Training sparse natural image models with a fast Gibbs sampler of an extended state space

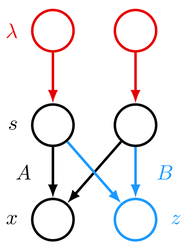
Lucas Theis[1], Jascha Sohl-Dickstein[2], Matthias Bethge[1,3,4]

[1] Werner Reichardt Centre for Integrative Neuroscience, Tübingen [2] Redwood Center for Theoretical Neuroscience, Berkeley
[3] Max Planck Institute for Biological Cybernetics, Tübingen [4] Bernstein Center for Computational Neuroscience, Tübingen

Introduction

We present an efficient inference and optimization algorithm for maximum likelihood learning in the **overcomplete linear model** (OLM). Using a persistent variant of expectation maximization, we find that using overcomplete representations significantly improves the performance of the linear model when applied to natural images while most previous studies were unable to detect an advantage [1, 2, 3, 4].

Overcomplete linear model



$$x = As$$

$$p(s) = \prod_i f_i(s_i)$$

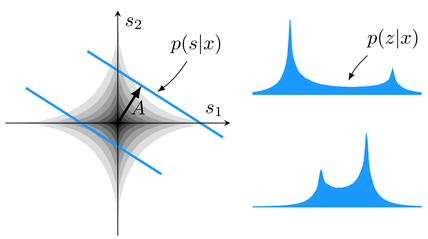
Here, $A \in \mathbb{R}^{M \times N}$ and $M \leq N$. Note that we don't assume any additive noise on the visible units.

We use **Gaussian scale mixtures** (GSMs) to represent the source distributions. This family contains the Laplace, Student-t, Cauchy and exponential power distribution as a special case.

$$f_i(s_i) = \int_0^\infty g_i(\lambda_i) \mathcal{N}(s_i; 0, \lambda_i^{-1})\, d\lambda_i$$

Inference

The posterior distribution over latent source variables is constrained to a linear subspace and can have multiple modes with heavy tails.



Blocked Gibbs sampling

We introduce two sets of auxiliary variables. One set, z, represents the missing visible variables [5].

$$\begin{bmatrix} x \\ z \end{bmatrix} = \begin{bmatrix} A \\ B \end{bmatrix} s \qquad s = A^+ x + B^+ z$$

The other set, λ, represents the precisions of the GSM source distributions. Our blocked Gibbs sampler alternately samples z and λ.

$$p(z \mid x, \lambda) = \mathcal{N}(z; \mu_{z|x}, \Sigma_{z|x})$$

$$p(\lambda \mid x, z) \propto \prod_i \mathcal{N}(s_i; 0, \lambda_i^{-1}) g_i(\lambda_i)$$

Persistent EM

For maximum likelihood learning, we use a Monte Carlo variant of **expectation maximization** (EM) where in each E-step the data is completed by sampling from the posterior. To further speed up learning, we initialize the Markov chain with the samples of the previous iteration.

```
1. initialize z
2. repeat
3.     z ~ T(·; z, x)
4.     maximize log p(x, z | θ)
```

This algorithm works and will converge because each iteration can only improve a lower bound on the log-likelihood.

$$F[q, \theta] = \log p(\mathbf{x} \mid \theta) - D_{\mathrm{KL}}\left[q(\mathbf{z} \mid \mathbf{x}) \,||\, p(\mathbf{z} \mid x, \theta)\right]$$
$$F[Tq, \theta] \geq F[q, \theta]$$

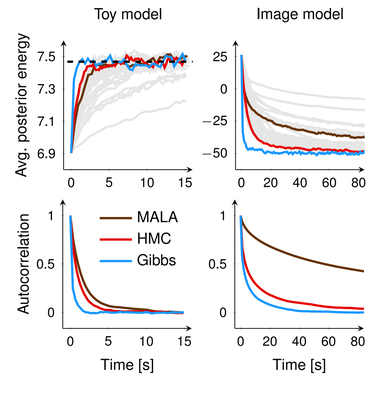
Likelihood estimation

We use a form of importance sampling (**annealed importance sampling**) to estimate the likelihood of the model. This yields an unbiased estimator of the likelihood and a conservative estimator of the log-likelihood.

