Select and Sample - A Model of Efficient Neural Inference and Learning

Jacquelyn A. Shelton[1], Jörg Bornschein[1], Abdul-Saboor Sheikh[1], Pietro Berkes[2] and Jörg Lücke[1]
[1]Frankfurt Institute for Advanced Studies, Germany; [2]Volen Center for Complex Systems. Brandeis University, USA

FIAS Frankfurt Institute for Advanced Studies

Bernstein Focus: Neurotechnology Frankfurt

DFG

Highlights

Introduction:

- Experimental evidence – perception encodes and maintains posterior probability distributions over possible causes of sensory stimuli
- Full posterior representation costly/complex – very high-dimensional, multi-modal, possibly highly correlated
- But, the brain can nevertheless perform rapid learning and inference
- Evidence for fast feed-forward processing and recurrent processing

Goals:

[1] Can we find rich representation of the posterior for very high-dimensional spaces?

[2] This goal believed to be shared by the brain, can find a biologically plausible solution reaching it?

- Want: method to combine feed-forward processing and recurrent stages of processing
- Idea: approximate inference and learning with good posterior representation → use pre-selection of most relevant latent variables and sample from this selection

Results:

- Experiments on image patches with $H = 1600$ hidden dimensions
- Method scales well – applicable to high dimensional data while maintaining sampling-based representation of posterior
- All model parameters learnable
- Combined approach formulates pre-selection and sampling as approximations to exact inference in a probabilistic framework for perception

The Setting

- Generative model for sensory data $\vec{y} = (y_1, \ldots, y_D)$ with hidden causes/objects $\vec{s} = (s_1, \ldots, s_H)$ and parameters Θ:

$$p(\vec{y} \,|\, \Theta) = \sum_{\vec{s}} p(\vec{y} \,|\, \vec{s}, \Theta)\, p(\vec{s} \,|\, \Theta)$$

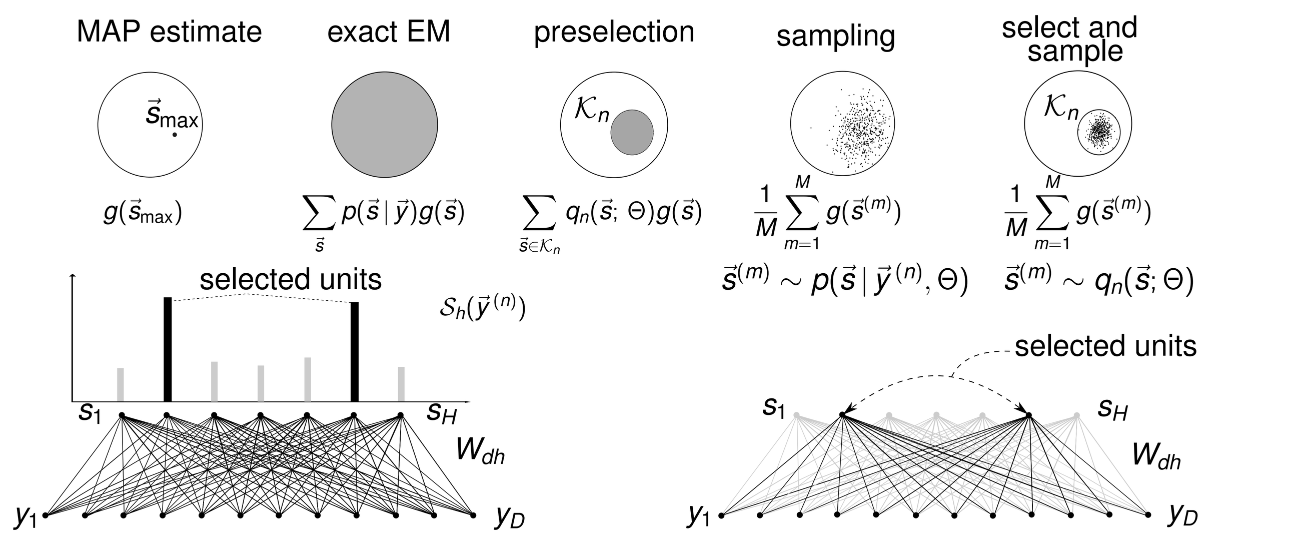
- Given data set $Y = \{\vec{y}_1, \ldots, \vec{y}_N\}$ find maximum likelihood parameters $\Theta^* = \mathrm{argmax}_{\Theta}\, p(Y \,|\, \Theta)$ using EM.
- M-step usually involves a small number of expected values w.r.t. the posterior distribution:

$$\langle g(\vec{s}) \rangle_{p(\vec{s}\,|\,\vec{y}^{(n)},\Theta)} = \sum_{\vec{s}} p(\vec{s} \,|\, \vec{y}^{(n)}, \Theta)\, g(\vec{s})$$

where $g(\vec{s})$ is usually an elementary function of the hidden variables (e.g. $g(\vec{s}) = \vec{s}$ or $g(\vec{s}) = \vec{s}\vec{s}^T$ for standard sparse coding)

- Computation of expectations is usually the computationally demanding part

Select and Sample Approach



Expectation Truncation (ET)

- Restrict approximate posterior to pre-selected states

$$p(\vec{s} \,|\, \vec{y}^{(n)}, \Theta) \approx q_n(\vec{s}; \Theta) = \frac{p(\vec{s} \,|\, \vec{y}^{(n)}, \Theta)}{\sum_{\vec{s}' \in \mathcal{K}_n} p(\vec{s}' \,|\, \vec{y}^{(n)}, \Theta)}\, \delta(\vec{s} \in \mathcal{K}_n)$$

- Set $\mathcal{K}_n$ chosen using a *selection function* $\mathcal{S}_h(\vec{y}, \Theta)$; efficiently selects candidates s_h with most posterior mass:

$$\mathcal{K}_n = \{\vec{s} \,|\, \text{for all } h \notin \mathcal{I}_n :\; s_h = 0\}$$

where $\mathcal{I}_n$ contains the H' indices h with the highest values of $\mathcal{S}_h(\vec{y}^{(n)}, \Theta)$, most likely contributors

- Can be seen as variational approximation to posterior
- Efficiently computable expectations in $\mathcal{O}(|\mathcal{K}_n|)$:

$$\langle g(\vec{s}) \rangle_{p(\vec{s}\,|\,\vec{y}^{(n)},\Theta)} \approx \langle g(\vec{s}) \rangle_{q_n(\vec{s};\Theta)} = \frac{\sum_{\vec{s} \in \mathcal{K}_n} p(\vec{s}, \vec{y}^{(n)} \,|\, \Theta)\, g(\vec{s})}{\sum_{\vec{s}' \in \mathcal{K}_n} p(\vec{s}', \vec{y}^{(n)} \,|\, \Theta)}$$

Sampling

- Alternative: approximate expectations using samples from the posterior distribution:

$$\langle g(\vec{s}) \rangle_{p(\vec{s}\,|\,\vec{y}^{(n)},\Theta)} \approx \tfrac{1}{M} \textstyle\sum_{m=1}^{M} g(\vec{s}^{(m)}) \text{ with } \vec{s}^{(m)} \sim p(\vec{s} \,|\, \vec{y}, \Theta)$$

- Obtaining samples from true posterior often difficult

Combining ET & Sampling

- Approximate expectations using samples from the truncated distribution:

$$\langle g(\vec{s}) \rangle_{q_n(\vec{s};\Theta)} \approx \tfrac{1}{M} \textstyle\sum_{m=1}^{M} g(\vec{s}^{(m)}) \quad \text{with} \quad \vec{s}^{(m)} \sim q_n(\vec{s}; \Theta)$$

- Subspace $\mathcal{K}_n$ is small, allowing MCMC algorithms to operate more efficiently, i.e. shorter burn-in times, reduced number of required samples

Example Application: Binary Sparse Coding

- Model: sparse coding with binary latent variables

$$p(\vec{s}|\pi) = \prod_{h=1}^{H} \pi^{s_h} \big(1 - \pi\big)^{1 - s_h}$$

$$p(\vec{y}|\vec{s}, W, \sigma) = \mathcal{N}(\vec{y}; W\vec{s}, \sigma^2 I)$$

$\vec{y} \in \mathbb{R}^D$	observed variables	π	prior parameter
$\vec{s} \in \{0,1\}^H$	hidden variables	σ	noise level
$W \in \mathbb{R}^{D \times H}$	basis functions		

