Closed-Form Entropy Limits – A Tool to Monitor Likelihood Optimization of Probabilistic Generative Models

Jörg Lücke and Marc Henniges
FIAS, Goethe-University Frankfurt, 60438 Frankfurt, Germany

Abstract

The maximization of the data likelihood under a given probabilistic generative model is the essential goal of many algorithms for unsupervised learning. If expectation maximization is used for optimization, a lower bound on the data likelihood, the free-energy, is optimized. The parameter-dependent part of the free-energy (the difference between free-energy and posterior entropy) is the essential entity in the derivation of learning algorithms. Here we show that for many common generative models the optimal values of the parameter-dependent part can be derived in closed-form. These closed-form expressions are hereby given as sums of the negative (differential) entropies of the individual model distributions. We apply our theoretical results to derive such closed-form expressions for a number of common and recent models, including probabilistic PCA, factor analysis, different versions of sparse coding, and Linear Dynamical Systems. The main contribution of this work is theoretical but we show how the derived results can be used to efficiently compute free-energies, and how they can be used for consistency checks of learning algorithms.

1 Introduction

A standard way to derive algorithms for unsupervised learning is based on probabilistic generative models. They represent parameterized models of the data generation process. Learning algorithms derived from generative models seek those parameters of the model that best match the distribution of a given set of data points. A very common criteria for the quality of such a match is the likelihood of the data under the generative model. To derive maximum likelihood algorithms, expectation maximization (EM) is one of the most widely used approaches. Instead of maximizing the likelihood directly, EM iteratively maximizes a lower bound, the free-energy. In this way parameter update equations can conveniently be derived, and approximations, e.g., in the form of variational EM (Jordan et al., 1999; Jaakkola, 2000), can be introduced. The crucial function for the derivation of parameter update-rules is hereby the parameter-dependent part of the free-energy. Prominent learning algorithms based on generative models using EM include probabilistic PCA (p-PCA; Roweis, 1998; Tipping and Bishop, 1999) and factor analysis (FA; see, e.g. Everitt, 1984), different versions of sparse coding (SC; e.g. Olshausen and Field, 1996) or Linear Dynamical Systems (LDS; see, e.g., Bishop, 2006). The goal of all these algorithms is the maximization of the data likelihood. Unfortunately, it is in general not possible to know the optimally achievable likelihood values for a given model. The same applies for optimal values of the free-energy. In this work, we show, however, that for many common models it is possible to analytically derive closed-form expressions for the parameter dependent part of the free-energy. These expressions will still not allow to compute the optimal possible likelihood values but they can be used as a tool to check for model consistencies and to efficiently compute free-energies.

We will first show how closed-form expressions are obtained in the limit of large sample sizes and at global optima, and we will list formulas for a number of common models. For a large class of models, we will then show that the same expressions provide the convergence values also for finite sample sizes, at local optima, model/data mismatches, and for approximate posterior distributions. Finally, we give examples for the concrete applicability of the results.

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

2 Maximum Likelihood and EM

Let us consider N data points, $\{\vec{y}^{(n)}\}_{n=1,\ldots,N}$, independently drawn from the same underlying distribution $p(\vec{y})$. A generative model of the data is given by a parameterized distribution $p(\vec{y}\,|\,\Theta)$ with Θ as the set of parameters. Elementary generative models usually take the form of latent variable models with distributions $p(\vec{y}\,|\,\Theta)$ consisting of a prior distribution over latent variables $p(\vec{z}\,|\,\Theta)$ and a distribution over observed variables given the latents $p(\vec{y}\,|\,\vec{z},\Theta)$: $p(\vec{y}\,|\,\Theta) = \int p(\vec{y}\,|\,\vec{z},\,\Theta)\,p(\vec{z}\,|\,\Theta)\,d\vec{z}$. By repeating this step for the latent variable distribution $p(\vec{z}\,|\,\Theta)$ itself, more complex hierarchical models can be obtained.

Many learning algorithms seek parameters $\Theta^{\dagger}$ that maximize the data likelihood $\mathcal{L}(\Theta)$ under a given generative model,

$$\Theta^{\dagger} = \mathrm{argmax}_{\Theta}\{\mathcal{L}(\Theta)\} \tag{1}$$

$$\text{where} \quad \mathcal{L}(\Theta) = \frac{1}{N}\sum_{n=1}^{N}\log\big(p(\vec{y}^{(n)}\,|\,\Theta)\big). \tag{2}$$

Note that we use the data likelihood normalized by the number of data points. Maximizing this form of the likelihood is equivalent to maximizing $N\mathcal{L}(\Theta)$, which is usually used.

To find the parameters $\Theta^{\dagger}$ at least approximately, a wide range of algorithms rely on the EM formalism as it appears, e.g., in (Neal and Hinton, 1998). That is, instead of maximizing $\mathcal{L}(\Theta)$ directly, a lower-bound, the so-called free-energy $\mathcal{F}(\Theta',\Theta)$, is optimized:

$$\mathcal{L}(\Theta) \geq \mathcal{F}(\Theta',\Theta) = \mathcal{Q}(\Theta',\Theta) + \tfrac{1}{N}\sum_{n=1}^{N}\mathcal{H}(p(\vec{z}\,|\,\vec{y}^{(n)},\Theta')),$$

$$\text{where} \quad \mathcal{Q}(\Theta',\Theta) = \tfrac{1}{N}\textstyle\sum_{n=1}^{N}\int p(\vec{z}\,|\,\vec{y}^{(n)},\,\Theta') \times \log\big(p(\vec{y}^{(n)},\,\vec{z}\,|\,\Theta)\big)\,d\vec{z}, \tag{3}$$

$$\mathcal{H}(p) = -\int p(\vec{x})\log(p(\vec{x}))\,d\vec{x}\,. \tag{4}$$

where $\mathcal{H}(p)$ is the (differential) entropy of a distribution p. Note that, more generally, the free-energy depends on a distribution q and parameters Θ, $\mathcal{F}(q,\Theta)$ (see, e.g., Neal and Hinton, 1998). This allows for the introduction of approximate EM, and we will come back to this formulation later on.

The dependency on two sets of parameters, Θ' and Θ, distinguishes the free-energy from the likelihood. It allows for an iterative optimization with respect to the different sets, which facilitates the derivation of parameter update rules. A resulting algorithm for likelihood optimization can formally be stated as:

$$\Theta' = \Theta^{\text{new}} \quad \text{and} \quad \Theta^{\text{new}} = \mathrm{argmax}_{\Theta}\{\mathcal{F}(\Theta',\Theta)\} \tag{5}$$

The first step merely consists of setting the old parameters to the new ones and the second maximizes the free-energy with respect to its second argument. The entropy term in the free-energy does not depend on the parameters that are optimized, however. The function important for the optimization is thus given by $\mathcal{Q}(\Theta',\Theta)$ in (3), which we will refer to here (in analogy to 'free-energy') as the *inner-energy*. Furthermore, we define by $\mathcal{Q}(\Theta) = \mathcal{Q}(\Theta,\Theta)$ the inner-energy with equal arguments—an expression that will be of convenience later on. Using the inner-energy (3) we can now simplify algorithm (5) which becomes:

$$\Theta' = \Theta^{\text{new}} \quad \text{and} \quad \Theta^{\text{new}} = \mathrm{argmax}_{\Theta}\{\mathcal{Q}(\Theta',\Theta)\} \tag{6}$$

Algorithm (6) guarantees that the free-energy as well as the likelihood is never decreased. In practice, it increases the data likelihood at least to a local optimum. An optimum is hereby reached when the parameters Θ have converged (in practice this means after the changes of the parameters in any further EM iteration are negligible: $\Theta \approx \Theta'$).

