Learning Sparse Codes with Entropy-Based ELBOs

Dmytro Velychko
Machine Learning Lab
University of Oldenburg, Germany

Simon Damm
Faculty of Computer Science
Ruhr University Bochum, Germany

Asja Fischer
Faculty of Computer Science
Ruhr University Bochum, Germany

Jörg Lücke
Machine Learning Lab
University of Oldenburg, Germany

Abstract

Standard probabilistic sparse coding assumes a Laplace prior, a linear mapping from latents to observables, and Gaussian observable distributions. We here derive a solely entropy-based learning objective for the parameters of standard sparse coding. The novel variational objective has the following features: (A) unlike MAP approximations, it uses non-trivial posterior approximations for probabilistic inference; (B) the novel objective is fully analytic; and (C) the objective allows for a novel principled form of annealing. The objective is derived by first showing that the standard ELBO objective converges to a sum of entropies, which matches similar recent results for generative models with Gaussian priors. The conditions under which the ELBO becomes equal to entropies are then shown to have analytic solutions, which leads to the fully analytic objective. Numerical experiments are used to demonstrate the feasibility of learning with such entropy-based ELBOs. We investigate different posterior approximations including Gaussians with correlated latents and deep amortized approximations. Furthermore, we numerically investigate entropy-based annealing which results in improved learning. Our main contributions are theoretical, however, and they are twofold: (1) we provide the first demonstration on how a recently shown convergence of the ELBO to entropy sums can be used for learning; and (2) using the entropy objective, we derive a fully analytic ELBO objective for the standard sparse coding generative model.

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

1 INTRODUCTION AND RELATED WORK

Sparse coding seeks to represent data vectors $\mathbf{x}$ by latent vectors $\mathbf{z}$. Sparse coding requires the vectors $\mathbf{z}$ to be *sparse*, i.e., on average only few of the values z_h significantly contribute in representing any given vector $\mathbf{x}$. Our main focus will be the (by far) most standard data model for probabilistic sparse coding (Williams, 1995; Olshausen and Field, 1996; Seeger et al., 2007). The model assumes a Laplacian (a.k.a. double-exponential) prior distribution for latents $\mathbf{z} \in \mathbb{R}^H$, and a Gaussian noise distribution for observables $\mathbf{x} \in \mathbb{R}^D$,

$$\begin{aligned} p(\mathbf{z}) &= \prod_{h=1}^{H} \frac{1}{2} \exp\big(-|z_h|\big) \quad \text{and} \\ p_\Theta(\mathbf{x}\,|\,\mathbf{z}) &= \mathcal{N}\big(\mathbf{x}\,|\,W\mathbf{z}, \sigma^2\mathbb{I}\big) \ , \end{aligned} \tag{1}$$

where weight matrix $W \in \mathbb{R}^{D\times H}$ and observation noise $\sigma^2 > 0$ are the model parameters $\Theta = (W, \sigma^2)$. The sparse coding model, and in particular the Laplace prior distribution, are closely related to deterministic sparse coding approaches that use the l_1-objective (e.g., Hastie et al., 2015). A standard form of deterministic sparse coding addresses the optimization problem

$$\min_{\mathbf{z}^{(1)},\ldots,\mathbf{z}^{(N)}} \Big\{ \underbrace{\sum_{n=1}^{N} \big\|\mathbf{x}^{(n)} - \tilde{W}\mathbf{z}^{(n)}\big\|^2}_{\text{reconstruction}} + \tilde{\gamma} \underbrace{\sum_{n=1}^{N}\sum_{h=1}^{H} \big|z_h^{(n)}\big|}_{\text{sparsity}} \Big\} \ , \tag{2}$$

where $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}$ are deterministic latent vectors corresponding to data vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$, and where $\tilde{W} \in \mathbb{R}^{D\times H}$ with columns of unit length. The constant $\tilde{\gamma}$ (often also denoted λ) weights the sparsity term vs. the reconstruction term of the objective.

For more than two decades, sparse coding approaches have been very thoroughly investigated with large numbers of papers dedicated to theoretical investigations of the respective optimization problems, and with many papers using different (including deep) forms of sparse coding for numerous tasks. Such tasks included (to name a few) denoising, inpainting, compression, disentanglement, or super-resolution (e.g. Mairal et al., 2014; Yao et al., 2022; Cheng et al., 2022; Drefs et al., 2023).

For the standard probabilistic data model, given in Eq. (1), the presumably most common way to derive algorithms for parameter optimization is maximum likelihood (ML) estimation. That is, we seek those parameters of the model that maximize the (marginal) log-likelihood $\mathcal{L}^{\mathrm{LL}}(\Theta)$ in dependence of the likelihood parameters $\Theta = (W, \sigma^2)$ with

$$\mathcal{L}^{\mathrm{LL}}(\Theta) = \frac{1}{N}\sum_{n=1}^{N} \log\Big(\int p_\Theta(\mathbf{x}^{(n)}|\mathbf{z})\, p(\mathbf{z})\, \mathrm{d}\mathbf{z}\Big) \ . \quad (3)$$

In order to facilitate the challenging problem of maximizing $\mathcal{L}^{\mathrm{LL}}$, approximations to ML optimization are very commonly applied. One of the most common approximation methods applied for probabilistic sparse coding (and probabilistic generative models in general) is variational approximation (e.g. Jaakkola and Jordan, 1997). Concretely, instead of maximizing the likelihood directly, a lower bound of the log-likelihood is maximized, which is referred to as free-energy or ELBO (e.g., Neal and Hinton, 1998; Jordan et al., 1999):

$$\mathcal{L}^{\mathrm{EL}}(\Phi,\Theta) = \frac{1}{N}\sum_{n=1}^{N}\Big[\int q_\Phi^{(n)}(\mathbf{z}) \log p_\Theta(\mathbf{x}^{(n)}\,|\,\mathbf{z})\mathrm{d}\mathbf{z} - D_{\mathrm{KL}}\big(q_\Phi^{(n)}(\mathbf{z})\,\|\,p(\mathbf{z})\big)\Big] \ . \quad (4)$$

Given the data model in Eq. (1), the ELBO is defined by the family of variational distributions $q_\Phi^{(n)}(\mathbf{z})$ used to approximate the true posteriors of a given model. The standard choice for probabilistic sparse coding are Gaussian variational distributions to approximate the analytically intractable posteriors of the model. For sparse coding as in Eq. (1), the true posteriors are known to be mono-modal (Olshausen and Field, 1996; Seeger et al., 2007) due to log-concavity. Therefore, Gaussian approximations (by matching mode and correlations) can be considered as capturing the most essential structure of the model's true posteriors.

The optimization of lower bounds such as the ELBO usually represents an easier optimization problem than optimizing the likelihood itself. However, the crucial challenge for both of these optimizations is posed by the integrals over potentially high-dimensional latent spaces. For the standard sparse coding model, no analytic solutions have been reported, so far. In particular, no analytic solutions have been reported for the common case of using Gaussians as the family of variational distributions.

It could be argued that deterministic algorithms are, nonetheless, available if much more simplifying approximations than Gaussians are used for optimization. The arguably most common approach is given by maximum a-posteriori (MAP) training (Olshausen and Field, 1996). From a probabilistic perspective, MAP approximations may be interpreted as a limit case of variational approximations in which the family of variational distributions are delta-distributions. The high-dimensional integrations over latent space are then trivially solved. As a source for its high popularity, MAP approximations allow for linking the standard probabilistic model in Eq. (1) to the deterministic l_1-sparse coding objective in Eq. (2). That is, the sparse coding objective in Eq. (2) can be recovered if the MAP approximation is applied. Another source of the ongoing popularity of MAP (also in general) is the resulting closed-form objective (cf. Eq. 2), i.e., no high-dimensional integrals have to be numerically estimated.

However, from a probabilistic machine learning perspective, delta-distributions do not represent theoretically well-grounded approximations. One consequence of using MAP is, for instance, that the ELBO objective is rendered non-finite and thus cannot be considered as a learning objective anymore (see, e.g., Barello et al., 2018, for a discussion); also the meaning of the ELBO as a lower bound of the log-likelihood ceases to provide meaning in the MAP case. Moreover, severe degeneracies are introduced: optimization of W tends to yield infinite entries (Olshausen and Field, 1996), which has to be manually corrected, and data noise σ^2 and sparsity are not learnable independently of one another. More generally, no probabilistic encoding is provided, i.e., neither can a probabilistic objective be used for tasks such as model selection nor is there uncertainty information available for data encoding (with all the negative consequences for downstream tasks one may seek to address). Such major drawbacks have, consequently, resulted in substantial research efforts to allow for appropriate uncertainty estimation in sparse coding. Strategies that were followed include (i) the application expectation propagation (Seeger et al., 2007), (ii) sampling-based fully Bayesian approaches (Mohamed et al., 2012), (iii) amendments of the original data model (Berkes et al., 2007; Sheikh et al., 2014) such that non-trivial variational optimizations could be applied, (iv) the use of amortized variational distributions for the original data model (Barello et al., 2018), or (v) amendments of both data model and variational distributions (Tonolini et al., 2020; Drefs et al., 2023).

2 ELBO CONVERGENCE TO ENTROPY SUMS

There are many ways to rewrite the ELBO and relate it to Kullback-Leibler divergence, entropies, cross-entropies, mutual information, and expected reconstruction error (e.g. Alemi et al., 2018; Hoffman and Johnson, 2016; Zhao et al., 2017). In contrast, we in this work seek to rewrite the ELBO as a sum of entropies. The reformulation is obtained through assuming convergence of a subset of the model parameters, i.e., our reformulation is valid on a submanifold in the space of all model parameters. Our reformulation is consequently different from previously known reformulations that are valid for the entire space of parameters.

We will first show that the ELBO for the sparse coding model given in Eq. (1) converges to a sum of three entropies given optimization of specific model parameters. The derived results will apply to general variational distributions, the specific variational family of Gaussian distributions will only be used later. Our derivations are based on recent results for variational autoencoders (VAEs) that show convergence of standard (Gaussian) VAEs to entropy sums (Damm et al., 2023). These results have deeper roots in the exponential family property of Gaussians (Lücke and Warnken, 2023) and we here, for the first time, show that the ELBO of standard sparse coding converges to entropy sums.

Our main focus will be on learning, which contrasts with previous work (Damm et al., 2023; Lücke and Warnken, 2023) that investigated the properties of the ELBO at stationary points. The focus on learning means that we will exploit properties the ELBO in Eq. (4) attains if a subset of model parameters have converged. To specify those parameters we will first reparameterize the sparse coding model introduced in Eq. (1) before we investigate convergence to entropy sums.

2.1 Reparameterization of Sparse Coding

Consider an elementary Bayesian network (Fig. 1, left) for probabilistic latent variable models, that covers models such as sparse coding, as in Eq. (1), probabilistic PCA (Tipping and Bishop, 1999), and VAEs (Kingma and Welling, 2014). For our derivations of the entropy-based ELBO, we will use a slightly altered form of the model with learnable prior parameters (Fig. 1, right). Concretely, we constrain the columns of the weight matrix (now termed $\tilde{W}$) to be of unit length but we use parameterized Laplace distributions for the prior (instead of the parameterless standard choice).

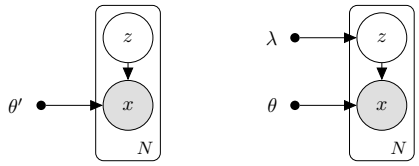


Figure 1: **Latent variable model.** *Left:* graphical model representation corresponding to many popular latent variable models, including VAEs. *Right:* graphical model with learnable prior parameters and constrained likelihood parameters as used in this work.

The sparse coding model is thus given by

$$\begin{aligned} p_\Theta(\mathbf{z}) &= \prod_{h=1}^{H} \frac{1}{2\lambda_h} \exp\left(-\frac{|z_h|}{\lambda_h}\right) \quad \text{and} \\ p_\Theta(\mathbf{x}|\mathbf{z}) &= \mathcal{N}\big(\mathbf{x} \,|\, \tilde{W}\mathbf{z}, \sigma^2\mathbb{I}\big) \ , \end{aligned} \tag{5}$$

where $\forall h: \ \sum_i \big(\tilde{W}_{ih}\big)^2 = 1$, such that Θ now reads $\Theta = (\boldsymbol{\lambda}, \tilde{W}, \sigma^2) \in \mathbb{R}_+^H \times \mathbb{R}_{\text{norm}}^{D\times H} \times \mathbb{R}_+$. The prior parameters λ_h are commonly referred to as scales.

It is straightforward to show that the reparameterized model in Eq. (5) parameterizes the same family of distributions $p_\Theta(\mathbf{x})$ as the original model in Eq. (1), with a one-to-one mapping between their respective parameters. This parameterization is important in order to show that the ELBO of sparse coding becomes equal to entropy sums under certain conditions.

2.2 Equality of ELBO and Entropy Sums

Motivated by previous work (Damm et al., 2023), we now investigate if the ELBO of the model in Eq. (5) becomes equal to entropy sums during learning. Damm et al. (2023) did show equality of ELBO and entropy sums for Gaussian models (Gaussian prior and Gaussian noise model) at all stationary points. For this work, we will show equality to entropy sums for the ELBO of sparse coding. But, furthermore, it will be important for this work to explicitly note that only the parameters $\boldsymbol{\lambda}$ and σ^2 have to be at stationary points in order to realize equality to entropy sums.

Theorem 1 (ELBO converges to a sum of entropies)**.** *Consider the ELBO in Eq.* (4) *for the sparse coding model in Eq.* (5) *with parameters* $\Theta = (\boldsymbol{\lambda}, \tilde{W}, \sigma^2)$*. If the parameters* $\boldsymbol{\lambda}$ *and* σ^2 *are at a stationary point, i.e.,*

$$\tfrac{\partial}{\partial \boldsymbol{\lambda}} \mathcal{L}^{\text{EL}}(\Phi, \Theta) = 0 \quad \textit{and} \quad \tfrac{\partial}{\partial \sigma^2} \mathcal{L}^{\text{EL}}(\Phi, \Theta) = 0 \quad , \tag{6}$$

then it applies for any variational distributions $q_\Phi(\mathbf{z})$ *and for any matrix* $\tilde{W}$ *(with unit column lengths) that:*

$$\begin{aligned} &\mathcal{L}^{\text{EL}}(\Phi, \Theta) \\ &= \tfrac{1}{N} \textstyle\sum_n \mathcal{H}[q_\Phi^{(n)}(\mathbf{z})] - \mathcal{H}[p_\Theta(\mathbf{z})] - \mathcal{H}[p_\Theta(\mathbf{x}|\mathbf{z})] \ . \end{aligned} \tag{7}$$

Proof. The ELBO objective in Eq. (4) can be rewritten to consist of three summands, i.e.,

$$\mathcal{L}^{\mathrm{EL}}(\Phi,\Theta) = \frac{1}{N}\sum_n \mathcal{H}[q_\Phi^{(n)}(\mathbf{z})] + \overbrace{\tfrac{1}{N}\textstyle\sum_n \int q_\Phi^{(n)}(\mathbf{z}) \log p_\Theta(\mathbf{z})\mathrm{d}\mathbf{z}}^{\mathcal{L}_1^{\mathrm{EL}}(\Phi,\Theta)} + \underbrace{\tfrac{1}{N}\textstyle\sum_n \int q_\Phi^{(n)}(\mathbf{z}) \log p_\Theta(\mathbf{x}^{(n)}|\mathbf{z})\mathrm{d}\mathbf{z}}_{\mathcal{L}_2^{\mathrm{EL}}(\Phi,\Theta)} .$$

The first summand is already in the form of an (average) entropy. The last summand, $\mathcal{L}_2^{\mathrm{EL}}(\Phi,\Theta)$, has the form

$$\mathcal{L}_2^{\mathrm{EL}}(\Phi,\Theta) = \tfrac{1}{N}\sum_n \Big(-\frac{D}{2}\log\big(2\pi\sigma^2\big) - \frac{1}{2\sigma^2}\int q_\Phi^{(n)}(\mathbf{z})\,\|\mathbf{x}^{(n)} - \tilde{W}\mathbf{z}\|^2 \mathrm{d}\mathbf{z}\Big) .$$

If $\frac{\partial}{\partial\sigma^2}\mathcal{L}(\Phi,\Theta) = 0$, we get $\mathcal{L}_2^{\mathrm{EL}}(\Phi,\Theta) = -\mathcal{H}[p_\Theta(\mathbf{x}|\mathbf{z})]$, which can be shown analogously to the Gaussian noise distribution used in Gaussian variational autoencoders (Damm et al., 2023).[1]

To show that Eq. (7) holds, it is therefore left to show that the summand $\mathcal{L}_1^{\mathrm{EL}}(\Phi,\Theta)$ has the form of an entropy under the conditions of the theorem. We invoke the factorization $q_\Phi^{(n)}(\mathbf{z}) = q_\Phi^{(n)}(\mathbf{z}_{/h}|z_h)q_\Phi^{(n)}(z_h)$ to simplify the integral $\mathcal{L}_1^{\mathrm{EL}}(\Phi,\Theta)$ such that $\mathcal{L}_1^{\mathrm{EL}}(\Phi,\Theta) =$

$$\sum_h \Big(-\log(2\lambda_h) - \frac{1}{N}\sum_n \frac{1}{\lambda_h}\int |z_h|\, q_\Phi^{(n)}(z_h)\,\mathrm{d}z_h\Big). \tag{8}$$

As only the term $\mathcal{L}_1^{\mathrm{EL}}(\Phi,\Theta^*)$ of the ELBO depends on $\boldsymbol{\lambda}$, we obtain at stationary points of Eq. (8) w.r.t λ_h:

$$\begin{aligned} 0 &= \frac{\partial}{\partial\lambda_h}\mathcal{L}^{\mathrm{EL}}(\Phi,\Theta) = \frac{\partial}{\partial\lambda_h}\mathcal{L}_1^{\mathrm{EL}}(\Phi,\Theta) \\ &= -\frac{1}{\lambda_h} + \frac{1}{N}\sum_n \frac{1}{\lambda_h^2}\int |z_h|\, q_\Phi^{(n)}(z_h)\,\mathrm{d}z_h \\ &= \frac{1}{\lambda_h}\left(-1 + \frac{1}{N}\sum_n \frac{1}{\lambda_h}\int |z_h|\, q_\Phi^{(n)}(z_h)\,\mathrm{d}z_h\right), \end{aligned}$$

for all h. As $\lambda_h \neq 0$, it follows that

$$\frac{1}{N}\sum_n \frac{1}{\lambda_h}\int |z_h|\, q_\Phi^{(n)}(z_h)\,\mathrm{d}z_h = 1. \tag{9}$$

Now we insert Eq. (9) into Eq. (8) and obtain:

$$\mathcal{L}_1^{\mathrm{EL}}(\Phi,\Theta) = -\sum_h \log(2e\lambda_h) = -\mathcal{H}[p_\Theta(\mathbf{z})]. \qquad \square$$

[1] For completeness, we reiterate the derivation for our case in Appendix A.3.

In Appendix A.1 we present a simple but more general theorem that *constructively* proves convergence to entropies for a small class of exponential family distributions. More general convergence criteria were presented by Lücke and Warnken (2023), see Appendix A.2.

3 ENTROPY-BASED ELBOS AS LEARNING OBJECTIVES

The entropy sum expression in Theorem 1 does by itself *not* represent a learning objective because it requires the conditions in Eq. (6); and these are usually not satisfied during optimization. However, do note that the conditions only concern a subset of the parameters of the ELBO, i.e., $\boldsymbol{\lambda}$ and σ^2. No conditions have to be fulfilled for the parameters $\tilde{W}$ and the variational parameters Φ. Importantly, this means that the expression in Eq. (7) can potentially be used as a learning objective if we can derive solutions for $\boldsymbol{\lambda}$ and σ^2 that satisfy the conditions stated in Eq. (6). For our specific choice of variational distributions $q_\Phi^{(n)}(\mathbf{z})$ we can, notably, find analytic such solutions.

