theta}) = \frac{1}{N}\sum_{n=1}^{N}\mathcal{H}[q_{\Phi}^{(n)}(\vec{z})] - \mathcal{H}[\,p_{\vec{\Psi}}(\vec{z})] - \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{q_{\Phi}^{(n)}}\big\{\mathcal{H}[\,p_{\vec{\Theta}}(\vec{x}\,|\,\vec{z})]\big\}, \tag{76}$$

from which (40) follows by using (30) for the last summand.
□

4 Sigmoid Belief Networks

Let us consider as the first concrete generative model the Sigmoid Belief Network (SBN; Neal, 1992; Hinton et al., 2006). SBNs are amongst the first and most well-known generative models with an extensive body of literature; SBNs are also often considered as prototypical examples for Bayes Nets. Here we study SBNs in with one set of latents and one set of observables in their standard form. Deep SBNs can be treated analogously (see Sec. 7). We will formulate the theorem in standard notation to be more accessible without requiring the special notation we introduced in Sec. 2, i.e., we use one set Θ for parameters of prior and noise model. The separate parameter vectors $\vec{\Psi}$ and $\vec{\Theta}$ will only be used for the proof.

Definition D. (Sigmoid Belief Nets)

The generative model of an SBN is defined as follows:

$$\vec{z} \;\sim\; p_{\Theta}(\vec{z}) = \prod_{h=1}^{H}\mathrm{Bern}(z_h;\,\pi_h),\;\text{ with } 0<\pi_h<1, \tag{77}$$

$$\vec{x} \;\sim\; p_{\Theta}(\vec{x}\,|\,\vec{z}) = \prod_{d=1}^{D}\mathrm{Bern}\Big(x_d;\,\mathcal{S}\big(\sum_{h=1}^{H}W_{dh}z_h+\mu_d\big)\Big),\;\text{ where } \mathcal{S}(a)=\frac{1}{1+e^{-a}}\,, \tag{78}$$

and where $\mathrm{Bern}(z;\,\pi) = \pi^z(1-\pi)^{1-z}$ is again the standard parameterization of the Bernoulli distribution, and the same applies for $\mathrm{Bern}(x;\,\cdot)$ for the observables. The set of all parameters is $\Theta = (\vec{\pi}, W, \vec{\mu})$.
□

Proposition 1. (Sigmoid Belief Nets)

A Sigmoid Belief Net (Definition D) is an EF generative model which satisfies the parameterization criterion (Definition C). It therefore applies at all stationary points:

$$\mathcal{F}(\Phi,\Theta) \;=\; \frac{1}{N}\sum_{n=1}^{N}\mathcal{H}[q^{(n)}_{\Phi}(\vec{z})] \quad-\; \mathcal{H}[\,p_{\Theta}(\vec{z})] \quad-\; \mathbb{E}_{\,\overline{q}_{\Phi}}\big\{\mathcal{H}[\,p_{\vec{\Theta}}(\vec{x}\,|\,\vec{z})]\big\}\,. \tag{79}$$

Proof

The generative model (77) and (78) has by definition two different sets of parameters for prior and noise model $\Theta=(\vec{\Psi},\vec{\Theta})$, which we arrage into column vectors as follows:

$$\vec{\Psi}=\vec{\pi}=\begin{pmatrix}\pi_1\\ \vdots\\ \pi_H\end{pmatrix},\;\vec{\Theta}=\begin{pmatrix}\vec{w}_1\\ \vdots\\ \vec{w}_H\\ \vec{\mu}\end{pmatrix},\text{ with }\vec{w}_h=\begin{pmatrix}W_{1h}\\ \vdots\\ W_{Dh}\end{pmatrix}, \tag{80}$$

where $\vec{\Psi}$ has H entries and $\vec{\Theta}$ has $HD+D$ entries (for all weights and offsets). The generative model (77) and (78) is an EF generative model because it can be reparameterized as follows:

$$\vec{z}\sim p_{\vec{\zeta}(\vec{\pi})}(\vec{z}) \qquad\text{where}\qquad \zeta_h(\vec{\pi})=\log\big(\frac{1}{1-\pi_h}\big),, \tag{81}$$

$$\vec{x}\sim p_{\vec{\eta}(\vec{z};\,\vec{\Theta})}(\vec{x}) \qquad\text{where}\qquad \vec{\eta}(\vec{z};\,\vec{\Theta})=\textstyle\sum_h \vec{w}_h\,z_h\;+\;\vec{\mu}\,, \tag{82}$$

and where the distributions $p_{\vec{\zeta}}(\vec{z})$ and $p_{\vec{\eta}}(\vec{x})$ are now Bernoulli distributions in the exponential family form. For the parameterization criterion, the Jacobian matrices are given by:

$$\frac{\partial\vec{\zeta}(\vec{\Psi})}{\partial\vec{\Psi}^{\mathrm{T}}}=\begin{pmatrix}(1-\pi_1)^{-1} & & \\ & \ddots & \\ & & (1-\pi_H)^{-1}\end{pmatrix}\quad\text{and}\quad\frac{\partial\vec{\eta}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}^{\mathrm{T}}}=\big(z_1\mathbb{1}_{D\times D},\ldots,z_H\mathbb{1}_{D\times D},\mathbb{1}_{D\times D}\big)\,, \tag{83}$$

where we have chosen all parameters for $\vec{\theta}$ (i.e., $\vec{\theta}=\vec{\Theta}$) for $\frac{\partial\vec{\eta}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}^{\mathrm{T}}}$. As all $\pi_h\in(0,1)$, the diagonal entries of the $(H\times H)$-matrix $\frac{\partial\vec{\zeta}(\vec{\Psi})}{\partial\vec{\Psi}^{\mathrm{T}}}$ are non-zero such that $\frac{\partial\vec{\zeta}(\vec{\Psi})}{\partial\vec{\Psi}^{\mathrm{T}}}$ is invertible. Part A of the criterion is therefore fulfilled by choosing $\vec{\alpha}(\vec{\Psi})=\Big(\frac{\partial\vec{\zeta}(\vec{\Psi})}{\partial\vec{\Psi}^{\mathrm{T}}}\Big)^{-1}\vec{\zeta}(\Psi)$. For Part B, it is straightforward to see that

$$\vec{\beta}(\vec{\Theta})=\begin{pmatrix}\vec{w}_1\\ \vdots\\ \vec{w}_H\\ \vec{\mu}\end{pmatrix} \tag{84}$$

fulfills equation (11). Hence, part B of the parameterization criterion is also satisfied. By virtue of Theorem 1 it therefore applies:

$$\mathcal{F}(\Phi,\vec{\Psi},\vec{\Theta}) \;=\; \tfrac{1}{N}\textstyle\sum_{n=1}^{N}\mathcal{H}[q^{(n)}_{\Phi}(\vec{z})] \;-\; \mathcal{H}[\,p_{\vec{\Psi}}(\vec{z})] -\; \tfrac{1}{N}\textstyle\sum_{n=1}^{N}\mathbb{E}_{q^{(n)}_{\Phi}}\big\{\mathcal{H}[\,p_{\vec{\Theta}}(\vec{x}\,|\,\vec{z})]\big\}\,, \tag{85}$$

and (79) is just (85) expressed using the original parameter notation of Definition D.
□

For Sigmoid Belief Nets it is also common to consider more than one set of latent variables. Such deep Sigmoid Belief Nets are defined as follows:

Definition E. (Deep Sigmoid Belief Nets)
The generative model of an deep SBN is given by:

$$\vec{z}^{(L)} \sim p_\Theta(\vec{z}^{(L)}) = \prod_{h=1}^{H_L} \mathrm{Bern}(z_h^{(L)}; \pi_h), \text{ with } 0 < \pi_h < 1, \tag{86}$$

$$\vec{z}^{(l)} \sim p_\Theta(\vec{z}^{(l)} \mid \vec{z}^{(l+1)}) = \prod_{h=1}^{H_l} \mathrm{Bern}\Big(z_h^{(l)}; \mathcal{S}\big(\sum_{h'=1}^{H_{l+1}} W_{hh'}^{(l)} z_{h'}^{(l+1)} + \mu_h^{(l)}\big)\Big), \quad l = 1, \dots, L-1, \tag{87}$$

$$\vec{x} \sim p_\Theta(\vec{x} \mid \vec{z}^{(1)}) = \prod_{d=1}^{D} \mathrm{Bern}\Big(x_d; \mathcal{S}\big(\sum_{h=1}^{H} W_{dh}^{(0)} z_h^{(1)} + \mu_d^{(0)}\big)\Big), \tag{88}$$

where $\mathcal{S}$ is again the function $\mathcal{S}(a) = \frac{1}{1+\mathrm{e}^{-a}}$ and $\mathrm{Bern}(z, \pi) = \pi^z(1-\pi)^{1-z}$ the standard parameterization of the Bernoulli distribution. The set of all parameters is

$$\Theta = \big(\vec{\pi}, W^{(0)}, \dots, W^{(L-1)}, \vec{\mu}^{(0)}, \dots, \vec{\mu}^{(L-1)}\big). \tag{89}$$