3 Entropy Limits of the Inner-Energy

Algorithm (6) shows that the inner-energy $\mathcal{Q}(\Theta',\Theta)$ emerges as the crucial function to derive learning algorithms. By executing an EM-based learning algorithm, $\mathcal{Q}(\Theta',\Theta)$ will change until it converges to a limit value. For most generative models the maximization of the likelihood is a non-convex problem and the convergence point of the parameters Θ may represent a local optimum. The best possible outcome of the algorithm is thus $\Theta = \Theta^{\dagger}$, where $\Theta^{\dagger}$ are the global maximum likelihood parameters in (1). The parameters $\Theta^{\dagger}$ are optimal, given a sample of the underlying data distribution. For a finite number of data points N, the parameters $\Theta^{\dagger}$ depend on the sample. However, in the limit of infinitely many data points, $\Theta^{\dagger}$ becomes independent of the sample. We will refer to the global maximum likelihood parameters in the limit $N \to \infty$ as Θ^{*} (note that there may be multiple solutions).

Our first observation is that in the limit we can derive formulas for the value $\mathcal{Q}(\Theta',\Theta)$ converges to if the parameters converge to a maximum likelihood solution Θ^{*}. That is, given a generative model, we can derive an analytical expression that provides the convergence value of $\mathcal{Q}(\Theta',\Theta)$ if the algorithm reaches its goal. While such limits of convergence can formally be stated for all generative models, a main result of this work is that we can obtain closed-form formulas for these limit values. For specific models such as sparse coding, we will further see that the convergence values will only depend on a subset of model parameters for many models. For standard sparse coding they are, e.g., independent of the basis functions.

To obtain closed-form expressions, an important observation is that many standard generative models have

a specific structure: the entropy of their noise distributions is independent of the values of the latent variables. As example consider the standard sparse coding model given by:

$$p(\vec{z}\,|\,\Theta) \;=\; \prod_{h=1}^{H} \tfrac{1}{2\gamma} \exp\big(-\tfrac{1}{\gamma}|z_h|\big) \qquad (7)$$

$$p(\vec{y}\,|\,\vec{z},\Theta) \;=\; \mathcal{N}(\vec{y}; W\vec{z}, \sigma^2\,\mathbb{1}) \qquad (8)$$

The model (with $\gamma = 1$) has famously been applied to patches of images in (Olshausen and Field, 1996). By considering the entropy of the noise model (8), we can instantly observe that it is independent of the latent vector $\vec{z}$. Other well-known models that satisfy this property are p-PCA (Roweis, 1998; Tipping and Bishop, 1999) or factor analysis and a large set of further models including almost all sparse coding variants. Using this property, we can re-express the values of $\mathcal{Q}(\Theta) = \mathcal{Q}(\Theta,\Theta)$ in the likelihood optimum and in the limit $N \to \infty$ by the (differential) entropies of the constituting distributions:

Proposition 1
Consider a generative model with prior $p(\vec{z}\,|\,\Theta)$ and noise distribution $p(\vec{y}\,|\,\vec{z},\Theta)$. For this model let $\mathcal{Q}(\Theta',\Theta)$ be the function defined by (3), and let $\{\vec{y}^{(n)}\}_{n=1,\ldots,N}$ be the data sample to which the model is applied.

If there exist parameters Θ such that the underlying data distribution $p(\vec{y})$ can exactly be matched by the generative distribution $p(\vec{y}\,|\,\Theta)$, and if $\mathcal{H}\big(p(\vec{y}\,|\,\vec{z},\Theta)\big)$ is independent of $\vec{z}$, it applies in the limit of $N\to\infty$ that

$$\mathcal{Q}(\Theta^*) \;=\; -\,\mathcal{H}\big(p(\vec{y}\,|\,\vec{z},\,\Theta^*)\big) \;-\; \mathcal{H}\big(p(\vec{z}\,|\,\Theta^*)\big) \qquad (9)$$

for any global maximum likelihood parameters Θ^*.

Proof
Let us first reiterate the standard procedure of taking the limit of infinitely many data points (also compare alternative proof, see Supplement): If the data points $\vec{y}^{(n)}$ are identically and independently drawn from the distribution $p(\vec{y})$, we get:

$$\lim_{N\to\infty} \tfrac{1}{N}\sum_{n=1}^{N} f(\vec{y}^{(n)}) \;=\; \int p(\vec{y})\, f(\vec{y})\, d\vec{y}\,, \quad (10)$$

for any well-behaved function $f(\vec{y})$. We first show that parameters Θ satisfying $p(\vec{y}\,|\,\Theta) = p(\vec{y})$ (at least in the distribution sense) represent maximum likelihood solutions in the limit of $N \to \infty$. By applying the well-known formula (10) we can (in the limit $N \to \infty$) rewrite the likelihood (2) as:

$$\mathcal{L}(\Theta) \;=\; -\,\mathrm{KL}(p(\vec{y}), p(\vec{y}\,|\,\Theta)) \;-\; \mathcal{H}(p(\vec{y})), \qquad (11)$$

which is also a standard result. The maximum of (11) is given by the minimum of $\mathrm{KL}(p(\vec{y}), p(\vec{y}\,|\,\Theta)) \geq 0$. The KL-divergence is zero, however, if and only if $p(\vec{y}\,|\,\Theta) = p(\vec{y})$. The existence of parameters satisfying $p(\vec{y}\,|\,\Theta) = p(\vec{y})$ thus implies that for any maximum likelihood solution Θ^* in (1) applies $p(\vec{y}\,|\,\Theta^*) = p(\vec{y})$.

For the inner-energy $\mathcal{Q}$ it then follows for Θ^*:

$$\begin{aligned}
&\mathcal{Q}(\Theta^*)\\
&= \lim_{N\to\infty} \tfrac{1}{N}\textstyle\sum_{n=1}^{N} \int p(\vec{z}\,|\,\vec{y}^{(n)},\Theta^*) \log\big(p(\vec{y}^{(n)},\,\vec{z}\,|\,\Theta^*)\big)\, d\vec{z}\\
&= \textstyle\int\int p(\vec{y})\, p(\vec{z}\,|\,\vec{y},\Theta^*)\, \log\big(p(\vec{y},\,\vec{z}\,|\,\Theta^*)\big)\, d\vec{z}\, d\vec{y}\\
&= \textstyle\int\int p(\vec{y}\,|\,\Theta^*)\, p(\vec{z}\,|\,\vec{y},\Theta^*)\, \log\big(p(\vec{y},\,\vec{z}\,|\,\Theta^*)\big)\, d\vec{z}\, d\vec{y}\\
&= \textstyle\int\int p(\vec{y},\,\vec{z}\,|\,\Theta^*)\, \log\big(p(\vec{y},\,\vec{z}\,|\,\Theta^*)\big)\, d\vec{z}\, d\vec{y}\\
&= -\mathcal{H}(p(\vec{y},\,\vec{z}\,|\,\Theta^*)). \qquad (12)
\end{aligned}$$

Inserting prior and noise distributions into this formula, we obtain[1]:

$$\begin{aligned}
&Q(\Theta^*) = \textstyle\int\int p(\vec{y},\,\vec{z}\,|\,\Theta^*)\, \log\big(p(\vec{y},\,\vec{z}\,|\,\Theta^*)\big)\, d\vec{z}\, d\vec{y}\\
&= \textstyle\int p(\vec{z}\,|\,\Theta^*)\Big(-\mathcal{H}\big(p(\vec{y}\,|\,\vec{z},\Theta^*)\big)\Big)\, d\vec{z} \;-\; \mathcal{H}\big(p(\vec{z}\,|\,\Theta^*)\big)\\
&= -\,\mathcal{H}\big(p(\vec{y}\,|\,\vec{z},\Theta^*)\big) \;-\; \mathcal{H}\big(p(\vec{z}\,|\,\Theta^*)\big),
\end{aligned}$$

where we have used the assumed property that the entropy of the noise distribution is independent of $\vec{z}$.
□

The main content of Proposition 1 should not be confused with the well-know result of equality of likelihood and negative entropy in the limit of infinitely many data points (compare alternative proof of Prop. 1 which actually starts with this fact, see Supplement). The main point is the relatively straight-forward observation that the inner-energy at the optimum is a simple sum of two (negative) entropies if the entropy of the observed variables is independent of the hidden variables (Eqn.9). If the entropy of the observed variables does depend on the hidden variables (e.g., for a Bernoulli noise model), Eqn. 9 does not apply. However, for many models it does. In this case note that the entropies of the individual model distributions are often given in closed-form. It thus follows that closed-form expressions for the limits of $\mathcal{Q}$ can be derived in many cases while closed-form solutions for the likelihood and the free-energy at the same time both do not exist. An example will be discussed later.

