Stochastic Spectral Descent for Restricted Boltzmann Machines: Supplemental Material

David Carlson[1] **Volkan Cevher**[2] **Lawrence Carin**[1]

[1] Department of Electrical and Computer Engineering, Duke University

[2]Laboratory for Information and Inference Systems (LIONS), EPFL

A Theorem proofs

Proof. Proof of Theorem 1.
The Hessian of the lse function is given by

$$\nabla^2 lse_{\boldsymbol{\omega}}(\boldsymbol{u}) = \frac{\text{diag}(\boldsymbol{\omega} \odot \exp(\boldsymbol{u}))}{\boldsymbol{\omega}^T \exp(\boldsymbol{u})} - \frac{(\boldsymbol{\omega} \odot \exp(\boldsymbol{u}))(\boldsymbol{\omega} \odot \exp(\boldsymbol{u}))^T}{(\boldsymbol{\omega}^T \exp(\boldsymbol{u}))^2} \quad \text{(A.1)}$$

There are two terms in the Hessian matrix. The first term is

$$\frac{\text{diag}(\boldsymbol{\omega} \odot \exp(\boldsymbol{u}))}{\boldsymbol{\omega}^T \exp(\boldsymbol{u})}$$

This is a diagonal matrix where the diagonal entries are nonnegative and sum to one. The second term is

$$-\frac{(\boldsymbol{\omega} \odot \exp(\boldsymbol{u}))(\boldsymbol{\omega} \odot \exp(\boldsymbol{u}))^T}{(\boldsymbol{\omega}^T \exp(\boldsymbol{u}))^2}$$

This term is a rank-one matrix with a negative eigenvalue.

Writing Taylor's theorem:

$$lse_{\omega}(\boldsymbol{v}) = lse_{\boldsymbol{\omega}}(\boldsymbol{u}) + \langle \nabla lse_{\omega}(\boldsymbol{u}), \boldsymbol{v} - \boldsymbol{u} \rangle + \int_0^1 (1-t)(\boldsymbol{v}-\boldsymbol{u})^T \nabla^2 lse_{\omega}(\boldsymbol{u} + t(\boldsymbol{v}-\boldsymbol{u}))(\boldsymbol{v}-\boldsymbol{u}) dt$$

The terms in the integral can be bound

$$\begin{aligned}
&(\boldsymbol{v}-\boldsymbol{u})^T \nabla^2 lse_{\boldsymbol{\omega}}(\boldsymbol{u} + t(\boldsymbol{v}-\boldsymbol{u}))(\boldsymbol{v}-\boldsymbol{u}) \\
\leq\ & (\boldsymbol{v}-\boldsymbol{u}) \tfrac{\text{diag}(\boldsymbol{\omega} \odot \exp(\boldsymbol{u} + t(\boldsymbol{v}-\boldsymbol{u})))}{\boldsymbol{\omega}^T \exp(\boldsymbol{u} + t(\boldsymbol{v}-\boldsymbol{u}))} (\boldsymbol{v}-\boldsymbol{u}) && \text{(A.2)} \\
=\ & \textstyle\sum_{j=1}^J \tfrac{\omega_j \exp(u_j + t(v_j - u_j))}{\boldsymbol{\omega}^T \exp(\boldsymbol{u} + t(\boldsymbol{v}-\boldsymbol{u}))} (v_j - u_j)^2 \\
\leq\ & \max_{\boldsymbol{c} \geq 0, ||\boldsymbol{c}||_1 = 1} \textstyle\sum_{j=1}^J c_j (v_j - u_j)^2 && \text{(A.3)} \\
=\ & ||\boldsymbol{v}-\boldsymbol{u}||_\infty^2 && \text{(A.4)}
\end{aligned}$$

Eq. A.2 follows because the second term in the Hessian will give a nonpositive value and Eq. A.3 follows because the diagonal entries are nonnegative and sum to 1. The integral has an upper bound of $\frac{1}{2}||\boldsymbol{v}-\boldsymbol{u}||_\infty^2$. □

