Restricted Boltzmann Machine Derivations

Jan Schlüter

last change: March 26, 2014

Abstract

This document gives detailed derivations for the central quantities of Restricted Boltzmann Machines (RBMs): The conditional distributions of visible and hidden units, and the log likelihood gradient with respect to the model parameters. It handles the standard Bernoulli-Bernoulli RBM (with binary visible and hidden units) as well as different formulations of the Gaussian-Bernoulli RBM (with real-valued visible units). It is not meant as a general introduction to RBMs, but as a supplement helping to follow the mathematics. If you are not familiar with RBMs, introductions can be found in [3] (mathematical), [2] (practical) or [7, Section 4.4] (intuitive).

1 General equations

An RBM is defined by its energy function

$$E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{\theta}) \tag{1}$$

defining the joint probability distribution

$$p(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta}) = \frac{\exp\big(-E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{\theta})\big)}{Z(\boldsymbol{\theta})} = \frac{\exp\big(-E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{\theta})\big)}{\sum_{\boldsymbol{u},\boldsymbol{g}} \exp\big(-E(\boldsymbol{u}, \boldsymbol{g}, \boldsymbol{\theta})\big)}, \tag{2}$$

where $\boldsymbol{v}$ is the vector of visible units, $\boldsymbol{h}$ is the vector of hidden units, and $\boldsymbol{\theta}$ is the set of model parameters (i.e., the connection weights and biases). In sums over configurations, we use $\boldsymbol{u}$ for visible and $\boldsymbol{g}$ for hidden unit states. For continuous visible units, all sums over visible configurations $\boldsymbol{u}$ in this section need to be replaced by integrals.

1.1 Marginal distribution of visible and hidden units

From (2), we can derive the marginal distribution of the visible units

$$p(\boldsymbol{v}; \boldsymbol{\theta}) = \sum_{\boldsymbol{g}} p(\boldsymbol{v}, \boldsymbol{g}; \boldsymbol{\theta}) = \frac{\sum_{\boldsymbol{g}} \exp\big(-E(\boldsymbol{v}, \boldsymbol{g}, \boldsymbol{\theta})\big)}{Z(\boldsymbol{\theta})} \tag{3}$$

and the marginal distribution of the hidden units

$$p(\boldsymbol{h}; \boldsymbol{\theta}) = \sum_{\boldsymbol{u}} p(\boldsymbol{u}, \boldsymbol{h}; \boldsymbol{\theta}) = \frac{\sum_{\boldsymbol{u}} \exp\big(-E(\boldsymbol{u}, \boldsymbol{h}, \boldsymbol{\theta})\big)}{Z(\boldsymbol{\theta})}. \tag{4}$$

1.2 Conditional distribution of visible and hidden units

From (2) and (4), we can derive the conditional distribution of the visible units

$$p(\boldsymbol{v}|\boldsymbol{h}; \boldsymbol{\theta}) = \frac{p(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})}{p(\boldsymbol{h}; \boldsymbol{\theta})} = \frac{\frac{1}{Z(\theta)} \exp\big(-E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{\theta})\big)}{\frac{1}{Z(\theta)} \sum_{\boldsymbol{u}} \exp\big(-E(\boldsymbol{u}, \boldsymbol{h}, \boldsymbol{\theta})\big)} = \frac{\exp\big(-E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{\theta})\big)}{\sum_{\boldsymbol{u}} \exp\big(-E(\boldsymbol{u}, \boldsymbol{h}, \boldsymbol{\theta})\big)}, \tag{5}$$

and from (2) and (3), we can derive the conditional distribution of the hidden units

$$p(\boldsymbol{h}|\boldsymbol{v}; \boldsymbol{\theta}) = \frac{p(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})}{p(\boldsymbol{v}; \boldsymbol{\theta})} = \frac{\frac{1}{Z(\theta)} \exp\big(-E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{\theta})\big)}{\frac{1}{Z(\theta)} \sum_{\boldsymbol{g}} \exp\big(-E(\boldsymbol{v}, \boldsymbol{g}, \boldsymbol{\theta})\big)} = \frac{\exp\big(-E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{\theta})\big)}{\sum_{\boldsymbol{g}} \exp\big(-E(\boldsymbol{v}, \boldsymbol{g}, \boldsymbol{\theta})\big)}. \tag{6}$$