□

As can be observed when considering Theorem 1 and Proposition 1, the proof idea is to show convergence to an entropy for each summand of the ELBO individually. It is consequently straightforward to generalize the proof of Theroem 1 and Proposition 1 to deep Sigmoid Belief Nets. For this we write the ELBO as follows:

$$\mathcal{F}(\Phi, \Theta) = \mathcal{F}_1(\Phi) - \sum_{l=1}^{L} \mathcal{F}_2^{(l)}(\Phi, \Theta) - \mathcal{F}_3(\Phi, \Theta), \quad \text{where} \tag{90}$$

$$\mathcal{F}_1(\Phi) = -\frac{1}{N} \sum_n \int q_\Phi^{(n)}(\vec{z}) \log\big(q_\Phi^{(n)}(\vec{z})\big) \mathrm{d}\vec{z} \tag{91}$$

$$\mathcal{F}_2^{(l)}(\Phi, \Theta) = -\frac{1}{N} \sum_n \int q_\Phi^{(n)}(\vec{z}) \log\big(p_\Theta(\vec{z}^{(l)} \mid \vec{z}^{(l+1)})\big) \mathrm{d}\vec{z} \tag{92}$$

$$\mathcal{F}_3(\Phi, \Theta) = -\frac{1}{N} \sum_n \int q_\Phi^{(n)}(\vec{z}) \log\big(p_\Theta(\vec{x} \mid \vec{z}^{(1)})\big) \mathrm{d}\vec{z}, \tag{93}$$

where we used the convention $p_\Theta(\vec{z}^{(L)} \mid \vec{z}^{(L+1)}) = p_\Theta(\vec{z}^{(L)})$ to denote the prior. Following the proofs of Theorem 1 and Proposition 1, each summand $\mathcal{F}_2^{(l)}(\Phi, \Theta)$ can be observed to converge to the expected entropy

$$\frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_{q_\Phi^{(n)}}\{\mathcal{H}[p_\Theta(\vec{z}^{(l)} \mid \vec{z}^{(l+1)})]\}. \tag{94}$$

The generalization of Proposition 1 for deep SBNs, therefore, reads:

Proposition 2. (Deep Sigmoid Belief Nets)

The following applies to a deep Sigmoid Belief Net at all stationary points:

$$\mathcal{F}(\Phi,\Theta) = \frac{1}{N}\sum_{n=1}^{N}\mathcal{H}[q_\Phi^{(n)}(\vec{z})] - \frac{1}{N}\sum_{l=1}^{L}\sum_{n=1}^{N}\mathbb{E}_{q_\Phi^{(n)}}\{\mathcal{H}[p_\Theta(\vec{z}^{(l)}\mid\vec{z}^{(l+1)})]\} - \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{q_\Phi^{(n)}}\{\mathcal{H}[p_\Theta(\vec{x}\mid\vec{z}^{(1)})]\}, \quad (95)$$

where we again used the convention $p_\Theta(\vec{z}^{(L)}\mid\vec{z}^{(L+1)}) = p_\Theta(\vec{z}^{(L)})$ to denote the prior.

5 Gaussian Observables and Probabilistic PCA

While Theorem 1 applies for a broad range of generative models, those models with Gaussian observable distributions are very common and, therefore, represent an important case. At the same time, the properties of Gaussian distributions will further simplify the result of Theorem 1. For these reasons, we will treat Gaussian observables explicitly in the following.

Definition F. (Gaussian observables with scalar variance)

We consider generative models as in Definition A where the observable distribution $p_{\vec{\Theta}}(\vec{x}\,|\,\vec{z})$ is a Gaussian. Concretely, we consider the following class of Generative models:

$$\vec{z} \sim p_{\vec{\Psi}}(\vec{z}) \quad (96)$$

$$\vec{x} \sim p_{\vec{\Theta}}(\vec{x}\,|\,\vec{z}) = \mathcal{N}(\vec{x};\vec{\mu}(\vec{z};\vec{w}),\sigma^2\mathbb{1}),\ \text{where}\ \ \sigma^2>0, \quad (97)$$

and where $\vec{\mu}(\vec{z};\vec{w})$ is a well-behaved function (with parameters $\vec{w}$) from the latents $\vec{z}$ to the mean of the Gaussian.
□

For linear functions $\vec{\mu}$ the vector $\vec{w}$ can be thought of as containing all entries of a weight matrix W and potentially all offsets.

Proposition 3. (Gaussian observables with scalar variance)

Consider the generative model of Definition F. If the prior $p_{\vec{\zeta}(\vec{\Psi})}(\vec{z}) = p_{\vec{\Psi}}(\vec{z})$ is an exponential family distribution which satisfies part A of the parameterization criterion (Definition C) then at all stationary point applies:

$$\begin{aligned}\mathcal{F}(\Phi,\vec{\Psi},\vec{\Theta}) &= \frac{1}{N}\sum_{n=1}^{N}\mathcal{H}[q_\Phi^{(n)}(\vec{z})] - \mathcal{H}[p_{\vec{\Psi}}(\vec{z})] - \mathcal{H}[p_{\vec{\Theta}}(\vec{x}\,|\,\vec{z})]\\ &= \frac{1}{N}\sum_{n=1}^{N}\mathcal{H}[q_\Phi^{(n)}(\vec{z})] - \mathcal{H}[p_{\vec{\Psi}}(\vec{z})] - \frac{D}{2}\log\left(2\pi e\sigma^2\right)\end{aligned} \quad (98)$$

Proof

We need to show that part B of the parameterization criterion is fulfilled. We abbreviate the variance σ^2 again by $\tilde{\sigma}$, and we choose $\vec{\theta} = \tilde{\sigma}$ as subset of $\vec{\Theta}$ (i.e., we chose a single-valued vector $\vec{\theta}$). The Jacobian $\frac{\partial \vec{\eta}(\vec{z};\vec{\Theta})}{\partial \vec{\theta}^{\mathrm{T}}}$ of Exn. 9 is then a column vector. Now consider the natural parameters of a scalar Gaussian $\vec{\eta} = \begin{pmatrix} \frac{\mu}{\tilde{\sigma}} \\ -\frac{1}{2\tilde{\sigma}} \end{pmatrix}$. The form of the natural parameters means that we can rewrite $\vec{\eta}(\vec{z};\vec{\Theta}) = \frac{1}{\tilde{\sigma}}\, \tilde{\vec{\eta}}(\vec{z};\vec{w})$, where the second factor does not depend on $\tilde{\sigma}$. The Jacobian $\frac{\partial \vec{\eta}(\vec{z};\vec{\Theta})}{\partial \vec{\theta}^{\mathrm{T}}}$ now becomes:

$$\frac{\partial \vec{\eta}(\vec{z};\vec{\Theta})}{\partial \vec{\theta}^{\mathrm{T}}} \;=\; \frac{\partial}{\partial \tilde{\sigma}}\, \vec{\eta}^{\mathrm{T}}_{(\vec{z};\vec{\Theta})} \;=\; \big(\frac{\partial}{\partial \tilde{\sigma}}\frac{1}{\tilde{\sigma}}\big)\tilde{\vec{\eta}}^{\mathrm{T}}(\vec{z};\vec{w}) \;=\; -\frac{1}{\tilde{\sigma}}\,\frac{1}{\tilde{\sigma}}\;\tilde{\vec{\eta}}^{\mathrm{T}}(\vec{z};\vec{w}) \;=\; -\frac{1}{\tilde{\sigma}}\;\vec{\eta}^{\mathrm{T}}_{(\vec{z};\vec{\Theta})}\,. \tag{99}$$

Therefore, the scalar function $\beta(\tilde{\sigma}) = -\tilde{\sigma}$ satisfies Eqn. 11, which shows that part B of the parameterization criterion is fulfilled. Consequently, Theorem 1 applies for the generative model.