Proof. Proof of Theorem 2.
The log partition function can be written as a sum over only the hidden units to give a similar form to Theorem 1. Define the set $\{h_i\}_{i=1}^{2^J}$ as the set of unique binary vectors $\{0,1\}^J$, and let $\mathbf{H} \in \{0,1\}^{J \times 2^J}$ be the matrix form of this set.

$$f(\boldsymbol{\theta}) = \log \sum_{i=1}^{2^J} \omega_i \exp(\boldsymbol{h}_i^T \boldsymbol{b}) \quad \text{(A.5)}$$

$$\omega_i = \sum_{m=1}^{M} \log(1 + \exp(\mathbf{W}_{m,\cdot} \boldsymbol{h}_i + c_m)) \quad \text{(A.6)}$$

Equation A.5 can be equivalently written as

$$f(\boldsymbol{\theta}) = \log \boldsymbol{\omega}^T \exp(\mathbf{H}^T \boldsymbol{b}) \quad \text{(A.7)}$$

with $\boldsymbol{\omega}$ not dependent on $\boldsymbol{b}$. Plugging into Equation 17,

$$\begin{aligned}
f(\{\boldsymbol{b}, \boldsymbol{c}^k, \mathbf{W}^k\}) \leq\ & f(\boldsymbol{\theta}^k) \\
&+ \langle \nabla_{\mathbf{H}^T\boldsymbol{b}} lse_{\boldsymbol{\omega}}(\mathbf{H}^T \boldsymbol{b}^k), \mathbf{H}^T(\boldsymbol{b} - \boldsymbol{b}^k) \rangle \\
&+ \frac{1}{2} ||\mathbf{H}^T(\boldsymbol{b} - \boldsymbol{b}^k)||_\infty^2 \quad \text{(A.8)}
\end{aligned}$$

To rewrite the inner product term, note that

$$\begin{aligned}
\nabla_{\mathbf{H}^T\boldsymbol{b}} lse_{\boldsymbol{\omega}}(\mathbf{H}^T \boldsymbol{b}^k) &= \mathbf{H}^T \nabla_{\boldsymbol{b}} f(\boldsymbol{\theta}^k) \quad \text{(A.9)} \\
(\nabla_{\mathbf{H}^T\boldsymbol{b}} lse_{\boldsymbol{\omega}}(\mathbf{H}^T \boldsymbol{b}^k))^T \mathbf{H}(\boldsymbol{b} - \boldsymbol{b}^k) &= (\nabla_{\boldsymbol{b}} f(\boldsymbol{\theta}^k))^T (\boldsymbol{b} - \boldsymbol{b}^k)
\end{aligned}$$

The bound is simplified as

$$||\mathbf{H}^T(\boldsymbol{b} - \boldsymbol{b}^k)||_\infty = \max_i |\boldsymbol{h}_i^T(\boldsymbol{b} - \boldsymbol{b}^k)| \leq J||\boldsymbol{b} - \boldsymbol{b}^k||_\infty$$

Alternatively, this could be bound as

$$\begin{aligned}
||\mathbf{H}^T(\boldsymbol{b} - \boldsymbol{b}^k)||_\infty &\leq \sqrt{J}||\boldsymbol{b} - \boldsymbol{b}_k||_2 \quad \text{(A.10)} \\
||\mathbf{H}^T(\boldsymbol{b} - \boldsymbol{b}^k)||_\infty &\leq ||\boldsymbol{b} - \boldsymbol{b}_k||_1 \quad \text{(A.11)}
\end{aligned}$$