1.3 Log likelihood gradient

We will derive the general form of the gradient of the log likelihood w.r.t. an arbitrary model parameter θ for a single training sample $\boldsymbol{v}$ here and use it to obtain the gradients for specific types of RBMs and model parameters in the following sections. Some steps are marked with a number and explained below.

$$
\begin{aligned}
&\frac{\partial}{\partial\theta}\log p(\boldsymbol{v};\boldsymbol{\theta}) \\
&\overset{1}{=} \frac{\partial}{\partial\theta}\log\left(\frac{\sum_{\boldsymbol{g}}\exp\big(-E(\boldsymbol{v},\boldsymbol{g},\boldsymbol{\theta})\big)}{Z(\boldsymbol{\theta})}\right) \\
&= \frac{\partial}{\partial\theta}\log\left(\sum_{\boldsymbol{g}}\exp\big(-E(\boldsymbol{v},\boldsymbol{g},\boldsymbol{\theta})\big) - \log Z(\boldsymbol{\theta})\right) \\
&= \frac{\partial}{\partial\theta}\log\sum_{\boldsymbol{g}}\exp\big(-E(\boldsymbol{v},\boldsymbol{g},\boldsymbol{\theta})\big) - \frac{\partial}{\partial\theta}\log Z(\boldsymbol{\theta}) \\
&\overset{2}{=} \frac{\frac{\partial}{\partial\theta}\sum_{\boldsymbol{g}}\exp\big(-E(\boldsymbol{v},\boldsymbol{g},\boldsymbol{\theta})\big)}{\sum_{\boldsymbol{g}}\exp\big(-E(\boldsymbol{v},\boldsymbol{g},\boldsymbol{\theta})\big)} - \frac{\frac{\partial}{\partial\theta}Z(\boldsymbol{\theta})}{Z(\boldsymbol{\theta})} \\
&\overset{3}{=} \frac{\sum_{\boldsymbol{g}}\exp\big(-E(\boldsymbol{v},\boldsymbol{g},\boldsymbol{\theta})\big)\cdot\frac{\partial}{\partial\theta}\big(-E(\boldsymbol{v},\boldsymbol{g},\boldsymbol{\theta})\big)}{\sum_{\boldsymbol{g}}\exp\big(-E(\boldsymbol{v},\boldsymbol{g},\boldsymbol{\theta})\big)} - \frac{\sum_{\boldsymbol{u},\boldsymbol{g}}\exp\big(-E(\boldsymbol{u},\boldsymbol{g},\boldsymbol{\theta})\big)\cdot\frac{\partial}{\partial\theta}\big(-E(\boldsymbol{u},\boldsymbol{g},\boldsymbol{\theta})\big)}{Z(\boldsymbol{\theta})} \\
&= \sum_{\boldsymbol{g}}\left(\frac{\exp\big(-E(\boldsymbol{v},\boldsymbol{g},\boldsymbol{\theta})\big)}{\sum_{\boldsymbol{g'}}\exp\big(-E(\boldsymbol{v},\boldsymbol{g'},\boldsymbol{\theta})\big)}\cdot\frac{\partial}{\partial\theta}\Big(-E(\boldsymbol{v},\boldsymbol{g},\boldsymbol{\theta})\Big)\right) - \sum_{\boldsymbol{u},\boldsymbol{g}}\left(\frac{\exp\big(-E(\boldsymbol{u},\boldsymbol{g},\boldsymbol{\theta})\big)}{Z(\boldsymbol{\theta})}\cdot\frac{\partial}{\partial\theta}\Big(-E(\boldsymbol{u},\boldsymbol{g},\boldsymbol{\theta})\Big)\right) \\
&\overset{4}{=} \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v};\boldsymbol{\theta})\cdot\frac{\partial}{\partial\theta}\Big(-E(\boldsymbol{v},\boldsymbol{g},\boldsymbol{\theta})\Big) - \sum_{\boldsymbol{u},\boldsymbol{g}} p(\boldsymbol{u},\boldsymbol{g};\boldsymbol{\theta})\cdot\frac{\partial}{\partial\theta}\Big(-E(\boldsymbol{u},\boldsymbol{g},\boldsymbol{\theta})\Big) \\
&= \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v};\boldsymbol{\theta})\cdot\frac{\partial}{\partial\theta}\Big(-E(\boldsymbol{v},\boldsymbol{g},\boldsymbol{\theta})\Big) - \sum_{\boldsymbol{u}} p(\boldsymbol{u};\boldsymbol{\theta})\sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{u};\boldsymbol{\theta})\cdot\frac{\partial}{\partial\theta}\Big(-E(\boldsymbol{u},\boldsymbol{g},\boldsymbol{\theta})\Big) \\
&\overset{5}{=} G_\theta(\boldsymbol{v},\boldsymbol{\theta}) - \langle G_\theta(\boldsymbol{u},\boldsymbol{\theta})\rangle_{p(\boldsymbol{u};\boldsymbol{\theta})} \qquad \text{with } G_\theta(\boldsymbol{x},\boldsymbol{\theta}) := \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{x};\boldsymbol{\theta})\cdot\frac{\partial}{\partial\theta}\Big(-E(\boldsymbol{x},\boldsymbol{g},\boldsymbol{\theta})\Big)
\end{aligned}
\tag{7}
$$

