Scalable Model Selection for Belief Networks
Supplemental Materials

Zhao Song†, Yusuke Muraoka∗, Ryohei Fujimaki∗, Lawrence Carin†
†Department of ECE, Duke University
Durham, NC 27708, USA
{zhao.song, lcarin}@duke.edu
∗NEC Data Science Research Laboratories
Cupertino, CA 95014, USA
{ymuraoka, rfujimaki}@nec-labs.com

1 Proof of Theorem 1

Proof. We first compute the negative Hessian as

\[ \Psi^m = -\frac{1}{N} \frac{\partial}{\partial W_m^T} \frac{\partial}{\partial W_m} \sum_n \ln p(v_n, h_n | \theta) = \frac{1}{N} \sum_n \sigma(W_m h_n) \sigma(-W_m h_n) h_n^T h_n. \]

From Assumption 1, \( \Psi^m \) has full rank, since \( \sigma(x) \in (0, 1) \), \( \forall x \in \mathbb{R} \). Furthermore, the determinant of \( \Psi^m \) is bounded, since \( \Psi^m_{ij} \in (0, 1) \), \( \forall i, j \). Next, we define the following diagonal matrix

\[ \Lambda = \text{diag} \left[ \left( \frac{\sum_n h_{n,1}}{N} \right), \ldots, \left( \frac{\sum_n h_{n,J}}{N} \right) \right]. \]

From Assumption 2, \( \lim_{N \to \infty} \text{Pr}[\sum_n h_{n,j} = 0] = 0 \). Therefore, \( \Lambda \) is full-rank and its determinant is bounded, when \( N \to \infty \). Subsequently, we can decompose it as

\[ \Psi^m = \Lambda F \tag{A1} \]

where \( F \) also has full rank and bounded determinant. Finally, applying the log determinant operator to the right side of (A1) leads to our conclusion.

2 Proof of Corollary 1

Proof. The conclusion holds as a direct extension of the consistency results in Hayashi and Fujimaki [1].

3 Derivation of Minibatch gFIC in (6)

We first note that an unbiased estimator for (4) is

\[ \text{gFIC}_{SBN} = \max_q \mathbb{E}_q \left[ -\frac{M}{2} \sum_j \ln \left( \frac{N_{T+1}}{N_{\text{mini}}} \sum_{i=1}^{N_{\text{mini}}} h_{i+N_T,j} \right) 
+ \frac{N_{T+1}}{N_{\text{mini}}} \sum_{i=1}^{N_{\text{mini}}} \ln \frac{p(v_{i+N_T,j} | h_{i+N_T,j} \theta)}{q(h_{i+N_T,j} | \phi)} + \frac{MJ - D_\theta}{2} \ln N_{T+1} \right] \tag{A2} \]

Simplifying and ignoring constant terms in (A2) leads to \( \text{gFIC}_{SBN} \) in (6) of the main text.

4 Additional Results

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
Figure A1: Test log-likelihood and the number of nodes in FABIA, as a function of CPU time on the MNIST dataset, for an SBN with initial size as (a) 200 (b) 500.

Figure A2: Test log-likelihood and the number of nodes in FABIA, as a function of the number of iterations on the MNIST dataset, for an SBN with initial size as (a) 200 (b) 500.

Figure A3: Test log-likelihood and the number of nodes in FABIA, as a function of CPU time on the MNIST dataset, for an SBN with initial size as (a) 200-200 (b) 500-500.
Figure A4: Test log-likelihood and the number of nodes in FABIA, as a function of the number of iterations on the MNIST dataset, for an SBN with initial size as (a) 200-200 (b) 500-500.

Figure A5: Test log-likelihood and the number of nodes in FABIA, as a function of the number of iterations on the MNIST dataset, for an SBN with initial size as (a) 200-200-200 (b) 500-500-500.

Figure A6: Test perplexities as a function of number of nodes in the first layer, in the three-layer case.

Table 1: Test perplexities and model size on the benchmarks, for NVIL and FABIA with three layers. FABIA starts from a model initialized with 400 hidden units in each layer.

<table>
<thead>
<tr>
<th></th>
<th>RCV1</th>
<th>Wikipedia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>Perplexity</td>
<td>Size</td>
</tr>
<tr>
<td>NVIL</td>
<td>400-400-400</td>
<td>859</td>
</tr>
<tr>
<td>FABIA</td>
<td>398-46-10</td>
<td>860</td>
</tr>
</tbody>
</table>
References