The proof on $\boldsymbol{c}$ follows with the same techniques. □

Proof. Proof of Theorem 3.
As in the proof for Theorem 2, let $\mathbf{H} \in \{0,1\}^{J \times 2^J}$

and $\mathbf{V} \in \{0,1\}^{M \times 2^M}$, where each column is an unique binary vector. Define $\mathbf{U} = \mathbf{V}^T \mathbf{W} \mathbf{H}$ and $\mathbf{\Omega}_{ij} = \boldsymbol{v}_i^T \boldsymbol{c} + \boldsymbol{h}_j^T \boldsymbol{b}$. Let $\boldsymbol{u} = \text{vec}(\mathbf{U})$ and $\boldsymbol{\omega} = \text{vec}(\mathbf{\Omega})$. The log partition function is equivalently written

$$f(\boldsymbol{\theta}) = \log \sum_{i=1}^{2^M} \sum_{j=1}^{2^J} \mathbf{\Omega}_{ij} \exp \mathbf{U}_{ij} \quad \text{(A.12)}$$

$$f(\boldsymbol{\theta}) = \log \left(\boldsymbol{\omega}^T \exp \boldsymbol{u} \right) \quad \text{(A.13)}$$

Plugging this form into Equation 17:

$$\begin{aligned} lse_{\boldsymbol{\omega}}(\boldsymbol{u}) \geq\ & lse_{\boldsymbol{\omega}}(\boldsymbol{u}^k) + \langle \nabla_{\boldsymbol{u}} lse_{\boldsymbol{\omega}}(\boldsymbol{u}^k), \boldsymbol{u} - \boldsymbol{u}^k \rangle \\ & + \frac{1}{2} ||\text{vec}(\mathbf{U} - \mathbf{U}^k)||_\infty^2 \end{aligned} \quad \text{(A.14)}$$

Note that

$$\begin{aligned} \langle \nabla_{\boldsymbol{u}} lse_{\boldsymbol{\omega}}(\boldsymbol{u}), \boldsymbol{u} - \boldsymbol{u}^k \rangle &= \text{tr}((\nabla_{\mathbf{U}} lse_{\mathbf{\Omega}}(\mathbf{U}))^T (\mathbf{U} - \mathbf{U}^k)) \\ \mathbf{V} \nabla_{\mathbf{U}} lse_{\mathbf{\Omega}}(\mathbf{U}) \mathbf{H}^T &= \nabla_{\mathbf{W}} f(\boldsymbol{\theta}) \end{aligned} \quad \text{(A.15)}$$

Writing the inner product in terms of $\mathbf{W}$ gives

$$\text{tr}((\nabla_{\mathbf{U}} lse_{\mathbf{\Omega}}(\mathbf{U}))^T (\mathbf{U} - \mathbf{U}^k)) = \text{tr}((\nabla_{\mathbf{W}})^T (\mathbf{W} - \mathbf{W}^k)) \quad \text{(A.16)}$$

The bound is simplified:

$$\begin{aligned} ||\text{vec}(\mathbf{U} - \mathbf{U}^k)||_\infty =\ & \max_{i,j} |\boldsymbol{v}_i^T (\mathbf{W} - \mathbf{W}^k) \boldsymbol{h}_j| \\ \leq\ & \sqrt{MJ} ||\mathbf{W} - \mathbf{W}^k||_{S^\infty} \end{aligned} \quad \text{(A.17)}$$

Combining these two elements proves Theorem 3. □

B Derivation of optimal steps

Proof. Proof of $\boldsymbol{b}^*$ in Equation 25.
We want to find the minimizer of

$$\min_{\boldsymbol{b}} \langle \nabla_{\boldsymbol{b}} F(\boldsymbol{\theta}^k), \boldsymbol{b} - \boldsymbol{b}^k \rangle + \frac{J}{2} ||\boldsymbol{b} - \boldsymbol{b}^k||_\infty^2$$

First, add an additional variable a such that the minimizer of the expanded problem is the same as the original problem

$$= \min_{\boldsymbol{b}, a, |b_j| \leq a, a \geq 0} \langle \nabla_{\boldsymbol{b}} F(\boldsymbol{\theta}^k), \boldsymbol{b} - \boldsymbol{b}^k \rangle + \frac{J}{2} a^2 \quad \text{(B.1)}$$