Explanation of steps marked with a number:

1. We substitute the definition of $p(\boldsymbol{v};\boldsymbol{\theta})$ (3).
2. We derive the log functions, applying the chain rule of differentiation.
3. We substitute the definition of the partition function $Z(\boldsymbol{\theta})$ (see denominators of (2)), then derive the exponential functions, again applying the chain rule.
4. We back-substitute equations (6) and (2), respectively.
5. We note that the second term computes the expectation of the first term under the model's marginal distribution of visible units $p(\boldsymbol{v};\boldsymbol{\theta})$.

2 Bernoulli-Bernoulli RBM

Energy function:

$$E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{\theta}) = -\sum_i v_i a_i - \sum_{i,j} v_i W_{ij} h_j - \sum_j h_j b_j \tag{8}$$

2.1 Conditional distribution of the visible units

In the binary RBM, the conditional probability of a visible unit being active given the hidden unit states is given as:

$$p(v_k = 1|\boldsymbol{h}; \boldsymbol{\theta}) = \sigma\Big(a_k + \sum_j W_{kj} h_j\Big), \tag{9}$$

where $\sigma(x) = (1 + \exp(-x))^{-1}$ is the logistic sigmoid function. The derivation is symmetric to the one for the hidden units discussed in the next section.

2.2 Conditional distribution of the hidden units

This derivation is copied from [7, p. 143]. Some steps are marked with a number and explained below.

$$\begin{aligned}
p(h_k = 1|\boldsymbol{v}; \boldsymbol{\theta}) &\overset{1}{=} \frac{p(\boldsymbol{v}, h_k = 1; \boldsymbol{\theta})}{p(\boldsymbol{v}; \boldsymbol{\theta})} \\
&= \frac{\sum_{\{\boldsymbol{g}|g_k=1\}} p(\boldsymbol{v}, \boldsymbol{g}; \boldsymbol{\theta})}{\sum_{\boldsymbol{g}} p(\boldsymbol{v}, \boldsymbol{g}; \boldsymbol{\theta})} \\
&\overset{2}{=} \frac{\sum_{\{\boldsymbol{g}|g_k=1\}} \exp\big(-E(\boldsymbol{v}, \boldsymbol{g}, \boldsymbol{\theta})\big)}{\sum_{\boldsymbol{g}} \exp\big(-E(\boldsymbol{v}, \boldsymbol{g}, \boldsymbol{\theta})\big)} \\
&\overset{3}{=} \frac{\sum_{\{\boldsymbol{g}|g_k=1\}} \exp\big(-E(\boldsymbol{v}, \boldsymbol{g}, \boldsymbol{\theta})\big)}{\sum_{\{\boldsymbol{g}|g_k=1\}} \exp\big(-E(\boldsymbol{v}, \boldsymbol{g}, \boldsymbol{\theta})\big) + \sum_{\{\boldsymbol{g}|g_k=0\}} \exp\big(-E(\boldsymbol{v}, \boldsymbol{g}, \boldsymbol{\theta})\big)} \\
&= \frac{1}{1 + \frac{\sum_{\{\boldsymbol{g}|g_k=0\}} \exp\big(-E(\boldsymbol{v}, \boldsymbol{g}, \boldsymbol{\theta})\big)}{\sum_{\{\boldsymbol{g}|g_k=1\}} \exp\big(-E(\boldsymbol{v}, \boldsymbol{g}, \boldsymbol{\theta})\big)}} \\
&\overset{4}{=} \frac{1}{1 + \frac{\sum_{\{\boldsymbol{g}|g_k=0\}} \exp\big(\sum_{i,j} v_i W_{ij} g_j + \sum_i v_i a_i + \sum_j g_j b_j\big)}{\sum_{\{\boldsymbol{g}|g_k=1\}} \exp\big(\sum_{i,j} v_i W_{ij} g_j + \sum_i v_i a_i + \sum_j g_j b_j\big)}} \\
&\overset{5}{=} \frac{1}{1 + \frac{\sum_{\{\boldsymbol{g}|g_k=0\}} \exp\big(\sum_i \sum_{j\neq k} v_i W_{ij} g_j + \sum_i v_i a_i + \sum_{j\neq k} g_j b_j\big)}{\sum_{\{\boldsymbol{g}|g_k=1\}} \exp\big(\sum_i \sum_{j\neq k} v_i W_{ij} g_j + \sum_i v_i a_i + \sum_{j\neq k} g_j b_j\big)} \cdot \frac{\exp\big(\sum_i v_i W_{ik} \cdot 0 + 0 \cdot b_k\big)}{\exp\big(\sum_i v_i W_{ik} \cdot 1 + 1 \cdot b_k\big)}} \\
&\overset{6}{=} \frac{1}{1 + 1 \cdot \frac{1}{\exp\big(\sum_i v_i W_{ik} + b_k\big)}} \\
&= \frac{1}{1 + \exp\big(-(b_k + \sum_i v_i W_{ik})\big)} \\
&= \sigma\Big(b_k + \sum_i v_i W_{ik}\Big)
\end{aligned} \tag{10}$$

