A Theorem proofs

Proof. Proof of Theorem 1.

The Hessian of the lse function is given by

\[
\nabla^2 lse_\omega(u) = \frac{\text{diag}(\omega \circ \exp(u))}{\omega^T \exp(u)} - \frac{(\omega \circ \exp(u))(\omega \circ \exp(u))^T}{(\omega^T \exp(u))^2}
\]

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

\[
\frac{\text{diag}(\omega \circ \exp(u))}{\omega^T \exp(u)}
\]

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

\[
-\frac{(\omega \circ \exp(u))(\omega \circ \exp(u))^T}{(\omega^T \exp(u))^2}
\]

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

Writing Taylor’s theorem:

\[
lse_\omega(v) = lse_\omega(u) + \langle \nabla lse_\omega(u), v - u \rangle + \int_0^1 (1 - t)(v - u)^T \nabla^2 lse_\omega(u + t(v - u))(v - u)dt
\]

The terms in the integral can be bound

\[
(v - u)^T \nabla^2 lse_\omega(u + t(v - u))(v - u) \leq (v - u)^T \frac{\text{diag}(\omega \circ \exp(u + t(v - u)))}{\omega^T \exp(u + t(v - u))}(v - u)
\]

\[
= \sum_{j=1}^J \frac{\omega_j \exp(u_j + t(v_j - u_j))}{\omega^T \exp(u + t(v - u))} (v_j - u_j)^2
\]

\[
\leq \max_{c \geq 0, ||c||_1 = 1} \sum_{j=1}^J c_j (v_j - u_j)^2
\]

\[
= ||v - u||_\infty^2
\]

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 \(\frac{1}{2}||v - u||_\infty^2\).

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 \(\{h_i\}_{i=1}^2\) as the set of unique binary vectors \(\{0,1\}^J\), and let \(H \in \{0,1\}^{J \times 2^J}\) be the matrix form of this set.

\[
f(\theta) = \log \sum_{i=1}^{2^J} \omega_i \exp(h_i^T b)
\]

Equation A.5 can be equivalently written as

\[
f(\theta) = \log \omega^T \exp(H^T b)
\]

with \(\omega\) not dependent on \(b\). Plugging into Equation 17,

\[
f(\{b, c^k, W^k\}) \leq f(\theta^k) + \langle \nabla_{H^T b} lse_\omega(H^T b^k), H^T (b - b^k) \rangle + \frac{1}{2} ||H^T(b - b^k)||^2_\infty
\]

To rewrite the inner product term, note that

\[
\nabla_{H^T b} lse_\omega(H^T b^k) = H^T \nabla_b f(\theta^k)
\]

The bound is simplified as

\[
||H^T(b - b^k)||_\infty = \max_i |h_i^T(b - b^k)| \leq J||b - b^k||_\infty
\]

Alternatively, this could be bound as

\[
||H^T(b - b^k)||_\infty \leq \sqrt{J}||b - b^k||_2
\]

\[
||H^T(b - b^k)||_\infty \leq ||b - b^k||_1
\]

The proof on \(c\) follows with the same techniques.

Proof. Proof of Theorem 3.

As in the proof for Theorem 2, let \(H \in \{0,1\}^{J \times 2^J}\)
and \( V \in \{0, 1\}^{M \times 2^H} \), where each column is an unique binary vector. Define \( U = V^T W H \) and \( \Omega_{ij} = u_i^T c + h_j^T b \). Let \( u = \text{vec}(U) \) and \( \omega = \text{vec}(\Omega) \). The log partition function is equivalently written

\[
  f(\theta) = \log \sum_{i=1}^{2^H} \sum_{j=1}^{2^H} \Omega_{ij} \exp U_{ij} \tag{A.12}
\]

\[
  f(\theta) = \log (\omega^T \exp u) \tag{A.13}
\]

Plugging this form into Equation 17:

\[
  lse_\omega(u) \geq lse_\omega(u^k) + \langle \nabla_u lse_\omega(u^k), u - u^k \rangle + \frac{1}{2} ||\text{vec}(U - U^k)||_\infty^2 \tag{A.14}
\]

Note that

\[
  \langle \nabla_u lse_\omega(u), u - u^k \rangle = \text{tr}((\nabla_U lse_\omega(U))^T (U - U^k)) = \text{tr}(W) f(\theta) \tag{A.15}
\]

Writing the inner product in terms of \( W \) gives

\[
  \text{tr}((\nabla_U lse_\omega(U))^T (U - U^k)) = \text{tr}((W)^T (W - W^k)) \tag{A.16}
\]

The bound is simplified:

\[
  ||\text{vec}(U - U^k)||_\infty = \max_{i,j} ||u_i^T (W - W^k) h_j|| \leq \sqrt{MJ} ||W - W^k||_{S^\infty} \tag{A.17}
\]

Combining these two elements proves Theorem 3.