The entropy of the Gaussian $p_{\vec{\Theta}}(\vec{x}\,|\,\vec{z})$ is given by $\frac{D}{2}\log\big(2\pi e\tilde{\sigma}\big)$, i.e., it does not depend on $\vec{z}$. The last term of (40) consequently simplifies to:

$$\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{q_{\Phi}^{(n)}}\big\{\mathcal{H}[\,p_{\vec{\Theta}}(\vec{x}\,|\,\vec{z})]\big\} \;=\; \frac{D}{2}\log\big(2\pi e\tilde{\sigma}\big) \;=\; \mathcal{H}[\,p_{\vec{\Theta}}(\vec{x}\,|\,\vec{z})]\,. \tag{100}$$

□

The maybe most prominent generative model covered by Proposition 3 is probabilistic PCA (p-PCA; Roweis, 1998; Tipping and Bishop, 1999), which has a Gaussian prior and a Gaussian noise model, i.e., both distributions are in the exponential family. Given Proposition 3, we only have to show that the parameterization criterion applies for the prior distribution. This can be shown similarly as was done for the noise distribution above. However, p-PCA first has to be parameterized similarly to Example 2 as we require a parameterized prior. In the standard parameterization of p-PCA (i.e., $p_\Theta(\vec{z}) = \mathcal{N}(\vec{z};0,\mathbb{1})$ and $p_\Theta(\vec{x}\,|\,\vec{z}) = \mathcal{N}(\vec{x};W\vec{z}+\vec{\mu},\sigma^2\mathbb{1})$) and for the case when full posteriors are used as variational distributions, we then obtain:

$$\mathcal{L}(\Theta) \;=\; \mathcal{H}[\,p_\Theta(\vec{z}\,|\,\vec{x})] \;-\; \mathcal{H}[\,p_\Theta(\vec{z})] \;-\; \mathcal{H}[\,p_\Theta(\vec{x}\,|\,\vec{z})] \tag{101}$$

$$\;=\; -\tfrac{1}{2}\log\big(\det(\tfrac{1}{\sigma^2}W^{\mathrm{T}}W \;+\; \mathbb{1})\big) \;-\; \tfrac{D}{2}\log\big(2\pi e\sigma^2\big)\,, \tag{102}$$

at all stationary points. The result can be shown to be consistent with the maximum likelihood solution of Tipping and Bishop (1999) by expressing $M = W^TW + \sigma^2\mathbb{1}$ and σ^2 in terms of eigenvalues of the data covariance matrix. Also note the relation to variational autoencoders with linear decoder as discussed by (Damm et al., 2023).

While Definition F already covers many well-known models including probabilistic PCA, other models such as Factor Analysis require a further generalization.

Definition G. (Gaussian observables with diagonal covariance)
We consider generative models as in Definition A where the observable distribution $p_{\vec{\Theta}}(\vec{x}\,|\,\vec{z})$ is a Gaussian with diagonal covariance matrix:

$$\vec{z} \;\sim\; p_{\vec{\Psi}}(\vec{z}) \tag{103}$$

$$\vec{x} \;\sim\; p_{\vec{\Theta}}(\vec{x}\,|\,\vec{z}) = \mathcal{N}(\vec{x};\vec{\mu}(\vec{z};\vec{w}),\Sigma),\ \text{where}\ \Sigma = \begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_D^2 \end{pmatrix},\ \sigma_d^2 > 0, \tag{104}$$

where $\vec{\mu}(\vec{z};\vec{w})$ is again a well-behaved function (with parameters $\vec{w}$) from the latents $\vec{z}$ to the set the mean of the Gaussian.
□

Proposition 4. (Gaussian observables with diagonal covariance)
Consider the generative model of Definition G. If the prior $p_{\vec{\zeta}(\vec{\Psi})}(\vec{z}) = p_{\vec{\Psi}}(\vec{z})$ is an exponential family distribution which satisfies part A of the parameterization criterion (see Definition C) then at all stationary point applies:

$$\begin{aligned} \mathcal{F}(\Phi,\vec{\Psi},\vec{\Theta}) &= \frac{1}{N}\sum_{n=1}^{N}\mathcal{H}[q_{\Phi}^{(n)}(\vec{z})] \;-\; \mathcal{H}[p_{\vec{\Psi}}(\vec{z})] \;-\; \mathcal{H}[p_{\vec{\Theta}}(\vec{x}\,|\,\vec{z})] \\ &= \frac{1}{N}\sum_{n=1}^{N}\mathcal{H}[q_{\Phi}^{(n)}(\vec{z})] \;-\; \mathcal{H}[p_{\vec{\Psi}}(\vec{z})] \;-\; \sum_{d=1}^{D}\frac{1}{2}\log\left(2\pi e\sigma_d^2\right) \end{aligned} \tag{105}$$

Proof

To show that part B the parameterization criterion is fulfilled, we consider as subset $\vec{\theta}$ of $\vec{\Theta}$ the vector of all variances σ_d^2 (with abbreviation $\tilde{\sigma}_d = \sigma_d^2$) and the following natural parameters:

$$\vec{\theta} = \begin{pmatrix} \tilde{\sigma}_1 \\ \vdots \\ \tilde{\sigma}_D \end{pmatrix},\ \vec{\eta}(\vec{z};\vec{\Theta}) = \begin{pmatrix} \vec{a}(\vec{z};\vec{\Theta}) \\ \vec{b}(\vec{z};\vec{\Theta}) \end{pmatrix},\ \vec{a}(\vec{z};\vec{\Theta}) = \begin{pmatrix} \frac{\mu_1(\vec{z};\vec{w})}{\tilde{\sigma}_1} \\ \vdots \\ \frac{\mu_D(\vec{z};\vec{w})}{\tilde{\sigma}_D} \end{pmatrix},\ \vec{b}(\vec{z};\vec{\Theta}) = \begin{pmatrix} -\frac{1}{2\,\tilde{\sigma}_1} \\ \vdots \\ -\frac{1}{2\,\tilde{\sigma}_D} \end{pmatrix}. \tag{106}$$

Given $\vec{\theta}$ as above, the Jacobian (9) is a $(2D \times D)$-matrix given as follows:

$$\frac{\partial\vec{\eta}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}^{\mathrm{T}}} \;=\; \begin{pmatrix} \dfrac{\partial\vec{a}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}^{\mathrm{T}}} \\ \dfrac{\partial\vec{b}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}^{\mathrm{T}}} \end{pmatrix} = \begin{pmatrix} A \\ B \end{pmatrix}, \tag{107}$$

where A and B are $(D \times D)$-diagonal matrices with entries:

$$A = \begin{pmatrix} -\frac{\mu_1(\vec{z};\vec{w})}{\tilde{\sigma}_1^2} & & \\ & \ddots & \\ & & -\frac{\mu_D(\vec{z};\vec{w})}{\tilde{\sigma}_D^2} \end{pmatrix} \ \text{and}\ \ B = \begin{pmatrix} \frac{1}{2\tilde{\sigma}_1^2} & & \\ & \ddots & \\ & & \frac{1}{2\tilde{\sigma}_D^2} \end{pmatrix}. \tag{108}$$

By choosing $\vec{\beta}(\vec{\theta}) = -\left(\tilde{\sigma}_1, \ldots, \tilde{\sigma}_D\right)^{\mathrm{T}}$, equation (11) und hence the second part of the parameterization criterion is fullfilled and Theorem 1 applies.

The entropy of the Gaussian $p_{\vec{\Theta}}(\vec{x} \,|\, \vec{z})$ is finally given by $\frac{1}{2}\sum_d \log\left(2\pi e \tilde{\sigma}_d\right)$, i.e., it again does not depend on $\vec{z}$. The last term of (40) consequently simplifies to:

$$\frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{q_\Phi^{(n)}}\{\mathcal{H}[p_{\vec{\Theta}}(\vec{x} \,|\, \vec{z})]\} \;=\; \frac{1}{2}\sum_{d} \log\left(2\pi e \tilde{\sigma}_d\right) \;=\; \mathcal{H}[p_{\vec{\Theta}}(\vec{x} \,|\, \vec{z})]\,. \tag{109}$$

□

Notably, one has to consider the cases with scalar variance and diagonal covariance separately because scalar variance can in our setting not be treated as a special case of diagonal covariance. Using Proposition 4, we could now treat factor analysis (e.g. Everitt, 1984; Bartholomew and Knott, 1987) as a model with Gaussian prior and a linear mapping to observables.

6 Mixture Models

As the last example we consider mixture models with exponential family distributions as noise distributions. Such models are very closely related to Bregman clustering (e.g. McLachlan and Peel, 2004; Banerjee et al., 2005), and mixtures are a tool that is for machine learning and data science of similar importance as PCA-like approaches. In this section we follow the notational convention for mixture models and replace the latent variable $\vec{z}$ by c, where c takes on integer values. More concretely, we here consider the class of mixutre models defined as follows.

Definition H. (EF Mixture)

We consider generative models with a latent variable c and observables $\vec{x}$. The latent can take on one of C different values, i.e. $c \in \{1, \ldots, C\}$, each with probability $\pi_c > 0$ where $\sum_{c=1}^{C} \pi_c = 1$. The observable $\vec{x}$ is distributed according to a distribution $p(\vec{x} \,|\, \vec{\Theta}_c)$ with parameters depending on the value of the latent c. More formally, the generative model is given by:

$$c \sim \mathrm{Cat}(c\,;\vec{\pi}), \tag{110}$$

$$\vec{x} \sim p(\vec{x} \,|\, \vec{\Theta}_c), \tag{111}$$

where $\mathrm{Cat}(c\,;\vec{\pi})$ is the categorical distribution with standard parameterization. The set of all model parameters Θ is given by $\Theta = \left(\vec{\pi}, \vec{\Theta}_1, \vec{\Theta}_2, \ldots, \vec{\Theta}_C\right)$.