4 Generalization to Graphical Models

Many standard generative models take the form of a single prior and a single noise distribution. Many others, however, show more complex dependencies amongst observed and hidden variables. This raises the question of whether the result of Proposition 1 can be generalized. For this, first consider the directed acyclic graph G in Fig. 1A which is an

[1]Note that we can alternatively derive this result via the data likelihood (see Supplement).

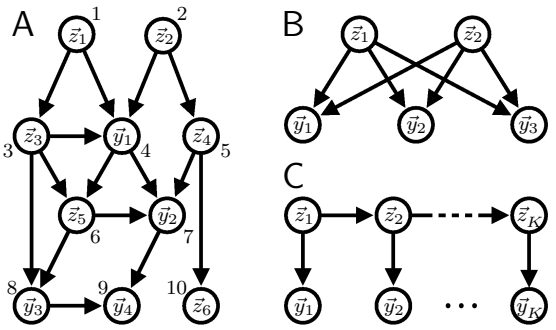

Figure 1: Bayesian networks. **A** Network with complex dependencies between hidden and observed units. Numbering outside the nodes reflects partial order of the graph. **B** Network for p-PCA, FA, and SC. **C** Network for LDS.

example of such a more complex generative model. Let $\vec{Y} = (\vec{y}_1, \ldots, \vec{y}_D)$ denote the observed variables of the model and let $\vec{Z} = (\vec{z}_1, \ldots, \vec{z}_H)$ denote the hidden variables. The joint distribution of the generative model can then be written as

$$p(\vec{Y}, \vec{Z} \,|\, \Theta) = \Big(\prod_{d=1}^{D} p(\vec{y}_d \,|\, \mathrm{pa}_d, \Theta)\Big)\Big(\prod_{h=1}^{H} p(\vec{z}_h \,|\, \mathrm{pa}_h, \Theta)\Big), \quad (13)$$

where pa_k denote the parents of node k according to the graph G. The data generation process is fully defined if the conditional probabilities in (13) are given. Let us denote the data that is generated by the model by $\{\vec{Y}^{(n)}\}_{n=1,\ldots,N}$ where $\vec{Y}^{(n)} = (\vec{y}_1^{\,(n)}, \ldots, \vec{y}_D^{\,(n)})$ is a generated data point. Note that the set of models given by (13) and directed acyclic graphs contains as a subset hierarchical models with layers of latents and one layer of observed variables.

Like for the elementary generative models considered in Sec. 3, we can define data likelihood, free-energy, and inner-energy for the models defined by (13). The definitions are simply given by replacing $\vec{y}$ by $\vec{Y}$ and $\vec{z}$ by $\vec{Z}$ in equations (2) to (3), respectively. Typical learning algorithms based on the generalized models again seek parameters Θ that maximize the data likelihood.

To obtain closed-form entropy limits of the inner-energy (3), we would like to express the limits in terms of entropies of their constituting distributions $p(\vec{y}_d \,|\, \mathrm{pa}_d, \Theta))$ and $p(\vec{z}_h \,|\, \mathrm{pa}_h, \Theta))$ for which closed-form expressions may be known. Such expressions again require a specific structure similar to the structure of models Proposition 1 is applicable to: For the generalized models we have to assume that the entropies of the nodes' distributions remain unaltered by changing the values of the corresponding parent nodes. If a model fulfills this assumption, the limit of the inner-energy is given by the negative sum of the entropies of the model's constituting distributions:

Proposition 2

Consider a generative model with D observed variables $\vec{Y} = (\vec{y}_1, \ldots, \vec{y}_D)^T$ and H hidden variables $\vec{Z} = (\vec{z}_1, \ldots, \vec{z}_H)^T$, and let $\{\vec{Y}^{(n)}\}_{n=1,\ldots,N}$ be the data sample the model is applied to. For the model, let there exist a Bayesian network representation G with joint probability (13) such that the entropies of the constituting distributions $p(\vec{y}_d \,|\, \mathrm{pa}_d, \Theta)$ and $p(\vec{z}_h \,|\, \mathrm{pa}_h, \Theta)$ are independent of the values of the corresponding parents.

If there now exist parameters Θ such that the underlying data distribution $p(\vec{Y})$ can exactly be matched by the generative distribution $p(\vec{Y} \,|\, \Theta)$, the following applies in the limit of $N \to \infty$ and at any global maximum likelihood solution $\Theta = \Theta^*$:

$$\mathcal{Q}(\Theta^*) = \overline{\mathcal{Q}}(\Theta^*) \quad \text{where} \quad (14)$$

$$\overline{\mathcal{Q}}(\Theta) = -\sum_{d=1}^{D} \mathcal{H}\big(p(\vec{y}_d | \,\mathrm{pa}_d, \Theta)\big) - \sum_{h=1}^{H} \mathcal{H}\big(p(\vec{z}_h | \,\mathrm{pa}_h, \Theta)\big).$$

Proof

First note that for the proof of Proposition 1 we have only used the specific dependency structure of a given generative model for the last steps (after Eqn. 12). Thus, Eqn. 12 also applies for the general case. By inserting the joint (13) given by the Bayesian network G, we have to evaluate:

$$\mathcal{Q}(\Theta^*, \Theta^*)$$
$$= -\mathcal{H}\Big(\big(\prod_{d=1}^{D} p(\vec{y}_d \,|\, \mathrm{pa}_d, \Theta^*)\big)\,\big(\prod_{h=1}^{H} p(\vec{z}_h \,|\, \mathrm{pa}_h, \Theta^*)\big)\Big) \quad (15)$$

The details of the further steps are given in the appendix. In brief, we redefine indices and variables to take the structures of a graphical model into account (compare Fig. 1A). We can then show that the product in the argument of the entropy in (15) results in a sum of individual entropies. By back-inserting the original indices and variable notations, we obtain Proposition 2.
□

In order to find a closed-form entropy limit for a given generative model, we would first write down a suitable Bayesian network representation of the model. We can then apply Proposition 2 if there exist closed-form entropy expressions of all network nodes and if the parents of each node do not change these entropies. Note that for a given generative model some Bayesian network representations may fulfill these requirements while others may not. For a generative model there may, of course, not exist a suitable Bayesian network representation at all but if it does, its entropy limit is given by Eqn. 14.

5 Application to Generative Models

As an application for the theoretical results in Propositions 1 and 2, let us derive closed-form expressions for entropy limits for a number of well-known and for recent generative models.