- Selection function: cosine similarity

$$\mathcal{S}_h(\vec{y}^{(n)}) = \frac{\vec{W}_h^{\mathrm{T}} \vec{y}^{(n)}}{||\vec{W}_h||}$$

- Inference: ET with Gibbs sampling; ET posterior equivalent to full post. with only selected dims

$$p(s_h = 1 \,|\, \vec{s}_{\setminus h}, \vec{y}) = \frac{p(s_h = 1, \vec{s}_{\setminus h}, \vec{y})^{\beta}}{p(s_h = 0, \vec{s}_{\setminus h}, \vec{y})^{\beta} + p(s_h = 1, \vec{s}_{\setminus h}, \vec{y})^{\beta}}$$

- Complexity of E-step (all 4 BSC cases):

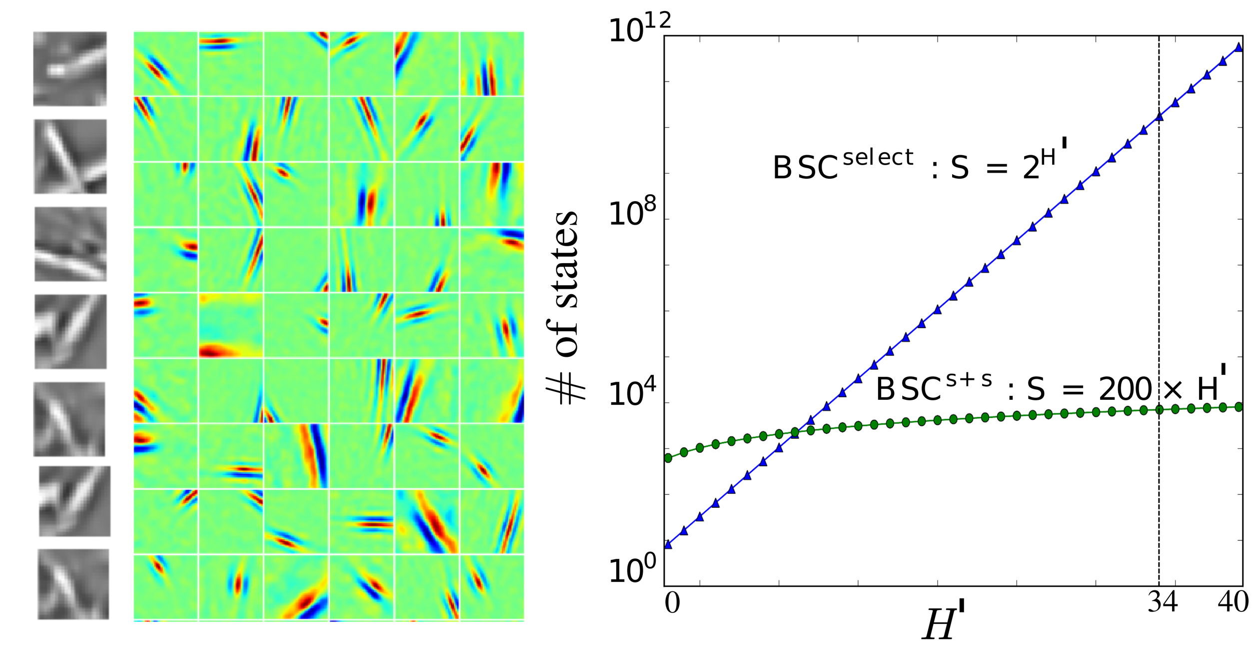
$$\mathcal{O}\big(NS(\underbrace{D}_{p(\vec{s},\vec{y})} + \underbrace{1}_{\langle \vec{s} \rangle} + \underbrace{H}_{\langle \vec{s}\vec{s}^T \rangle})\big)$$