We refer to the model (110) and (111) as an *exponential family mixture (EF mixture)* if the observable distribution $p(\vec{x} \,|\, \vec{\Theta}_c)$ can be written as an exponential family distribution $p_{\vec{\eta}(\vec{\Theta}_c)}(\vec{x})$ (with constant base measure) where $\vec{\eta}(\cdot)$ maps an L-dim vector $\vec{\Theta}_c$ of standard parameters to an L-dim vector of natural parameters.

□

For a mixture of scalar Gaussian distributions, $\vec{\Theta}_c$ would be an $L = 2$ dimensional vector equal to $\begin{pmatrix} \mu_c \\ \sigma_c^2 \end{pmatrix}$, where μ_c and σ_c^2 are, respectively, mean and variance of cluster c. The mapping to natural parameters $\vec{\eta}(\cdot)$ would be given by $\vec{\eta}(\vec{\Theta}_c) = \begin{pmatrix} \frac{\mu_c}{\sigma_c^2} \\ -\frac{1}{2\sigma_c^2} \end{pmatrix}$. For a mixture of gamma distributions, $\vec{\Theta}_c$ would be an $L = 2$ dimensional vector equal to $\begin{pmatrix} \alpha_c \\ \beta_c \end{pmatrix}$, where α_c and β_c are the standard parameters (shape and rate, respectively). The mapping to natural parameters $\vec{\eta}(\cdot)$ would be given by $\vec{\eta}(\vec{\Theta}_c) = \begin{pmatrix} \alpha_c - 1 \\ -\beta_c \end{pmatrix}$. Similarly for other standard distributions. For the proof of the proposition on EF mixtures below, we will again use vectors for the parameters and will distinguish between prior parameters $\vec{\Psi} = (\pi_1, \ldots, \pi_{C-1})^{\mathrm{T}}$ and noise model parameters $\vec{\Theta} = (\vec{\Theta}_1, \ldots, \vec{\Theta}_C)^{\mathrm{T}}$. Note that we will only consider $(C-1)$ prior parameters as π_C can be taken to be determined by π_1 to π_{C-1}.

We remark the technical but important difference between the mapping $\vec{\eta}(\cdot)$ of Definition H and the mapping $\vec{\eta}(c;\vec{\Theta})$ which takes the role of $\vec{\eta}(\vec{z};\vec{\Theta})$ as used for Definitions B and C. $\vec{\eta}(\cdot)$ is the usual mapping from standard parameters of an exponential family distribution to its natural parameters (see the just discussed examples Gaussian and gamma distributions). $\vec{\eta}(c;\vec{\Theta})$ also maps standard parameters to natural parameters of the observable distributions (hence the same symbol) but, in contrast to $\vec{\eta}(\cdot)$, the mapping $\vec{\eta}(c;\vec{\Theta})$ has the vector of *all* noise model parameters as input. The Jacobians of the mappings are consequently different. The Jacobian of $\vec{\eta}(c;\vec{\Theta})$ (computed w.r.t. all parameters $\vec{\Theta}$) is a $(L \times CL)$-matrix, while the Jacobian of $\vec{\eta}(\cdot)$ is a square $(L \times L)$-matrix. This smaller $(L \times L)$-Jacobian is important to state a sufficient condition for the following proposition. The large Jacobian of $\vec{\eta}(c;\vec{\Theta})$ is important for the proposition's proof.

Proposition 5. (Mixture Models)
Consider an EF mixture model of Definition H where $\vec{\eta}(\cdot)$ maps the standard parameters of the noise model to the distribution's natural parameters. If the Jacobian of $\vec{\eta}(\cdot)$ is everywhere invertible, then the model is an EF generative model (Definition B) and the parameterization criterion (Definition C) is fulfilled. It then applies at all stationary points:

$$\mathcal{F}(\Phi,\Theta) \;=\; \frac{1}{N}\sum_{n=1}^{N}\mathcal{H}[q^{(n)}_{\Phi}(c)] \;-\; \mathcal{H}[\,\mathrm{Cat}(c\,;\vec{\pi})\,] \;-\; \mathbb{E}_{\,\overline{q}_{\Phi}}\big\{\mathcal{H}[\,p(\vec{x}\,|\,\vec{\Theta}_c)]\big\}\,. \tag{112}$$

Proof

We change, as for previous proofs, to vectorial notation with $\vec{\Psi} = (\pi_1, \pi_{C-1})^{\mathrm{T}}$ denoting the prior parameters and with $\vec{\Theta} = (\vec{\Theta}_1, \ldots, \vec{\Theta}_C)^{\mathrm{T}}$ denoting the noise model parameters. The categorical distribution of the prior can be rewritten in exponential family form as follows:

$$p_{\vec{\zeta}(\vec{\Psi})}(c) \;=\; h(c)\exp\big(\vec{\zeta}^{\mathrm{T}}(\vec{\Psi})\,\vec{T}(c) \;-\; A(\vec{\Psi})\big) \tag{113}$$

where the natural parameters $\vec{\zeta}(\vec{\Psi})$ have length $(C-1)$ and are defined for $i = 1, \ldots, (C-1)$ by $\zeta_i(\vec{\Psi}) = \log\big(\frac{\pi_i}{1-\sum_{i'=1}^{C-1}\pi_{i'}}\big)$. The sufficient statistics $\vec{T}(c)$ are also of length $(C-1)$ and are defined for $i = 1, \ldots, (C-1)$ by $T_i(c) = \delta_{ci}$ (i.e., $T_i(c) = 1$ if $c = i$ and zero otherwise). The partition function is $A(\vec{\Psi}) = -\log\big(1 - \sum_{c=1}^{C-1} \pi_c\big)$ and the base measure is a constant, $h(c) = 1$. For the observable distribution, we assumed in Def. H that it can be written as an exponential family distribution, $p_{\vec{\eta}(\vec{\Theta}_c)}(\vec{x})$, with constant base measure. With this assumption, the mixture model of Def. H is an EF generative model (Def. B). $\vec{\zeta}(\vec{\Psi})$ as defined above are the natural parameters of the prior; and $\vec{\eta}(c; \vec{\Theta}) = \vec{\eta}(\vec{\Theta}_c)$ are the natural parameters of the observable distribution (for mixtures we use c for the latent instead of $\vec{z}$ for previous models). It remains to be shown that the parameterization criterion holds (Def. C).

Regarding part A of the parameterization criterion, the Jacobian matrix $\frac{\partial \vec{\zeta}(\vec{\Psi})}{\partial \vec{\Psi}^{\mathrm{T}}}$ of $\vec{\zeta}(\vec{\Psi})$ is a $(C-1) \times (C-1)$-matrix. The elements of the Jacobian can be derived to be:

$$\left(\frac{\partial \vec{\zeta}(\vec{\Psi})}{\partial \vec{\Psi}^{\mathrm{T}}}\right)_{ic} \;=\; \frac{\partial}{\partial \pi_c} \zeta_i(\vec{\Psi}) \;=\; \frac{\partial}{\partial \pi_c} \log\Big(\frac{\pi_i}{1 - \sum_{i'=1}^{C-1} \pi_{i'}}\Big) \;=\; \delta_{ic}\,\frac{1}{\pi_c} + \frac{1}{\pi_C}, \tag{114}$$

where we abbreviated $\pi_C = \sum_{c=1}^{C-1} \pi_c$. We now consider the i-th component of the right-hand side of (10) which is given by:

$$\left(\frac{\partial \vec{\zeta}(\vec{\Psi})}{\partial \vec{\Psi}^{\mathrm{T}}} \vec{\alpha}(\vec{\Psi})\right)_i = \sum_c \Big(\frac{\partial \vec{\zeta}(\vec{\Psi})}{\partial \vec{\Psi}^{\mathrm{T}}}\Big)_{ic} \alpha_c(\vec{\Psi}) = \frac{1}{\pi_i} \alpha_i(\vec{\Psi}) + \frac{1}{\pi_C} \sum_c \alpha_c(\vec{\Psi}) \stackrel{!}{=} \zeta_i(\vec{\Psi}). \tag{115}$$

By multiplying with π_i and summing over all components $i = 1, \ldots, C-1$, we can conclude from (115) that:

$$\sum_i \pi_i \zeta_i(\vec{\Psi}) = \sum_i \alpha_i(\vec{\Psi}) + \frac{1}{\pi_C}\Big(\sum_i \pi_i\Big)\Big(\sum_c \alpha_c(\vec{\Psi})\Big) = 2 \sum_i \alpha_i(\vec{\Psi}) \tag{116}$$

$$\Rightarrow \frac{1}{\pi_C} \sum_i \alpha_i(\Psi) = \frac{1}{2\pi_C} \sum_i \pi_i \zeta_i(\vec{\Psi}) =: \rho(\vec{\Psi}), \tag{117}$$

where we introduced the function $\rho(\vec{\Psi})$ as an abbreviation. By using c as the summation index and substituting (117) in (115), we get

$$\alpha_i(\vec{\Psi}) = \pi_i\big(\zeta_i(\vec{\Psi}) - \rho(\vec{\Psi})\big), \quad i = 1, \ldots, C-1. \tag{118}$$

The corresponding $\vec{\alpha}(\vec{\Psi})$ satisfies equation (10). Part A of the parameterization criterion is therefore fulfilled.