Probabilistic PCA and Factor Analysis. Consider the generative model for p-PCA given by

$$p(\vec{z}\,|\,\Theta) = \mathcal{N}(\vec{z};\vec{0},\mathbb{1}) \quad (16)$$

$$p(\vec{y}\,|\,\vec{z},\Theta) = \mathcal{N}(\vec{y};W\vec{z}+\vec{\mu},\sigma^2\,\mathbb{1}) \quad (17)$$

where $\mathcal{N}$ denotes a Gaussian distribution. The differential entropy of the Gaussian is easily derivable and well-known. For the prior (16) it is given by $\mathcal{H}\big(p(\vec{z}\,|\,\Theta^*)\big) = \frac{H}{2}\log(2e\pi)$, for the noise distribution (17) it is given by $\frac{D}{2}\log(2\pi e\sigma^2)$, where H is the number of hidden variables, D the number of observed variables, and where e is the Euler number (introduced to abbreviate the expressions). By applying Proposition 1 to the model in (16) and (17) the entropy limit of p-PCA is given by:

$$\overline{\mathcal{Q}}(\Theta) = -\tfrac{D}{2}\,\log(2\pi e\sigma^2) - \tfrac{H}{2}\log(2\pi e). \quad (18)$$

Note that the limit can also be obtained by applying Proposition 2 to the p-PCA graphical model with nodes containing scalar variables (compare Fig. 1B). In that case we only require the entropy of a scalar Gaussian given by $\frac{1}{2}\log(2e\pi\sigma^2)$. Using Proposition 2 it is also straight-forward to obtain the entropy limit for factor analysis (FA). FA only differs from p-PCA by having different noise variances per observed dimension (see, e.g., Bishop, 2006, for further references). The entropy limit for FA is given in Tab. 1.

Considering (18) observe that the limit for p-PCA only depends on a single parameter of the generative model: the noise variance σ^2. It is independent of the basis functions W and of the offset vector $\vec{\mu}$. Thus, for a given level of data noise, the value $\mathcal{Q}(\Theta)$ converges to is independent of the parameters defining position, orientation and parameterization of the hyperplane (the same applies for FA).

The p-PCA case is instructive because the underlying generative model is simple enough to also allow for analytical expressions for the inner-energy $\mathcal{Q}(\Theta)$. Using definition (16) and (17) we can show that:

$$\mathcal{Q}(\Theta) = -\tfrac{D}{2}\log(2\pi\sigma^2) - \tfrac{H}{2}\log(2\pi e) - \tfrac{1}{2}\mathrm{Tr}(C^{-1}S), \quad (19)$$

where S is the data covariance matrix $S = \frac{1}{N}\sum_n(\vec{y}^{\,(n)}-\vec{\mu})(\vec{y}^{\,(n)}-\vec{\mu})^T$ and where C is a function of the weights W and the noise variance σ^2. $C = W\,W^T + \sigma^2\mathbb{1}$. Two things can be observed considering expression (19): First, $\mathcal{Q}(\Theta)$ does depend on W through the matrix C. Second, if N goes to infinity, the data covariance matrix S converges to the covariance matrix of the true underlying distribution. If the data is indeed distributed according to the p-PCA generative model, this covariance matrix is given by $C = (W^*)\,(W^*)^T + (\sigma^*)^2\mathbb{1}$ (see, e.g. Tipping and Bishop, 1999) where W^* and σ^* are the maximum likelihood parameters. Thus, if Θ converges to Θ^*, expression (19) converges for $N\to\infty$ to the entropy limit $\overline{\mathcal{Q}}(\Theta^*)$ in (18) because $\mathrm{Tr}(C^{-1}S) = \mathrm{Tr}(C^{-1}C) = D$ (note the Euler number "e" in Eqn. 18).

In the case of p-PCA we have in this way confirmed Proposition 1 and 2 by deriving the result in an alternative way. Note however, that first this derivation is much less straight forward (involving lengthy matrix algebra manipulations to obtain Eqn. 19), and second, it is only possible due to the simplicity of the p-PCA generative model that has closed-from posteriors. For the sparse coding models discussed below, $Q(\Theta)$ can not be obtained in closed-form anymore while closed-form entropy limits $\overline{\mathcal{Q}}(\Theta)$ still exist.

Sparse Coding. As an example of a more complex model consider again the sparse coding generative model (7) and (8). As already mentioned earlier the (differential) entropy of the noise distribution is independent of $\vec{z}$ and given by $\frac{D}{2}\log(2\pi e\sigma^2)$ as for p-PCA. The entropy $\mathcal{H}\big(p(\vec{z}\,|\,\Theta)\big)$ is the entropy of the Laplace distribution given by $\mathcal{H}\big(p(\vec{z}\,|\,\Theta)\big) = H\,\log(2e\gamma)$, where H is the number of hidden dimensions and where γ parameterizes sparseness in the prior (7). By applying Proposition 1 (or Proposition 2) to the model (7) and (8), the entropy limit is given by:

$$\overline{\mathcal{Q}}(\Theta) = -\tfrac{D}{2}\,\log(2\pi e\sigma^2) - H\,\log(2e\gamma)\,. \quad (20)$$

Observe that the limit now depends on two parameters, the variance of the generative model, σ^2, and the prior parameter γ. Again, the entropy limit is independent of the weight matrix W, however.

Because of its popularity in Computational Neuroscience and Machine Learning, many different versions of sparse coding exist. In particular, different priors were investigated. Recent choices have, for instance, been the Student-t distribution (Osindero et al., 2006; Berkes et al., 2008),

$$p(\vec{z}\,|\,\Theta) = \prod_{h=1}^{H}\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\Big(1+\frac{z_h^2}{\nu}\Big)^{-\frac{\nu+1}{2}}, \quad (21)$$

and the Bernoulli prior for binary hidden units (Haft et al., 2004; Henniges et al., 2010),

$$p(\vec{z}\,|\,\Theta) = \prod_{h=1}^{H}\big(\lambda^{z_h}(1-\lambda)^{1-z_h}\big)\,. \quad (22)$$

While the entropy of the Bernoulli distribution is well-known, $(-(1-\lambda)\log(1-\lambda) - \lambda\log(\lambda))$, the entropy of the Student-t distribution is more intricate. However, it has been derived earlier, e.g., in (Lazo and Rathie, 1978), and is (in scalar form) given by:

Tab. 1: Entropy Limits $\overline{\mathcal{Q}}(\Theta)$		
p-PCA	$-\frac{D}{2}\,\log(2\pi e\sigma^2)\,-\,\frac{H}{2}\log(2\pi e)$	Roweis, 1998 Tipping & Bishop, 1999
Factor Analysis	$-\frac{1}{2}\sum_{d=1}^{D}\log(2\pi e\sigma_d^2)\,-\,\frac{H}{2}\log(2\pi e)$	e.g., Everitt, 1984
$\mathrm{SC}_{\mathrm{Cauchy}}$	$-\frac{D}{2}\,\log(2\pi e\sigma^2)\,-\,H\,\log(4\pi\gamma)$	Olshausen & Field, 1996
$\mathrm{SC}_{\mathrm{Laplace}}$	$-\frac{D}{2}\,\log(2\pi e\sigma^2)\,-\,H\,\log(2e\gamma)$	Olshausen & Field, 1996
$\mathrm{SC}_{\mathrm{student-t}}$	$-\frac{D}{2}\log(2\pi e\sigma^2) - H\frac{\nu+1}{2}\Big(\psi\big(\frac{\nu+1}{2}\big) - \psi\big(\frac{\nu}{2}\big)\Big) - H\log\Big(\sqrt{\nu}B(\frac{\nu}{2},\frac{1}{2})\Big)$	Osindero et al., 2006 Berkes et al., 2008
Binary SC	$-\frac{D}{2}\log(2\pi e\sigma^2) + H(1-\lambda)\log(1-\lambda) + H\lambda\log(\lambda)$	Haft et al., 2004 Henniges et al., 2010
MCA	$-\frac{D}{2}\log(2\pi e\sigma^2) + H(1-\lambda)\log(1-\lambda) + H\lambda\log(\lambda)$	Lücke & Eggert, 2010
LDS	$-\frac{1}{2}\,\log(\lvert 2\pi eV\rvert)\,-\,\frac{K-1}{2}\,\log(\lvert 2\pi e\Sigma\rvert)\,-\,\frac{K}{2}\,\log(\lvert 2\pi e\Lambda\rvert)$	Bishop, 2006 for refs