This is straightforward to solve:

$$\begin{aligned} &= \min_{a, a \geq 0} \langle \nabla_{\boldsymbol{b}} F(\boldsymbol{\theta}^k), -a \times \text{sign}(\nabla_{\boldsymbol{b}} F(\boldsymbol{\theta}^k)) \rangle + \frac{J}{2} a^2 \\ a^* &= \frac{1}{J} ||\nabla_{\boldsymbol{b}} F(\boldsymbol{\theta}^k)||_1 \quad \text{(B.2)} \\ \boldsymbol{b}^* &= \boldsymbol{b} - \frac{1}{J} ||\nabla_{\boldsymbol{b}} F(\boldsymbol{\theta}^k)||_1 \times \text{sign}(\nabla_{\boldsymbol{b}} F(\boldsymbol{\theta}^k)) \quad \text{(B.3)} \end{aligned}$$

□

Proof. Proof of $\mathbf{W}^*$ in Equation 28.
Let $\mathbf{D} = \mathbf{W} - \mathbf{W}^k$, and decompose $\mathbf{D} = \mathbf{A} \mathbf{R} \mathbf{B}^T$, with $\mathbf{A}$ and $\mathbf{B}$ denoting the left and right singular vectors of $\nabla_{\mathbf{W}} F(\boldsymbol{\theta}^k)$. Then we want to minimize the quantity

$$\min_{\mathbf{D}} \text{tr}(\nabla_{\mathbf{W}} F(\boldsymbol{\theta}^k) \mathbf{D}) + \frac{MJ}{2} ||\mathbf{D}||_{S^\infty}^2$$

As in the proof on the biases, add an additional variable that will give the same minimizer and solve for the solution.

$$\begin{aligned} &= \min_{\mathbf{D}, a, ||\mathbf{D}||_{S_\infty} < a} \text{tr}(\nabla_{\mathbf{W}} F(\boldsymbol{\theta}^k) \mathbf{D}) + \frac{MJ}{2} a^2 \\ &= \min_{\mathbf{D}, a, ||\mathbf{D}||_{S_\infty} < a} \text{tr}(\nabla_{\mathbf{W}} F(\boldsymbol{\theta}^k) \mathbf{D}) + \frac{MJ}{2} a^2 \\ &= \min_{a, \mathbf{F}, ||\mathbf{F}||_{S_\infty} < a} \boldsymbol{\lambda}^T \text{diag}(\mathbf{R}) + \frac{MJ}{2} a^2 \end{aligned}$$

Letting $\mathbf{I}_M$ denote the M-dimensional identity matrix, this gives:

$$\begin{aligned} \mathbf{R}^* &= \frac{-a}{MJ} \mathbf{I}_M \quad \text{(B.4)} \\ a &= ||\boldsymbol{\lambda}||_1 \quad \text{(B.5)} \\ \mathbf{R}^* &= \left(\frac{-1}{MJ} ||\boldsymbol{\lambda}||_1 \times \mathbf{I}_M \right) \quad \text{(B.6)} \end{aligned}$$

□

C Discussion of using ℓ_2 bound instead of ℓ_∞ bound on lse function

[Böhning, 1992] introduces a bound on the lse function

$$\begin{aligned} lse_{\mathbf{1}}(\boldsymbol{v}) \leq\ & lse_{\mathbf{1}}(\boldsymbol{u}) + \langle \nabla_{\boldsymbol{u}} lse_{\mathbf{1}}(\boldsymbol{u}), \boldsymbol{v} - \boldsymbol{u} \rangle \\ & + \frac{1}{2} (\boldsymbol{v} - \boldsymbol{u})^T \mathbf{B} (\boldsymbol{v} - \boldsymbol{u}) \quad \text{(C.1)} \\ \mathbf{B} =\ & \frac{1}{2} \left[\mathbf{I}_J - \frac{1}{J} \mathbf{1}_J \mathbf{1}_J^T \right] \quad \text{(C.2)} \end{aligned}$$