Theorem 2 (Optimal scales and variance)**.** *For the sparse coding model in Eq.* (5) *consider the ELBO in Eq.* (4) *defined with Gaussian distributions* $q_\Phi^{(n)}(\mathbf{z}) = \mathcal{N}(\mathbf{z}\,|\,\boldsymbol{\nu}^{(n)},\mathcal{T}^{(n)})$*, for* $n = 1,\ldots,N$*, as family of variational distributions. The variational parameters are consequently given by* $\Phi = (\boldsymbol{\nu}^{(1)},\ldots,\boldsymbol{\nu}^{(N)},\mathcal{T}^{(1)},\ldots,\mathcal{T}^{(N)})$ *with* $\boldsymbol{\nu}^{(n)} \in \mathbb{R}^H$ *and positive semi-definite matrices* $\mathcal{T}^{(n)} \in \mathbb{R}^{H\times H}$*. For arbitrary such variational distributions and for an arbitrary matrix* $\tilde{W}$ *(with unit length columns), we can then find the values for* $\boldsymbol{\lambda}$ *and* σ^2 *that satisfy Eq.* (6)*. The solutions for* $\boldsymbol{\lambda}$ *and* σ^2 *are unique and are given by*

$$\sigma^2_{\mathrm{opt}}\big(\Phi,\tilde{W}\big) = \frac{1}{N}\sum_n \frac{1}{D}\Big[\mathrm{tr}(\tilde{W}^{\mathrm{T}}\tilde{W}\mathcal{T}^{(n)}) + (\tilde{W}\boldsymbol{\nu}^{(n)} - \mathbf{x}^{(n)})^{\mathrm{T}}(\tilde{W}\boldsymbol{\nu}^{(n)} - \mathbf{x}^{(n)})\Big] , \tag{10}$$

$$\forall h : \lambda_h^{\mathrm{opt}}\big(\Phi\big) = \frac{1}{N}\sum_n \sqrt{\mathcal{T}_{hh}^{(n)}}\,\mathcal{M}\left(\frac{\nu_h^{(n)}}{\sqrt{\mathcal{T}_{hh}^{(n)}}}\right) \tag{11}$$

$$\textit{with } \mathcal{M}(a) = \sqrt{\frac{2}{\pi}}\exp\left(-\frac{1}{2}\,a^2\right) + a\,\mathrm{erf}\left(\frac{a}{\sqrt{2}}\right). \tag{12}$$

Proof sketch. We solve the arising integrals analytically to find the corresponding parameters at stationary points. Appendix B.1 contains the full derivations. $\square$

3.1 Entropy-based Learning Objective for Standard Sparse Coding

We can now consider the subspace of all parameters Θ with optimal $\boldsymbol{\lambda}$ and σ^2. These are all parameters that can be obtained from parameters Φ and $\tilde{W}$ through the following function:

$$\Theta_{\text{opt}}(\Phi, \tilde{W}) = \big(\boldsymbol{\lambda}_{\text{opt}}(\Phi), \tilde{W}, \sigma^2_{\text{opt}}(\Phi, \tilde{W})\big) \;, \tag{13}$$

where $\boldsymbol{\lambda}_{\text{opt}}(\Phi)$ and $\sigma^2_{\text{opt}}(\Phi, \tilde{W})$ are provided by Theorem 2. As for all $\Theta_{\text{opt}}(\Phi, \tilde{W})$ the conditions for Theorem 1 are fulfilled, it applies for all Φ and $\tilde{W}$ that

$$\begin{aligned}\mathcal{L}^{\text{EL}}\big(\Phi, \Theta_{\text{opt}}(\Phi, \tilde{W})\big) &= \frac{1}{N}\sum_n \mathcal{H}[q^{(n)}_\Phi(\mathbf{z})] \\ &\quad - \mathcal{H}[p_{\Theta_{\text{opt}}(\Phi,\tilde{W})}(\mathbf{z})] - \mathcal{H}[p_{\Theta_{\text{opt}}(\Phi,\tilde{W})}(\mathbf{x}|\mathbf{z})] \;.\end{aligned} \tag{14}$$

The entropy-based right-hand-side of Eq. (14) only depends on Φ and $\tilde{W}$, and it suggests itself as a novel objective for these remaining parameters. Importantly, as the entropies in Eq. (14) are all given in closed-form, and as $\boldsymbol{\lambda}_{\text{opt}}(\Phi)$ and $\sigma^2_{\text{opt}}(\Phi, \tilde{W})$ are analytic functions, the novel objective is an analytic function as well. Using the expressions for the entropies in Eq. (14) and the solutions $\boldsymbol{\lambda}_{\text{opt}}(\Phi)$ and $\sigma^2_{\text{opt}}(\Phi, \tilde{W})$, the objective is given by (see Appendix B.1 for intermediate steps):

$$\begin{aligned}\mathcal{L}^{\mathcal{H}}(\Phi, \tilde{W}) &= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}\log\big(\,|\,2\pi\, e\,\mathcal{T}^{(n)}\,|\,\big) \\ &\quad - \sum_{h=1}^{H}\log\left(2\, e\,\frac{1}{N}\sum_{n=1}^{N}\sqrt{\mathcal{T}^{(n)}_{hh}}\,\mathcal{M}\left(\frac{\nu^{(n)}_h}{\sqrt{\mathcal{T}^{(n)}_{hh}}}\right)\right) \\ &\quad - \frac{D}{2}\log\left(2\pi\, e\,\frac{1}{N}\sum_{n=1}^{N}\frac{1}{D}\Big[\,\text{tr}(\tilde{W}^{\text{T}}\tilde{W}\mathcal{T}^{(n)}) \right. \\ &\qquad \left. + \big(\tilde{W}\boldsymbol{\nu}^{(n)} - \mathbf{x}^{(n)}\big)^{\text{T}}\big(\tilde{W}\boldsymbol{\nu}^{(n)} - \mathbf{x}^{(n)}\big)\Big]\right) \;,\end{aligned} \tag{15}$$

where $\mathcal{M}\big(\,\cdot\,\big)$ is the analytic function from Eq. (12).

Considering the new objective $\mathcal{L}^{\mathcal{H}}$, there is, however, a subtle but important difference compared to $\mathcal{L}^{\text{EL}}$: the solutions for $\boldsymbol{\lambda}$ and σ^2 introduce dependencies between model parameters Θ and variational parameters Φ. As a consequence, the standard lower-bound relation between log-likelihood (only depending on Θ) and $\mathcal{L}^{\mathcal{H}}$ becomes more intricate away from stationary points. We can, however, show that the objective $\mathcal{L}^{\mathcal{H}}(\Phi, \tilde{W})$ has the same stationary points (together with $\boldsymbol{\lambda}_{\text{opt}}(\Phi)$ and $\sigma^2_{\text{opt}}(\Phi, \tilde{W})$) as the original ELBO $\mathcal{L}^{\text{EL}}(\Phi, \Theta)$.

Theorem 3. *Consider the sparse coding model formulated in Eq.* (5) *with model parameters* $\Theta = (\boldsymbol{\lambda}, \tilde{W}, \sigma^2) \in \mathbb{R}^H_+ \times \mathbb{R}^{D\times H}_{\text{norm}} \times \mathbb{R}_+$*, and variational parameters* $\Phi = (\Phi_\nu, \Phi_\mathcal{T})$ *that parameterize mean* $\boldsymbol{\nu}^{(n)} \in \mathbb{R}^H$ *and covariance* $\mathcal{T}^{(n)} \in \mathbb{R}^{H\times H}$ *(in amortized or non-amortized fashion) where* $\Phi_\nu \cap \Phi_\mathcal{T} = \emptyset$*. Then, the set of stationary points of the original objective* $\mathcal{L}^{\text{EL}}(\Phi, \Theta)$*, given in Eq.* (4)*, and of the entropy-based objective* $\mathcal{L}^{\mathcal{H}}(\Phi, \tilde{W})$*, given in Eq.* (15)*, coincide. Furthermore, at any stationary point it holds*

$$\mathcal{L}^{\text{EL}}(\Phi^\star, \Theta^\star) = \mathcal{L}^{\mathcal{H}}(\Phi^\star, \tilde{W}^\star) \;. \tag{16}$$

Proof sketch. As all stationary points must satisfy Eq. (6), Eq. (16) holds directly by Theorem 1. To show that any stationary point of one objective is also a stationary point of the other we show that the gradients of both objectives coincide whenever Eq. (6) holds. The full proof is deferred to Appendix C. □

In virtue of Theorem 3, $\mathcal{L}^{\mathcal{H}}(\Phi, \tilde{W})$, given in Eq. (15), can be used as novel objective for standard sparse coding with Laplace prior. Note that also the error function is known to be an analytic function, which can be seen, e.g., by considering its representation by the Bürmann series (Schöpf and Supancic, 2014). The series representations also highlight that for all practical reasons, very accurate (and readily available) closed-form approximations of the error function can be used for optimization (see Appendix D.1).

3.2 Properties of the New Objective

Equation (15) represents the most general form of the novel objective. In case of diagonal covariance matrices, $q^{(n)}(\mathbf{z}) = \mathcal{N}\big(\mathbf{z}\,|\,\boldsymbol{\nu}^{(n)}, \text{diag}\,\big(\big(\tau^{(n)}_1\big)^2 \ldots \big(\tau^{(n)}_H\big)^2\big)\big)$, Eq. (15) simplifies significantly as $\text{tr}(\tilde{W}^{\text{T}}\tilde{W}\mathcal{T}^{(n)})$ becomes $\sum_h \big(\tau^{(n)}_h\big)^2$, and the log-determinant of $\mathcal{T}^{(n)}$ is easy to compute (see Appendix B.2 for the details).

Also note that the entropy objective is fully compatible with amortized inference, i.e., if the variational parameters are functions (usually deep neural networks) of data points: $\boldsymbol{\nu}^{(n)} = \text{DNN}_\nu(\mathbf{x}^{(n)}; \Phi)$ and $\mathcal{T}^{(n)} = \text{DNN}_\mathcal{T}(\mathbf{x}^{(n)}; \Phi)$. In this context, the functions can map to diagonal covariance matrices (as is standard), to full rank covariance matrices, or to intermediate low-rank versions. The entropy-ELBO remains an analytic function in all these cases. As a consequence, we can use standard gradient-based approaches for analytic functions to optimize all parameters. Without an analytic ELBO, sampling-based estimation of integrals and the reparameterization trick, or similar approaches to estimate ELBO gradients are required (Sec. 4 and Appendix D.3 for details and experiments).

3.3 Entropy Annealing

Direct optimizations of ELBO objectives often result in locally optimal solutions. This observation is a main

motivation to use *annealed* versions of the ELBO objective. A very prominent example is β-annealing (Higgins et al., 2017; Huang et al., 2018). In β-annealing, the KL-divergence term is weighted: $\mathcal{L}(\Phi,\Theta) = \int q_\Phi(\mathbf{z}) \log p_\Theta(\mathbf{x}|\mathbf{z}) \mathrm{d}\mathbf{z} - \beta D_{\mathrm{KL}}\big(q_\Phi(\mathbf{z}) \,\|\, p_\Theta(\mathbf{z})\big)$.[2] The here derived entropy-based ELBOs invite to new types of annealing. As all terms of the ELBO are of the same principled type (i.e., entropy) it is straightforward to reweight the entropy contributions for annealing. The annealed objective thus becomes:

$$\begin{aligned}\mathcal{L}^{\mathcal{H}}_{\gamma,\delta}(\Phi,\Theta) = &\frac{1}{N}\sum_n \mathcal{H}[q^{(n)}_\Phi(\mathbf{z})] \\ &- \gamma\,\mathcal{H}[p_\Theta(\mathbf{z})] - \delta\,\mathcal{H}[p_\Theta(\mathbf{x}|\mathbf{z})] \;.\end{aligned} \tag{17}$$

Notice that we can effectively anneal equivalently to β-annealing by setting $\delta = \frac{1}{\beta}$ and $\gamma = 1$. By using $\gamma = \delta \geq 1$, we recover *energy tempering* (a.k.a., α-annealing) (Katahira et al., 2008; Huang et al., 2018). Equation (17) suggests a third alternative using $\gamma \geq 1$ and $\delta = 1$, which we will call *prior annealing*.

3.4 Relation to l_1-Sparse Coding

Eq. (15) as well as the ELBOs for less general Gaussian distributions (see Appendix B.2), represent analytic learning objectives for sparse coding. Therefore, it may be of interest to study the relation of entropy-ELBOs to objectives for standard l_1 sparse coding, which are likewise analytic functions. And l_1-objectives have extensively been researched (Daubechies et al., 2004; Lee et al., 2006; Beck and Teboulle, 2009; Gregor and LeCun, 2010; Hastie et al., 2015). At first sight, the similarity between entropy-ELBOs and l_1-objectives does not seem to go very far because the intricate ELBO in Eq. (15) seems very different from objectives like Eq. (2). At closer inspection, the similarity to l_1-objectives is higher than it first seems, however. In this context, consider the entropy-based ELBO for Gaussian distributions with diagonal covariance matrix (derived in Appendix B.2). If we use an annealed version in analogy to Eq. (17), the resulting objective function is given by Eq. (109) in the appendix. We set $\delta = 1$ to use prior annealing. If we now focus on the optimization of variational parameters $\boldsymbol{\nu}^{(n)}$, then just the latter two of the three entropies are relevant (the first is independent of $\boldsymbol{\nu}^{(n)}$). Removing constant terms of these remaining entropies then results in the following objective for $\boldsymbol{\nu}^{(n)}$ that has to be minimized:[3]

$$\underbrace{\frac{D}{2}\log\big(\sigma^2_{\mathrm{opt}}(\Phi,\tilde{W})\big)}_{\text{reconstruction}} + \gamma \underbrace{\sum_{h=1}^{H}\log\big(\lambda^{\mathrm{opt}}_h(\Phi)\big)}_{\text{sparsity}} \;. \tag{18}$$

Considering the result of Theorem 2 for $\sigma^2_{\mathrm{opt}}(\Phi,\tilde{W})$ for diagonal covariances, we obtain

$$\sigma^2_{\mathrm{opt}}(\Phi,\tilde{W}) \;=\; \frac{1}{D}\frac{1}{N}\sum_{n=1}^{N}\|\tilde{W}\boldsymbol{\nu}^{(n)} - \mathbf{x}^{(n)})\|^2 + \frac{H}{D}\,\bar{\tau}^2,$$

where $\bar{\tau}^2$ is just the average of $\big(\tau^{(n)}_h\big)^2$ across data points and latent dimensions (see Eq. (105)). Hence, the first term of Eq. (18) is (a monotonic function of) a mean squared reconstruction error. Using again Theorem 2, this time for $\lambda^{\mathrm{opt}}_h(\Phi)$, we can now more closely inspect the second term. For this, we define a smoothed magnitude function $|\cdot|^*$ using the function $\mathcal{M}(a)$ of Eq. (12) such that the solutions for the prior parameters become:

$$\lambda^{\mathrm{opt}}_h(\Phi) = \frac{1}{N}\sum_n \big|\nu^{(n)}_h\big|^\star \text{ with } \big|\nu^{(n)}_h\big|^\star = \tau^{(n)}_h \mathcal{M}\Big(\frac{\nu^{(n)}_h}{\tau^{(n)}_h}\Big).$$

Observe that for small $\tau^{(n)}_h$ compared to $\nu^{(n)}_h$, we indeed obtain that $\big|\nu^{(n)}_h\big|^\star \approx \big|\nu^{(n)}_h\big|$, see Appendix E. Hence, the second term of Eq. (18) penalizes large values of $\nu^{(n)}_h$ according to a (logarithmic) l_1 sparsity penalty.

Taking gradients w.r.t. $\boldsymbol{\nu}^{(n)}$ of the objective in Eq. (18) makes the similarity to classical l_1-objectives like Eq. (2) still more salient because the logarithms disappear as well as the $\bar{\tau}^2$-offset in $\sigma^2_{\mathrm{opt}}(\Phi,\tilde{W})$. However, gradients of the reconstruction and sparsity term will be weighted by $1/\sigma^2_{\mathrm{opt}}(\Phi,\tilde{W})$ and $1/\lambda^{\mathrm{opt}}_h(\Phi)$, respectively.

In the next section, we will use different values of $\gamma \geq 1$, i.e., prior annealing to investigate resulting encodings, e.g., for image patches. This will allow us to numerically investigate the similarity of (annealed) entropy-ELBOs and l_1-objectives. From a theoretical perspective, a notable difference to standard l_1-objectives is, however, that Eq. (18) is derived from the standard ELBO objective. That ELBO is itself an approximation for maximum likelihood parameter estimation. As a consequence, the optimal γ is known in our case ($\gamma = 1$), while for classical l_1-sparse coding the weighting factor is an important free parameter that has to be tuned.

4 EXPERIMENTS

We use numerical experiments to verify the feasibility of the novel entropy-based objectives. Our main interests

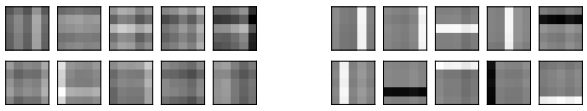
[2] In terms of entropies the KL-divergence corresponds – at optimality – to the gap between prior entropy and average variational entropy.

[3] Minimization is the convention for l_1-objectives.

will be convergence speed and insights into the effects of entropy annealing on sparsity. The source code for the experiments is available at `https://github.com/Learning-with-Entropies/sparse-coding.git`

4.1 Verification on Artificial Data

We first investigated learning based on entropy-based ELBOs using artificial data with ground-truth. Data consisted of $N = 1000$ data points with horizontal and vertical bars (Földiák, 1990; Hoyer, 2002) on a 5×5 grid ($D = 25$). We used (unamortized) Gaussian variational distributions with full covariance matrix (i.e., variational parameters for all data samples) to optimize the model given in Eq. (5) with $H = 10$. Using Eq. (15), the variational parameters were then optimized jointly with $\tilde{W} \in \mathbb{R}^{D\times H}$ by applying L-BFGS (Liu and Nocedal, 1989), which is readily available in PyTorch (Paszke et al., 2017). After convergence, each recovered generative field (GF; i.e., each column of $\tilde{W}$) contained one bar (Fig. 2). Convergence was fast: after approximately 100 L-BFGS calls to optimize the entropy-ELBO, ELBO values were already close to values of the ELBO for (ground-truth) generating parameters (for details, see Appendix D.2).



(a) Training data samples (b) Learned generative fields

Figure 2: **Artificial sparse bars dataset.** (a) Example data which is constructed by Laplace-distributed activation of horizontal and vertical bars. (b) Optimizing $\mathcal{L}^{\mathcal{H}}$ recovers the bars and their activations (up to signs of generative fields).

4.2 Natural Image Patches, Sparsity and Entropy Annealing

After verification on artificial data, we investigated entropy-ELBOs using the presumably most standard application of sparse coding models: natural image patches. Learning sparse codes for image patches based on Eq. (1) is also the most common approach to explain neuronal receptive fields in cortical area V1 (Olshausen and Field, 1996). Originally estimated with MAP approximation, further extensions were developed to employ, e.g., VAE-style training with stochastic estimation of the ELBO gradient using the "reparameterization trick" (e.g. Barello et al., 2018; Tonolini et al., 2020). Here we used analytic entropy-ELBOs to optimize the model in Eq. (5) on whitened images (Olshausen and Field, 1996) with $N = 204\,800$, $D = 16\times16$ and $H = 100$ and 400 (Appendix D.5 for experiments with $H = 400$). We explored different versions of the entropy-ELBO, Eq. (15), in order to investigate the effect of entropy-based annealing and to compare it to amortized optimization. To allow for sufficient computational efficiency, we used variational distributions with diagonal (see Eq. (108)) or low-rank covariance matrices. Concretely, we used (A) an entropy-ELBO without annealing; (B) an entropy-ELBO (Eq. (109)) with prior annealing ($\gamma \geq 1$, $\delta = 1$); (C) an entropy-ELBO using amortized variational distributions with diagonal covariance; and (D) the same amortized entropy-ELBO as in (C) but with low-rank approximation of the covariance matrices. For (D) we also use prior annealing ($\gamma \geq 1$, $\delta = 1$). For (A) and (B) we used EM-like updates: for every minibatch, we optimized variational parameters with L-BFGS and then took a gradient step to update $\tilde{W}$. For (C) and (D) neural networks were used to map data to means and to (diagonal or low-rank) covariances, and for parameter optimization we used Adam-based gradient ascent provided by the standard PyTorch implementation (Paszke et al., 2017).

For all four different versions of entropy-ELBOs, optimization ultimately resulted in the familiar Gabor-like generative fields: Figure 3 shows ELBO-optimization for the different versions, Appendix D.4 shows final GFs for $H = 100$ and $H = 400$. However, salient quantitative differences could be observed (see Fig. 3 and Fig. 4). Prior annealing of non-amortized entropy-ELBOs resulted in the fastest convergence and the highest ELBO values. Without annealing, non-amortized entropy-ELBOs finally resulted in very similar ELBO values but required longer to converge. Using also diagonal encoder covariances but amortized optimization, entropy-ELBOs converged more slowly and showed lower final ELBO values (see Fig. 3). Final ELBOs improved using low-rank covariance approximations and prior annealing (again Fig. 3). Low-rank covariances had a stronger effect on improvements (Wipf, 2023, for a related analysis) than prior annealing. In general, we observed that annealing has a comparably smaller effect for the amortized ELBO versions, which may be due to an interaction between annealing and standard Adam optimizers (Appendix D.3 for details).

Next, we were interested in the different types of anneal-

Table 1: **Different entropy annealings**. No annealing (top), prior annealing (middle) and β-annealing (bottom) are compared (Appendix D.4 for details).

ANNEALING		$\mathcal{H}[p_\Theta(\mathbf{z})]$	$\mathcal{H}[p_\Theta(\mathbf{x}\|\mathbf{z})]$	$\mathcal{H}[q_\Phi(\mathbf{z})]$	ELBO	Gini($\mathbf{z}$)±SD
No annealing		32.40	-234.70	-98.97	103.33	0.47 ± 0.04
$\mathcal{H}[p_\Theta(\mathbf{z})]$	$\gamma = 10.0$	-218.72	10.71	-365.22	-157.21	0.59 ± 0.09
	$\gamma = 2.0$	-17.29	-144.81	-112.68	49.41	0.58 ± 0.10
	$\gamma = 1.0$	29.65	-235.50	-99.67	106.18	0.48 ± 0.05
$\mathcal{H}[p_\Theta(\mathbf{x}\|\mathbf{z})]$	$\delta = 0.14$	-228.33	30.59	-234.02	-36.27	0.55 ± 0.03
	$\delta = 0.50$	-17.77	-115.24	-72.77	60.24	0.62 ± 0.04
	$\delta = 1.0$	29.74	-234.98	-99.43	105.80	0.47 ± 0.05

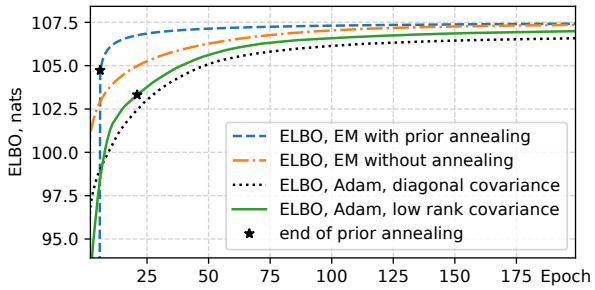
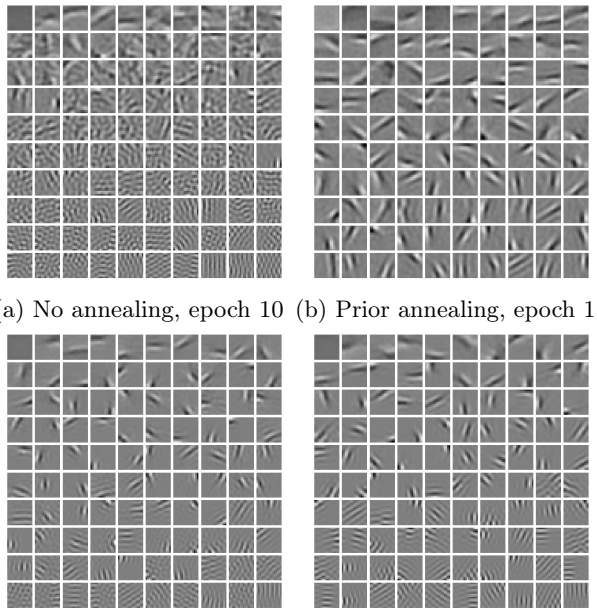


Figure 3: **Optimization of entropy-ELBOs**. Two non-amortized optimizations and two amortized optimizations are shown. Two optimizations use annealing.