where S is # of evaluated hidden states

Experiments

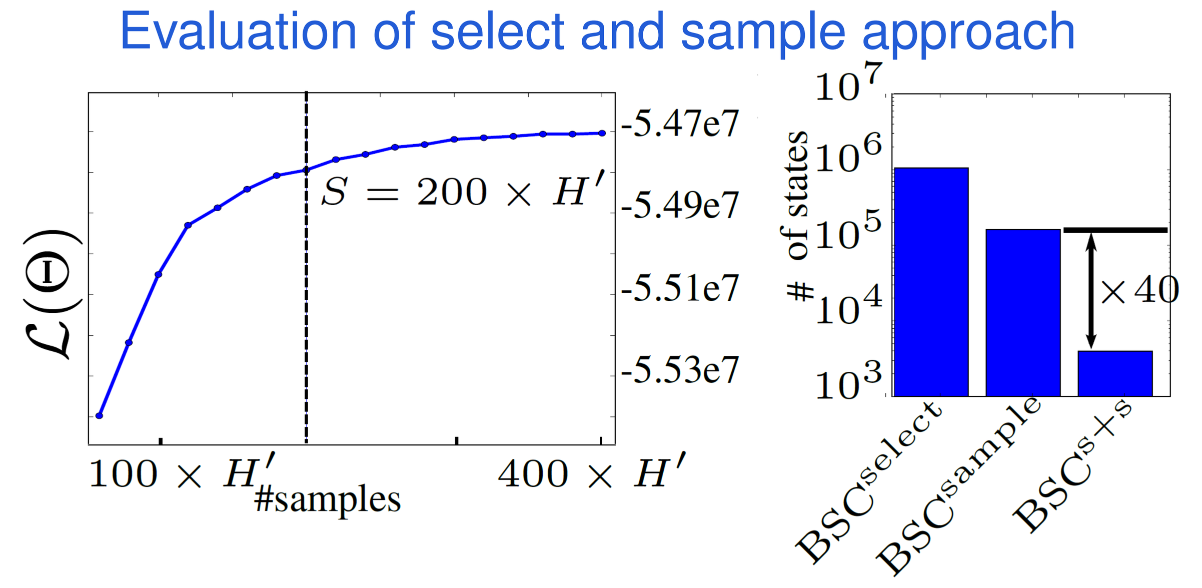
Natural image patches

1600 latent dimensions with sampling-based posterior



- Large-scale application of select and sample ($\mathrm{BSC}^{\mathrm{s+s}}$) to $N = 500,000$ image patches with $H = 1600$, $H' = 34$, $D = 40 \times 40 = 1600$ pixels
- Shown: data, handful of the inferred basis functions W_h and comparison the of computational complexity
- $\mathrm{BSC}^{\mathrm{select}}$ scales exponentially with H' whereas $\mathrm{BSC}^{\mathrm{s+s}}$ scales linearly. Note the large difference at $H' = 34$, used in obtaining the W