Explanation of steps marked with a number:

1. We start from the definition of conditional probability.
2. We substitute the definition of the joint probability density function (2).
3. We split the sum over all configurations $\boldsymbol{g}$ into two sums over configurations with an active and inactive hidden unit g_k, respectively.
4. We substitute the energy function (8).
5. From the sums over j, we move out the terms for $j = k$, for which we substitute the known values of $g_j = g_k$.

6. As the terms depending on g_k have been moved out of the two long exponentials, the sums over $\{\boldsymbol{g}|g_k = 0\}$ and $\{\boldsymbol{g}|g_k = 1\}$ are equal so their quotient evaluates to 1.

2.3 Log likelihood gradient

Following the derivation of Section 1.3, we just need to compute $G_\theta(\boldsymbol{v}, \boldsymbol{\theta})$ for the model parameters θ we are interested in: a connection weight W_{ij}, visible bias a_i and hidden bias b_j. The derivations are based on [3, p. 26]. As always, some steps are marked with a number and explained in the end of the section. We start with the connection weight:

$$\begin{aligned}
G_{W_{ij}}(\boldsymbol{v}, \boldsymbol{\theta}) &= \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v}; \boldsymbol{\theta}) \cdot \frac{\partial}{\partial W_{ij}} \Big(-E(\boldsymbol{v}, \boldsymbol{g}, \boldsymbol{\theta}) \Big) \\
&\overset{1}{=} \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v}; \boldsymbol{\theta}) \cdot v_i g_j \\
&\overset{2}{=} v_i \cdot \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v}; \boldsymbol{\theta}) g_j \\
&\overset{3}{=} v_i \cdot \sum_{g_j} \sum_{\boldsymbol{g_{\neg j}}} p(g_j|\boldsymbol{v}; \boldsymbol{\theta}) p(\boldsymbol{g_{\neg j}}|\boldsymbol{v}; \boldsymbol{\theta}) g_j \\
&\overset{4}{=} v_i \cdot \sum_{g_j} p(g_j|\boldsymbol{v}; \boldsymbol{\theta}) g_j \sum_{\boldsymbol{g_{\neg j}}} p(\boldsymbol{g_{\neg j}}|\boldsymbol{v}; \boldsymbol{\theta}) \\
&\overset{5}{=} v_i \cdot \sum_{g_j} p(g_j|\boldsymbol{v}; \boldsymbol{\theta}) g_j \\
&\overset{6}{=} v_i \cdot \Big(p(h_j = 1|\boldsymbol{v}; \boldsymbol{\theta}) \cdot 1 + p(h_j = 0|\boldsymbol{v}; \boldsymbol{\theta}) \cdot 0 \Big) \\
&= v_i \cdot p(h_j = 1|\boldsymbol{v}; \boldsymbol{\theta}) \\
&\overset{7}{=} v_i \cdot \sigma\Big(b_j + \sum_k v_k W_{kj} \Big)
\end{aligned}$$