$$p(x) = \int q(z \mid x) \frac{p(x, z)}{q(z \mid x)}\, dz \approx \frac{1}{N} \sum_n \frac{p(x, z_n)}{q(z_n \mid x)}$$

Performance of the blocked Gibbs sampler

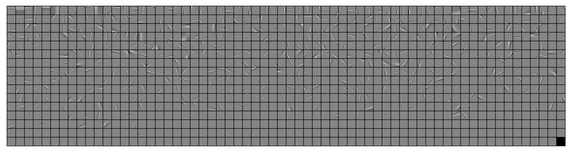
We compare the performance of the Gibbs sampler to the performance of **Hamiltonian Monte Carlo** (HMC) sampling.



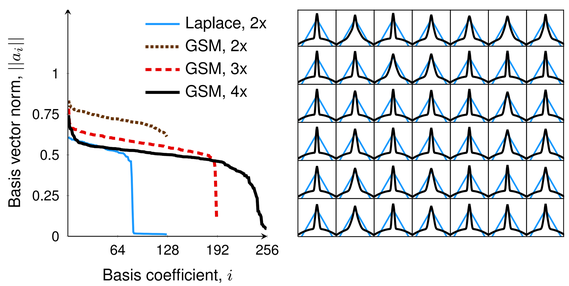
The upper plots show trace plots of posterior density, the lower plots show autocorrelation functions averaged over several data points.

$$R(\tau) = \frac{E\left[(z_t - \mu)^\top (z_{t+\tau} - \mu) \mid x\right]}{E\left[(z_t - \mu)^\top (z_t - \mu) \mid x\right]}$$

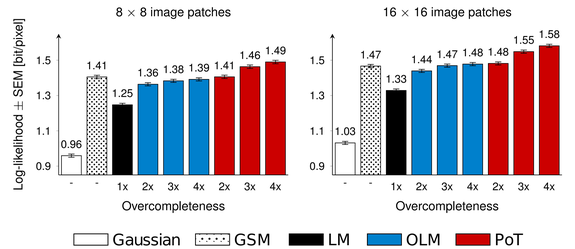
Overcompleteness and sparsity



When applied to natural image patches, the model learns highly sparse source distributions and three to four times overcomplete representations.



Model comparison



We compared the performance of the OLM with a **product of experts** (PoE) model, which represents another generalization of the linear model to overcomplete representations.

$$s = Wx, \quad p(x) \propto \prod_i f_i(s_i)$$

Here, we use the product of Student-t distributions.

Conclusions and remarks

- Efficient maximum likelihood learning in overcomplete linear models is possible and enables us to **jointly optimize filters and sources**.
- A linear model for natural images can benefit from overcomplete representations if the source distributions are **highly leptokurtotic**.
- The presented algorithm can easily be extended to more powerful models of natural images such as subspace or bilinear models.

Resources

Code for training and evaluating overcomplete linear models:

http://bethgelab.org/code/theis2012d/

References

[1] M. Lewicki and B. A. Olshausen, JOSA A, 1999
[2] P. Berkes, R. Turner, and M. Sahani, NIPS 21, 2007
[3] M. Seeger, JMLR, 2008
[4] D. Zoran and Y. Weiss, NIPS 25, 2012
[5] R.-B. Chen and Y. N. Wu, Computational Statistics & Data Analysis, 2007

Acknowledgements

This study was financially supported through the Bernstein award (BMBF; FKZ: 01GQ0601), the German Research Foundation (DFG; priority program 1527, Sachbeihilfe BE 3848/2-1), and a DFG-NSF collaboration grant (TO 409/8-1).