(a) No annealing, epoch 10 (b) Prior annealing, epoch 1

(c) Prior annealing, epoch 10 (d) Amortized, epoch 200

Figure 4: **Learned generative fields** on natural image patches. Without annealing, the convergence is slow (a). With the prior entropy annealing, even after one epoch, we observe the familiar localized Gabor filters (b). The final GFs (c) comprise a set of Gabors and higher frequency texture-like images. Learning with amortized posterior results in similar GFs (d).

ing suggested by entropy-ELBOs, see Eq. (17). Prior annealing ($\gamma \geq 1$, $\delta = 1$) very quickly resulted in a sparse encoding and localized GFs (see Fig. 4). This is consistent with the role of γ in weighting a sparsity penalty term, see Eq. (18). Hence, it is *prior annealing* with $\gamma \geq 1$ which is analogous to high weights for the sparsity penalty in l_1 sparse coding. That prior annealing results in sparser encodings is also confirmed when, e.g., using the Gini index (Hurley and Rickard, 2009) as a measure of sparsity (both Gini values and ELBO values are high, see Table 1). In contrast, β-annealing (used in β-VAEs), represents a type of regularization different from prior annealing resulting in much less localized GFs (Appendix D.4 and Fig. 11).

Finally, our analytic objective also allows for easy estimation of ELBO values for sparse coding models optimized using standard MAP-based approaches. To show this, we used the original "sparsenet" code of Olshausen and Field (Olshausen, 1996). After optimization, we used the resulting W matrix, normalized its columns, and optimized only the variational parameters of Gaussians (means and diagonal covariances). The obtained ELBO values of 94.719 with Gini(z) $= 0.464 \pm 0.041$ indicates underfitting due to a manually selected weighting of the sparsity penalty.

5 DISCUSSION

Our main contributions are Theorems 1 to 3. Taken together, these three theorems ensured that the here derived analytic objective in Eq. (15) can be used to optimize model parameters of standard probabilistic sparse coding. Apart from MAP-based approximations with known shortcomings for uncertainty encoding (cf. Section 1), there are many other approaches for sparse coding that maintain non-trivial posterior approximations. It could, of course, be argued that for those approaches at least the optimization algorithms (if not the objectives) are described by analytic or closed-form equations. Examples are work by Seeger (2008), who used expectation propagation to derive a learning algorithm, or by Berkes et al. (2007), who used a student-t prior and Gaussian scale mixture ideas to facilitate variational optimization. The approach by Sheikh et al. (2014) also provides an analytic objective for probabilistic sparse coding but at the cost of a combinatorial discrete optimization. We also state work by Challis and Barber (2013). The contribution focuses theoretically and empirically on fully Bayesian models in which weights are sparse, and the optimization bound is therefore different. But models and bound are closely related to probabilistic sparse coding (we elaborate in Appendix B.1).

In contrast to these and other previous approaches, we here remain with the most standard choices for probabilistic sparse coding. And it is for this setting that we show the ELBO to have an analytic solution. Concretely, we remain (A) with the (by far) most standard model, Eq. (1); we use (B) the presumably most standard optimization framework (ELBOs for approximate maximum likelihood); and we use (C) the most standard posterior approximations (Gaussians). The here derived objective, presented in Eq. (15), then shows that all (high dimensional) integrals that emerge can be solved analytically. To the knowledge of the authors,

this has previously not been shown and/or empirically used. However, we remark the similarity to problems emerging for probabilistic inference using sparse weights (Challis and Barber, 2013) (Appendix B.1), and we remark that analytic solutions for standard ELBOs can also be derived without knowledge of entropy convergence (see Appendix B.3).

The results here derived do, notably, apply very generally. We have, for instance, numerically verified that potentially intricate deep neural networks (DNNs) can be used as encoders. The analytic objective then represents a deterministic DNN objective, and such objectives can conveniently be optimized with standard DNN tools. Equation (7) of Theorem 1 is still more general by applying for any decoder (linear or nonlinear) with Gaussian observables. Theorem 1 thus extends to sparse VAEs which are of recent interest (Fallah and Rozell, 2022; Drefs et al., 2023; Chen et al., 2023). Future work can consequently investigate the here presented approaches like entropy annealing for such deep sparse coding models.

Conceptually maybe most relevantly, we here for the first time investigated how an ELBO objective can be reformulated as a solely entropy-based objective. From a theoretical perspective, entropies are more deeply rooted in the foundations of probabilistic machine learning, mathematical statistics, and information theory. Furthermore, for the class of distributions usually used to define generative models (exponential family, constant base measure), entropies are closed-form and are equipped with potentially convenient properties (via their log-partition function). Also, the derivatives of entropies (that are used for learning) have similarly convenient properties, and future work can link those to information geometry. Entropy convergence has previously only been considered for analysis (Lücke and Henniges, 2012; Damm et al., 2023), and, so far, it has been unclear if or how entropy convergence can be used for learning. In this work, we provided the first demonstration that solely entropy-based objectives *can* be used for learning, and there is no principled obstacle to extending this general approach to further generative models in the future.

Acknowledgments

This work was funded by the German Research Foundation (DFG) within the priority program SPP 2298 "Theoretical Foundations of Deep Learning" - project 464104047 (FI 2583/1-1 and LU 1196/9-1). Asja Fischer also acknowledges support by the DFG under Germany's Excellence Strategy – EXC-2092 CASA – 390781972.

References

A. Alemi, B. Poole, I. Fischer, J. Dillon, R. A. Saurous, and K. Murphy. Fixing a broken elbo. In *International conference on machine learning*, pages 159–168. PMLR, 2018.

D. Barber and C. M. Bishop. Ensemble learning in bayesian neural networks. *Nato ASI Series F Computer and Systems Sciences*, 168:215–238, 1998.

G. Barello, A. S. Charles, and J. W. Pillow. Sparse-Coding Variational Auto-Encoders. *bioRxiv preprint*, 2018. doi: 10.1101/399246.

A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

P. Berkes, R. Turner, and M. Sahani. On Sparsity and Overcompleteness in Image Models. In *Advances in Neural Information Processing Systems*, 2007.

E. Challis and D. Barber. Gaussian Kullback-Leibler approximate inference. *Journal of Machine Learning Research*, 14(8), 2013.

J. Chen, R. Wang, J. He, and M. J. Li. Encouraging Sparsity in Neural Topic Modeling with Non-Mean-Field Inference. In *Machine Learning and Knowledge Discovery in Databases: Research Track*, pages 142–158, 2023.

L. Cheng, F. Yin, S. Theodoridis, S. Chatzis, and T.-H. Chang. Rethinking Bayesian Learning for Data Analysis: The art of prior and inference in sparsity-aware modeling. *IEEE Signal Processing Magazine*, 39(6):18–52, 2022.

S. Damm, D. Forster, D. Velychko, Z. Dai, A. Fischer, and J. Lücke. The ELBO of Variational Autoencoders Converges to a Sum of Entropies. In *Proc. AISTATS*, volume 206, pages 3931–3960. PMLR, 2023.

I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004.

J. Drefs, E. Guiraud, F. Panagiotou, and J. Lücke. Direct Evolutionary Optimization of Variational Autoencoders with Binary Latents. In *Proc. ECML 2022*, volume 13715 of *LNCS/LNAI*, pages 357–372. Springer, 2023.

K. Fallah and C. J. Rozell. Variational Sparse Coding with Learned Thresholding. In *Proc. ICML*, pages 6034–6058. PMLR, 2022.

P. Földiák. Forming sparse representations by local anti-Hebbian learning. *Biological Cybernetics*, 64(2): 165–170, 1990.

K. Gregor and Y. LeCun. Learning fast approximations of sparse coding. In *Proc. ICML*, page 399–406, 2010.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC press, 2015.

I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *Proc. ICLR*, 2017.

M. D. Hoffman and M. J. Johnson. ELBO surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, 2016.

P. Hoyer. Non-negative sparse coding. In *Proceedings of the 12th IEEE workshop on neural networks for signal processing*, pages 557–565. IEEE, 2002.

C.-W. Huang, S. Tan, A. Lacoste, and A. C. Courville. Improving Explorability in Variational Inference with Annealed Variational Objectives. In *Advances in Neural Information Processing Systems*, 2018.

N. Hurley and S. Rickard. Comparing Measures of Sparsity. *IEEE Transactions on Information Theory*, 55(10):4723–4741, 2009.

T. S. Jaakkola and M. I. Jordan. A Variational Approach to Bayesian Logistic Regression Models and their Extensions. In *Sixth International Workshop on Artificial Intelligence and Statistics*, pages 283–294. PMLR, 1997.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An Introduction to Variational Vethods for Graphical Models. *Machine learning*, 37:183–233, 1999.

K. Katahira, K. Watanabe, and M. Okada. Deterministic annealing variant of variational Bayes method. *Journal of Physics: Conference Series*, 95(1):012015, 2008.

D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *ICLR*, 2014.

N. Korotkov and A. Korotkov. *Integrals Related to the Error Function*. CRC Press, 2020.

N. E. Korotkov. *Integrals for applications of integral of probabilities*. 2002. ISBN 5-900777-10-3. (in Russian).

N. E. Korotkov and A. N. Korotkov. *Integrals Related to the Integrals of Probability*. 2012. ISBN 078-5-900777-18-4. (in Russian).

M. Kuss and C. Rasmussen. Assessing approximations for gaussian process classification. *Advances in Neural Information Processing Systems*, 18, 2005.

H. Lee, A. Battle, R. Raina, and A. Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems*, 2006.

D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1):503–528, 1989.

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 and J. Warnken. On the Convergence of the ELBO to Entropy Sums. *arXiv preprint arXiv:2209.03077*, 2023.

J. Mairal, F. Bach, and J. Ponce. Sparse Modeling for Image and Vision Processing. *Foundations and Trends in Computer Graphics and Vision*, 2014.

S. Mohamed, K. A. Heller, and Z. Ghahramani. Bayesian and L1 approaches for sparse unsupervised learning. In *Proc. ICML*, pages 683–690, 2012.

R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.

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

B. A. Olshausen. Sparse coding simulation software, 1996. URL `https://www.rctn.org/bruno/sparsenet/`.

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *Neural Information Processing Systems Workshop on Autodiff*, 2017.

C. J. Rozell, D. H. Johnson, R. G. Baraniuk, and B. A. Olshausen. Sparse Coding via Thresholding and Local Competition in Neural Circuits. *Neural Computation*, 20(10):2526–2563, 2008.

H. Schöpf and P. Supancic. On Bürmann's Theorem and Its Application to Problems of Linear and Nonlinear Heat Transfer and Diffusion. *The Mathematica Journal*, 16, 2014.

M. Seeger, F. Steinke, and K. Tsuda. Bayesian Inference and Optimal Design in the Sparse Linear Model. In *Proc. AISTATS*, pages 444–451. PMLR, 2007.

M. W. Seeger. Bayesian Inference and Optimal Design for the Sparse Linear Model. *Journal of Machine Learning Research*, 9(26):759–813, 2008.

A.-S. Sheikh, J. A. Shelton, and J. Lücke. A Truncated EM Approach for Spike-and-Slab Sparse Coding. *Journal of Machine Learning Research*, 15(77): 2653–2687, 2014.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3): 611–622, 1999.

F. Tonolini, B. S. Jensen, and R. Murray-Smith. Variational Sparse Coding. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, pages 690–700. PMLR, 2020.

P. M. Williams. Bayesian Regularization and Pruning Using a Laplace Prior. *Neural Computation*, 7(1): 117–143, 1995.

D. Wipf. Marginalization is not marginal: No bad vae local minima when learning optimal sparse representations. In *Proc. ICML*, 2023.

D. Yao, S. McLaughlin, and Y. Altmann. Patch-based image restoration using expectation propagation. *SIAM Journal on Imaging Sciences*, 15(1):192–227, 2022.

S. Zhao, J. Song, and S. Ermon. Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator if your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Learning Sparse Codes with Entropy-Based ELBOs: Supplementary Materials

Organization of the Appendix

A CONVERGENCE CRITERIA AND PROOFS

This section contains an additional theorem, which allows a quick check of whether a model ELBO possesses the convergence to entropies property. Additionally, we provide a more detailed proof of convergence to entropies sums for the sparse coding models using the general conditions derived in (Lücke and Warnken, 2023).

A.1 Factorization Criteria for Natural Parameters

Here, we provide a simple theorem that gives a set of *sufficient* conditions, under which the ELBO converges to a sum of entropies.

Theorem 4. *Consider a model $p_\Theta(\mathbf{x},\mathbf{z}) = p_\Theta(\mathbf{x}|\mathbf{z},\theta)p_\Theta(\mathbf{z}|\boldsymbol{\lambda})$. We assume that prior and observable (likelihood) distribution belong to an exponential family with constant base measure. If the model can be stated with the following factorization of the natural parameters,*

$$p_\Theta(\mathbf{z}|\boldsymbol{\lambda}) = \frac{1}{Z(\boldsymbol{\lambda})}\exp(-T^{\mathrm{T}}(\mathbf{z})\eta_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}))) \ , \tag{19}$$

$$p_\Theta(\mathbf{x}|\mathbf{z},\theta) = \frac{1}{Z(\theta)}\exp(-T^{\mathrm{T}}(\mathbf{x})(\eta_z(\mathbf{z})\odot\eta_\theta(\theta))) \ , \tag{20}$$

and if the Jacobians $\frac{\partial\eta_\theta(\theta)}{\partial\theta}$ and $\frac{\partial\eta_{\boldsymbol{\lambda}}(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}}$ of the parameter mappings are invertible, then the model ELBO converges to the following sum of entropies:

$$\mathcal{L}^{\mathcal{H}} = \frac{1}{N}\sum_{n=1}^{N}\mathcal{H}[q_\Phi^{(n)}(\mathbf{z})] - \mathcal{H}[p_\Theta(\mathbf{z}|\boldsymbol{\lambda})] - \mathcal{H}[p_\Theta(\mathbf{x}|\mathbf{z},\theta)] \ . \tag{21}$$

We emphasize that, unlike the general conditions (Lücke and Warnken, 2023), this theorem, although being more restrictive, allows us to not only *check*, but also to easily *construct* probability distributions for models that converge to entropies sums. If conditions from the Theorem 4 do not hold, one still has to check the more general conditions (Lücke and Warnken, 2023).

Proof. Here we spell out the proof by introducing the above requirements to the model distributions and checking the convergence at stationary points. First, we write the ELBO of such models with approximate posterior $q_\Phi^{(n)}(\mathbf{z})$ and prove the convergence for the $\mathcal{L}_2^{\mathrm{EL}}(\Phi,\Theta)$ term:

$$\mathcal{L}(\Phi,\Theta) = \tfrac{1}{N}\textstyle\sum_{n=1}^{N}\mathcal{H}[q_\Phi^{(n)}(\mathbf{z})] + \underbrace{\tfrac{1}{N}\sum_{n=1}^{N}\int q_\Phi^{(n)}(\mathbf{z})\log p_\Theta(\mathbf{z})d\mathbf{z}}_{\mathcal{L}_1^{\mathrm{EL}}(\Phi,\Theta)} + \underbrace{\tfrac{1}{N}\sum_{n=1}^{N}\int q_\Phi^{(n)}(\mathbf{z})\log p_\Theta(\mathbf{x}^{(n)}|\mathbf{z})d\mathbf{z}}_{\mathcal{L}_2^{\mathrm{EL}}(\Phi,\Theta)} \ .$$

We consider a class of distributions that belong to the exponential family with constant base measure $Z(\mathbf{z},\theta)$ and a factorizable negative energy term $E(\mathbf{x}^{(n)};\mathbf{z},\theta)$. It can be written as follows:

$$p_\Theta(\mathbf{x}^{(n)}|\mathbf{z},\theta) = \frac{1}{Z(\mathbf{z},\theta)}\exp(E(\mathbf{x}^{(n)};\mathbf{z},\theta)) \ , \tag{22}$$

$$E(\mathbf{x}^{(n)};\mathbf{z},\theta) = \left\langle T(\mathbf{x}^{(n)}),\eta(\mathbf{z},\theta)\right\rangle \ . \tag{23}$$

Here we introduce the following assumptions from the theorem:

$$Z(\mathbf{z},\theta) = Z(\theta) \ , \tag{24}$$

$$\eta(\mathbf{z},\theta) = \eta_z(\mathbf{z})\odot\eta_\theta(\theta) \ . \tag{25}$$

Then the $\mathcal{L}_2^{\mathrm{EL}}(\Phi,\Theta)$ term reads:

$$\mathcal{L}_2^{\mathrm{EL}}(\Phi,\Theta) = \frac{1}{N}\sum_{n=1}^{N}\int q_\Phi^{(n)}(\mathbf{z})\left\langle T(\mathbf{x}^{(n)}),\eta_z(\mathbf{z})\odot\eta_\theta(\theta)\right\rangle d\mathbf{z} - \log Z(\theta) \tag{26}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left\langle \eta_\theta(\theta),\int q_\Phi^{(n)}(\mathbf{z})T(\mathbf{x}^{(n)})\odot\eta_z(\mathbf{z})d\mathbf{z}\right\rangle - \log Z(\theta) \tag{27}$$

$$= \left\langle \eta_\theta(\theta),\frac{1}{N}\sum_{n=1}^{N}\int q_\Phi^{(n)}(\mathbf{z})T(\mathbf{x}^{(n)})\odot\eta_z(\mathbf{z})d\mathbf{z}\right\rangle - \log Z(\theta) \ . \tag{28}$$

We are interested in stationary points of $\mathcal{L}_2^{\mathrm{EL}}(\Phi, \Theta)$ w.r.t. θ, which means that:

$$0 = \frac{\partial \mathcal{L}_2^{\mathrm{EL}}(\Phi, \Theta^*)}{\partial \theta} = \frac{1}{N} \sum_{n=1}^{N} \int q_\Phi^{(n)}(\mathbf{z}) \frac{\partial E(\mathbf{x}^{(n)}; \mathbf{z}, \theta^*) - \log Z(\mathbf{z}, \theta^*)}{\partial \theta} d\mathbf{z} \tag{29}$$

$$= \frac{1}{N} \sum_{n=1}^{N} \int q_\Phi^{(n)}(\mathbf{z}) \left(\frac{\partial \left\langle T(\mathbf{x}^{(n)}), \eta(\mathbf{z}, \theta^*) \right\rangle}{\partial \theta} - \frac{\partial \log Z(\mathbf{z}, \theta^*)}{\partial \theta} \right) d\mathbf{z} \ . \tag{30}$$

Applying the assumptions (24) and (25), it can be rewritten as follows:

$$0 = \frac{\partial \mathcal{L}_2^{\mathrm{EL}}(\Phi, \Theta^*)}{\partial \theta} = \frac{1}{N} \sum_{n=1}^{N} \int q_\Phi^{(n)}(\mathbf{z}) \left(\frac{\partial \left\langle T(\mathbf{x}^{(n)}), \eta_z(\mathbf{z}) \odot \eta_\theta(\theta^*) \right\rangle}{\partial \theta} - \frac{\partial \log Z(\theta^*)}{\partial \theta} \right) d\mathbf{z} \tag{31}$$

$$= \frac{\partial \eta_\theta(\theta^*)}{\partial \theta} \frac{1}{N} \sum_{n=1}^{N} \int q_\Phi^{(n)}(\mathbf{z}) T(\mathbf{x}^{(n)}) \odot \eta_z(\mathbf{z}) d\mathbf{z} - \frac{\partial \log Z(\theta^*)}{\partial \theta} \ . \tag{32}$$

If the $\frac{\partial \eta_\theta(\theta^*)}{\partial \theta}$ Jacobian is invertible, it follows that:

$$\frac{1}{N} \sum_{n=1}^{N} \int q_\Phi^{(n)}(\mathbf{z}) T(\mathbf{x}^{(n)}) \odot \eta_z(\mathbf{z}) dz = \left[\frac{\partial \eta_\theta(\theta^*)}{\partial \theta} \right]^{-1} \frac{\partial \log Z(\theta^*)}{\partial \theta} \tag{33}$$

$$= \frac{\partial \log Z(\theta^*)}{\partial \eta_\theta(\theta)} \ . \tag{34}$$

Plugging (33-34) into (28) we get the final entropy representation:

$$\mathcal{L}_2^{\mathrm{EL}}(\Phi, \Theta^*) = \left\langle \eta_\theta(\theta^*), \frac{\partial \log Z(\theta^*)}{\partial \eta_\theta(\theta)} \right\rangle - \log Z(\theta^*) = -H[p_\Theta(\mathbf{x}|\mathbf{z}, \theta^*)] \ . \tag{35}$$

The last equation is a common form of entropies for exponential family distributions, see e.g. (Nielsen and Nock, 2010).