To construct the Jacobian $\frac{\partial \vec{\eta}(c;\vec{\Theta})}{\partial \vec{\theta}^{\mathrm{T}}}$ for part B of the criterion, we use all parameters $\vec{\Theta}$, i.e. $\vec{\theta} = \vec{\Theta}$. The Jacobian is then the $(L \times CL)$-matrix given by:

$$\frac{\partial \vec{\eta}(c; \vec{\Theta})}{\partial \vec{\theta}^{\mathrm{T}}} = \left(\frac{\partial \vec{\eta}(c; \vec{\Theta})}{\partial \vec{\theta}_1^{\mathrm{T}}}, \ldots, \frac{\partial \vec{\eta}(c; \vec{\Theta})}{\partial \vec{\theta}_C^{\mathrm{T}}}\right) = \left(0_{L\times L}, \ldots, 0_{L\times L}, \frac{\partial \vec{\eta}(\vec{\Theta}_c)}{\partial \vec{\theta}_c^{\mathrm{T}}}, 0_{L\times L}, \ldots, 0_{L\times L}\right). \tag{119}$$

Here $0_{L\times L}$ denotes the $(L \times L)$-matrix where all entries are zero. The only non-vanishing components are those with derivative w.r.t. the parameters $\vec{\theta}_c$. Since the Jacobian $\frac{\partial \vec{\eta}(\vec{\Theta}_c)}{\partial \vec{\theta}_c^{\mathrm{T}}}$ is invertible, the vector valued function

$$\vec{\beta}(\vec{\Theta}) = \left(\Big(\frac{\partial\vec{\eta}(\vec{\Theta}_1)}{\partial\vec{\theta}_1^{\mathrm{T}}}\Big)^{-1}\vec{\eta}(\vec{\Theta}_1), \ldots, \Big(\frac{\partial\vec{\eta}(\vec{\Theta}_C)}{\partial\vec{\theta}_C^{\mathrm{T}}}\Big)^{-1}\vec{\eta}(\vec{\Theta}_C) \right)^{\mathrm{T}} \tag{120}$$

is well-defined, independent of c and satisfies equation (11) because of $\vec{\eta}(c,\vec{\Theta}) = \vec{\eta}(\vec{\Theta}_c)$. Therefore, part B of the parameterization criterion is also fulfilled.

As the model of Def. H with the assumption of Prop. 5 is an EF generative model that fulfills the parameterization criterion, (112) follows from Theorem 1.
□

The assumptions made in Proposition 5 are sufficient for (112) to apply. The assumptions are also sufficiently broad to apply for most mixtures but they can presumably be weakened further. For a given mixture model, Proposition 5 is relatively easy to apply because the mapping from standard parameterization to natural parameters is usually known.

As example let us use a mixture of Gamma distribtuions. A mixture component c can then be written as exponential family distribution with $L=2$ parameters given by:

$$p_{\vec{\eta}(\vec{\Theta}_c)}(x) \;=\; h(x)\exp\big(\vec{\eta}^{\mathrm{T}}(\vec{\Theta}_c)\,\vec{T}(\vec{x}) \;-\; A(\vec{\Theta}_c)\big) \tag{121}$$

with $\vec{\Theta}_c = \begin{pmatrix} \alpha_c \\ \beta_c \end{pmatrix}$ and $\vec{\eta}(\vec{\Theta}_c) = \begin{pmatrix} \alpha_c - 1 \\ -\beta_c \end{pmatrix}$. The base measure is constant, $h(x) = 1$, and $A(\vec{\Theta}_c) = \log\big(\Gamma(\alpha_c)\big) - \alpha_c \log\big(\beta_c\big)$ is the partition function. The Jacobian of $\vec{\eta}(\cdot)$ is the (2×2)-matrix $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ which is everywhere invertible. Similarly, mixture models based on other concrete observable distributions can be treated.

Variational optimization can be used to accelerate training of mixture models (Neal and Hinton, 1998; Lücke and Forster, 2019) and (112) applies for the corresponding variational distributions. However, it is more common for mixture models to be trained without approximation usually by directly using expectation maximization (EM; Dempster et al., 1977). In our setting, the exact (i.e. full) posterior can be considered as a special case of a variational distribution. Furthermore, using full posteriors means that the ELBO becomes equal to the likelihood $\mathcal{L}(\Theta)$ at all stationary points. For mixture models optimized using EM, we can therefore conclude that at all stationary points of learning applies:

$$\mathcal{L}(\Theta) \;=\; \frac{1}{N}\sum_{n=1}^{N}\mathcal{H}[\,p(c\,|\,\vec{x}^{(n)},\Theta)] - \;\mathcal{H}[\,\mathrm{Cat}(c\,;\vec{\pi}\,)\,] - \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{p(c\,|\,\vec{x}^{(n)},\Theta)}\big\{\mathcal{H}[\,p(\vec{x}\,|\,\vec{\Theta}_c)]\big\}\,. \tag{122}$$

As a small final consistency result, suppose that we train a mixture of C scalar Gaussians on data generated from a single scalar Gaussian distribution. Then optimization using EM can result, e.g., in the mixing proportion of one cluster to converge to one, while the mixing proportions of all other clusters converge to zero. In that case, prior entropy and posterior entropy in (122) converge to zero. Furthermore, as the posterior becomes one for always the same cluster, we obtain $\mathcal{L}(\Theta) = -\mathcal{H}[\,p(\vec{x}\,|\,\vec{\Theta}_1)]$, where we assumed w.l.o.g. that it is the first cluster ($c=1$) for which the mixing proportion converges to one ($\pi_1 \to 1$). More concretely: $\mathcal{L}(\Theta) = -\frac{1}{2}\log\left(2\pi e\sigma_1^2\right)$ at stationary points. This boundary case connects to the relation of maximum likelihood parameter estimation and (negative) entropy of exponential family distributions (as briefly discussed in the introduction). Of course, in practice EM for mixtures of Gaussians will not precisely recover values of zero or one for the mixing proportions and, e.g., also two essentially identical clusters or other results are possible (depending on initial values). But the boundary case may nevertheless serve to illustrate consistency of (122) with conventional maximum likelihood estimation.

7 Discussion and Outlook

By considering the used definitions and the proof of Theorem 1, a number of aspects can be noted.

Theorem 1 and other standard generative models. Based on Theorem 1 further examples of standard generative models can be considered and treated similarly as shown for the three here considered types of generative models (Sections 4 to 6). For instance, other models with Gaussian observables can be considered with probabilistic sparse coding (e.g. Olshausen and Field, 1996) or factor analysis (e.g. Everitt, 1984; Bartholomew and Knott, 1987) being just two of many other examples. For each model, we have to show that all conditions required for Theorem 1 are fulfilled. The main challenge is usually the parameterization criterion but also the generative model first has to take on an appropriate form (such as being equipped with a parameterized prior). Further models that can be treated are specific mixture models (McLachlan and Peel, 2004; Banerjee et al., 2005), i.e., mixtures with concrete examples for observable distributions such as Bernoulli, exponential or beta-distributions. In the case of mixtures, no further analytical steps are required at least for models that fulfill the requirements of Proposition H. Hence, concrete expressions for the ELBO (or for the log-likelihood) can be stated with only minor effort as entropies of the most common exponential family distributions are known, or can relatively easily be obtained (e.g. Nielsen and Nock, 2010). More efforts may be required for less conventional mixtures or heterogeneous mixtures but applications of Theorem 1 will presumably be similar to the application to standard mixtures as treated here.

Non-constant base measures. By considering the proof of the main result (Theorem 1) the assumption of a constant base measure is not essential for the proof but the result would have to be modified. For instance, we could introduced an amended entropy $\tilde{\mathcal{H}}$ which includes base measure terms. However, we could also resort to a more abstract notation using Lebesgue integrals. In doing so, as it is common for exponential families, the base measure can become part of the Lebesgue measure, and proofs can essentially be maintain as here described. For our purposes, we considered it more suitable to keep the more basic notation (cf. Bickel and Doksum, 2015) in order to connect to the standard notational conventions as used in most of the Machine Learning literature.

Deep generative models. We have here focused on elementary generative models based on bipartite graphical models, which represent the natural first examples. But the treatment can, of course, in principle be further extended, e.g., towards deep generative models that are of considerable interest (see Bond-Taylor et al., 2022, for a review). Two examples of such deep models are variational autoencoders (VAEs; Kingma and Welling, 2014; Rezende et al., 2014) or deep SBNs (Neal, 1992; Hinton et al., 2006). Standard Gaussian VAEs have been treated by (Damm et al., 2023) using earlier versions of the here investigated properties specific to Gaussian distributions. Deep SBNs have here been treated as generalization of shallow SBNs. The study of a general treatment of deep generative models represents a natural next future research effort.