Where $\mathbf{I}$ is the J-dimensional identity matrix and $\mathbf{1}_J$ is a J-dimensional ones vector. This is trivially extended to use a nonnegative vector $\boldsymbol{\omega}$ in place of $\mathbf{1}_J$. The quadratic term is equivalently written

$$\frac{1}{2} (\boldsymbol{v} - \boldsymbol{u})^T \mathbf{B} (\boldsymbol{v} - \boldsymbol{u}) = \frac{1}{4} ||\boldsymbol{v} - \boldsymbol{u}||_2^2 - \frac{1}{4} \text{mean}(\boldsymbol{v} - \boldsymbol{u})^2 \quad \text{(C.3)}$$

Because of the differences of logsumexp functions, the mean term drops out and so this bound gives

$$\begin{aligned} lse_{\boldsymbol{\omega}}(\boldsymbol{v}) \leq\ & lse_{\boldsymbol{\omega}}(\boldsymbol{u}) + \langle \nabla_{\boldsymbol{u}} lse_{\boldsymbol{\omega}}(\boldsymbol{u}), \boldsymbol{v} - \boldsymbol{u} \rangle \\ & + \frac{1}{2 \times 2} ||\boldsymbol{v} - \boldsymbol{u}||_2^2 \end{aligned} \quad \text{(C.4)}$$

Using Equation C.4 instead of Equation 17 in the proofs in Supplemental Section A leads to looser bounds due to the high-dimensional nature of the observation space. However, it should be noted that it may be possible to bound this more tightly.

First, examining the bound on the matrix $\mathbf{W}$,

$$\begin{aligned}
& \frac{1}{4}||\mathrm{vec}(\mathbf{U}-\mathbf{U}^k)||_2^2 && \text{(C.5)}\\
= & \frac{1}{4}\sum_{i=1}^{2^M}\sum_{j=1}^{2^J}(\boldsymbol{v}_i^T(\mathbf{W}-\mathbf{W}^k)\boldsymbol{u}_j)^2 && \text{(C.6)}\\
\leq & \frac{1}{4}\sum_{i=1}^{2^M}\sum_{j=1}^{2^J}\boldsymbol{v}_i^T((\mathbf{W}-\mathbf{W}^k)\odot(\mathbf{W}-\mathbf{W}^k))\boldsymbol{u}_j && \text{(C.7)}\\
= & \frac{1}{4}\mathrm{tr}(((\mathbf{W}-\mathbf{W}^k)\odot(\mathbf{W}-\mathbf{W}^k))\sum_{i=1}^{2^M}\sum_{j=1}^{2^J}\boldsymbol{h}_j\boldsymbol{v}_i^T) && \\
= & \frac{1}{4}\mathrm{tr}(((\mathbf{W}-\mathbf{W}^k)\odot(\mathbf{W}-\mathbf{W}^k))(\frac{2^{M+J}}{4}\mathbf{1}_{J\times M})) && \\
= & \frac{2^{M+J}}{16}||\mathbf{W}-\mathbf{W}||_F^2 && \text{(C.8)}
\end{aligned}$$

For realistic problems sizes of RBMs, the bound that comes out of the logsumexp ∞-norm bound is exponentially tighter than the bound using logsumexp ℓ_2 norm bound.

Similar analysis on the bias terms reveals a bounding term equations

$$\begin{aligned}
f(\{\boldsymbol{b},\boldsymbol{c}^k,\mathbf{W}^k\}) \leq & f(\boldsymbol{\theta}^k)+\langle\nabla_{\boldsymbol{b}}f(\boldsymbol{\theta}^k),\boldsymbol{b}-\boldsymbol{b}^k\rangle\\
& +\frac{2^J}{8}||\boldsymbol{b}-\boldsymbol{b}^k||_\infty^2 \quad \text{(C.9)}\\
f(\{\boldsymbol{b}^k,\boldsymbol{c},\mathbf{W}^k\}) \leq & f(\boldsymbol{\theta}^k)+\langle\nabla_{\boldsymbol{c}}f(\boldsymbol{\theta}^k),\boldsymbol{c}-\boldsymbol{c}^k\rangle\\
& +\frac{2^M}{8}||\boldsymbol{c}-\boldsymbol{c}^k||_\infty^2 \quad \text{(C.10)}
\end{aligned}$$