Now we take the parameterized prior distribution $p_\theta(\mathbf{z})$ and consider the $\mathcal{L}_1^{\mathrm{EL}}(\Phi, \Theta)$ term:

$$\mathcal{L}_1^{\mathrm{EL}}(\Phi, \Theta) = \frac{1}{N} \sum_{n=1}^{N} \int q_\Phi^{(n)}(\mathbf{z}) \log p_\theta(\mathbf{z}) d\mathbf{z} \ . \tag{36}$$

Similarly to the likelihood distribution, we require it to belong to the following class of exponential family distributions:

$$p_{\boldsymbol{\lambda}}(\mathbf{z}) = \exp(\eta_{\boldsymbol{\lambda}}^{\mathrm{T}}(\boldsymbol{\lambda}) \eta(\mathbf{z}) - \log Z(\boldsymbol{\lambda})) \ . \tag{37}$$

Then we can rewrite the integral over the prior (36) as:

$$\mathcal{L}_1^{\mathrm{EL}}(\Phi, \Theta) = \frac{1}{N} \sum_{n=1}^{N} \int q_\Phi^{(n)}(\mathbf{z}) \log p_{\boldsymbol{\lambda}}(\mathbf{z}) d\mathbf{z} = \left\langle \eta_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}), \frac{1}{N} \sum_{n=1}^{N} \int q_\Phi^{(n)}(\mathbf{z}) \eta(\mathbf{z}) d\mathbf{z} \right\rangle - \log Z(\boldsymbol{\lambda}) \ . \tag{38}$$

We are interested in stationary points w.r.t. the $\boldsymbol{\lambda}$ parameters. Thus, setting the derivative to zero:

$$0 = \frac{\partial \mathcal{L}_1^{\mathrm{EL}}(\Phi, \Theta^*)}{\partial \boldsymbol{\lambda}} = \frac{\partial \eta_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}^*)}{\partial \boldsymbol{\lambda}} \frac{1}{N} \sum_{n=1}^{N} \int q_\Phi^{(n)}(\mathbf{z}) \eta(\mathbf{z}) d\mathbf{z} - \frac{1}{Z(\boldsymbol{\lambda})} \frac{\partial Z(\boldsymbol{\lambda}^*)}{\partial \boldsymbol{\lambda}} \ \Rightarrow \tag{39}$$

$$\frac{1}{N} \sum_{n=1}^{N} \int q_\Phi^{(n)}(\mathbf{z}) \eta(\mathbf{z}) d\mathbf{z} = \left[\frac{\partial \eta_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}^*)}{\partial \boldsymbol{\lambda}} \right]^{-1} \frac{\partial \log Z(\boldsymbol{\lambda}^*)}{\partial \boldsymbol{\lambda}} = \frac{\partial \log Z(\boldsymbol{\lambda}^*)}{\partial \eta_{\boldsymbol{\lambda}}(\boldsymbol{\lambda})} \ . \tag{40}$$

Inserting it into (38) we get:

$$\mathcal{L}_1^{\mathrm{EL}}(\Phi,\Theta) = \left\langle \eta_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}^*), \frac{\partial \log Z(\boldsymbol{\lambda}^*)}{\partial \boldsymbol{\lambda}} \right\rangle - \log Z(\boldsymbol{\lambda}^*) = -H[p_\Theta(\mathbf{z}|\boldsymbol{\lambda}^*)] \ , \tag{41}$$

which completes the proof. □

Now we show that the Laplace prior sparse coding model fulfills these sufficient conditions. The mapping functions read:

$$\eta_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}) = \mathrm{vec}[-\frac{1}{\lambda_1}, \cdots, -\frac{1}{\lambda_H}] \ , \tag{42}$$

$$\eta_{\mathbf{z}}(\mathbf{z}) = \mathrm{vec}[\tilde{W}\mathbf{z}, -\frac{1}{2}] \ , \tag{43}$$

$$\eta_\theta(\theta) = \mathrm{vec}[-\frac{1}{\sigma^2}] \ , \tag{44}$$

$$T(\mathbf{x}) = \mathrm{vec}[\mathbf{x}, \mathbf{x}\mathbf{x}^T] \ . \tag{45}$$

Jacobians $\frac{\partial \eta_\theta(\theta)}{\partial \theta}$ and $\frac{\partial \eta_{\boldsymbol{\lambda}}(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}$ are diagonal (with non-zero elements) and thus clearly invertible, which satisfies the conditions.

A.2 Proof of Convergence to Three Entropies Using the General Conditions

Here we prove the convergence to three entropies for the sparse coding model with Laplace prior, using the general conditions (Lücke and Warnken, 2023). For this, we have to show that the parametrizations of the prior and the likelihood distributions fulfill the corresponding criteria.

Prior distribution parametrization criterion. Let $\zeta(\Psi)$ be a function that maps model parameters to the natural parameters of $p_\Theta(z)$, $\mathcal{I}_{(\Psi)} = [\frac{\partial \zeta_i(\Psi)}{\Psi_j}]$ is the Jacobian matrix of the mapping function. Then the following criterion should be met for the ELBO integral to be equal to entropy at convergence, for any function $f(\Phi,\Psi)$:

$$\mathcal{I}_{(\Psi)}^{\mathrm{T}} f(\Phi,\Psi) = 0 \implies \zeta(\Psi)^{\mathrm{T}} f(\Phi,\Psi) = 0 \ . \tag{46}$$

The Laplace prior has a very simple mapping into the natural parameter space: $\zeta(\psi) = -\frac{1}{\boldsymbol{\lambda}}$. The Jacobian is a diagonal matrix and reads:

$$\mathcal{I}_{(\Psi)}^{\mathrm{T}} = \begin{bmatrix} \frac{1}{\lambda_1^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\lambda_H^2} \end{bmatrix} \ . \tag{47}$$

Now we can check that the parametrization criterion holds:

$$\left[\frac{1}{\lambda_1^2} \quad \cdots \quad \frac{1}{\lambda_H^2}\right] f(\Phi,\Psi) = 0 \tag{48}$$

$$\implies f(\Phi,\Psi) = 0 \tag{49}$$

$$\implies \zeta(\Psi)^{\mathrm{T}} f(\Phi,\Psi) = 0 \ . \tag{50}$$

Likelihood distribution parametrization criterion. Let $\eta(\mathbf{z},\Theta)$ be a function that maps model parameters to the natural parameters of $p_\Theta(\mathbf{x}|\mathbf{z})$, $\mathcal{J}_{(\mathbf{z},\Theta)} = [\frac{\partial \eta_i(\mathbf{z},\Theta)}{\Theta_j}]$ is the Jacobian matrix of the mapping function. Then the following criterion should be met for the ELBO integral to be equal to entropy at convergence for any function $g(\mathbf{z},\Phi,\Theta)$ and a subset of parameters $\theta \in \Theta$:

$$\int \mathcal{J}_{(\mathbf{z},\theta)}^{\mathrm{T}} g(\mathbf{z},\Phi,\Theta) d\mathbf{z} = 0 \implies \int \eta(\mathbf{z},\Theta)^{\mathrm{T}} g(\mathbf{z},\Phi,\Theta) = 0 \ . \tag{51}$$

Let's check if Gaussian likelihood in the sparse coding model fulfills this criterion. The function to map z and $\Theta = \{\tilde{W}, \sigma^2\}$ to natural parameters reads:

$$\eta(z, \Theta) = \begin{bmatrix} \frac{\tilde{W}\mathbf{z}}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix} \tag{52}$$

We choose the subset $\theta = \{\sigma^2\}$. Then the Jacobian of the mapping w.r.t. θ reads:

$$\mathcal{J}^{\mathrm{T}}_{(\mathbf{z},\theta)} = \left[-\tfrac{\tilde{W}_{1,:}\mathbf{z}}{\sigma^4}, \quad \dots, \quad -\tfrac{\tilde{W}_{H,:}\mathbf{z}}{\sigma^4}, \quad \tfrac{1}{2\sigma^4} \right] \ . \tag{53}$$

Check the criterion:

$$\begin{aligned} 0 &= \int \mathcal{J}^{T}_{(\mathbf{z},\theta)} g(\mathbf{z}, \Phi, \Theta) d\mathbf{z} && (54) \\ &= \int \left[-\tfrac{\tilde{W}_{1,:}\mathbf{z}}{\sigma^4}, \quad \dots, \quad -\tfrac{\tilde{W}_{H,:}\mathbf{z}}{\sigma^4}, \quad \tfrac{1}{2\sigma^4} \right] g(\mathbf{z}, \Phi, \Theta) d\mathbf{z} && (55) \\ &= -\frac{1}{\sigma^2} \int \left[\tfrac{\tilde{W}_{1,:}\mathbf{z}}{\sigma^2}, \quad \dots, \quad \tfrac{\tilde{W}_{H,:}\mathbf{z}}{\sigma^2}, \quad -\tfrac{1}{2\sigma^2} \right] g(\mathbf{z}, \Phi, \Theta) d\mathbf{z} \ . && (56) \end{aligned}$$

We can recognize the entries of the mapping function $\eta(\mathbf{z}, \Theta)$ in the argument of the integral. We can, therefore, rewrite and conclude:

$$\begin{aligned} \frac{1}{\sigma^2} \int \eta(\mathbf{z}, \Theta)^{\mathrm{T}} g(\mathbf{z}, \Phi, \Theta) d\mathbf{z} = 0 && (57) \\ \implies \int \eta(\mathbf{z}, \Theta)^{\mathrm{T}} g(\mathbf{z}, \Phi, \Theta) d\mathbf{z} = 0 \ , && (58) \end{aligned}$$

where the last step follows from σ^2 being unequal to zero. Thus, the parametrization criterion is fulfilled.

A.3 Likelihood Convergence Proof (Gaussian)

To have this paper self-contained we here reiterate the argument that the log-likelihood term $\mathcal{L}_2(\Phi, \Theta)$ in the ELBO (see Theorem 1) with optimal observation noise reduces to the negative entropy of the likelihood (see, e.g., Damm et al. (2023), their Theorem 1 for a slightly different derivation).

The expectation of the log-likelihood under the variational posterior, denoted as $\mathcal{L}_2(\Phi, \Theta)$ in Theorem 1, reads

$$\mathcal{L}_2(\Phi, \Theta) = -\frac{1}{N} \sum_n \Big(\frac{1}{2\sigma^2} \int q^{(n)}_\Phi(\mathbf{z}) \, \|\mathbf{x}^{(n)} - \tilde{W}\mathbf{z}\|^2 \mathrm{d}\mathbf{z} \Big) - \frac{D}{2} \log\big(2\pi\sigma^2\big) \ . \tag{59}$$

Note that this is the only term in the ELBO, given in Eq. (4), that depends on the observation noise σ^2. Consequently, the derivative of the ELBO, denoted as $\mathcal{L}$, w.r.t. σ^2 is given by

$$\frac{\mathrm{d}\mathcal{L}(\Phi, \Theta)}{\mathrm{d}\sigma^2} = \frac{\mathrm{d}\mathcal{L}_2(\Phi, \Theta)}{\mathrm{d}\sigma^2} = -\frac{1}{N} \sum_n \left(\frac{1}{2\sigma^4} \int q^{(n)}_\Phi(\mathbf{z}) \|\mathbf{x}^{(n)} - \tilde{W}\mathbf{z}\|^2 \mathrm{d}\mathbf{z} \right) - \frac{D}{2\sigma^2} \ . \tag{60}$$

As $\sigma^2 > 0$, we conclude the whenever $\frac{\mathrm{d}\mathcal{L}(\Phi, \Theta)}{\mathrm{d}\sigma^2} = 0$ the following holds

$$\begin{aligned} \frac{1}{N} \sum_n \left(\frac{1}{2\sigma^2} \int q^{(n)}_\Phi(\mathbf{z}) \|\mathbf{x}^{(n)} - \tilde{W}\mathbf{z}\|^2 \mathrm{d}\mathbf{z} \right) - \frac{D}{2} = 0 && (61) \\ \implies \frac{1}{N} \sum_n \left(\frac{1}{2\sigma^2} \int q^{(n)}_\Phi(\mathbf{z}) \|\mathbf{x}^{(n)} - \tilde{W}\mathbf{z}\|^2 \mathrm{d}\mathbf{z} \right) = \frac{D}{2} \ . && (62) \end{aligned}$$

So, at optimality the high-dimensional integral in Eq. (59) has a particularly simple solution and by plugging Eq. (62) into Eq. (59) we obtain

$$\mathcal{L}_2(\Phi, \Theta) = -\frac{D}{2} - \frac{D}{2} \log\big(2\pi\sigma^2\big) = -\frac{D}{2} \log\big(2e\pi\sigma^2\big) = -\mathcal{H}[p_\Theta(\mathbf{x}|\mathbf{z})] \ , \tag{63}$$

that is, $\mathcal{L}_2$ becomes equal to the (negative) entropy of $p_\Theta(\mathbf{x}|\mathbf{z})$, which concludes the argument.

B ELBOS FOR LAPLACE PRIOR SPARSE CODING

B.1 Deriving Analytic ELBO for Laplace Prior Sparse Coding Model

Here we discuss sparse coding defined as a linear latent variable model with Laplace prior defined as:

$$p_\Theta(\mathbf{z}) = \prod_{h=1}^{H} \frac{1}{2\lambda_h} \exp\left(-\frac{|z_h|}{\lambda_h}\right) \ , \tag{64}$$

$$p_\Theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\tilde{W}\mathbf{z}, \sigma^2\mathbb{I}) \ , \tag{65}$$

$$||\tilde{W}_{:,h}|| = 1 \ . \tag{66}$$

For convenience we will use a more extended notation $p_\Theta(\mathbf{x}^{(n)}|\mathbf{z}, \tilde{W}, \sigma^2)$ for the noise distribution in Eq. (65). The ELBO with variational distribution $q_\Phi^{(n)}(\mathbf{z})$ for N data points reads:

$$\log p_\Theta(\mathbf{x}) \geq \mathcal{L}^{\mathrm{EL}} = \frac{1}{N}\sum_{n=1}^{N} \int q_\Phi^{(n)}(\mathbf{z}) \log \frac{p_\Theta(\mathbf{x}^{(n)}|\mathbf{z}, \tilde{W}, \sigma^2)\, p_\Theta(\mathbf{z})}{q_\Phi^{(n)}(\mathbf{z})} d\mathbf{z} \ . \tag{67}$$

We can rewrite it as:

$$\mathcal{L}^{\mathrm{EL}} = \frac{1}{N}\sum_{n=1}^{N} \int q_\Phi^{(n)}(\mathbf{z}) \log p_\Theta(\mathbf{x}^{(n)}|\mathbf{z}, \tilde{W}, \sigma^2) d\mathbf{z} \tag{68}$$

$$+ \frac{1}{N}\sum_{n=1}^{N} \int q_\Phi^{(n)}(\mathbf{z}) \log p_\Theta(\mathbf{z}) d\mathbf{z} \tag{69}$$

$$- \frac{1}{N}\sum_{n=1}^{N} \int q_\Phi^{(n)}(\mathbf{z}) \log q_\Phi^{(n)} d\mathbf{z} \ . \tag{70}$$

At stationary points for parameters $\{\boldsymbol{\lambda}, \sigma^2\}$, it converges to the following expression of entropy terms:

$$\mathcal{L}^{\mathcal{H}} = -\mathcal{H}[p_\Theta(\mathbf{x}|\mathbf{z}, \sigma^2)] - \mathcal{H}[p_\Theta(\mathbf{z}|\{\lambda_i\})] + \frac{1}{N}\mathcal{H}[q_\Phi^{(n)}(\mathbf{z})] \ . \tag{71}$$

For the sparse coding model, it reads:

$$\mathcal{L}^{\mathcal{H}} = -\frac{D}{2}\log(2\pi e\sigma^2) - \sum_{h=1}^{H} \log(2\lambda_h e) + \frac{1}{N}\sum_{n} \mathcal{H}[q_\Phi^{(n)}(\mathbf{z})] \ . \tag{72}$$

Detailed, from the derivation of the convergence to entropies, recall that at stationary points it holds that:

$$\frac{1}{N}\sum_{n=1}^{N} \int q_\Phi^{(n)}(\mathbf{z}) \log p_\Theta(\mathbf{x}^{(n)}|\mathbf{z}, \tilde{W}, \sigma^2) d\mathbf{z} = \frac{D}{2}\log(2\pi e\sigma^2) \tag{73}$$

$$\frac{1}{N}\sum_{n=1}^{N} \int q_\Phi^{(n)}(\mathbf{z}) \log p_\Theta(\mathbf{z}|\boldsymbol{\lambda}) d\mathbf{z} = \sum_{h=1}^{H} \log(2\lambda_h e) \ . \tag{74}$$

Now we can analytically solve the integrals and obtain expressions for optimal λ_h and σ^2 at stationary points.

We use a Gaussian distribution with full covariance as the variational distribution for each data point $\mathbf{x}^{(n)}$:

$$q_\Phi^{(n)}(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\nu}^{(n)}, \mathcal{T}^{(n)}) \ . \tag{75}$$

Let us start to with Eq. (73) and solve the integral analytically to obtain σ^2:

$$\frac{1}{N}\sum_{n=1}^{N}\int q_\Phi^{(n)}(\mathbf{z}) \log p_\Theta(\mathbf{x}^{(n)}|\mathbf{z}, \tilde{W}, \sigma^2) d\mathbf{z} \tag{76}$$

$$= -\log[Z(\mathbb{I}\sigma^2)] - \frac{1}{N}\sum_{n=1}^{N}\frac{1}{2\sigma^2}\,\mathbb{E}_{q^{(n)}(\mathbf{z})}\left[(\tilde{W}\mathbf{z} - \mathbf{x}^{(n)})^{\mathrm{T}}(\tilde{W}\mathbf{z} - \mathbf{x}^{(n)})\right] \tag{77}$$

$$= -\frac{D}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\frac{1}{N}\sum_{n=1}^{N}\left[\mathrm{tr}(\tilde{W}^{\mathrm{T}}\tilde{W}\mathcal{T}^{(n)}) + (\tilde{W}\boldsymbol{\nu}^{(n)} - \mathbf{x}^{(n)})^{\mathrm{T}}(\tilde{W}\boldsymbol{\nu}^{(n)} - \mathbf{x}^{(n)})\right] \ . \tag{78}$$

The last equation can be obtained by carefully expanding the quadratic form and taking the corresponding expectations w.r.t. the Gaussian density $q^{(n)}(\mathbf{z})$.

Taking derivative w.r.t. σ^2 and setting it to zero:

$$0 = \frac{\partial \frac{1}{N}\sum_{n=1}^{N}\int q_\Phi^{(n)}(\mathbf{z}) \log p_\Theta(\mathbf{x}^{(n)}|\mathbf{z}, \tilde{W}, \sigma^2) d\mathbf{z}}{\partial \sigma^2} \tag{79}$$

$$= -\frac{D}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\frac{1}{N}\sum_{n=1}^{N}\left[\mathrm{tr}(\tilde{W}^{\mathrm{T}}\tilde{W}\mathcal{T}^{(n)}) + (\tilde{W}\boldsymbol{\nu}^{(n)} - \mathbf{x}^{(n)})^{\mathrm{T}}(\tilde{W}\boldsymbol{\nu}^{(n)} - \mathbf{x}^{(n)})\right] \ . \tag{80}$$

Solve it w.r.t. σ^2:

$$\sigma^2 = \frac{1}{DN}\sum_{n=1}^{N}\left[\mathrm{tr}(\tilde{W}^{\mathrm{T}}\tilde{W}\mathcal{T}^{(n)}) + (\tilde{W}\boldsymbol{\nu}^{(n)} - \mathbf{x}^{(n)})^{\mathrm{T}}(\tilde{W}\boldsymbol{\nu}^{(n)} - \mathbf{x}^{(n)})\right] \ . \tag{81}$$

Next, we solve for the optimal prior scales λ_h. As Eq. (69) is the only term of the ELBO that depends on $\boldsymbol{\lambda}$, the condition for a stationary point for λ_h yields:

$$0 = \frac{\partial \frac{1}{N}\sum_{n=1}^{N}\int q_\Phi^{(n)}(\mathbf{z}) \log p_\Theta(\mathbf{z}|\boldsymbol{\lambda}) d\mathbf{z}}{\partial \lambda_h} \tag{82}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\int q_\Phi^{(n)}(\mathbf{z})\left(-\frac{1}{\lambda_h} + \frac{|z_h|}{\lambda_h^2}\right) d\mathbf{z} \tag{83}$$

$$= \frac{1}{\lambda_h^2}\frac{1}{N}\sum_{n=1}^{N}\int q_\Phi^{(n)}(\mathbf{z})(|z_h| - \lambda_h) d\mathbf{z} \tag{84}$$

$$\implies \ 0 = \frac{1}{N}\sum_{n=1}^{N}\int q_\Phi^{(n)}(\mathbf{z})|z_h| d\mathbf{z} - \frac{1}{N}\sum_{n=1}^{N}\int q_\Phi^{(n)}(\mathbf{z})\lambda_h d\mathbf{z} \tag{85}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\int q_\Phi^{(n)}(\mathbf{z})|z_h| d\mathbf{z} - \lambda_h \tag{86}$$

$$\implies \ \lambda_h = \frac{1}{N}\sum_{n=1}^{N}\int \mathcal{N}(\mathbf{z}\,|\,\boldsymbol{\nu}^{(n)}, \mathcal{T}^{(n)})\,|z_h|\,d\mathbf{z} \tag{87}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\int \mathcal{N}(z_h\,|\,\nu_h^{(n)}, \mathcal{T}_{hh}^{(n)})\,|z_h|\,dz_h \ . \tag{88}$$

The h-dimensional integral (87) w.r.t. $\mathbf{z}$ can be simplified to (88) because we can rewrite the Gaussian distribution as a product of a marginal and a conditional Gaussian $q_\Phi^{(n)}(\mathbf{z}) = q_\Phi^{(n)}(\mathbf{z}_{\setminus h}|z_h)q_\Phi^{(n)}(z_h)$:

$$= \int \mathcal{N}(z_{\setminus h}|\boldsymbol{\nu}_{\setminus h}^{(n)}(z_h), \mathcal{T}_{\setminus h}^{(n)}(z_h))\mathcal{N}(z_h|\nu_h^{(n)}, \mathcal{T}_{hh}^{(n)})|z_h|d\mathbf{z} \tag{89}$$

$$= \int \mathcal{N}(z_h|\nu_h^{(n)}, \mathcal{T}_{hh}^{(n)})|z_h| \int \mathcal{N}(z_{\setminus h}|\boldsymbol{\nu}_{\setminus h}^{(n)}(z_h), \mathcal{T}_{\setminus h}^{(n)}(z_h))d\mathbf{z}_{\setminus h}dz_h \tag{90}$$

$$= \int \mathcal{N}(z_h|\nu_h^{(n)}, \mathcal{T}_{hh}^{(n)})|z_h|dz_h \ . \tag{91}$$