Previous work. As detailed in the introduction, our investigation of entropy convergence for models based on exponential family distributions (Theorem 1) has its roots in two earlier papers (Lücke and Henniges, 2012; Damm et al., 2023). Both of those contributions focus on models based on Gaussian distributions while the here presented results show convergence to entropy sums more generally: Theorem 1 shows that convergence to entropy sums is a very common feature of generative models, which does not require properties specific to the Gaussian distribution. Given the general result of Theorem 1, we can treat the Gaussian case as a specialization, of course. For Gaussian VAEs (especially VAEs with DNNs also for decoder variances) convergence to entropy sums can in this way be proven by showing that the parameterization criterion (Definition C) is fulfilled for such VAEs (see Damm et al., 2023). In an earlier approach for proving entropy convergence for Gaussian VAEs (Lücke et al., 2021), the properties of DNNs and Gaussians were much more entangled[1].

In contrast to VAEs, the case of deep SBNs differs as deep SBNs consist of stochastic units in each layer (while layers of VAE generative models consist of the deterministic units of standard DNNs). Hence, for deep SBNs more than one set of latent variables has to be considered, which would require an extension of the models treated here. By considering the proof of the main result (Theorem 1), one observes that there is no need to restrict oneself to just one set of stochastic/latent variables, however. Indeed, it is relatively straight-forward to consider models with several sets of latents. Deep SBNs (but also other similar types of deep generative model) can, therefore, be expected to converge to sums of entropies with more than three entropy terms.

Conclusion. Future work will further investigate the theoretical implications of the result, and the classes of generative models for which results like Theorem 1 can be derived. Furthermore, different applications of the theorem will be investigated that leverage the convergence of the ELBO to entropy sums. The previously mentioned study (Damm et al., 2023) takes first steps in this direction by analyzing phenomena like mode collapse or by using entropy sums for fast model selection. Further steps currently considered do make use of the fact that equality to entropy sums applies also when only subsets of the model parameters are at stationary points. Following such considerations, more global properties of the optimization landscape can be investigated and learning itself could be improved.

Acknowledgement. Funding by the German Research Foundation (DFG) in the project 'On the Convergence of Variational Deep Learning to Sums of Entropies', LU 1196/9-1, within the priority program 'Theoretical Founda-

[1]The papers (Lücke et al., 2021) and (Damm et al., 2023) are two arXiv versions of the same paper.

tions of Deep Learning' (SPP 2298) is acknowledged. Thanks go to Asja Fischer of the same project for fruitful discussions on the topic. Furthermore, many thanks go to Simon Damm and Dmytro Velychko for giving very valuable feedback to earlier versions of this manuscript, and for pointing out that factorization can be used to show the parameterization criterion to hold. Thanks also go to an anonymous reviewer of the paper (Lücke et al., 2021) who pointed out factorization for the Gaussian case.

Appendix

Here we show that the two statements in Lemma 1 are actually equivalent to the parameterization criterion given in Definition C if we assume a few mild additional conditions. First, we claim that part A of Lemma 1 is equivalent to part A of the parameterization criterion. For this we do not need any additional conditions as we can see in the following lemma:

Lemma 2. (Equivalence for the Prior)

Consider an EF generative model as given in Definition B with a mapping to the natural parameters of the prior given by $\vec{\zeta}(\vec{\Psi})$. The dimensionalities of the vectors $\vec{\zeta}$ and $\vec{\Psi}$ are K and R, respectively. Then the following two statements are equivalent:

(A) There exists a vectorial function $\vec{\alpha}(\cdot)$ from the set of parameters $\vec{\Psi}$ to $\mathbb{R}^R$ that satisfies the equation

$$\vec{\zeta}(\vec{\Psi}) = \frac{\partial \vec{\zeta}(\vec{\Psi})}{\partial \vec{\Psi}^{\mathrm{T}}} \vec{\alpha}(\vec{\Psi}). \qquad (123)$$

(B) For any vectorial function $\vec{f}(\Phi, \vec{\Psi})$ from the sets of parameters Φ and $\vec{\Psi}$ to the (K-dim) space of natural parameters of the prior it holds:

$$\frac{\partial \vec{\zeta}^{\mathrm{T}}(\vec{\Psi})}{\partial \vec{\Psi}} \vec{f}(\Phi, \vec{\Psi}) \;=\; \vec{0} \;\;\Rightarrow\;\; \vec{\zeta}^{\mathrm{T}}(\vec{\Psi}) \vec{f}(\Phi, \vec{\Psi}) \;=\; 0. \qquad (124)$$

Proof

We may rewrite part B using linear algebra. In fact, for fixed Φ and $\vec{\Psi}$, the first equation in (124) is equivalent to the statement that the vector $\vec{f}(\Phi, \vec{\Psi})$ is an element of the kernel of the Jacobian matrix $\frac{\partial \zeta(\vec{\Psi})}{\partial \vec{\Psi}^{\mathrm{T}}}$. So, because $\vec{\zeta}^{\mathrm{T}}(\vec{\Psi})\vec{f}(\vec{\Psi})$ is the standard Euclidean inner product, part B is equivalent to

$$\vec{\zeta}(\vec{\Psi}) \perp \ker \frac{\partial \vec{\zeta}^{\mathrm{T}}(\vec{\Psi})}{\partial \vec{\Psi}}. \qquad (125)$$

In other words, $\vec{\zeta}(\vec{\Psi})$ is in the orthogonal complement of the Jacobian, i.e.

$$\vec{\zeta}(\vec{\Psi}) \in \left(\ker \frac{\partial \vec{\zeta}^{\mathrm{T}}(\vec{\Psi})}{\partial \vec{\Psi}}\right)^{\perp}. \qquad (126)$$

Denoting the image of a matrix A by $\operatorname{im} A$ and using the well known identities $\ker A = (\operatorname{im} A^{\mathrm{T}})^{\perp}$ and $((\operatorname{im} A)^{\perp})^{\perp} = \operatorname{im} A$ for an arbitrary matrix A, we can rewrite Eqn. 126 and hence part B in the form

$$\vec{\zeta}(\vec{\Psi}) \in \operatorname{im} \frac{\partial \vec{\zeta}(\vec{\Psi})}{\partial \vec{\Psi}^{\mathrm{T}}}. \tag{127}$$

By definition of $\operatorname{im} \frac{\partial \vec{\zeta}(\vec{\Psi})}{\partial \vec{\Psi}^{\mathrm{T}}}$, this is equivalent to part A and the lemma is proved.
□

The equivalence of part B of Lemma 1 and part B of the parameterization criterion (Definition C) is somewhat more difficult to show. For the proof we interpret the integral in (34) as a bounded linear operator between two suitable Hilbert spaces and compute its adjoint. More precisely, if we denote the latent space by $\Omega_{\vec{z}}$ and the dimensionality of the natural parameter vector $\vec{\eta}$ by L, we can define for any variational parameter Φ and any $n \in \{1, \dots, N\}$ the Hilbert space[2]

$$L^2_{n,\Phi}(\Omega_{\vec{z}}, \mathbb{R}^L) := \{\vec{g}_n \colon \Omega_{\vec{z}} \to \mathbb{R}^L \mid \int q^{(n)}_{\Phi}(\vec{z})\, |\vec{g}_n(\vec{z})|^2 \,\mathrm{d}\vec{z} < \infty\} \tag{128}$$

of square integrable functions w.r.t. the weight $q^{(n)}_{\Phi}(\vec{z})$ and equip it with the inner product

$$\langle \vec{f}_n, \vec{g}_n \rangle_{n,\Phi} := \int q^{(n)}_{\Phi}(\vec{z})\, \vec{f}^{\mathrm{T}}_n(\vec{z}) \vec{g}_n(\vec{z}) \,\mathrm{d}\vec{z}, \quad \vec{f}_n, \vec{g}_n \in L^2_{n,\Phi}(\Omega_{\vec{z}}, \mathbb{R}^L). \tag{129}$$