$$\mathcal{H}\big(p(z_h\,|\,\Theta)\big) \;=\; \tfrac{\nu+1}{2}\Big(\psi\big(\tfrac{\nu+1}{2}\big) - \psi\big(\tfrac{\nu}{2}\big)\Big) + \log\Big(\sqrt{\nu}B(\tfrac{\nu}{2},\tfrac{1}{2})\Big), \qquad (23)$$

where ψ is the digamma function and B the beta function. By applying Proposition 2 to the graphical model representation of SC in Fig. 1B, we can thus obtain the entropy limits for both of these sparse coding versions. The obtained expressions for $\overline{\mathcal{Q}}(\Theta)$ are listed in Tab. 1. For the BSC limit, note that the same limit is also obtained for a recently studied model of non-linear sparse coding (MCA; Lücke and Eggert, 2010). This shows that two different models can have the same entropy limit.

LDS. Finally, let us consider Linear Dynamical Systems (LDS) as an example for a generative model with dependent hidden variables. An LDS model generates data sequences $\vec{Y} = (\vec{y}_1, \ldots, \vec{y}_K)^T$. Its graphical model representation is shown in Fig. 1C, with the conditional distributions given by:

$$p(\vec{z}_1\,|\,\Theta) \;=\; \mathcal{N}(\vec{z}_1; \vec{r}, V) \qquad (24)$$

$$p(\vec{z}_k\,|\,\vec{z}_{k-1}, \Theta) \;=\; \mathcal{N}(\vec{z}_k; A\vec{z}_{k-1} + \vec{a}, \Sigma) \qquad (25)$$

$$p(\vec{y}_k\,|\,\vec{z}_k, \Theta) \;=\; \mathcal{N}(\vec{y}_k; C\vec{z}_k + \vec{c}, \Lambda) \qquad (26)$$

These distributions and the graphical model fully describe the data generation process. Considering the nodes of the Bayesian network representation in Fig. 1C, we observe that the entropy of each node is independent of the values of the parent nodes. This can instantly be seen by noting that each variable only changes the means of the childs' distributions. The assumptions of Proposition 2 are thus fulfilled. Furthermore, the entropies of the constituting distributions are known. By applying Proposition 2 we thus obtain a closed-form entropy limit $\overline{\mathcal{Q}}(\Theta)$ given by:

$$-\tfrac{1}{2}\log(\lvert 2\pi eV\rvert) - \tfrac{K-1}{2}\log(\lvert 2\pi e\Sigma\rvert) - \tfrac{K}{2}\,\log(\lvert 2\pi e\Lambda\rvert).$$

Note for this example that we have to apply Proposition 2 because of the dependencies between the latents $\vec{z}_k$. Also note that the Bayesian network representation Fig. 1C is a kind of minimal graphical model for which the assumptions of Proposition 2 are still fulfilled. Introducing more nodes by substituting $\vec{z}_k$ or $\vec{y}_k$ by graphs with nodes of their scalar entries would result in a Bayesian network violating the assumptions.

Tab. 1 summarizes the results obtained in this section and lists some more results not explicitly mentioned. The most frequent distribution used for our models is the Gaussian. In fact, all used noise distributions are Gaussian. However, the existence of closed-form entropy limits is not exclusive to this type of noise distribution. Any other distribution for which the parent variables do not change the entropy have closed-form entropy limits. An example would be the Laplace distribution (Eqn. 7). All models of Tab. 1 would still have closed-form entropy limits simply given by replacing $\frac{1}{2}\log(2\pi e\sigma^2)$ with $\log(2e\gamma)$ (the entropy of the Laplace distribution).

6 Entropy Limits for Finite Data Sets, Local Optima, and Approximate EM

Once the existence of closed-form entropy limits of the inner-energy has been observed, convergence results under milder assumptions can be investigated. It could thus be asked if statements about the limit values of $\mathcal{Q}(\Theta)$ at local optima, finite sample sizes, or for approximate EM can be made. For this consider again the free- and inner-energy, this time in their formulation for variational EM:

$$\mathcal{L}(\Theta) \geq \mathcal{F}(q,\Theta) = \mathcal{Q}(q,\Theta) + \tfrac{1}{N}\textstyle\sum_n \mathcal{H}(q_n(\vec{z};\,\Theta')), \quad (27)$$

$$\mathcal{Q}(q,\Theta) = \tfrac{1}{N}\textstyle\sum_{n=1}^{N}\int q_n(\vec{z};\,\Theta')\,\log\big(p(\vec{y}^{\,(n)},\,\vec{z}\,|\,\Theta)\big)\,d\vec{z}.$$

For the derivation of entropy limits in Propositions 1 and 2 we explicitly used properties of global optima and consistency of distributions. At first, it therefore seems unlikely that entropy limits can be obtained without these assumptions. However, we may be able to exploit specific properties of some generative models that go beyond an independence of the entropy from parent node values. As an example, let us consider the most common sparse coding model given by a Laplace prior and a Gaussian noise model (Eqns. 7 and

8). By applying EM, we obtain the M-step equations (see Supplement for the derivation):

$$\sigma^2_{\text{new}} = \frac{1}{D\,N}\sum_{n=1}^{N}\int q_n(\vec{z};\,\Theta')\,\|\vec{y}^{\,(n)} - W\vec{s}\,\|^2 \mathrm{d}\vec{z}, \quad (28)$$

$$\gamma_{\text{new}} = \frac{1}{N}\sum_{n=1}^{N}\int q_n(\vec{z};\,\Theta')\,\|\vec{z}\|_1\,\mathrm{d}\vec{z}, \quad (29)$$

and a corresponding equation for W. In the global optimum we know that $Q(\Theta)$ in Eqn. 3 converges to the entropy limit given in Eqn. 20. We therefore rewrite $Q(q,\Theta)$ in (27) as a sum of its entropy limit and additional terms:

$$\mathcal{Q}(q,\Theta) = \overline{\mathcal{Q}}(\Theta) + A(q,W,\sigma) + B(q,\gamma), \text{with} \quad (30)$$

$$A(q,W,\sigma) = -\frac{1}{2\sigma^2 N}\sum_{n=1}^{N}\int q_n(\vec{z};\,\Theta') \times\|\vec{y}^{\,(n)} - W\vec{z}\|^2\,d\vec{z} + \frac{D}{2}$$

$$B(q,\gamma) = -\frac{H}{\gamma N}\sum_{n=1}^{N}\int q_n(\vec{z};\,\Theta')\,\|\vec{z}\|_1\,d\vec{z} + H$$

If the assumptions of Proposition 1 are fulfilled, the terms $A(q,W,\sigma)$ and $B(q,\gamma)$ consequently vanish at the global optimum. However, nothing about the behavior of A and B can be concluded from Proposition 1 if the assumptions are not fulfilled, e.g., in case of finite sample sizes, local optima, or model/data mismatches.