To solve this integral we now make use of another integral that is known to have an analytic solution:

$$\int_0^{+\infty} z\, \exp\big(-(az+b)^2\big)\, dz = \frac{\sqrt{\pi}\, b}{2a^2}\big(\operatorname{erf}(b) - 1\big) + \frac{(-b^2)}{2a^2} \text{ for } a > 0 \ . \tag{92}$$

The analytic solution of the integral is, e.g., stated in Eq. 2.1.2 by (Korotkov and Korotkov, 2020). The book is itself based on two earlier books by the same authors (Korotkov and Korotkov, 2012) and (Korotkov, 2002). Rewriting the integral to a proper Gaussian integral by substituting mean and covariance and by multiplying by the normalizing coefficient gives:

$$\int_0^{+\infty} z\, \mathcal{N}(z|\nu, \sigma^2)\, dz = \frac{\sigma}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{\nu^2}{\sigma^2}\right) - \frac{\nu}{2}\left[\operatorname{erf}\left(-\frac{\nu}{\sqrt{2}\sigma}\right) - 1\right] \ . \tag{93}$$

The integral over the complementary set of the support reads:

$$\int_{-\infty}^0 (-z)\, \mathcal{N}(z|\nu, \sigma^2)\, dz = \int_0^{+\infty} z\, \mathcal{N}(z|-\nu, \sigma^2)\, dz \tag{94}$$

$$= \frac{\sigma}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{\nu^2}{\sigma^2}\right) + \frac{\nu}{2}\left[\operatorname{erf}\left(\frac{\nu}{\sqrt{2}\sigma}\right) - 1\right] \ . \tag{95}$$

The full integral over the magnitude of z therefore reads:

$$\int \mathcal{N}(z|\nu, \sigma^2)\, |z|\, dz = \frac{2\sigma}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{\nu^2}{\sigma^2}\right) + \frac{\nu}{2}\left[\operatorname{erf}\left(\frac{\nu}{\sqrt{2}\sigma}\right) - \operatorname{erf}\left(-\frac{\nu}{\sqrt{2}\sigma}\right)\right] \tag{96}$$

$$= \frac{2\sigma}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{\nu^2}{\sigma^2}\right) + \nu \operatorname{erf}\left(\frac{\nu}{\sqrt{2}\sigma}\right) \ . \tag{97}$$

Therefore we obtain for λ_h the expression:

$$\lambda_h = \frac{1}{N}\sum_{n=1}^N \int \mathcal{N}(z_h|\nu_h^{(n)}, \mathcal{T}_{hh}^{(n)})\, |z_h|\, dz_h \tag{98}$$

$$= \frac{1}{N}\sum_{n=1}^N \left[\frac{2\sqrt{\mathcal{T}_{hh}^{(n)}}}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(\nu_h^{(n)})^2}{\mathcal{T}_{hh}^{(n)}}\right) + \nu_h^{(n)} \operatorname{erf}\left(\frac{\nu_h^{(n)}}{\sqrt{2\mathcal{T}_{hh}^{(n)}}}\right)\right] \tag{99}$$

$$= \frac{1}{N}\sum_{n=1}^N \sqrt{\mathcal{T}_{hh}^{(n)}}\, \mathcal{M}\left(\frac{\nu_h^{(n)}}{\sqrt{\mathcal{T}_{hh}^{(n)}}}\right) \ , \tag{100}$$

where $\mathcal{M}(a) = \sqrt{\frac{2}{\pi}} \exp\left(-\frac{1}{2}a^2\right) + a \operatorname{erf}\left(\frac{a}{\sqrt{2}}\right)$ as defined in the main text, Eq. (12). So the important observation is that the integral Eq. (87) has an analytic solution, Eq. (100). In this context, we remark that integrals such as Eq. (87) emerged in other contexts of probabilistic machine learning. Concretely, Challis and Barber (2013) investigated integrals of Gaussians with different 'potential functions' including integrals with potential functions $\exp(-|x|)$ (while also other potential functions were treated). The same authors point

out (Barber and Bishop, 1998; Kuss and Rasmussen, 2005) for procedures to reduce high-dimensional to one-dimensional integrals analogously to how Eq. (88) is obtained from Eq. (87) (but marginalizations involving Gaussians are also generally well-known).

All models treated in (Challis and Barber, 2013) (theoretically and empirically) consider fully Bayesian learning using sparse weight matrices. Consequently, the used approximation bound is different from the here considered ELBOs for sparse coding (for both the entropy-based version Eq. (72) as well as the classical ELBO Eq. (111) to Eq. (113)). The emerging problems are closely related, however, and the bound treated by Challis and Barber (2013) could be reformulated to relate to the probablistic sparse coding problem using the classical ELBO Eq. (111) to Eq. (113) (which is also explicitly stated by the authors in the introduction). The integral that emerges for Laplace potentials in their site potential term of the bound is the same as the integral required to solve for λ_h in our context (see Eq. (87)); and Challis and Barber (2013) also provide an analytic soltuion for the integral. In the context of fully Bayesian approaches, entropy convergence results could, visa versa, also be applied to the bound of Challis and Barber (2013) albeit some algebraic transformations would be required. Convergence to entropy sums could then potentially be useful for models such as Gaussian process regression etc.

In principle, the analytic integral solutions that emerge in standard probabilistic sparse coding Eq. (1) are known since still earlier. For instance, Korotkov (see 2002); Korotkov and Korotkov (see 2012) provided analytic solutions of integrals that can be used for the here emerging integrals (and we used these solutions in our derivation above). Hence, analytic solutions could have been used, e.g., for work by Seeger (2008) or (Barello et al., 2018), and may prove useful in future work in these direction.

B.2 Sparse Coding. ELBO for Other Versions of Gaussian Variational Distributions

If the variational posterior is an uncorrelated Gaussian $q^{(n)}(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\nu}^{(n)}, \mathrm{diag}((\tau_1^{(n)})^2, \ldots, (\tau_H^{(n)})^2))$, the entropy-based ELBO objective can be further simplified. First, we exploit that the columns of $\tilde{W}$ are normalized, and consider the term

$$\begin{aligned}
\mathrm{tr}(\tilde{W}^{\mathrm{T}}\tilde{W}\,\mathrm{diag}((\tau_1^{(n)})^2, \ldots, (\tau_H^{(n)})^2)) &= \mathrm{diag}(\tilde{W}^{\mathrm{T}}\tilde{W})^T\,\mathrm{vec}((\tau_1^{(n)})^2, \ldots, (\tau_H^{(n)})^2) && (101)\\
&= \sum_{h=1}^{H} \tilde{W}_{:,h}^{\mathrm{T}}\tilde{W}_{:,h}(\tau_h^{(n)})^2 && (102)\\
&= \sum_{h=1}^{H} (\tau_h^{(n)})^2 \ , && (103)
\end{aligned}$$

which removes the dependency on $\tilde{W}$ here. The optimal $\lambda_{\mathrm{opt},h}$ and σ^2_{opt} then read:

$$\lambda_{\mathrm{opt},h}(\Phi) = \frac{1}{N}\sum_{n=1}^{N}\tau_h^{(n)}\mathcal{M}\left(\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\right) \tag{104}$$

$$\sigma^2_{\mathrm{opt}}(\Phi, \tilde{W}) \;=\; \frac{H}{D}\,\bar{\tau}^2 \;+\; \frac{1}{D}\frac{1}{N}\sum_{n=1}^{N}(\tilde{W}\boldsymbol{\nu}^{(n)} - \mathbf{x}^{(n)})^{\mathrm{T}}(\tilde{W}\boldsymbol{\nu}^{(n)} - \mathbf{x}^{(n)}),\ \text{where}\ \bar{\tau}^2 = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{H}\sum_{h=1}^{H}(\tau_h^{(n)})^2 \ , \tag{105}$$

which gives us a simplified entropy-based objective:

$$\begin{aligned}
\mathcal{L}^{\mathcal{H}}(\Phi, \tilde{W}) &= \frac{1}{N}\sum_{n=1}^{N}\sum_{h=1}^{H}\frac{1}{2}\log\left(2\pi e(\tau_h^{(n)})^2\right) - \sum_{h=1}^{H}\log(2e\lambda_{\mathrm{opt},h}(\Phi)) - \frac{D}{2}\log(2\pi e\sigma^2_{\mathrm{opt}}(\Phi, \tilde{W})) && (106)\\
&= \frac{1}{N}\sum_{n=1}^{N}\sum_{h=1}^{H}\frac{1}{2}\log\left(2\pi e(\tau_h^{(n)})^2\right) - \sum_{h=1}^{H}\log\left(2e\frac{1}{N}\sum_{n=1}^{N}\tau_h^{(n)}\mathcal{M}\left(\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\right)\right) && (107)\\
&\quad - \frac{D}{2}\log\left(2\pi e\left[\frac{H}{D}\,\bar{\tau}^2 \;+\; \frac{1}{D}\frac{1}{N}\sum_{n=1}^{N}(\tilde{W}\boldsymbol{\nu}^{(n)} - \mathbf{x}^{(n)})^{\mathrm{T}}(\tilde{W}\boldsymbol{\nu}^{(n)} - \mathbf{x}^{(n)})\right]\right) \ . && (108)
\end{aligned}$$

If the objective is annealed as suggested in Eq. (17), then the objective reads:

$$\mathcal{L}^{\mathcal{H}}(\Phi,\tilde{W}) = \frac{1}{N}\sum_{n=1}^{N}\sum_{h=1}^{H}\frac{1}{2}\log\left(2\pi e\left(\tau_h^{(n)}\right)^2\right) - \gamma\sum_{h=1}^{H}\log\left(2e\lambda_{\mathrm{opt},h}(\Phi)\right) - \delta\,\frac{D}{2}\log\left(2\pi e\sigma^2_{\mathrm{opt}}(\Phi,\tilde{W})\right)\ . \tag{109}$$

B.3 Sparse Coding. Classical Variational Inference Objective

Having obtained the analytic solution of the ELBO in Eq. (15), it could be asked how much the results rely on the entropy convergence results. For this, we here consider the original ELBO of the sparse coding model defined as in Eq. (1). The ELBO with variational distribution $q^{(n)}(\mathbf{z})$ for N data points reads:

$$\log p_\Theta(\mathbf{x}) \geq \mathcal{L}^{\mathrm{EL}}(\Phi,\Theta) = \frac{1}{N}\sum_{n=1}^{N}\int q_\Phi^{(n)}(\mathbf{z})\log\frac{p_\Theta(\mathbf{x}^{(n)}|\mathbf{z},W,\sigma^2)p_\Theta(\mathbf{z})}{q_\Phi^{(n)}(\mathbf{z})}d\mathbf{z}\ . \tag{110}$$

We can rewrite it as:

$$\mathcal{L}^{\mathrm{EL}}(\Phi,\Theta) = \frac{1}{N}\sum_{n=1}^{N}\int q_\Phi^{(n)}(\mathbf{z})\log p_\Theta(\mathbf{x}^{(n)}|\mathbf{z},W,\sigma^2)d\mathbf{z} \tag{111}$$

$$+\frac{1}{N}\sum_{n=1}^{N}\int q_\Phi^{(n)}(\mathbf{z})\log p_\Theta(\mathbf{z})d\mathbf{z} \tag{112}$$

$$-\frac{1}{N}\sum_{n=1}^{N}\int q_\Phi^{(n)}(\mathbf{z})\log q_\Phi^{(n)}(\mathbf{z})d\mathbf{z}\ . \tag{113}$$

Similarly, as for the entropy-based ELBO, we use full covariance Gaussian $q^{(n)}(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\nu}^{(n)},\mathcal{T}^{(n)})$ as a variational posterior distribution. The integral over the likelihood function (Eq. 111) then reads:

$$\int q_\Phi^{(n)}(\mathbf{z})\log p_\Theta(\mathbf{x}^{(n)}|\mathbf{z},W,\sigma^2)d\mathbf{z} \tag{114}$$

$$= -\log[Z(\mathbb{I}\sigma^2)] - \frac{1}{2\sigma^2}\,\mathbb{E}_{q(\mathbf{z})}\left[(W\mathbf{z}-\mathbf{x}^{(n)})^{\mathrm{T}}(W\mathbf{z}-\mathbf{x}^{(n)})\right] \tag{115}$$

$$= -\frac{D}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left[\mathrm{tr}(W^{\mathrm{T}}W\mathcal{T}^{(n)}) + (W\boldsymbol{\nu}^{(n)}-\mathbf{x}^{(n)})^{\mathrm{T}}(W\boldsymbol{\nu}^{(n)}-\mathbf{x}^{(n)})\right]\ . \tag{116}$$

Now consider the integral in Eq. (112). We can rewrite it as follows:

$$\int q_\Phi^{(n)}(\mathbf{z})\log p_\Theta(\mathbf{z})d\mathbf{z} = \int\mathcal{N}(\mathbf{z}|\boldsymbol{\nu}^{(n)},\mathcal{T}^{(n)})\Big(H\log\left(\frac{1}{2}\right) - \sum_{h=1}^{H}|z_h|\Big)d\mathbf{z} \tag{117}$$

$$= H\log\left(\frac{1}{2}\right) - \int\mathcal{N}(\mathbf{z}|\boldsymbol{\nu}^{(n)},\mathcal{T}^{(n)})\sum_{h=1}^{H}|z_h|d\mathbf{z} \tag{118}$$

$$= H\log\left(\frac{1}{2}\right) - \sum_{h=1}^{H}\int\mathcal{N}(\mathbf{z}_h|\boldsymbol{\nu}_h^{(n)},\mathcal{T}_{hh}^{(n)})|z_h|d\mathbf{z}_h \tag{119}$$

$$= H\log\left(\frac{1}{2}\right) - \sum_{h=1}^{H}\left[\frac{2\sqrt{\mathcal{T}_{hh}^{(n)}}}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\frac{(\nu_h^{(n)})^2}{\mathcal{T}_{hh}^{(n)}}\right) + \nu_h^{(n)}\,\mathrm{erf}\left(\frac{\nu_h^{(n)}}{\sqrt{2\mathcal{T}_{hh}^{(n)}}}\right)\right]\ . \tag{120}$$

That is, we can use the result obtained for the entropy ELBO and also for the classical ELBO.

Equation (113) is just a Gaussian entropy:

$$-\int q_\Phi^{(n)}(\mathbf{z})\log q(\mathbf{z})d\mathbf{z} = \mathcal{H}[q_\Phi^{(n)}(\mathbf{z})] \tag{121}$$

$$= \frac{1}{2}\log(|2\pi e\mathcal{T}^{(n)}|)\ . \tag{122}$$

Thus, the classical ELBO objective can be reformulated as follows:

$$\mathcal{L}^{\mathrm{EL}}(\Phi,\Theta) = -\frac{D}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\frac{1}{N}\sum_{n=1}^{N}\Big[\mathrm{tr}(W^{\mathrm{T}}W\mathcal{T}^{(n)}) + (W\boldsymbol{\nu}^{(n)} - \mathbf{x}^{(n)})^{\mathrm{T}}(W\boldsymbol{\nu}^{(n)} - \mathbf{x}^{(n)})\Big] \tag{123}$$

$$+ H\log\Big(\frac{1}{2}\Big) - \frac{1}{N}\sum_{n=1}^{N}\sum_{h=1}^{H}\sqrt{\mathcal{T}_{hh}^{(n)}}\,\mathcal{M}\left(\frac{\nu_h^{(n)}}{\sqrt{\mathcal{T}_{hh}^{(n)}}}\right) \tag{124}$$

$$+ \frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}\log(|2\pi e\mathcal{T}^{(n)}|)\ . \tag{125}$$

C PROOF OF THEOREM 3

In this section, we lay out the details of the proof for Theorem 3. For completeness, we first restate the entropy-based ELBO (given equivalently in Eq. (15)), which reads

$$\begin{aligned}\mathcal{L}^{\mathcal{H}}(\Phi,\tilde{W}) = &\frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}\log\big(\,|\,2\pi\,e\,\tau_h^{(n)}\,|\,\big)\\ &-\sum_{h=1}^{H}\log\Big(2\,e\,\frac{1}{N}\sum_{n=1}^{N}\tau_h^{(n)}\mathcal{M}\Big(\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big)\Big)\\ &-\frac{D}{2}\log\Big(2\pi e\frac{1}{DN}\sum_{n=1}^{N}\mathbb{E}_{q_\Phi^{(n)}(\mathbf{z})}\,\|\mathbf{x}^{(n)} - \tilde{W}\mathbf{z}\|^2\Big)\ ,\end{aligned}$$

where we again use $\tau_h^2 = \mathcal{T}_{hh}$ as a short-hand for the diagonal elements of the covariance (and, accordingly, $\tau_h = \sqrt{\mathcal{T}_{hh}}$ for their positive square root). To prove Theorem 3 we rely on the following Lemma which establishes the equality of gradients on the manifold of optimal scales and variances, i.e., all points in parameter space that satisfy Eq. (6) for non-amortized and amortized parametrizations.

Lemma 1 (Equality of gradients on manifold of optimal scales and variance)**.** *Consider the learning objectives $\mathcal{L}^{\mathrm{EL}}(\Phi,\Theta)$, given in Eq. (4), and $\mathcal{L}^{\mathcal{H}}(\Phi,\tilde{W})$, given in Eq. (15), for the probabilistic sparse coding model formulated in Eq. (5) with $\Theta = (\tilde{W},\sigma^2,\boldsymbol{\lambda}) \in \mathbb{R}^{D\times H}_{\mathrm{norm}} \times \mathbb{R}_+ \times \mathbb{R}_+^H$, and variational parameters $\Phi = (\Phi_\nu,\Phi_\mathcal{T})$ that parameterize mean $\boldsymbol{\nu}^{(n)} \in \mathbb{R}^H$ and covariance $\mathcal{T}^{(n)} \in \mathbb{R}^{H\times H}$ (in amortized or non-amortized fashion). We assume $\Phi_\nu \cap \Phi_\mathcal{T} = \emptyset$.*

Then, whenever Eq. (6) holds, it holds that

$$\begin{aligned}\nabla_\Phi\mathcal{L}^{\mathrm{EL}}(\Phi,\Theta) &= \nabla_\Phi\mathcal{L}^{\mathcal{H}}(\Phi,\tilde{W})\ ,\\ \nabla_{\tilde{W}}\mathcal{L}^{\mathrm{EL}}(\Phi,\Theta) &= \nabla_{\tilde{W}}\mathcal{L}^{\mathcal{H}}(\Phi,\tilde{W})\ .\end{aligned} \tag{126}$$

Proof. We need to prove that the gradients for Θ and Φ for both objectives are equal at all points in parameter space whenever Eq. (6) is fulfilled, i.e., whenever the scales $\boldsymbol{\lambda}$ and the variance σ^2 are optimal.

Recall that the ELBO, given in Eq. (4), can be written as (using the notation of Theorem 1)

$$\mathcal{L}^{\mathrm{EL}}(\Phi,\Theta) = \overbrace{\frac{1}{N}\sum_{n=1}^{N}\mathcal{H}[q_\Phi^{(n)}(\mathbf{z})]}^{\text{regularization (neg. KL-divergence)}} + \overbrace{\mathcal{L}_1^{\mathrm{EL}}(\Phi,\Theta) + \mathcal{L}_2^{\mathrm{EL}}(\Phi,\Theta)}^{\text{reconstruction}} \tag{127}$$

and whenever the condition of Theorem 1 (i.e., Eq. (6)) is satisfied, we observe the term-wise convergence to entropies such that the ELBO decomposes into three entropies

$$= \underbrace{\frac{1}{N}\sum_{n=1}^{N}\mathcal{H}[q_\Phi^{(n)}(\mathbf{z})]}_{\text{regularization (neg. KL-divergence)}} \underbrace{-\mathcal{H}[p_\Theta(\mathbf{z})] - \mathcal{H}[p_\Theta(\mathbf{x}|\mathbf{z})]}_{\text{reconstruction}}\ . \tag{128}$$

Note that the entropy-based objective $\mathcal{L}^{\mathcal{H}}$, given in Eq. (15), is merely the sum of entropies above with analytically optimal scale and variance parameters obtained from Theorem 2.

We need to investigate the gradients with respect to the parameters of the model $\Theta = (\tilde{W}, \sigma^2, \boldsymbol{\lambda})$ and the variational parameters $\Phi = (\Phi_\nu, \Phi_{\mathcal{T}})$. We start by addressing the model parameters Θ.