### B Derivation of optimal steps

**Proof.** Proof of \( b^* \) in Equation 25.

We want to find the minimizer of

\[
  \min_b \langle \nabla_b F(\theta^k), b - b^k \rangle + \frac{J}{2} ||b - b^k||_\infty^2
\]

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

\[
  = \min_{b,a,|b| \leq a, a \geq 0} \langle \nabla_b F(\theta^k), b - b^k \rangle + \frac{J}{2} a^2 \tag{B.1}
\]

This is straightforward to solve:

\[
  a^* = \frac{1}{J} ||\nabla_b F(\theta^k)||_1 \tag{B.2}
\]

\[
  b^* = b - \frac{1}{J} ||\nabla_b F(\theta^k)||_1 \times \text{sign}(\nabla_b F(\theta^k)) \tag{B.3}
\]

**Proof.** Proof of \( W^* \) in Equation 28.

Let \( D = W - W^k \), and decompose \( D = ARB^T \), with \( A \) and \( B \) denoting the left and right singular vectors of \( \nabla_W F(\theta^k) \). Then we want to minimize the quantity

\[
  \min_D \text{tr}(\nabla_W F(\theta^k)D) + \frac{MJ}{2} ||D||_{S^\infty}^2
\]

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

\[
  = \min_{D,a} \text{tr}(\nabla_W F(\theta^k)D) + \frac{MJ}{2} a^2 \tag{B.4}
\]

\[
  a = ||\lambda||_1 \tag{B.5}
\]

\[
  R^* = \frac{1}{MJ} ||\lambda||_1 \times I_M \tag{B.6}
\]

**C Discussion of using \( \ell_2 \) bound instead of \( \ell_\infty \) bound on lse function**

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

\[
  \text{lse}_1(v) \leq \text{lse}_1(u) + \langle \nabla_u \text{lse}_1(u), v - u \rangle + \frac{1}{2} (v - u)^T B(v - u) \tag{C.1}
\]

\[
  B = \frac{1}{2} \left[ I_J - \frac{1}{J} 1_J 1_J^T \right] \tag{C.2}
\]

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

\[
  \frac{1}{2} (v - u)^T B(v - u) = \frac{1}{4} ||v - u||_2^2 - 4 \text{mean}(v - u)^2 \tag{C.3}
\]

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

\[
  \text{lse}_\omega(v) \leq \text{lse}_\omega(u) + \langle \nabla_u \text{lse}_\omega(u), v - u \rangle + \frac{1}{2} \times 2 ||v - u||_2^2 \tag{C.4}
\]

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 $W$,

$$
\frac{1}{4} \| \text{vec}(U - U^k) \|_2^2
$$

(C.5)

$$
= \frac{1}{4} \sum_{i=1}^{2^M} \sum_{j=1}^{2^J} \left( u_i^T (W - W^k) v_j \right)^2
$$

(C.6)

$$
\leq \frac{1}{4} \sum_{i=1}^{2^M} \sum_{j=1}^{2^J} v_i^T ((W - W^k) \odot (W - W^k)) u_j
$$

(C.7)

$$
= \frac{1}{4} \text{tr}( (W - W^k) \odot (W - W^k) ) \sum_{i=1}^{2^M} \sum_{j=1}^{2^J} h_j v_i^T
$$

(C.8)

$$
= \frac{1}{4} \text{tr}( (W - W^k) \odot (W - W^k) ) (2^{M+J} \frac{1}{4} 1_{J \times M})
$$

$$
= \frac{2^{M+J}}{16} \| W - W^k \|_F^2
$$

For realistic problem sizes of RBMs, the bound that comes out of the logsumexp $\infty$-norm bound is exponentially tighter than the bound using logsumexp $\ell_2$ norm bound.

Similar analysis on the bias terms reveals a bounding term equations

$$
f(b, c^k, W^k) \leq f(\theta^k) + \langle \nabla_b f(\theta^k), b - b^k \rangle + \frac{2^J}{8} \| b - b^k \|_\infty^2
$$

(C.9)

$$
f(b^k, c, W^k) \leq f(\theta^k) + \langle \nabla_c f(\theta^k), c - c^k \rangle + \frac{2^M}{8} \| c - c^k \|_\infty^2
$$

(C.10)