The crucial observation at this point is, that for the standard sparse coding model and after the convergence of parameters, $A(q,W,\sigma)$ and $B(q,\gamma)$ vanish also in the case of: local optima, finite sample sizes, and for any approximate distribution q. Also the data distribution does not have to be matched by the model distribution, which is related to the result being applicable for local optima. This remarkable weakening of the assumptions for entropy limits can be shown by using the update equations for σ and γ in Eqns. 28 and 29. By inserting these equations into the expression for $\mathcal{Q}(q,\Theta)$ in (30), we obtain:

$$\mathcal{Q}(q,\Theta) = \overline{\mathcal{Q}}(\Theta) + \underbrace{\frac{D}{2}\left(1 - \frac{\sigma^2_{\text{new}}}{\sigma^2}\right)}_{A(q,W,\sigma)} + \underbrace{H\left(1 - \frac{\gamma_{\text{new}}}{\gamma}\right)}_{B(q,\gamma)} \quad (31)$$

It follows that at any convergence point of the parameters, $\mathcal{Q}(q,\Theta)$ indeed converges to the entropy limit $\overline{\mathcal{Q}}(\Theta)$. Again, note that the entropy limit originally derived for infinite sample size, global optimum, and exact posterior is also describing the convergence value of the free-energy for finite sample sizes, local optima and even for approximate posteriors and model/data mismatches. But also note that the derivation of the result exploits more properties of the specific model distributions than were used for Propositions 1 and 2. However, these properties are provided by still a large class of distributions. Convergence to entropy limits can thus, for instance, be shown for any SC model with Gaussian noise model and a prior from the exponential family with $h(\vec{z})$=const (e.g., standard SC or the BSC model). Fig. 2B shows a numerical verification of this result ($\mathcal{Q}$ convergence to $\overline{\mathcal{Q}}$ also at the local optimum).

As a corollary with practical significance, consider the case when the sparse coding model (7) and (8) is trained by approximating its intractable exact posteriors by Gaussian distributions. That is, for each data point $\vec{y}^{\,(n)}$ the posterior $p(\vec{z}\,|\,\vec{y}^{\,(n)},\Theta)$ is approximated by a Gaussian $q_n(\vec{z};\Theta') = \mathcal{N}(\vec{z};\vec{\mu}^{\,(n)},\Sigma^{(n)})$ with appropriately chosen mean and covariance matrix (compare, e.g., Bishop, 2006; Seeger, 2008). By using that $Q(q,\Theta)$ converges to the entropy limit $\overline{\mathcal{Q}}(\Theta)$, we obtain with Eqn. 27:

$$\overline{\mathcal{F}}(q,\Theta) = -\frac{D}{2}\log(2\pi e\sigma^2) - H\log(2e\gamma) - \frac{1}{2\,N}\sum_{n=1}^{N}\log(|2\pi e\Sigma^{(n)}|)\,, \quad (32)$$

where the $\overline{\mathcal{F}}(q,\Theta)$ depends on both, model and approximation parameters, and is equal to $\mathcal{F}(q,\Theta)$ whenever the parameters have converged. Now note that the free-energy $\mathcal{F}(q,\Theta)$ is not computable in closed-form because it involves nested integrals over Gaussians with boundaries other than zero or $\pm$ infinity. In contrast, $\overline{\mathcal{F}}(q,\Theta)$ in Eqn. 32 is an elementary closed-form expression. Hence, to compute the free-energy, all that has to be done for a given run of the algorithm is to wait until the parameters have converged, and to use Eqn. 32 with the parameters at convergence. This provides the free-energy value without the necessity of any numerical integrations or look-up tables. As the free-energy is often computed or estimated to evaluate the quality of a given run of an algorithm, formula (32) is an example for the direct applicability of entropy limits in practice. As mentioned earlier, similar formulas can also be derived for other models or other approximation schemes including, for instance, mean-field variational approaches.

7 Numerical Consistency

Let us control for the consistency of our analytical results using numerical experiments on two generative models: p-PCA and Binary Sparse Coding (BSC). Both models have exact EM solutions but while parameter optimization in p-PCA is known to be convex, BSC has local likelihood optima.

P-PCA. In the first experiment we apply p-PCA to recover a one-dimensional subspace of a two dimensional observed space. The data have been generated according to the p-PCA model 16 and 17 with generating parameters Θ^* (note that the generating parameters are thus the maximum likelihood parameters for $N \to \infty$). As can be observed in Fig. 2A, the inner-energy $\mathcal{Q}(\Theta)$ converges to a value close to the entropy limit $\overline{\mathcal{Q}}(\Theta^*)$ of the generating parameters. The finite difference between $\mathcal{Q}(\Theta)$ and $\overline{\mathcal{Q}}(\Theta^*)$ is hereby due to

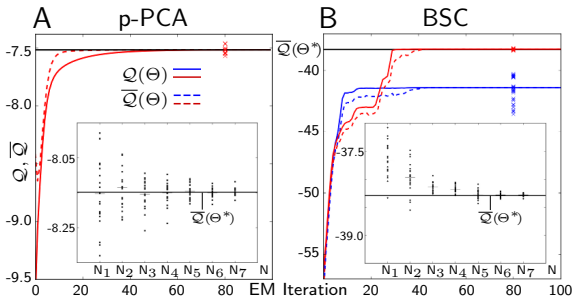

Figure 2: Development of the inner-energy during learning with Expectation Maximization for probabilistic PCA (**A**) and Binary Sparse Coding (**B**) for global (red) and local (blue) optima. Solid lines represent $\mathcal{Q}(\Theta)$, dashed lines mark $\overline{\mathcal{Q}}(\Theta)$. The black solid line represents the entropy limit, i.e. $\overline{\mathcal{Q}}(\Theta^*)$. The scattering of results for 20 runs with different data sets is indicated by red and blue x's. The sub-figures show values of $\mathcal{Q}(\Theta)$ for different data set sizes ($N_1 = 100$, $N_2 = 200$, $N_3 = 500$, $N_4 = 1000$, $N_5 = 2000$, $N_6 = 5000$, $N_7 = 10000$ data points).

the finite sample size. This is highlighted by the distribution of final $\mathcal{Q}(\Theta)$ for different samples (red x's). With increasing sample size, the differences become smaller (see sub-figure of Fig. 2A). During the development of a learning algorithm entropy limits can thus serve as a verification tool: In addition to likelihood increase and parameter recovery, an entropy limit provides an analytical expression for the inner-energy that an algorithm has to converge to for $N \to \infty$. Fig. 2A can thus be regarded as a consistency verification of the used p-PCA implementation.

BSC. In the second experiment we apply BSC to recover basis functions that are linearly mixed using binary factors. We use functions in the form of bars as, e.g., in (Spratling, 2006; Henniges et al., 2010). To optimize the parameters of BSC we use the learning algorithm described in (Henniges et al., 2010) but instead of variational EM we apply exact EM which is still tractable for low numbers of hidden dimensions. Fig 2B shows the convergence of the inner-energy $\mathcal{Q}(\Theta)$ to a limit value during EM optimization. In most cases Θ converges to the maximum likelihood solution but in some case to a local optimum. Note that also in the local optimum, $\mathcal{Q}(\Theta)$ converges to the entropy limit (compare Sec. 6). At the global optimum the limit value of $\mathcal{Q}(\Theta)$ is approximately given by $\overline{\mathcal{Q}}(\Theta^*)$. The finite difference is again due to a finite data sample. With increasing sample size, the differences again become increasingly small (see sub-figure of Fig. 2B). Also in this case the figure can be regarded as a consistency verification of the algorithm's implementation.

8 Discussion and Outlook

In this paper we have studied the behavior of a crucial entity emerging in the likelihood optimization using EM: the parameter dependent part of the free-energy which we have termed inner-energy. The free-energy and the likelihood of the data under a given model are important measures for the quality of an optimization result. However, they are in general not given in closed-form. This is the case for finite as well as for infinite sample sizes and at global as well as local maximum likelihood solutions. In this limit the likelihood $\mathcal{L}(\Theta^*)$ is given by the negative entropy of the modeled distribution, $\mathcal{L}(\Theta^*) = -\mathcal{H}(p(\vec{y}\,|\,\Theta^*))$ (compare Eqn. 11, and see Supplement Eqn. 33). The often rather intricate forms of the model distributions do not allow for closed-form solutions of the model entropy. For exact EM the same is true for the free-energy because it becomes equal to the likelihood at the optimum.