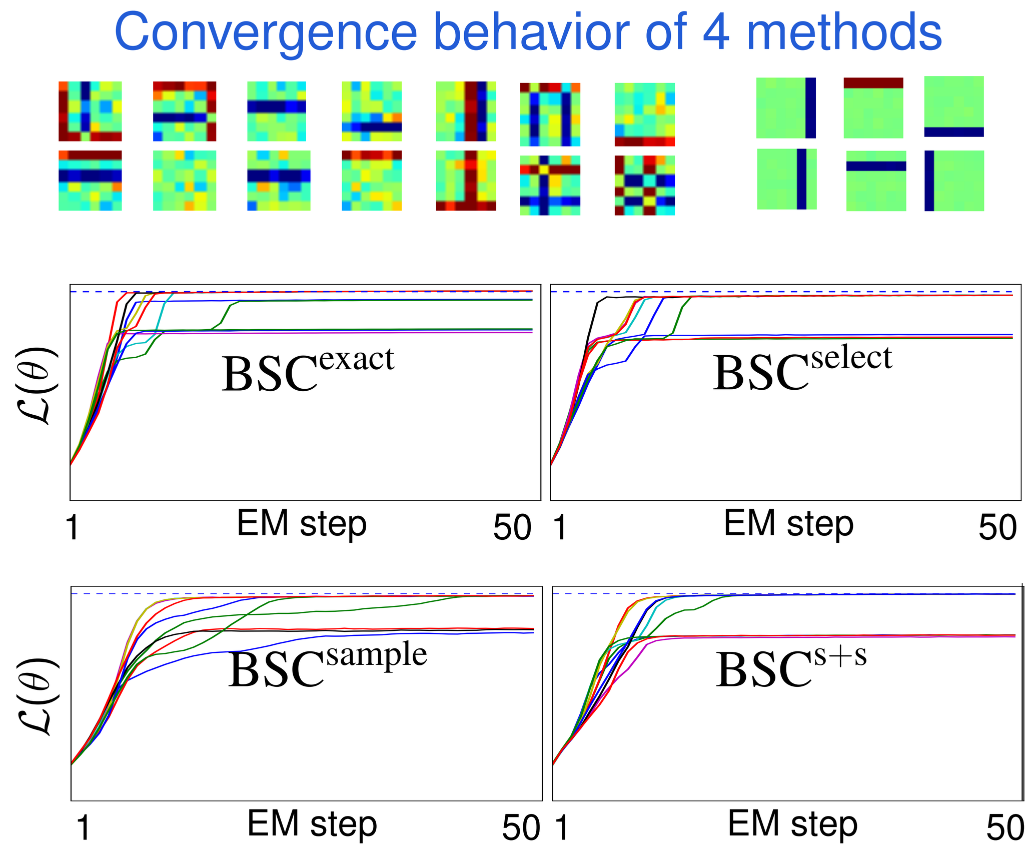
Evaluation of select and sample approach



- Experiments on $N = 40,000$ image patches with $D = 26 \times 26$, $H = 800$
- Goal: study effect of # samples on performance of $\mathrm{BSC}^{\mathrm{s+s}}$ across entire range of $12 \leq H' \leq 36$ and comparison of # states to be evaluated for $\mathrm{BSC}^{\mathrm{sample}}$, $\mathrm{BSC}^{\mathrm{select}}$, $\mathrm{BSC}^{\mathrm{s+s}}$
- Shown: end approximate log-likelihood after 100 EM-steps vs. number of samples per data point and # states that had to be evaluated for $H' = 20$ for the different approaches
- 200 samples per hidden dimension sufficient: drawing more helps likelihood less than 1%
- Select and sample approach is 40 times faster than sampling

Artificial data

Convergence behavior of 4 methods



- Experiments on artificial $N = 2000$ bars data with $H = 12$, $D = 6 \times 6$. Dotted line is $\mathcal{L}(\theta^{ground-truth})$
- Shown: data, basis functions W_h, and log-likelihood for multiple runs over 50 EM steps for all 4 methods

References & Acknowledgements

[1] J. Fiser, P. Berkes, G. Orban, and M. Lengye. (2010). Statistically optimal perception and learning: from behavior to neural representations. Trends in Cog. Sci., 14:119–130.
[2] W. J. Ma, J. M. Beck, P. E. Latham, and A. Pouget. (2006). Bayesian inference with probabilistic population codes. Nature Neuroscience, 9:1432–1438.
[3] P. Berkes, G. Orban, M. Lengyel, and J. Fiser. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. Science, 331(6013):83–87.
[4] P. O. Hoyer and A. Hyvarinen. Interpreting neural response variability as Monte Carlo sampling from the posterior. In Adv. Neur. Inf. Proc. Syst. 16, MIT Press, 2003.
[5] J. Lücke and J. Eggert. (2010). Expectation Truncation And the Benefits of Preselection in Training Generative Models. Journal of Machine Learning Research.
[6] B. A. Olshausen, D. J. Field. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381:607-609.

Work supported by the German Research Foundation (DFG) in the project LU 1196/4-1 (JL), the German Federal Ministry of Education and Research (BMBF), project 01GQ0840 (JAS, JB, ASS), the Swartz Foundation and the Swiss National Science Foundation (PB), the Physics Dept., and the Center for Scientific Computing (CSC) in Frankfurt.