In the case when $\vec{z}$ is discrete, we have to replace the integrals by sums again. Even then, $L^2_{n,\Phi}(\Omega_{\vec{z}}, \mathbb{R}^L)$ with the given inner product is still a Hilbert space. Alternatively, we could deal with Lebesgue integrals and general measures to summarise the discrete and continuous case. This would also have the advantage that we could absorb the weight $q^{(n)}_{\Phi}(\vec{z})$ into the measure. But for convenience and by following the standard notation used in Machine Learning, we do not work with general Lebesgue integrals. Either way, the Cartesian product

$$L^2_{1,\Phi}(\Omega_{\vec{z}}, \mathbb{R}^L) \times \cdots \times L^2_{N,\Phi}(\Omega_{\vec{z}}, \mathbb{R}^L) \tag{130}$$

together with the inner product

$$\langle (\vec{f}_1, \dots, \vec{f}_N), (\vec{g}_1, \dots, \vec{g}_n) \rangle_{\Phi} := \sum_{n=1}^{N} \langle \vec{f}_n, \vec{g}_n \rangle_{n,\Phi} \tag{131}$$

is also a Hilbert space. Since we have to deal with the Jacobian matrix $\frac{\partial \vec{\eta}^{\mathrm{T}}(\vec{z};\vec{\Theta})}{\partial \vec{\theta}}$, we also introduce for any variational parameter Φ and any $n \in \{1, \dots, N\}$ the Banach space

$$L^2_{n,\Phi}(\Omega_{\vec{z}}, \mathbb{R}^{S \times L}) := \{J \colon \Omega_{\vec{z}} \to \mathbb{R}^{S \times L} \mid \int q^{(n)}_{\Phi}(\vec{z})\, \|J(\vec{z})\|^2 \,\mathrm{d}\vec{z} < \infty\}, \tag{132}$$

where S is the dimension of the subset $\vec{\theta}$ of the noise parameters $\vec{\Theta}$ we use to compute the Jacobian, and $\|\cdot\|$ an arbitrary matrix norm on $\mathbb{R}^{S \times L}$.

[2] We should actually consider equivalence classes. But we will not do this for reasons of convenience.

From now on, for any fixed $\vec{\Theta}$, we will assume the mapping $\vec{\eta}(\vec{z};\Theta)$ to be an element of all Hilbert spaces $L^2_{n,\Phi}(\Omega_{\vec{z}},\mathbb{R}^L)$, and the transposed Jacobian $\frac{\partial\vec{\eta}^{\mathrm{T}}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}}$ to be an element of all Banach spaces $L^2_{n,\Phi}(\Omega_{\vec{z}},\mathbb{R}^{S\times L})$. This assumptions are quite mild, because often only some expectation values w.r.t. the variational distributions $q^{(n)}_{\Phi}(\vec{z})$ need to exist, which is usually the case. Especially, if the latent space $\Omega_{\vec{z}}$ is finite, since in this case the integrals are replaced by finite sums, which always exist.

After these preparations, we can define the linear bounded operator we want to use for rewriting part B of Lemma 1. Because the functions $\vec{g}_n(\vec{z};\vec{\Theta})$ are arbitrary, we can omit their parameter $\vec{\Theta}$ to get the same statement. Hence, for all parameters Φ and $\vec{\Theta}$ we define the linear Operator

$$A_{\Phi,\vec{\Theta}}\colon L^2_{1,\Phi}(\Omega_{\vec{z}},\mathbb{R}^L)\times\cdots\times L^2_{N,\Phi}(\Omega_{\vec{z}},\mathbb{R}^L)\to\mathbb{R}^S, \tag{133}$$

$$(\vec{g}_1,\ldots,\vec{g}_N)\mapsto\sum_{n=1}^{N}\int q^{(n)}_{\Phi}(\vec{z})\,\frac{\partial\vec{\eta}^{\mathrm{T}}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}}\vec{g}_n(\vec{z})\,\mathrm{d}\vec{z}. \tag{134}$$

Remember that the transposed Jacobian $\frac{\partial\vec{\eta}^{\mathrm{T}}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}}$ is a $(S\times L)$-matrix, where L is the dimension of the natural parameter vector $\vec{\eta}$ and S the number of parameters $\vec{\theta}$ we use for the computation of the Jacobian $\frac{\partial\vec{\eta}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}^{\mathrm{T}}}$. Using the triangle inequality und Hölder's inequality it is easy to see, that $A_{\Phi,\vec{\Theta}}$ is well-defined and bounded. The boundedness is important for the adjoint to be defined on the hole Hilbert space $\mathbb{R}^S$.

Lemma 3. (The Adjoint of $A_{\Phi,\vec{\Theta}}$)

The adjoint operator of $A_{\Phi,\vec{\Theta}}$ is given by

$$A^{\dagger}_{\Phi,\vec{\Theta}}\colon\mathbb{R}^S\to L^2_{1,\Phi}(\Omega_{\vec{z}},\mathbb{R}^L)\times\cdots\times L^2_{N,\Phi}(\Omega_{\vec{z}},\mathbb{R}^L),\quad\vec{\beta}\mapsto A^{\dagger}_{\Phi,\vec{\Theta}}\vec{\beta}, \tag{135}$$

in which $(A^{\dagger}_{\Phi,\vec{\Theta}}\vec{\beta})(\vec{z})=\left(\frac{\partial\vec{\eta}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}^{\mathrm{T}}}\vec{\beta},\ldots,\frac{\partial\vec{\eta}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}^{\mathrm{T}}}\vec{\beta}\right)$.

Proof

For any $\vec{g}=(\vec{g}_1,\ldots,\vec{g}_N)$, $\vec{g}_n\in L^2_{n,\Phi}(\Omega_{\vec{z}},\mathbb{R}^L)$, and any $\vec{\beta}\in\mathbb{R}^S$ we have to show that the equation

$$\langle A_{\Phi,\vec{\Theta}}\vec{g}\rangle_{\mathbb{R}^S}=\langle\vec{g},A^{\dagger}_{\Phi,\vec{\Theta}}\vec{\beta}\rangle_{\Phi} \tag{136}$$

is fulfilled. This can be done by a direct computation:

$$\langle A_{\Phi,\vec{\Theta}}\vec{g}\rangle_{\mathbb{R}^S} = \langle A_{\Phi,\vec{\Theta}}(\vec{g}_1,\dots,\vec{g}_N),\vec{\beta}\rangle_{\mathbb{R}^S} \tag{137}$$

$$= \langle\sum_{n=1}^{N}\int q^{(n)}(\vec{z})\,\frac{\partial\vec{\eta}^{\mathrm{T}}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}}\vec{g}_n(\vec{z})\,\mathrm{d}\vec{z},\vec{\beta}\rangle_{\mathbb{R}^S} \tag{138}$$

$$= \sum_{n=1}^{N}\int q^{(n)}(\vec{z})\,\langle\frac{\partial\vec{\eta}^{\mathrm{T}}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}}\vec{g}_n(\vec{z}),\vec{\beta}\rangle_{\mathbb{R}^S}\,\mathrm{d}\vec{z} \tag{139}$$

$$= \sum_{n=1}^{N}\int q^{(n)}(\vec{z})\,\langle\vec{g}_n(\vec{z}),\frac{\partial\vec{\eta}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}^{\mathrm{T}}}\vec{\beta}\rangle_{\mathbb{R}^L}\,\mathrm{d}\vec{z} \tag{140}$$

$$= \langle\vec{g},A^{\dagger}_{\Phi,\vec{\Theta}}\vec{\beta}\rangle_{\Phi}. \tag{141}$$

□

To show the equivalence of part B of Lemma 1 and part B of the parameterization criterion as given in Definition C, we need one more additional assumption. That this assumption is not actually necessary for the convergence of the ELBO to entropy sums is discussed below.

Lemma 4. (Equivalence for the Noise Model)

For an EF generative model as in Definition B, let the natural parameter vector $\vec{\eta}(\vec{z};\vec{\Theta})$ be in the Hilbert space $L^2_{n,\Phi}(\Omega_{\vec{z}},\mathbb{R}^L)$ and its Jacobian $\frac{\partial\vec{\eta}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}^{\mathrm{T}}}$ in the Banach space $L^2_{n,\Phi}(\Omega_{\vec{z}},\mathbb{R}^{S\times L})$ for all parameters Φ and $\vec{\Theta}$ and all $n\in\{1,\dots,N\}$. Moreover, assume that the variational distributions are strictly positive, i.e. $q^{(n)}_{\Phi}(\vec{z})>0$ for all $\vec{z}$. Then the following two statement are equivalent:

(A) There are a non-empty subset $\vec{\theta}$ of $\vec{\Theta}$, whose dimension we denote by S, and a vectorial function $\vec{\beta}(\vec{\Theta})$ from the set of hole parameters $\vec{\Theta}$ to the (S-dim) space of $\vec{\theta}$ that satisfies for almost every $\vec{z}$ the equation

$$\vec{\eta}(\vec{z};\vec{\Theta}) = \frac{\partial\vec{\eta}(\vec{z};\vec{\Theta})}{\partial\vec{\theta}^{\mathrm{T}}}\vec{\beta}(\vec{\Theta}). \tag{142}$$

Here "for almost every $\vec{z}$" means that the integral over the set of all $\vec{z}$ which do not fulfill the equation is zero.