Model Parameters Θ: Considering Θ, only the parameters $w \in \tilde{W}$ are of interest.[4] Thus, only the reconstruction terms, $\mathcal{L}_2^{\mathrm{EL}}$ and its counterpart $-\mathcal{H}[p_\Theta(\mathbf{x}|\mathbf{z})]$, contribute to the gradients for $\tilde{W}$. We consider a general parameter $w \in \tilde{W}$. Regarding the standard ELBO $\mathcal{L}^{\mathrm{EL}}$, the gradient is given by

$$\frac{\partial}{\partial w}\mathcal{L}^{\mathrm{EL}}(\Phi,\Theta) = \frac{\partial}{\partial w}\mathcal{L}_2^{\mathrm{EL}}(\Theta,\Phi) \tag{129}$$

$$= -\frac{1}{2\sigma^2}\frac{\partial}{\partial w}\left[\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{q_\Phi^{(n)}(\mathbf{z})}\left\|\mathbf{x}^{(n)} - \tilde{W}\mathbf{z}\right\|^2\right] \tag{130}$$

$$= -\frac{1}{\sigma^2}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{q_\Phi^{(n)}(\mathbf{z})}\left(\mathbf{x}^{(n)} - \tilde{W}\mathbf{z}\right)\frac{\partial}{\partial w}\tilde{W}\mathbf{z}\ . \tag{131}$$

Similarly, for the entropy-based ELBO $\mathcal{L}^{\mathcal{H}}$ we obtain

$$\frac{\partial}{\partial w}\mathcal{L}^{\mathcal{H}}(\Phi,\tilde{W}) = -\frac{\partial}{\partial w}\mathcal{H}[p_\Theta(\mathbf{x}|\mathbf{z})] \tag{132}$$

$$= -\frac{D}{2}\frac{\partial}{\partial w}\log\left(2\pi e\frac{1}{ND}\sum_{n=1}^{N}\mathbb{E}_{q_\Phi^{(n)}(\mathbf{z})}\left\|\mathbf{x}^{(n)} - \tilde{W}\mathbf{z}\right\|^2\right) \tag{133}$$

$$= -D\frac{\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{q_\Phi^{(n)}(\mathbf{z})}\left(\mathbf{x}^{(n)} - \tilde{W}\mathbf{z}\right)}{\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{q_\Phi^{(n)}(\mathbf{z})}\left\|\mathbf{x}^{(n)} - \tilde{W}\mathbf{z}\right\|^2}\frac{\partial}{\partial w}\tilde{W}\mathbf{z}\ , \tag{134}$$

and with $\sigma^2_{\mathrm{opt}}(\Phi,\tilde{W}) = \frac{1}{ND}\sum_{n=1}^{N}\mathbb{E}_{q_\Phi^{(n)}(\mathbf{z})}\left\|\mathbf{x}^{(n)} - \tilde{W}\mathbf{z}\right\|^2$ (as derived in Theorem 2) we conclude

$$= -\frac{1}{\sigma^2_{\mathrm{opt}}}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{q_\Phi^{(n)}(\mathbf{z})}\left(\mathbf{x}^{(n)} - \tilde{W}\mathbf{z}\right)\frac{\partial}{\partial w}\tilde{W}\mathbf{z}\ . \tag{135}$$

Observe that both objectives yield the same gradient information for any $w \in \tilde{W}$, just scaled by a ratio that reflects how far σ^2 is from its optimal value σ^2_{opt}, such that

$$\nabla_{\tilde{W}}\mathcal{L}^{\mathcal{H}}(\Phi,\tilde{W}) = \frac{\sigma^2}{\sigma^2_{\mathrm{opt}}}\nabla_{\tilde{W}}\mathcal{L}^{\mathrm{EL}}(\Phi,(\boldsymbol{\lambda},\tilde{W},\sigma^2))\ . \tag{136}$$

Importantly, both objectives give rise to the same gradients for $\tilde{W}$ whenever $\sigma^2 = \sigma^2_{\mathrm{opt}}$, i.e., when Eq. (6) is satisfied.[5]

Variational Parameters Φ: It remains to investigate how the gradients for the variational parameters Φ are affected when training with $\mathcal{L}^{\mathrm{EL}}$ or $\mathcal{L}^{\mathcal{H}}$. We consider $\phi \in \Phi$ for which the gradient decomposes into three terms

$$\frac{\partial}{\partial\phi}\mathcal{L}^{\mathrm{EL}}(\Phi,\Theta) = \frac{\partial}{\partial\phi}\frac{1}{N}\sum_{n=1}^{N}\mathcal{H}[q_\Phi^{(n)}(\mathbf{z})] + \frac{\partial}{\partial\phi}\mathcal{L}_1^{\mathrm{EL}}(\Phi,\Theta) + \frac{\partial}{\partial\phi}\sum_{n=1}^{N}\mathcal{L}_2^{\mathrm{EL}}(\Phi,\Theta)\ , \tag{137}$$

$$\frac{\partial}{\partial\phi}\mathcal{L}^{\mathcal{H}}(\Phi,\tilde{W}) = \frac{\partial}{\partial\phi}\frac{1}{N}\sum_{n=1}^{N}\mathcal{H}[q_\Phi^{(n)}(\mathbf{z})] - \frac{\partial}{\partial\phi}\mathcal{H}[p_\Theta(\mathbf{z})] - \frac{\partial}{\partial\phi}\mathcal{H}[p_\Theta(\mathbf{x}|\mathbf{z})]\ . \tag{138}$$

[4] The remaining parameters $\sigma, \boldsymbol{\lambda}$ will be learned in case of $\mathcal{L}^{\mathrm{EL}}$, or set to optimality in case of $\mathcal{L}^{\mathcal{H}}$. However, as Eq. (6) holds in both cases they always yield zero gradients.

[5] Note that the same constraints on $\tilde{W}$ are imposed for both objectives. That is, any additive regularization terms of the form $\mathcal{L} + R(\tilde{W})$ would consequently yield the very same gradient updates for $\mathcal{L} \in \{\mathcal{L}^{\mathrm{EL}}, \mathcal{L}^{\mathcal{H}}\}$.

The average encoder entropy (the first term in Eqs. (137) and (138), respectively) is part of both objectives and consequently provides the same gradient information for any $\phi \in \Phi$ (regardless of the concrete parametrization in terms of Φ). We continue with the gradient updates arising from the reconstruction score, i.e., $\mathcal{L}_2^{\mathrm{EL}}(\Phi, \Theta)$ vs. $-\mathcal{H}[p_\Theta(\mathbf{x}|\mathbf{z})]$, the last term in Eqs. (137) and (138), respectively. Starting with the latter we get

$$-\frac{\partial}{\partial \phi}\mathcal{H}[p_\Theta(\mathbf{x}|\mathbf{z})] = -\frac{D}{2}\frac{\partial}{\partial \phi}\log\left(2\pi e \sigma^2_{\mathrm{opt}}(\Phi, \tilde{W})\right) \tag{139}$$

$$= -\frac{D}{2}\frac{1}{\sigma^2_{\mathrm{opt}}(\Phi, \tilde{W})}\frac{\partial}{\partial \phi}\sigma^2_{\mathrm{opt}}(\Phi, \tilde{W}) \tag{140}$$

and by invoking $\sigma^2_{\mathrm{opt}}(\Phi, \tilde{W}) = \frac{1}{ND}\sum_{n=1}^N \mathbb{E}_{q_\Phi(\mathbf{z}|\mathbf{x}^{(n)})}\left\|\mathbf{x}^{(n)} - \tilde{W}\mathbf{z}\right\|^2$,

$$= -\frac{1}{2\sigma^2_{\mathrm{opt}}}\frac{\partial}{\partial \phi}\left[\frac{1}{N}\sum_{n=1}^N \mathbb{E}_{q(\mathbf{z}|\mathbf{x}^{(n)})}\left\|\mathbf{x}^{(n)} - \tilde{W}\mathbf{z}\right\|^2\right] \ . \tag{141}$$

Considering the corresponding counterpart, $\mathcal{L}_2^{\mathrm{EL}}(\Phi, \Theta)$ in the classical ELBO, the gradient is directly given by

$$\frac{\partial}{\partial \phi}\mathcal{L}_2^{\mathrm{EL}}(\Phi, \Theta) = -\frac{1}{2\sigma^2}\frac{\partial}{\partial \phi}\left[\frac{1}{N}\sum_{n=1}^N \mathbb{E}_{q(\mathbf{z}|\mathbf{x}^{(n)})}\left\|\mathbf{x}^{(n)} - \tilde{W}\mathbf{z}\right\|^2\right] \ , \tag{142}$$

such that gradient updates are again scaled depending on how close σ^2 is to σ^2_{opt}

$$-\frac{\partial}{\partial \phi}\mathcal{H}[p_\Theta(\mathbf{x}|\mathbf{z})] = \frac{\sigma^2}{\sigma^2_{\mathrm{opt}}}\frac{\partial}{\partial \phi}\mathcal{L}_2^{\mathrm{EL}}(\Phi, \Theta) \ . \tag{143}$$

We are left with the middle terms in Eqs. (137) and (138), i.e., $\frac{\partial}{\partial \phi}\mathcal{L}_1^{\mathrm{EL}}(\Phi, \Theta)$ vs. $-\frac{\partial}{\partial \phi}\mathcal{H}[p_\Theta(\mathbf{z})]$. To enable the gradient computations we first need to derive a closed-form expression for $\mathcal{L}_1^{\mathrm{EL}}(\Phi, \Theta)$ with $q_\Phi^{(n)}(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\nu}^{(n)}, \mathcal{T}^{(n)})$, again with diagonal elements $\tau_h^2 = \mathcal{T}_{hh}$, and a Laplace prior with learnable scales λ_h, i.e., $p_\Theta(\mathbf{z}) = \prod_{h=1}^H \frac{1}{2\lambda_h}\exp\left(-\frac{|z_h|}{\lambda_h}\right)$.

Recall that $\mathcal{L}_1^{\mathrm{EL}}(\Phi, \Theta) = \frac{1}{N}\sum_n \int q_\Phi^{(n)}(\mathbf{z}) \log p_\Theta(\mathbf{z}) \mathrm{d}\mathbf{z}$ for which the individual summands evaluate to

$$\int q_\Phi^{(n)}(\mathbf{z}) \log p_\Theta(\mathbf{z}) \mathrm{d}\mathbf{z} = \int \mathcal{N}(\mathbf{z}|\boldsymbol{\nu}^{(n)}, \mathcal{T}^{(n)})\Big(\sum_{h=1}^H \log\left(\frac{1}{2\lambda_h}\right) - \sum_{h=1}^H \frac{|z_h|}{\lambda_h}\Big)\mathrm{d}\mathbf{z} \tag{144}$$

$$= \sum_{h=1}^H \log\left(\frac{1}{2\lambda_h}\right) - \int \mathcal{N}(\mathbf{z}|\boldsymbol{\nu}^{(n)}, \mathcal{T}^{(n)}) \sum_{h=1}^H \frac{|z_h|}{\lambda_h}\mathrm{d}\mathbf{z} \tag{145}$$

$$= -\sum_{h=1}^H \log(2\lambda_h) - \sum_{h=1}^H \frac{1}{\lambda_h}\int \mathcal{N}(z_h|\nu_h^{(n)}, \mathcal{T}_{hh}^{(n)})|z_h|\mathrm{d}z_h \tag{146}$$

$$= -\sum_{h=1}^H \log(2\lambda_h) - \sum_{h=1}^H \frac{1}{\lambda_h}\left[\sqrt{\frac{2}{\pi}}\tau_h^{(n)}\exp\left(-\frac{1}{2}\left(\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\right)^2\right) + \nu_h^{(n)}\,\mathrm{erf}\left(\frac{\nu_h^{(n)}}{\sqrt{2}\tau_h^{(n)}}\right)\right] \ . \tag{147}$$

With help of the statistic $\mathcal{M}(a) = \sqrt{\frac{2}{\pi}}\exp\left(-\frac{1}{2}a^2\right) + a\,\mathrm{erf}\left(\frac{a}{\sqrt{2}}\right)$, introduced in Eq. (12), we get

$$= -\sum_{h=1}^H \log(2\lambda_h) - \sum_{h=1}^H \frac{\tau_h^{(n)}}{\lambda_h}\mathcal{M}\Big(\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big) \tag{148}$$

such that

$$\mathcal{L}_1^{\mathrm{EL}}(\Phi, \Theta) = -\frac{1}{N}\sum_{n=1}^N\sum_{h=1}^H\left[\log(\lambda_h) + \frac{\tau_h^{(n)}}{\lambda_h}\mathcal{M}\Big(\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big) + c\right] \tag{149}$$

for some constant c (that does not influence any gradients). Note that now the functional dependency for mean and variance parameters matters for the gradient calculations, such that we need to consider $\phi \in \Phi_\nu$ and $\phi \in \Phi_\mathcal{T}$ separately.

Before proceeding we need to address the different parametrization that arises for non-amortized vs. amortized approaches. In the non-amortized setting, the variational parameters $\Phi = (\Phi_\nu, \Phi_\mathcal{T})$ directly parameterize mean and covariance of $q_\Phi^{(n)}$ (per data point $\mathbf{x}^{(n)}$), i.e., $\Phi_\nu = (\boldsymbol{\nu}^{(1)}, \ldots, \boldsymbol{\nu}^{(N)})$ and $\Phi_\mathcal{T} = (\mathcal{T}^{(1)}, \ldots, \mathcal{T}^{(N)})$. In amortized approaches, we take $\Phi = (\Phi_\nu, \Phi_\mathcal{T})$ to parameterize the two functions[6] $\boldsymbol{\nu}_\Phi : \mathbb{R}^D \rightarrow \mathbb{R}^H$ and $\mathcal{T}_\Phi : \mathbb{R}^D \rightarrow \mathbb{R}^{H\times H}$ such that mean and covariance are given as the respective function outputs, i.e,

$$\boldsymbol{\nu}^{(n)} = \boldsymbol{\nu}_\Phi(\mathbf{x}^{(n)}) \ , \tag{150}$$

$$\mathcal{T}^{(n)} = \mathcal{T}_\Phi(\mathbf{x}^{(n)}) \ . \tag{151}$$

The derivations in the sequel cover both settings, as we only need to compare the resulting gradients in terms of functions of partial derivatives $\frac{\partial \nu_h^{(n)}}{\partial \phi_\nu}$ or $\frac{\partial \tau_h^{(n)}}{\partial \phi_\tau}$, which clearly differ in amortized vs. non-amortized parametrizations, but are the same for both objectives.

Variational Parameters: Mean

Let us continue with the gradient updates for the variational mean $\boldsymbol{\nu}_\Phi$, so just the mean parameters $\phi_\nu \in \Phi_\nu$ are of interest here. Considering $\mathcal{L}^{\mathrm{EL}}$ first, the updates for $\boldsymbol{\nu}$ from $\mathcal{L}_1^{\mathrm{EL}}(\Phi, \Theta)$, in the form of Eq. (149), result in

$$\frac{\partial}{\partial \phi_\nu}\mathcal{L}_1^{\mathrm{EL}}(\Phi,\Theta) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{h=1}^{H}\frac{\tau_h^{(n)}}{\lambda_h}\frac{\partial}{\partial \phi_\nu}\mathcal{M}\Big(\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big) \tag{152}$$

$$= -\frac{1}{N}\sum_{n=1}^{N}\sum_{h=1}^{H}\frac{1}{\lambda_h}\,\mathrm{erf}\Big(\frac{1}{\sqrt{2}}\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big)\frac{\partial \nu_h^{(n)}}{\partial \phi_\nu} \ , \tag{153}$$

where we made use of the following derivative which invokes the fact that $\frac{\partial \mathcal{M}(a)}{\partial a} = \mathrm{erf}\left(\frac{a}{\sqrt{2}}\right)$,

$$\frac{\partial}{\partial \phi_\nu}\mathcal{M}\Big(\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big) = \frac{1}{\tau_h^{(n)}}\,\mathrm{erf}\Big(\frac{1}{\sqrt{2}}\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big)\frac{\partial \nu_h^{(n)}}{\partial \phi_\nu} \ . \tag{154}$$

Now, the respective gradient updates from the prior entropy of the entropy-based objective $\mathcal{L}^\mathcal{H}$ are given as

$$-\frac{\partial}{\partial \phi_\nu}\mathcal{H}[p_\Theta(\mathbf{z})] = -\frac{\partial}{\partial \phi_\nu}\sum_{h=1}^{H}\log(2e\lambda_{\mathrm{opt},h}(\Phi)) \tag{155}$$

$$= -\sum_{h=1}^{H}\frac{1}{\lambda_{\mathrm{opt},h}}\frac{\partial}{\partial \phi_\nu}\lambda_{\mathrm{opt},h}(\Phi) \tag{156}$$

$$= -\frac{1}{N}\sum_{n=1}^{N}\sum_{h=1}^{H}\frac{1}{\lambda_{\mathrm{opt},h}(\Phi)}\,\mathrm{erf}\Big(\frac{1}{\sqrt{2}}\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big)\frac{\partial \nu_h^{(n)}}{\partial \phi_\nu} \ . \tag{157}$$

The line above makes use of the following derivation

$$\frac{\partial}{\partial \phi_\nu}\lambda_{\mathrm{opt},h}(\Phi) = \frac{\partial}{\partial \phi_\nu}\frac{1}{N}\sum_{n=1}^{N}\tau_h^{(n)}\mathcal{M}\Big(\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big) \tag{158}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\frac{\tau_h^{(n)}}{\tau_h^{(n)}}\,\mathrm{erf}\Big(\frac{1}{\sqrt{2}}\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big)\frac{\partial \nu_h^{(n)}}{\partial \phi_\nu} \tag{159}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\mathrm{erf}\Big(\frac{1}{\sqrt{2}}\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big)\frac{\partial \nu_h^{(n)}}{\partial \phi_\nu} \ . \tag{160}$$

[6] Commonly, artificial neural networks are utilized here, such that Φ_ν denotes the parameters of the neural net that predicts the mean, and $\Phi_\mathcal{T}$ the parameters of the neural net that predicts the covariance. The independence assumption in this Lemma does not allow for parameter sharing between those networks. Often, the covariance is restricted to be a diagonal matrix such that $\mathcal{T}_\Phi : \mathbb{R}^D \rightarrow \mathbb{R}^H$.

By comparing Eqs. (153) and (157) we again conclude that the gradients just differ in the scaling by $1/\lambda_h$ vs. $1/\lambda_{\text{opt},h}$. That is, for optimal scales $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{\text{opt}}$ the gradients of the regularization term for the variational mean $\boldsymbol{\nu}$ coincide.

Variational Parameters: Variance

A similar result holds for (the parameters of) the variational variances. We consider $\phi_\tau \in \Phi_\mathcal{T}$ and start with the prior entropy in $\mathcal{L}^\mathcal{H}$. The gradient w.r.t. ϕ_τ reads

$$-\frac{\partial}{\partial\phi_\tau}\mathcal{H}[p_\Theta(\mathbf{z})] = -\frac{\partial}{\partial\phi_\tau}\sum_{h=1}^{H}\log(2e\lambda_{\text{opt},h}(\Phi)) \tag{161}$$

$$\overset{(11)}{=} -\frac{\partial}{\partial\phi_\tau}\sum_{h=1}^{H}\log\left(2e\frac{1}{N}\sum_{n=1}^{N}\tau_h^{(n)}\mathcal{M}\Big(\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big)\right) \tag{162}$$

$$= -\sum_{h=1}^{H}\frac{1}{\lambda_{\text{opt},h}(\Phi)}\frac{\partial}{\partial\phi_\tau}\left(\frac{1}{N}\sum_{n=1}^{N}\tau_h^{(n)}\mathcal{M}\Big(\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big)\right) \tag{163}$$

$$= -\frac{1}{N}\sum_{n=1}^{N}\sum_{h=1}^{H}\frac{1}{\lambda_{\text{opt},h}(\Phi)}\sqrt{\frac{2}{\pi}}\exp\Big(-\frac{1}{2}\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big)\frac{\partial\tau_h^{(n)}}{\partial\phi_\tau}\ , \tag{164}$$

where the last line makes use of the following derivation

$$\frac{\partial}{\partial\phi_\tau}\left(\frac{1}{N}\sum_{n=1}^{N}\tau_h^{(n)}\mathcal{M}\Big(\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big)\right) = \frac{1}{N}\sum_{n=1}^{N}\left[\mathcal{M}\Big(\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big)\frac{\partial\tau_h^{(n)}}{\partial\phi_\tau} + \tau_h^{(n)}\frac{\partial}{\partial\phi_\tau}\mathcal{M}\Big(\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big)\right] \tag{165}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left[\mathcal{M}\Big(\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big) - \frac{\nu_h^{(n)}}{\tau_h^{(n)}}\operatorname{erf}\Big(\frac{1}{\sqrt{2}}\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big)\right]\frac{\partial\tau_h^{(n)}}{\partial\phi_\tau} \tag{166}$$

$$\overset{(12)}{=} \frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{2}{\pi}}\exp\Big(-\frac{1}{2}\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big)\frac{\partial\tau_h^{(n)}}{\partial\phi_\tau}\ . \tag{167}$$

Lastly, for $\mathcal{L}^{\text{EL}}$ we consider the remaining term $\mathcal{L}_1^{\text{EL}}(\Phi)$ which gradients evaluate to[7]

$$\frac{\partial}{\partial\phi_\tau}\mathcal{L}_1^{\text{EL}}(\Phi) = -\frac{\partial}{\partial\phi_\tau}\frac{1}{N}\sum_{n=1}^{N}\sum_{h=1}^{H}\left[\log(\lambda_h) + \frac{\tau_h^{(n)}}{\lambda_h}\mathcal{M}\Big(\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big) + c\right] \tag{168}$$

$$= -\frac{\partial}{\partial\phi_\tau}\frac{1}{N}\sum_{n=1}^{N}\sum_{h=1}^{H}\left[\frac{\tau_h^{(n)}}{\lambda_h}\mathcal{M}\Big(\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big)\right] \tag{169}$$

and with the derivations in Eq. (165) – Eq. (167) we get

$$= -\frac{1}{N}\sum_{n=1}^{N}\sum_{h=1}^{H}\frac{1}{\lambda_h}\sqrt{\frac{2}{\pi}}\exp\Big(-\frac{1}{2}\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big)\frac{\partial\tau_h^{(n)}}{\partial\phi_\tau}\ . \tag{170}$$

Again, the resulting gradients in Eqs. (164) and (170) just differ in the scaling $1/\lambda_h$ vs. $1/\lambda_{\text{opt},h}$.