For many generative models we do obtain closed-form solutions for the convergence value of the inner-energy, however. These closed-form solution are simply given by summing over the entropies of the constituting distributions. For some models, we have furthermore shown that the inner-energy converges to entropy limits also in the case of local optima, finite sample sizes, and approximate EM. For the example of the most standard sparse coding model and a standard approximation, we have shown that closed-form solutions for the free-energy can be derived using entropy limits. This formula for the free-energy (32) applies at any point of convergence, for finite sample sizes, and model/data mismatches. In order to provide a measure of the optimization quality after convergence of the algorithm, we can thus directly use a closed-form formula for the free-energy which otherwise would require numerical integration or a nested series of look-up tables. This example demonstrates, that the theoretical results obtained in this study can be very useful for practical applications. As discussed, similar formulas can be derived for other models and other approximation schemes. Moreover, entropy limits can serve as consistency check for learning algorithms as they provide theoretical and easy to compute values that a learning algorithm finally has to converge to (Fig. 2). As graphical models and training schemes tend to become increasingly more advanced and difficult to handle, theoretical results for checking the consistency of implementations or for monitoring the quality of a given run are bound to become more important. This contribution provides a new tool for these purposes that is applicable for a wide range of models, and we hope that its easy applicability will improve the process of developing new algorithms in the field.

References

Berkes, P., Turner, R., and Sahani, M. (2008). On sparsity and overcompleteness in image models. *NIPS*, 20.

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

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

Haft, M., Hofman, R., and Tresp, V. (2004). Generative binary codes. *Pattern Anal Appl*, 6:269–284.

Henniges, M., Puertas, G., Bornschein, J., Eggert, J., and Lücke, J. (2010). Binary Sparse Coding. LNCS 6365, pages 450–457. Springer.

Jaakkola, T. (2000). Tutorial on variational approximation methods. In Opper, M. and Saad, D., editors, *Advanced mean field methods: theory and practice*. MIT Press.

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. (1999). An introduction to variational methods for graphical models. *Machine Learn*, 37:183–233.

Lazo, A. and Rathie, P. (1978). On the entropy of continuous probability distributions. *IEEE Transactions on Information Theory*, 24:120 – 122.

Lücke, J. and Eggert, J. (2010). Expectation truncation and the benefits of preselection in training generative models. *JMLR*, 11:2855 – 2900.

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

Olshausen, B. and Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609.

Osindero, S., Welling, M., and Hinton, G. E. (2006). Topographic product models applied to natural scene statistics. *Neural Computation*, 18:381 – 414.

Roweis, S. (1998). EM algorithms for PCA and SPCA. *NIPS*, pages 626–632.

Seeger, M. (2008). Bayesian inference and optimal design for the sparse linear model. *JMLR*, 9:759–813.

Spratling, M. (2006). Learning image components for object recognition. *JMLR*, 7:793 – 815.

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

Acknowledgements. We acknowledge funding by the German Research Foundation (DFG) in the grant LU 1196/4-1 (JL) and by the Honda Research Institute, Europe (MH).

Appendix – Supplementary Material

Alternative Proof for Proposition 1

We first follow the first steps as the Proof in the main text (first paragraph). As a direct consequence of (11), we obtain at the maximum likelihood solution Θ^*:

$$\mathcal{L}(\Theta^*) \;=\; -\,\mathcal{H}(p(\vec{y}\,|\,\Theta^*)). \tag{33}$$

On the other hand, $\mathcal{L}(\Theta)$ is equal to the free-energy after each E-step, i.e., if $q_n(\vec{z};\Theta) = p(\vec{z}\,|\,\vec{y}^{\,(n)},\Theta)$ we get:

$$\mathcal{L}(\Theta) \;=\; Q(\Theta) - \frac{1}{N}\sum_{n=1}^{N}\int p(\vec{z}\,|\,\vec{y}^{\,(n)},\Theta) \times \log\big(p(\vec{z}\,|\,\vec{y}^{\,(n)},\Theta)\big)\,d\vec{z} \tag{34}$$

By applying (10) in the limit $N \to \infty$ and by using $p(\vec{y}) = p(\vec{y}\,|\,\Theta^*)$ we obtain at the optimum Θ^*:

$$\begin{aligned}\mathcal{L}(\Theta^*) &= Q(\Theta^*) - \int p(\vec{y}\,|\,\Theta^*)\,p(\vec{z}\,|\,\vec{y},\Theta^*) \times \log\big(p(\vec{z}\,|\,\vec{y},\Theta^*)\big)\,d\vec{z} \\ &= Q(\Theta^*) - \int p(\vec{z},\vec{y}\,|\,\Theta^*)\log\big(p(\vec{z},\vec{y}\,|\,\Theta^*)\big)d\vec{z}\,d\vec{y} \\ &\quad + \int p(\vec{z},\vec{y}\,|\,\Theta^*)\log\big(p(\vec{y}\,|\,\Theta^*)\big)d\vec{z}\,d\vec{y}\end{aligned}$$

It thus follows

$$\mathcal{L}(\Theta^*) = Q(\Theta^*) + \mathcal{H}(p(\vec{z},\vec{y}\,|\,\Theta^*)) - \mathcal{H}(p(\vec{y}\,|\,\Theta^*)), \tag{35}$$

and by combining (33) and (35) we thus get:

$$Q(\Theta^*) \;\;=\;\; -\mathcal{H}(p(\vec{z},\vec{y}\,|\,\Theta^*)). \tag{36}$$

The last steps are then equal to those of the original proof again.
□

Proof of Proposition 2

To simplify the steps following Eqn. 15, let us define

$$(\vec{x}_1,\ldots,\vec{x}_K) := (\vec{y}_1,\ldots,\vec{y}_D,\vec{z}_1,\ldots,\vec{z}_H). \tag{37}$$

The graph of a Bayesian network has a partial order and is finite. Without loss of generality, we can thus choose a numbering of the nodes $\vec{x}_k$ such that nodes with higher numbers can never be parents of nodes with lower numbers (compare Fig. 1A). With such a numbering we rewrite:

$$\begin{aligned}&\mathcal{H}\Big(\prod_{k=1}^{K} p(\vec{x}_k\,|\,\mathrm{pa}_k,\Theta)\Big) \\ &= \mathcal{H}\Big(\Big(\prod_{k\in B} p(\vec{x}_k\,|\,\mathrm{pa}_k,\Theta)\Big)\Big(\prod_{k\in A} p(\vec{x}_k\,|\,\mathrm{pa}_k,\Theta)\Big)\Big)\end{aligned}$$

with $A = \{1, \dots, k_o\}$ and $B = \{k_o + 1, \dots, K\}$. Because of the chosen numbering we now know that the second factor in the entropy is independent of the variables $\vec{x}_k$ with $k \in B$. For notational compactness let us further introduce $\mathrm{d}\vec{X}_A = \mathrm{d}\vec{x}_1 \dots \mathrm{d}\vec{x}_{k_o}$ and $\mathrm{d}\vec{X}_B = \mathrm{d}\vec{x}_{k_o+1} \dots \mathrm{d}\vec{x}_K$. It follows:

$$
\begin{aligned}
&\mathcal{H}\Big(\big(\prod_{k\in B} p(\vec{x}_k \,|\, \mathrm{pa}_k, \Theta) \big) \big(\prod_{k\in A} p(\vec{x}_k \,|\, \mathrm{pa}_k, \Theta) \big) \Big) \\
&= -\int \big(\prod_{k\in B} p(\vec{x}_k \,|\, \mathrm{pa}_k, \Theta) \big) \big(\prod_{k\in A} p(\vec{x}_k \,|\, \mathrm{pa}_k, \Theta) \big) \\
&\qquad\qquad \times \log \big(\prod_{k\in B} p(\vec{x}_k \,|\, \mathrm{pa}_k, \Theta) \big) \mathrm{d}\vec{X} \\
&\quad - \int \big(\prod_{k\in B} p(\vec{x}_k \,|\, \mathrm{pa}_k, \Theta) \big) \big(\prod_{k\in A} p(\vec{x}_k \,|\, \mathrm{pa}_k, \Theta) \big) \\
&\qquad\qquad \times \log \big(\prod_{k\in A} p(\vec{x}_k \,|\, \mathrm{pa}_k, \Theta) \big) \mathrm{d}\vec{X} \\
&= -\int \big(\prod_{k\in A} p(\vec{x}_k \,|\, \mathrm{pa}_k, \Theta) \big) \int \big(\prod_{k\in B} p(\vec{x}_k \,|\, \mathrm{pa}_k, \Theta) \big) \\
&\qquad\qquad \times \log \big(\prod_{k\in B} p(\vec{x}_k \,|\, \mathrm{pa}_k, \Theta) \big) \mathrm{d}\vec{X}_B \, \mathrm{d}\vec{X}_A \\
&\quad - \int \big(\prod_{k\in A} p(\vec{x}_k \,|\, \mathrm{pa}_k, \Theta) \big) \log \big(\prod_{k\in A} p(\vec{x}_k \,|\, \mathrm{pa}_k, \Theta) \big) \mathrm{d}\vec{X}_A \\
&= \int \big(\prod_{k\in A} p(\vec{x}_k \,|\, \mathrm{pa}_k, \Theta) \big) \, \mathcal{H}\big(\prod_{k\in B} p(\vec{x}_k \,|\, \mathrm{pa}_k, \Theta) \big) \, \mathrm{d}\vec{X}_A \\
&\quad + \mathcal{H}\big(\prod_{k\in A} p(\vec{x}_k \,|\, \mathrm{pa}_k, \Theta) \big) \\
&= \mathcal{H}\big(\prod_{k\in B} p(\vec{x}_k \,|\, \mathrm{pa}_k, \Theta) \big) \; + \; \mathcal{H}\big(\prod_{k\in A} p(\vec{x}_k \,|\, \mathrm{pa}_k, \Theta) \big)
\end{aligned}
$$

For the last step we used the assumption of the entropy being independent of the parent values.

The subgraphs defined amongst the nodes in A and B are partially ordered as well. We can thus recursively apply the result above and obtain:

$$
\mathcal{H}\big(\prod_{k=1}^{K} p(\vec{x}_k \,|\, \mathrm{pa}_k, \Theta) \big) \; = \; \sum_{k=1}^{K} \mathcal{H}\big(p(\vec{x}_k \,|\, \mathrm{pa}_k, \Theta) \big).
$$

If we back-replace $\vec{x}_k$ using the original variable names $\vec{y}_d$ and $\vec{z}_h$ and insert into Eqn. 15, we obtain the claim Eqn. 14 with the corresponding expression for $\overline{\mathcal{Q}}(\Theta^*)$.

Section 6 – Details of Derivations

To derive Eqn. 31, let us take a closer look at the behavior of the inner-energy at the optimum. The Sparse Coding generative model under consideration consists of a Laplace prior and a Gaussian noise model:

$$
\begin{aligned}
p\left(\vec{z} \,|\, \Theta\right) &= \prod_{h=1}^{H} \frac{1}{2\gamma} \exp\left(-\frac{|z_h|}{\gamma} \right) \\
p\big(\vec{y}^{\,(n)} \,|\, \vec{z}, \Theta\big) &= \mathcal{N}\big(\vec{y}^{\,(n)}; W\vec{z}, \sigma^2\big)
\end{aligned}
\tag{38}
$$

To find the optimal parameters for a given data set, we have to compute the derivatives of the inner-energy with respect to the corresponding parameters. The inner-energy is given by:

$$
\mathcal{Q}(q, \Theta) = \frac{1}{N} \sum_n \int q_n(\vec{z};\, \Theta') \, \log\left(p(\vec{y}^{\,(n)},\, \vec{z} \,|\, \Theta) \right) d\vec{z}
\tag{39}
$$

At the optimum, $\mathcal{Q}$ must be constant in the model parameters. The parameters that maximize $\mathcal{Q}$ are thus the ones for which the derivative of $\mathcal{Q}$ becomes zero:

$$
\frac{\partial \mathcal{Q}}{\partial \Theta} \stackrel{!}{=} 0
$$

To obtain the update rules, we first calculate the derivative of $\mathcal{Q}$ with respect to γ. Splitting up the joint in (39) and omitting the noise term which does not depend on γ we get:

$$
\begin{aligned}
\frac{\partial \mathcal{Q}}{\partial \gamma} &= \frac{1}{N} \sum_n \int q_n(\vec{z};\, \Theta') \\
&\qquad \times \frac{\partial}{\partial \gamma} \log \left[\prod_{h=1}^{H} \frac{1}{2\gamma} \exp\left(-\frac{|z_h|}{\gamma} \right) \right] d\vec{z} \\
&= \frac{1}{N} \sum_n \int q_n(\vec{z};\, \Theta') \\
&\qquad \times \left[-\frac{H}{\gamma} + \sum_{h=1}^{H} \frac{|z_h|}{\gamma^2} \right] d\vec{z} \stackrel{!}{=} 0
\end{aligned}
$$

This leads to:

$$
\gamma_{\mathrm{new}} = \frac{1}{N} \sum_n \int q_n(\vec{z};\, \Theta') \|\vec{z}\|_1 \, d\vec{z}
\tag{40}
$$

Now, we compute the derivative with respect to σ^2:

$$
\begin{aligned}
&\frac{\partial \mathcal{Q}}{\partial \sigma^2} = \frac{1}{N} \sum_n \int q_n(\vec{z};\, \Theta') \left[-\frac{D}{2\sigma^2} - \frac{1}{2\sigma^4} \|\vec{y}^{\,(n)} - W\vec{z}\|^2 \right] d\vec{z} \\
&\Rightarrow \sigma^2_{\mathrm{new}} = \frac{1}{ND} \sum_n \int q_n(\vec{z};\, \Theta') \, \|\vec{y}^{\,(n)} - W\vec{z}\|^2 d\vec{z}
\end{aligned}
\tag{41}
$$

We can now rewrite:

$$
\begin{aligned}
\mathcal{Q}(q, \Theta) &= \overline{\mathcal{Q}}(\Theta) + A(q, W, \sigma) + B(q, \gamma), \text{ where} \\
A(q, W, \sigma) &= -\frac{1}{2\sigma^2 N} \sum_{n=1}^{N} \int q_n(\vec{z};\, \Theta') \\
&\qquad\qquad \times \|\vec{y}^{\,(n)} - W\vec{z}\|^2 \, d\vec{z} + \frac{D}{2} \\
B(q, \gamma) &= -\frac{H}{\gamma N} \sum_{n=1}^{N} \int q_n(\vec{z};\, \Theta') \, \|\vec{z}\|_1 \, d\vec{z} + H
\end{aligned}
$$

By inserting the expressions in (40) and (41), we obtain (31). This can also be shown for any other SC model with Gaussian noise model and a prior from the exponential family provided it can be written as: $p(\vec{x}\,|\,\vec{\eta}) = h(\vec{x}\,)\, g(\vec{\eta}\,) \exp(\vec{\eta}^{\,T} \vec{u}(\vec{x}\,))$ with $h(\vec{x}\,) = \text{const}$.