(B) For any variational parameter Φ there is a non-empty subset $\vec{\theta}$ of $\vec{\Theta}$ such that the following applies for all vectorial functions $\vec{g}_1(\vec{z};\vec{\Theta}),\dots,\vec{g}_N(\vec{z};\vec{\Theta})$ from the parameter set $\vec{\Theta}$ and from the latent space $\Omega_{\vec{z}}$ to the (L-dim) space of natural parameters of the noise model:

$$\sum_{n=1}^{N}\int q^{(n)}_{\Phi}(\vec{z})\frac{\partial\vec{\eta}^{\mathrm{T}}(z;\vec{\Theta})}{\partial\vec{\theta}}\vec{g}_n(\vec{z};\vec{\Theta})\,\mathrm{d}\vec{z} = \vec{0} \tag{143}$$

$$\Rightarrow \sum_{n=1}^{N}\int q^{(n)}_{\Phi}(\vec{z})\vec{\eta}^{\mathrm{T}}(\vec{z};\vec{\Theta})\vec{g}_n(\vec{z};\vec{\Theta})\,\mathrm{d}\vec{z} = 0. \tag{144}$$

In the case when $\vec{z}$ is a discrete latent variable, the lemma is true as well if the integral is replaced by a sum over all discrete states of $\vec{z}$.

Proof

By using the operator $A_{\Phi,\vec{\Theta}}$ and the shorter notation $\vec{g} = (\vec{g}_1, ..., \vec{g}_N)$ Eqn. 143 can be rewritten as $\vec{g} \in \ker A_{\vec{\Theta}}$. Hence, because Eqn. 144 is an inner product in the Hilbert space

$$L^2_{1,\Phi}(\Omega_{\vec{z}}, \mathbb{R}^L) \times \cdots \times L^2_{N,\Phi}(\Omega_{\vec{z}}, \mathbb{R}^L), \tag{145}$$

statement B is equivalent to

$$\big(\vec{\eta}(\vec{z};\vec{\Theta}), \ldots, \vec{\eta}(\vec{z};\vec{\Theta})\big) \in \big(\ker A_{\Phi,\vec{\Theta}}\big)^{\perp}. \tag{146}$$

For a bounded linear operator between Hilbert spaces with a finite-dimensional image the two identities $\ker A = (\operatorname{im} A^\dagger)^\perp$ and $((\operatorname{im} A^\dagger)^\perp)^\perp = \overline{\operatorname{im} A^\dagger}$ hold. As before, we denote by $\operatorname{im} A^\dagger$ the image of the operator $A^\dagger$, which has a finite dimension and is hence closed. Therefore, (146) is equivalent to

$$\big(\vec{\eta}(\vec{z};\vec{\Theta}), \ldots, \vec{\eta}(\vec{z};\vec{\Theta})\big) \in \operatorname{im} A^\dagger_{\Phi,\vec{\Theta}} \tag{147}$$

By definition and by Lemma 3, there is a vetcor $\vec{\beta} \in \mathbb{R}^S$ such that

$$\big(\vec{\eta}(\vec{z};\vec{\Theta}), \ldots, \vec{\eta}(\vec{z};\vec{\Theta})\big) = \Big(\frac{\partial \vec{\eta}(\vec{z};\vec{\Theta})}{\partial \vec{\theta}^{\mathrm{T}}}\vec{\beta}, \ldots, \frac{\partial \vec{\eta}(\vec{z};\vec{\Theta})}{\partial \vec{\theta}^{\mathrm{T}}}\vec{\beta}\Big). \tag{148}$$

Because this is an equality in

$$L^2_{1,\Phi}(\Omega_{\vec{z}}, \mathbb{R}^L) \times \cdots \times L^2_{N,\Phi}(\Omega_{\vec{z}}, \mathbb{R}^L), \tag{149}$$

and all $q^{(n)}_\Phi(\vec{z})$ are non-zero, (148) is equivalent to

$$\vec{\eta}(\vec{z};\vec{\Theta}) = \frac{\partial \vec{\eta}(\vec{z};\vec{\Theta})}{\partial \vec{\theta}^{\mathrm{T}}}\vec{\beta}, \tag{150}$$

for almost every $\vec{z}$. Because the Operator $A_{\Phi,\vec{\Theta}}$ and therefore the adjoint $A^\dagger_{\Phi,\vec{\Theta}}$ depend on the parameters, the vector $\vec{\beta}$ may also depend on Φ and $\vec{\Theta}$. But since the natural parameter vector $\vec{\eta}(\vec{z};\vec{\Theta})$ is independent of Φ, we get statement A.

References

J. Aneja, A. Schwing, J. Kautz, and A. Vahdat. A contrastive learning approach for training variational autoencoder priors. *Advances in Neural Information Processing Systems*, 34:480–493, 2021.

A. Banerjee, S. Merugu, I. S. Dhillon, J. Ghosh, and J. Lafferty. Clustering with bregman divergences. *Journal of machine learning research*, 6(10), 2005.

D. Bartholomew and M. Knott. Latent variable analysis and factor analysis, 1987.

P. J. Bickel and K. A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*. Chapman and Hall/CRC, 2015.

C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks. Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7327–7347, 2022.

B. Dai, Y. Wang, J. Aston, G. Hua, and D. Wipf. Connections with robust PCA and the role of emergent sparsity in variational autoencoder models. *The Journal of Machine Learning Research*, 19(1):1573–1614, 2018.

S. Damm*, D. Forster, D. Velychko, Z. Dai, A. Fischer, and J. Lücke*. The ELBO of Variational Autoencoders Converges to a Sum of Entropies. *Proc. AISTATS*, 3931-3960, 2023. *joint main contributions

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

B. Efron. *Exponential Families in Theory and Practice*. Cambridge University Press, 2023.

B. Everitt. *An Introduction to Latent Variable Models*. Chapman and Hall, 1984.

W. Freeman and Y. Weiss. On the fixed points of the max-product algorithm. *IEEE Trans. on Info. Theory*, 2000.

B. Frey. Turbo factor analysis. *Adv. Neural Information Processing*, 1999.

G. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

M. D. Hoffman, D. M. Blei, C. Wang, and J. W. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

G. Holtzman, A. Soffer, and D. Vilenchik. A greedy anytime algorithm for sparse pca. In *Conference on Learning Theory*, pages 1939–1956. PMLR, 2020.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

J. Lucas, G. Tucker, R. B. Grosse, and M. Norouzi. Don't blame the ELBO! A linear VAE perspective on posterior collapse. In *Advances in Neural Information Processing Systems*, pages 9408–9418, 2019.

J. Lücke and D. Forster. k-means as a variational EM approximation of Gaussian mixture models. *Pattern Recognition Letters*, 125:349–356, 2019.

J. Lücke and M. Henniges. Closed-form entropy limits - A tool to monitor likelihood optimization of probabilistic generative models. In *Proc. AISTATS*, pages 731–740. PMLR, 2012.

J. Lücke, D. Forster, and Z. Dai. The evidence lower bound of variational autoencoders converges to a sum of three entropies. *Earlier version of Damm et al., 2023, arXiv:2010.14860*, 2021.

A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

G. McLachlan and D. Peel. *Finite mixture models*. John Wiley & Sons, 2004.

R. Neal and G. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*. Kluwer, 1998.

R. M. Neal. Connectionist learning of belief networks. *Artificial intelligence*, 56(1):71–113, 1992.

F. Nielsen and R. Nock. Entropies and cross-entropies of exponential families. In *2010 IEEE International Conference on Image Processing*, pages 3621–3624. IEEE, 2010.

B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–9, 1996.

B. Olshausen and K. Millman. Learning sparse codes with a mixture-of-Gaussians prior. *Advances in Neural Information Processing Systems*, 12:841–847, 2000.

M. Opper and D. Saad. *Advanced mean field methods: Theory and practice*. MIT press, 2001.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, 2014.

S. T. Roweis. EM algorithms for PCA and SPCA. In *Advances in Neural Information Processing Systems*, pages 626–632, 1998.

L. K. Saul, T. Jaakkola, and M. I. Jordan. Mean field theory for sigmoid belief networks. *Journal of artificial intelligence research*, 4(1):61–76, 1996.

M. Tipping and C. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B*, 61, 1999.

J. Tomczak and M. Welling. Vae with a vampprior. In *Proc. AISTATS*, pages 1214–1223. PMLR, 2018.

M. J. Wainwright and M. I. Jordan. *Graphical models, exponential families, and variational inference*. Foundations and trends in machine learning. Now Publisher, 2008.

Y. Weiss and W. T. Freeman. Correctness of belief propagation in Gaussian graphical models of arbitrary topology. In *Advances in Neural Information Processing Systems*, pages 673–679, 2000.

M. Welling and Y. W. Teh. Approximate inference in Boltzmann machines. *Artificial Intelligence*, 143(1):19–50, 2003.

H. Xu, C. Caramanis, and S. Mannor. Principal component analysis with contaminated data: The high dimensional case. In *Conference on Learning Theory*, 2010.

J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In *Advances in Neural Information Processing Systems*, pages 689–695, 2001.