Note that Theorem 3 can be generalized to allow for parameter sharing, i.e., the additional assumption $\Phi_\nu \cap \Phi_\mathcal{T} = \emptyset$ can be dropped. However, we invoked this additional assumption as it (slightly) simplifies and shortens the equations and the overall proof. With the assumption $\Phi_\nu \cap \Phi_\mathcal{T} = \emptyset$, we can now summarize the term-wise gradient

[7] Recall that for $\mathcal{L}^{\text{EL}}$, $\boldsymbol{\lambda}$ is just a learnable parameter and consequently no function of Φ (in contrast to $\mathcal{L}^\mathcal{H}$).

calculations by completing Eqs. (137) and (138)

$$\begin{aligned}\frac{\partial}{\partial\phi}\mathcal{L}^{\mathrm{EL}}(\Phi,\Theta) &= \frac{\partial}{\partial\phi}\frac{1}{N}\sum_{n=1}^{N}\mathcal{H}[q_\Phi^{(n)}(\mathbf{z})] + \frac{\partial}{\partial\phi}\mathcal{L}_1^{\mathrm{EL}}(\Phi,\Theta) + \frac{\partial}{\partial\phi}\sum_{n=1}^{N}\mathcal{L}_2^{\mathrm{EL}}(\Phi,\Theta)\\ &= \frac{\partial}{\partial\phi}\frac{1}{N}\sum_{n=1}^{N}\mathcal{H}[q_\Phi^{(n)}(\mathbf{z})] - \frac{1}{N}\sum_{n=1}^{N}\sum_{h=1}^{H}\frac{1}{\lambda_h}\,\mathrm{erf}\Big(\frac{1}{\sqrt{2}}\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big)\frac{\partial\nu_h^{(n)}}{\partial\phi}\\ &\quad - \frac{1}{N}\sum_{n=1}^{N}\sum_{h=1}^{H}\frac{1}{\lambda_h}\sqrt{\frac{2}{\pi}}\exp\Big(-\frac{1}{2}\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big)\frac{\partial\tau_h^{(n)}}{\partial\phi} - \frac{1}{2\sigma^2}\frac{\partial}{\partial\phi}\left[\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{q(\mathbf{z}|\mathbf{x}^{(n)})}\left\|\mathbf{x}^{(n)}-\tilde{W}\mathbf{z}\right\|^2\right]\,,\end{aligned} \tag{171}$$

$$\begin{aligned}\frac{\partial}{\partial\phi}\mathcal{L}^{\mathcal{H}}(\Phi,\tilde{W}) &= \frac{\partial}{\partial\phi}\frac{1}{N}\sum_{n=1}^{N}\mathcal{H}[q_\Phi^{(n)}(\mathbf{z})] - \frac{\partial}{\partial\phi}\mathcal{H}[p_\Theta(\mathbf{z})] - \frac{\partial}{\partial\phi}\mathcal{H}[p_\Theta(\mathbf{x}|\mathbf{z})]\\ &= \frac{\partial}{\partial\phi}\frac{1}{N}\sum_{n=1}^{N}\mathcal{H}[q_\Phi^{(n)}(\mathbf{z})] - \frac{1}{N}\sum_{n=1}^{N}\sum_{h=1}^{H}\frac{1}{\lambda_{\mathrm{opt},h}(\Phi)}\,\mathrm{erf}\Big(\frac{1}{\sqrt{2}}\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big)\frac{\partial\nu_h^{(n)}}{\partial\phi}\\ &\quad - \frac{1}{N}\sum_{n=1}^{N}\sum_{h=1}^{H}\frac{1}{\lambda_{\mathrm{opt},h}(\Phi)}\sqrt{\frac{2}{\pi}}\exp\Big(-\frac{1}{2}\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big)\frac{\partial\tau_h^{(n)}}{\partial\phi} - \frac{1}{2\sigma^2_{\mathrm{opt}}}\frac{\partial}{\partial\phi}\left[\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{q(\mathbf{z}|\mathbf{x}^{(n)})}\left\|\mathbf{x}^{(n)}-\tilde{W}\mathbf{z}\right\|^2\right]\,.\end{aligned} \tag{172}$$

Overall, at points in parameter space that satisfy Eq. (6), each pair of terms gives rise to the same gradients at stationary points as all scaling coefficients (highlighted in light red) coincide (or the respective terms already provide the very same gradient information regardless of whether Eq. (6) is satisfied). Consequently, the full gradients for all trainable parameters, i.e., the sum of all constitutive terms as given in Eqs. (171) and (172) for Φ and Eq. (136) for $\tilde{W}$, are equivalent whenever Eq. (6) holds, or simply: Eq. (6) implies

$$\nabla_\Phi\mathcal{L}^{\mathrm{EL}}(\Phi^\star,\Theta^\star) = \nabla_\Phi\mathcal{L}^{\mathcal{H}}(\Phi^\star,\tilde{W}^\star) \text{ and } \nabla_{\tilde{W}}\mathcal{L}^{\mathrm{EL}}(\Phi^\star,\Theta^\star) = \nabla_{\tilde{W}}\mathcal{L}^{\mathcal{H}}(\Phi^\star,\tilde{W}^\star)\ .$$

□

We are now ready to prove Theorem 3 from the main paper.

Theorem 3 (Restated from main paper)**.** *Consider the sparse coding model formulated in Eq.* (5) *with model parameters* $\Theta = (\tilde{W},\sigma^2,\boldsymbol{\lambda}) \in \mathbb{R}^{D\times H}_{\mathrm{norm}} \times \mathbb{R}_+ \times \mathbb{R}_+^H$*, and variational parameters* $\Phi = (\Phi_\nu,\Phi_\mathcal{T})$ *that parameterize mean* $\boldsymbol{\nu}^{(n)} \in \mathbb{R}^H$ *and covariance* $\mathcal{T}^{(n)} \in \mathbb{R}^{H\times H}$ *(in amortized or non-amortized fashion) where* $\Phi_\nu \cap \Phi_\mathcal{T} = \emptyset$*.*

Then, the set of stationary points of the original objective $\mathcal{L}^{\mathrm{EL}}(\Phi,\Theta)$*, given in Eq.* (4)*, and of the entropy-based objective* $\mathcal{L}^{\mathcal{H}}(\Phi,\tilde{W})$*, given in Eq.* (15)*, coincide. Furthermore, it applies at any stationary point of* $\mathcal{L}^{\mathrm{EL}}(\Phi,\Theta)$ *or* $\mathcal{L}^{\mathcal{H}}(\Phi,\tilde{W})$ *that*

$$\mathcal{L}^{\mathrm{EL}}(\Phi^\star,\Theta^\star) = \mathcal{L}^{\mathcal{H}}(\Phi^\star,\tilde{W}^\star)\ . \tag{173}$$

Proof. To prove that the sets of stationary points of $\mathcal{L}^{\mathrm{EL}}$ and $\mathcal{L}^{\mathcal{H}}$ are equal it suffices to show the following two statements:

(A) $(\Phi^\star,\Theta^\star)$ is a stationary point of $\mathcal{L}^{\mathrm{EL}}(\Phi,\Theta) \Rightarrow (\Phi^\star,\tilde{W}^\star)$ is a stationary point of $\mathcal{L}^{\mathcal{H}}(\Phi,\tilde{W})$,

(B) $(\Phi^\star,\tilde{W}^\star)$ is a stationary point of $\mathcal{L}^{\mathcal{H}}(\Phi,W) \Rightarrow (\Phi^\star,(\boldsymbol{\lambda}_{\mathrm{opt}},\tilde{W}^\star,\sigma^2_{\mathrm{opt}}))$ is a stationary point of $\mathcal{L}^{\mathrm{EL}}(\Phi,\Theta)$.

We start with statement (A). Let $(\Phi^\star,\Theta^\star)$ be an arbitrary stationary point of $\mathcal{L}^{\mathrm{EL}}(\Phi,\Theta)$. By definition of fixed points, Eq. (6) holds such that $\Theta^\star = (\tilde{W}^\star,\sigma^2_{\mathrm{opt}},\boldsymbol{\lambda}_{\mathrm{opt}})$.[8] As $(\Phi^\star,\Theta^\star)$ is a stationary point of $\mathcal{L}^{\mathrm{EL}}(\Phi,\Theta)$ we have

$$\begin{aligned}\nabla_{\tilde{W}}\mathcal{L}^{\mathcal{H}}(\Phi^\star,\tilde{W}^\star) = \nabla_{\tilde{W}}\mathcal{L}^{\mathrm{EL}}(\Phi^\star,\Theta^\star) = 0\ ,\\ \nabla_\Phi\mathcal{L}^{\mathcal{H}}(\Phi^\star,\tilde{W}^\star) = \nabla_\Phi\mathcal{L}^{\mathrm{EL}}(\Phi^\star,\Theta^\star) = 0\end{aligned}$$

[8] Recall that any local optima for $\boldsymbol{\lambda}$ and σ^2 are in fact the (respective) global optima as both problems are convex (see Theorem 2, which also provides the analytic solutions).

as the gradients w.r.t. Φ and $\tilde{W}$ for both objectives are equal by Lemma 1 (which accounts for the technicalities of different parameterizations, that arise from amortized vs. non-amortized approaches). Consequently, $(\Phi^\star, \tilde{W}^\star)$ must also a stationary point of the entropy-based objective $\mathcal{L}^{\mathcal{H}}(\Phi, \tilde{W})$.

To show the opposite direction, formulated in statement (B), we assume that $(\Phi^\star, \tilde{W}^\star)$ is a stationary point of $\mathcal{L}^{\mathcal{H}}(\Phi, \tilde{W})$. By design of the entropy-based objective, Eq. (6) is satisfied as scales and variance are chosen to be optimal for $\mathcal{L}^{\mathcal{H}}(\Phi, \tilde{W})$ in each iteration. We can therefore invoke Lemma 1 again and get

$$\nabla_{\tilde{W}} \mathcal{L}^{\mathrm{EL}}(\Phi^\star, \Theta^\star) = \nabla_{\tilde{W}} \mathcal{L}^{\mathcal{H}}(\Phi^\star, \tilde{W}^\star) = 0 \ ,$$
$$\nabla_{\Phi} \mathcal{L}^{\mathrm{EL}}(\Phi^\star, \Theta^\star) = \nabla_{\Phi} \mathcal{L}^{\mathcal{H}}(\Phi^\star, \tilde{W}^\star) = 0 \ .$$

Therefore, $(\Phi^\star, (\tilde{W}^\star, \sigma^2_{\mathrm{opt}}, \boldsymbol{\lambda}_{\mathrm{opt}}))$ must also be a stationary point of $\mathcal{L}^{\mathrm{EL}}(\Phi, \Theta)$.

Note that the objective functions $\mathcal{L}^{\mathrm{EL}}$ and $\mathcal{L}^{\mathcal{H}}$ are continuous and continuously differentiable functions. From Lemma 1 we can also conclude that the Hessians of $\mathcal{L}^{\mathrm{EL}}$ and $\mathcal{L}^{\mathcal{H}}$ in Φ and $\tilde{W}$ coincide at stationary points as they admit the very same functional dependencies in Φ and $\tilde{W}$. This implies the same convergence behavior in the vicinity (ϵ-ball) around the fixed points such that both objectives have the same stationary points (with same signature).

Eventually, by Theorem 1 also the function values coincide whenever Eq. (6) holds, which concludes the proof. □

D NUMERICAL RESULTS – DETAILS AND ADDITIONAL RESULTS

The numerical experiments were run on a desktop computer with Intel i9-9900k 3.6GHz CPU, 32GB RAM, and Nvidia GeForce GTX 1070 8GB. We used CUDA numerical backend for PyTorch whenever possible. The default floating point precision was set to float32. On average, optimization of one epoch of 204 800 image patches of size 16×16, with latent dimensionality of 100, by minibatches of 512 with EM-like updates took 156s. One epoch of optimization with stochastic updates by Adam took on average 12s.

D.1 Approximating the error function

While the exact $\mathrm{erf}(\cdot)$ evaluation requires the summing of an infinite number of terms, e.g. of its Taylor series expansion, its approximate computation is heavily optimized in common numerical libraries. To get a closed-form objective, we experimented with a simple second-order Bürmann approximation (Schöpf and Supancic, 2014):

$$\mathrm{erf}(x) \approx \frac{2}{\sqrt{\pi}} \sqrt{1 - e^{-x^2}} \left(\frac{\sqrt{\pi}}{2} + \frac{21}{200} e^{-kx^2} - \frac{341}{8000} e^{-2kx^2} \right) \ . \tag{174}$$

We did not find any significant difference in optimization results when compared to $\mathrm{erf}(\cdot)$ implemented in numerical libraries, but our naïve implementation led to 20-30% longer run time. In all our experiments we always used the $\mathrm{erf}(\cdot)$ implementation provided by numerical libraries.

D.2 Bars dataset

We generated the training data according to the model defined in Eq. (1). That is, we sampled activation vectors $\mathbf{z}^{(n)}$ from a Laplace distribution with $\lambda_h = 1$ (for all h), linearly combined the weighted generative fields, and added Gaussian noise with standard deviation $\sigma = 0.1$. Each ground truth generative field $W_{:,h}$ contained exactly one (horizontal or vertical) bar (value 1 for 'bar', value 0 as the background).

For this experiment, we used full covariance Gaussian variational posterior. We observed good convergence and complete recovery of the bars in approximately 70% of runs (7 out of 10). When the model converged to a local optimum, some of the recovered generative fields usually contained two bars, and the final ELBO was slightly lower. During the optimization ELBO values quickly approach the value computed with ground truth $\tilde{W}$, then ELBO values asymptotically converge. Fig. 5 illustrates the typical trajectories of the $\mathcal{L}^{\mathcal{H}}$ and different entropies during the optimization.

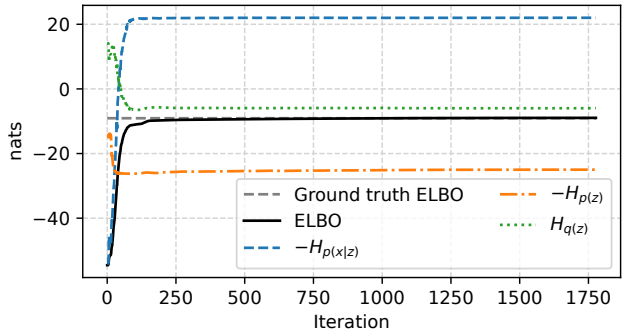


Figure 5: **Learning the artificial sparse bars dataset.** While the entropy-based ELBO is monotonously increasing, different entropy terms may undergo complex dynamics during the optimization.

D.3 Amortized learning

Our entropy-based objective can be combined with amortized inference and stochastic updates. We used a deep neural network that comprises two ResNet-like nonlinear mappings (parametric automorphisms) and separate linear readout maps for the mean and the diagonal covariance variational parameters of the posterior (Fig. 6), optimized by stochastic updates (Adam with $lr = 10^{-3}$). We compared the convergence speed to the previously suggested (non-amortized) EM-like updates, and considered cases with and without prior entropy annealing (Fig. 3). The EM updates with annealing allow the ELBO to be optimized faster and reach a better optimum. We also observe a minor gap (presumably an amortization gap) due to the limited neural network capacity. All three optimization methods finally result in a set of similar generative fields (Fig. 8).

Low-rank approximation of full covariance matrices for the variational posterior (Fig. 7) helps to diminish the amortization gap. To construct a low-rank convariance matrix, the DNN produces a set of r vectors $V \in \mathbb{R}^{H\times r}$, and a separate vecor of diagonal covariances σ^2. The covariance matrix is then computed as $\mathcal{T} = VV^{\mathrm{T}} + \mathrm{diag}(\sigma^2)$. We used $r = 5$ in our experiments.

We did not observe parameters convergence in reasonable time when we trained Laplace-prior sparse coding model with unamortized Gaussian posterior and used reparameterization trick and stochastic updates with even 100 samples (no analytic solutions of the ELBO integrals).

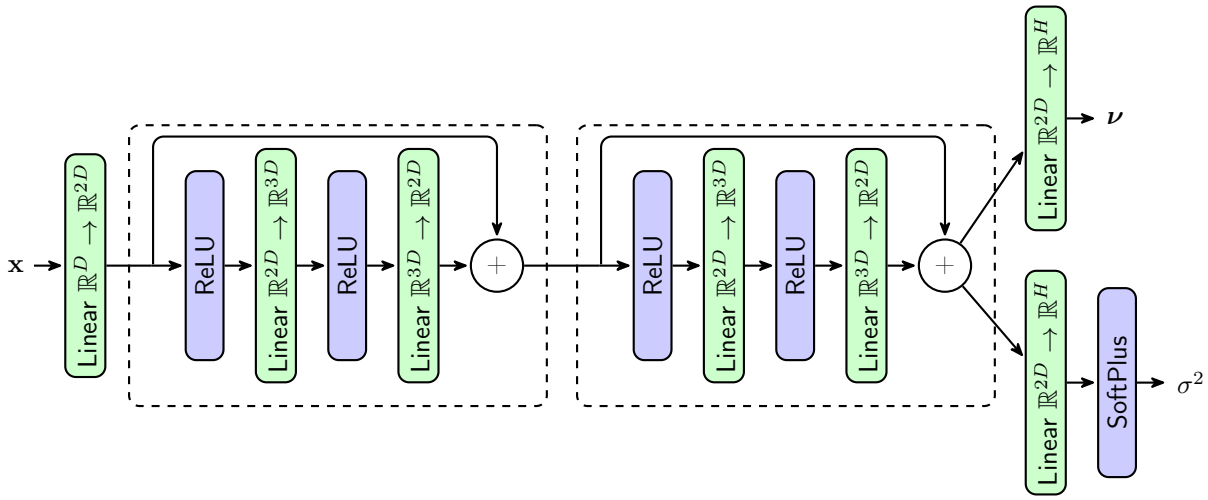


Figure 6: **Deep encoder architecture**. First, the input data $\mathbf{x}$ is linearly projected to a higher dimensional space, and then two ResNet-like transformations are applied. The variational parameters $\boldsymbol{\nu}$ and σ^2 are obtained by separate linear mappings. Posterior diagonal covariance is then constructed as $\mathcal{T} = \mathrm{diag}(\sigma^2)$.

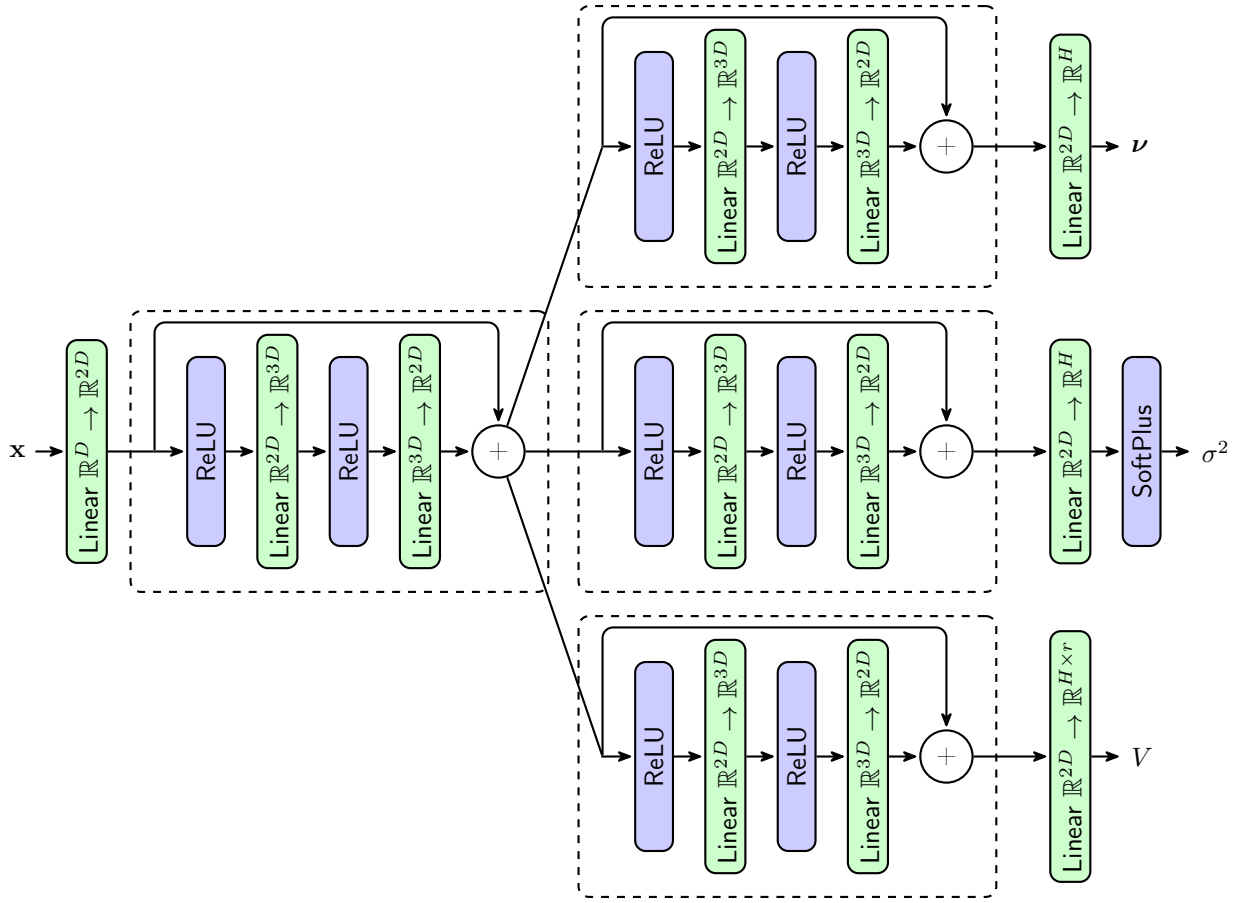
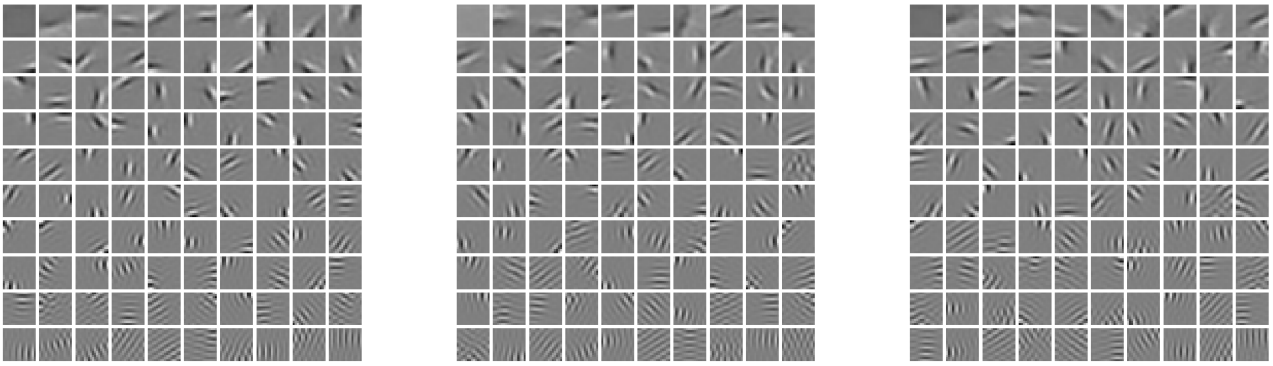


Figure 7: **Encoder architecture for variational posterior with a low-rank approximation of full covariance**. The covariance matrix is constructed as $\mathcal{T} = VV^{\mathrm{T}} + \mathrm{diag}(\sigma^2)$.



(a) EM, annealing (b) EM, no annealing (c) Adam, amortized

Figure 8: **Learned bases with different optimization methods** where the generative fields were obtained after 200 epochs of optimization. All methods result in practically the same set of filters, but the prior entropy annealing facilitates fast convergence.

D.4 Comparing annealing schemes

We used a basic linear annealing scheme for prior entropy annealing: $\gamma_i = \max(1.0, 2 * (5 - i))$ for epoch i. For the likelihood entropy annealing, we set $\delta_i = \min(1.0, 1/(7 - i))$. While prior (Fig. 10) and likelihood (Fig. 11) entropy annealing result in similar generative fields after convergence, the trajectories of the optimization of

generative fields and latent codes differ. With prior entropy annealing, all the latent dimensions are used for the encoding from the very beginning of the optimization. In the case of the likelihood annealing, the latent dimensions start contributing gradually to the reconstruction.

Table 1 provides numerical details and gives some insights into how the model parameters behave during the above-mentioned annealing. The table shows ELBOs, Gini coefficients, and contributions of different entropies to the entropy-based ELBO. To compute the ELBO, after every epoch, we evaluated the non-annealed ELBO for the learned parameters and the full dataset. Thus, the highest (and also the only proper) ELBO can be obtained only when a non-annealed objective is used for the optimization. We selected some of the annealing epochs, for which the contribution of the annealing coefficient causes a quantitatively similar balance of the contributing entropies to the annealed ELBO, that is, e.g., for the epochs when $\gamma = 2$ and $\delta = 0.5$ the corresponding contributing entropies are close. Despite the similarity in the values of the entropies, the learned generative fields are qualitatively very different (Fig. 10 and Fig. 11). Next, we provide an explanation of what causes such a qualitative difference.

Notice that the Gini coefficient of the latent codes is high during the annealing, which indicates high sparsity of the posterior. Here we have to remember that the likelihood annealing trades off the reconstruction quality to Kullback-Leibler divergence between the prior and the variational posterior. With small δ we can largely ignore the reconstruction term and focus only on the Kullback-Leibler divergence. Our reparameterized model allows two ways to minimize the Kullback-Leibler divergence term: by adjusting the variational posterior parameters, and by changing the prior scales. We observe both phenomena in the case of the likelihood entropy annealing, which leads to noisy and non-localized generative fields and posterior collapse of some of the latent dimensions. That is, some of the latent dimensions do not participate in the encoding, their corresponding scale parameters λ_h shrink to very small values, the corresponding contribution to the Kullback-Leibler divergence may become arbitrarily close to 0, and the corresponding generative fields do not contribute to the data reconstruction. Fig. 11 illustrates such noisy generative fields, which do not contribute to the reconstruction. Noisy and non-localized generative fields entirely disappear as the annealing ends.

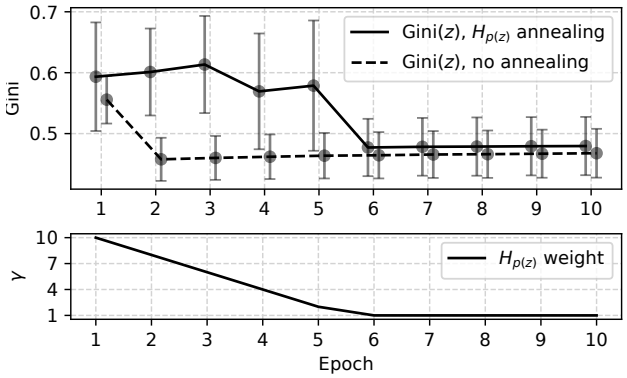
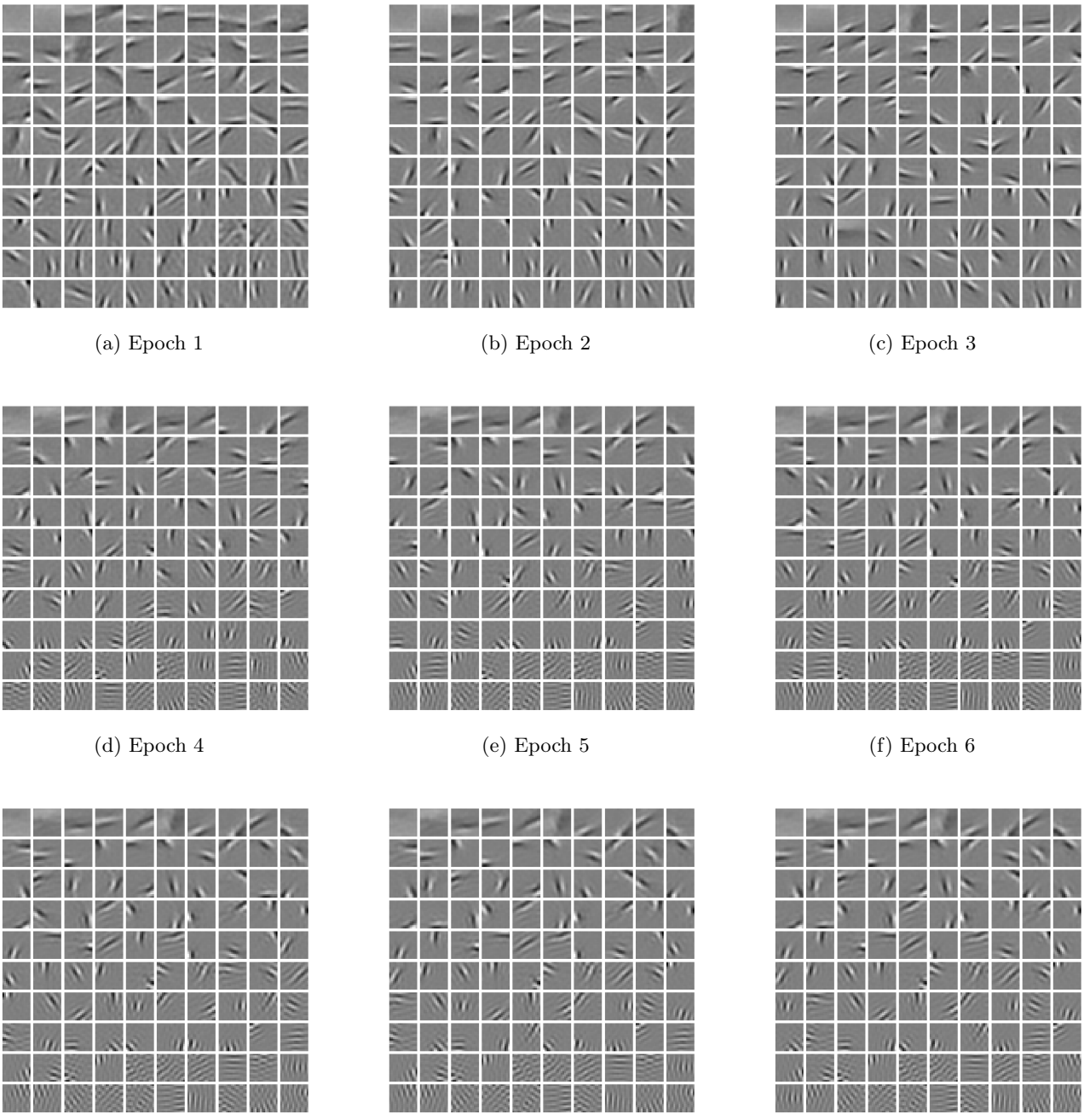


Figure 9: **Prior entropy annealing on natural image patches dataset**. Gini coefficients (mean ±SD bars) of the latent codes (Fig. 4 for example generative fields) stay marginally higher if the prior entropy annealing is used even after the annealing ends after epoch 5. The bottom plot shows the annealing schedule.

The qualitative difference of prior entropy annealing can be seen by comparing the generative fields we obtain during the optimization (Fig. 10). Even after one epoch with high weight on the prior entropy, the generative fields already resemble localized Gabor filters. As the annealing decays, more generative fields that represent high-frequency Gabors emerge. Fig. 9 shows the linear annealing schedule and how the Gini coefficient changes during the prior entropy annealing.

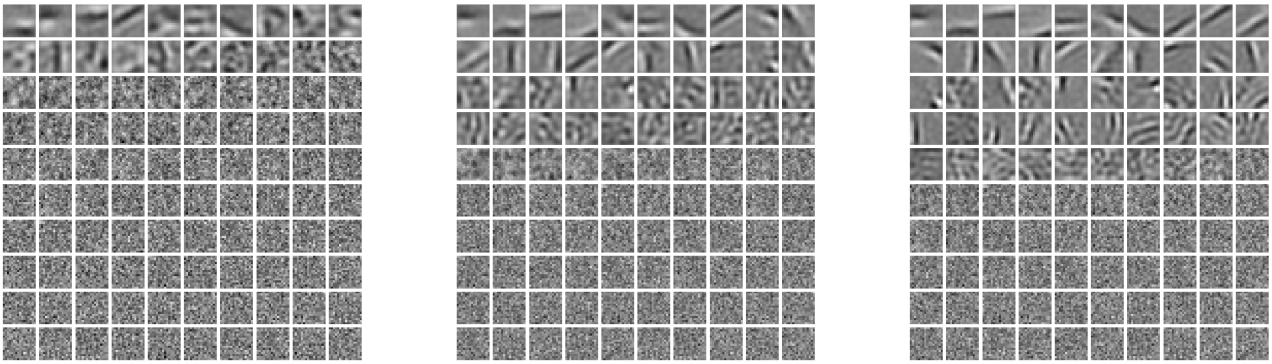


(a) Epoch 1 (b) Epoch 2 (c) Epoch 3

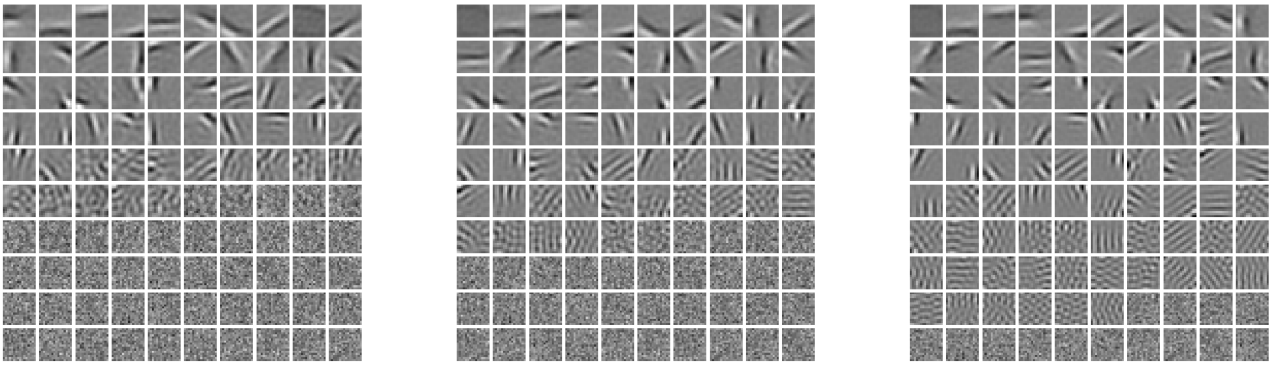
(d) Epoch 4 (e) Epoch 5 (f) Epoch 6

(g) Epoch 7 (h) Epoch 8 (i) Epoch 9

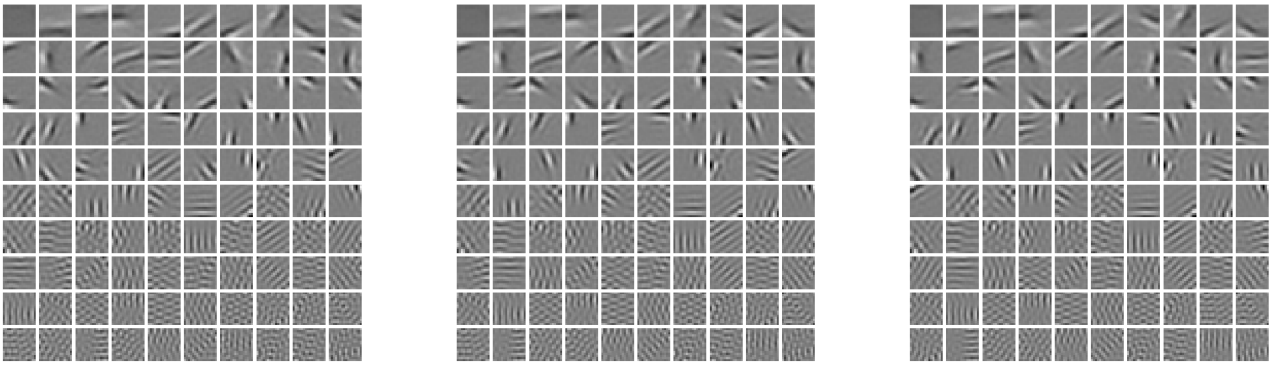
Figure 10: **Learned generative fields during optimization with prior entropy annealing**. The annealing stops after epoch 5.



(a) Epoch 1 (b) Epoch 2 (c) Epoch 3



(d) Epoch 4 (e) Epoch 5 (f) Epoch 6

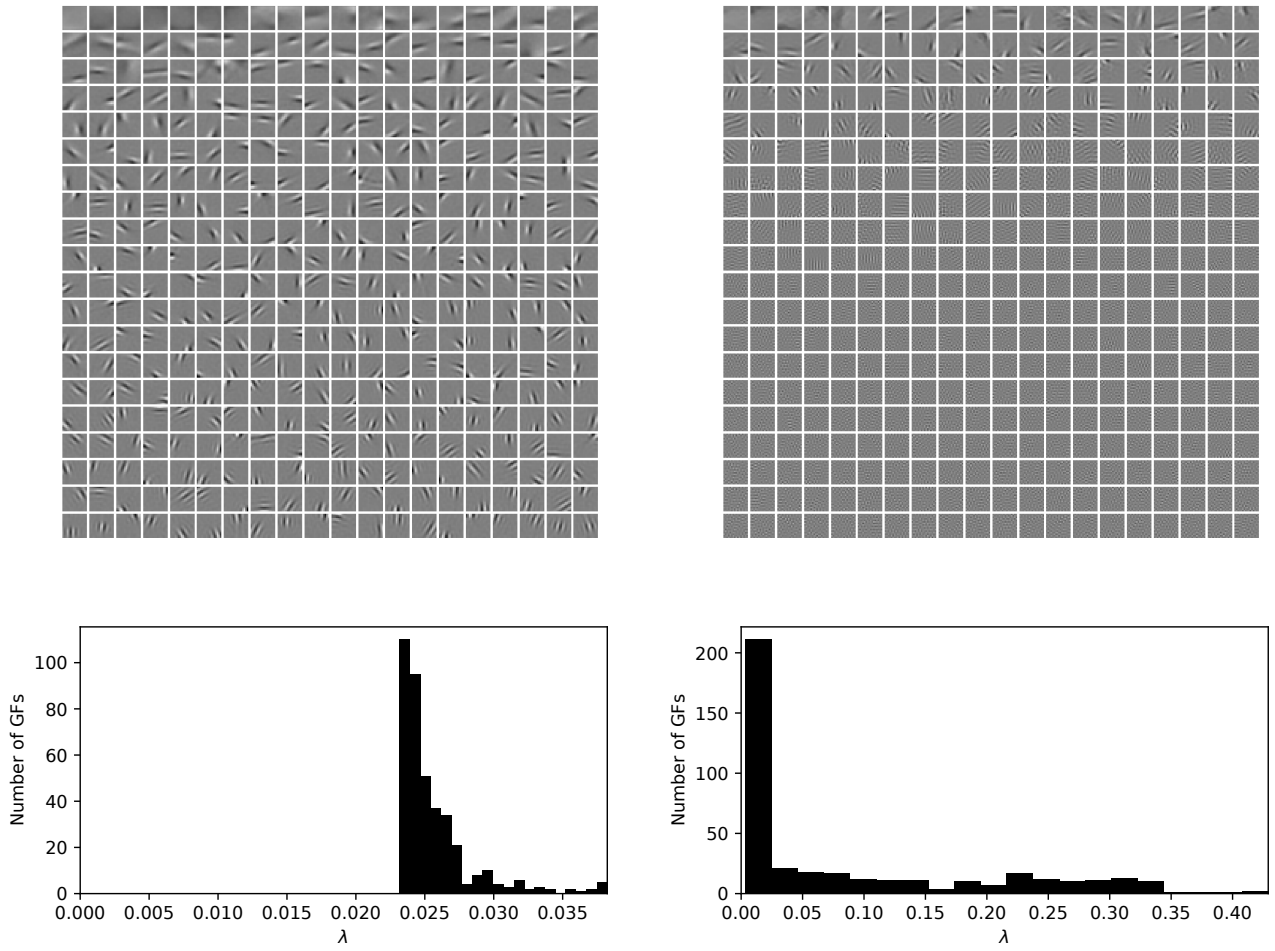


(g) Epoch 7 (h) Epoch 8 (i) Epoch 9

Figure 11: **Learned generative fields during optimization with likelihood annealing of entropy-ELBOs**. The annealing stops after epoch 6. It is equivalent to β-annealing, which is a popular scheme to tune reconstruction–embedding quality trade-off for VAE training.

D.5 Learning overcomplete basis

Prior entropy annealing significantly improves the quality of the learned dictionary and the sparseness of the latent codes (see Fig. 12). When after annealing the prior entropy weight is set to 1, approximately half of the generative fields converge to high-frequency textures and contribute only marginally to the reconstruction. Emphasizing the prior allows us to learn a rich dictionary of localized Gabor-like generative fields that span a wide range of frequencies, positions, and orientations, positively contributing to the sparsity of the latent codes.



(a) Prior entropy weight: 2

(b) Prior entropy weight: 1

Figure 12: **Learned overcomplete bases for the image patches dataset.** 400 generative fields are learned from 16×16 image patches. Different generative fields were obtained with prior entropy weight set to 2 (a), and with the original ELBO (b). The bottom histograms illustrate how this reweighting affects the prior scales λ_i. In (b) more than 200 prior coefficients are close to 0, which hints to the posterior collapse. The generative fields are sorted according to their λ_i scale.

E PROPERTIES OF THE FUNCTION $\mathcal{M}$ AND SOFTENED MAGNITUDE

We study the properties of $\mathcal{M}(a)$ in Eq. (12). We find for very small and for very large arguments a of the function that:

$$\lim_{a\to\infty} \mathcal{M}(a) = |a| \quad \text{and} \quad \lim_{a\to-\infty} \mathcal{M}(a) = |a| \ . \tag{175}$$

So $\mathcal{M}(a)$ approximates the "l_1" magnitude function $|a|$ if a has large or small values. Furthermore, the function upper-bounds the magnitude function everywhere, and the largest difference of $\mathcal{M}(a)$ compared to $|a|$ is at zero:

$$\text{for all } a \in \mathbb{R}\text{: } \mathcal{M}(a) > |a| \quad \text{and} \quad \mathcal{M}(0) = \sqrt{2/\pi} \ . \tag{176}$$

Turning back to the relatively intricate expression for λ_h^{opt} in Theorem 2, we can now define a 'softened' magnitude function (cf. Sec. 3.4) that formally simplifies the expression significantly:

$$\lambda_h^{\text{opt}} = \frac{1}{N}\sum_{n=1}^{N} \left|\nu_h^{(n)}\right|^\star \quad \text{where} \quad \left|\nu_h^{(n)}\right|^\star = \tau_h^{(n)} \mathcal{M}\Big(\frac{\nu_h^{(n)}}{\tau_h^{(n)}}\Big) \ . \tag{177}$$

Using the properties of $\mathcal{M}$, it can directly be observed that $|\nu_h^{(n)}|^\star \approx |\nu_h^{(n)}|$ whenever $\nu_h^{(n)} \gg \tau_h^{(n)}$, so for small $\tau_h^{(n)}$ the function $|\nu_h^{(n)}|^\star$ essentially represents the l_1 magnitude. We have to keep in mind, however, that $|\nu_h^{(n)}|^\star$ depends on $\tau_h^{(n)}$ (which we have omitted in the notation for convenience). As a principled difference between $|\nu_h^{(n)}|^\star$ and $|\nu_h^{(n)}|$ it remains that ultimately the derived function $|\nu_h^{(n)}|^\star$ does (in contrast to $|\nu_h^{(n)}|$) *not* vanish for vanishing ν_h. So while many entries $\nu_h^{(n)}$ will be pushed towards zero, the minimum of $|\nu_h^{(n)}|^\star$ will not be at zero. The derived objective is, therefore, genuinely different from l_1-sparse coding. It may be used, however, to relate to recent threshold-based variants of the sparse coding objectives (Rozell et al., 2008; Fallah and Rozell, 2022).

The derivative of the function $\mathcal{M}(a)$ has a particularly simple form, which we already discussed in Theorem 3:

$$\frac{\partial \mathcal{M}(a)}{\partial a} = \text{erf}\left(\frac{a}{\sqrt{2}}\right) \ . \tag{178}$$