Revisiting the derivation, note that we have established the following between the third and the previous to last line:

$$\sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v}; \boldsymbol{\theta}) g_j \quad = \quad p(h_j = 1|\boldsymbol{v}; \boldsymbol{\theta}) \tag{11}$$

This holds whenever $p(\boldsymbol{h}|\boldsymbol{v}; \boldsymbol{\theta})$ is factorial and will become useful in other derivations.
We continue with the gradient for the visible bias:

$$\begin{aligned}
G_{a_i}(\boldsymbol{v}, \boldsymbol{\theta}) &= \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v}; \boldsymbol{\theta}) \cdot \frac{\partial}{\partial a_i} \Big(-E(\boldsymbol{v}, \boldsymbol{g}, \boldsymbol{\theta}) \Big) \\
&\overset{1}{=} \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v}; \boldsymbol{\theta}) \cdot v_i \\
&\overset{2}{=} v_i \cdot \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v}; \boldsymbol{\theta}) \\
&\overset{5}{=} v_i
\end{aligned}$$

Finally, we compute the gradient for the hidden bias:

$$\begin{aligned}
G_{b_j}(\boldsymbol{v}, \boldsymbol{\theta}) &= \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v}; \boldsymbol{\theta}) \cdot \frac{\partial}{\partial b_j} \Big(-E(\boldsymbol{v}, \boldsymbol{g}, \boldsymbol{\theta}) \Big) \\
&\overset{1}{=} \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v}; \boldsymbol{\theta}) \cdot g_j \\
&\overset{8}{=} \sigma\Big(b_j + \sum_k v_k W_{kj} \Big)
\end{aligned}$$

Explanation of steps marked with a number:

1. We substitute the energy function (8) and differentiate.
2. The factor v_i is the same for every $\boldsymbol{g}$, so we can pull it out of the sum.
3. We split the sum over all configurations $\boldsymbol{g}$ into the sum over the two states of g_j and the sum over the states of the remaining hidden units $\boldsymbol{g}_{\boldsymbol{\neg j}}$. Furthermore, noting that the distribution $p(\boldsymbol{g}|\boldsymbol{v};\boldsymbol{\theta})$ is factorial (the hidden units are mutually independent given $\boldsymbol{v}$), we pull out the factor $p(g_j|\boldsymbol{v};\boldsymbol{\theta})$.
4. The factor $p(g_j|\boldsymbol{v};\boldsymbol{\theta})g_j$ is the same for every $\boldsymbol{g}_{\boldsymbol{\neg j}}$, so we can pull it out of the inner sum.
5. The sum over a probability distribution evaluates to 1 and can be dropped.
6. We write out the sum over the two possible values of g_j (and rename it to h_j).
7. We substitute the conditional hidden unit activation probability (10).
8. We substitute (11), then (10).

Inserting our results for $G_{W_{ij}}(\boldsymbol{v},\boldsymbol{\theta})$, $G_{a_i}(\boldsymbol{v},\boldsymbol{\theta})$, $G_{b_j}(\boldsymbol{v},\boldsymbol{\theta})$ into (7), we obtain:

$$\frac{\partial}{\partial W_{ij}}\log p(\boldsymbol{v};\boldsymbol{\theta}) = v_i \cdot \sigma\Big(b_j + \sum_k v_k W_{kj}\Big) - \left\langle u_i \cdot \sigma\Big(b_j + \sum_k u_k W_{kj}\Big)\right\rangle_{p(\boldsymbol{u};\boldsymbol{\theta})} \tag{12}$$

$$\frac{\partial}{\partial a_i}\log p(\boldsymbol{v};\boldsymbol{\theta}) = v_i - \langle u_i\rangle_{p(\boldsymbol{u};\boldsymbol{\theta})} \tag{13}$$

$$\frac{\partial}{\partial b_j}\log p(\boldsymbol{v};\boldsymbol{\theta}) = \sigma\Big(b_j + \sum_k v_k W_{kj}\Big) - \left\langle \sigma\Big(b_j + \sum_k u_k W_{kj}\Big)\right\rangle_{p(\boldsymbol{u};\boldsymbol{\theta})} \tag{14}$$

The first terms can be easily computed from the training example $\boldsymbol{v}$, the second terms require evaluating an expectation over $p(\boldsymbol{v};\boldsymbol{\theta})$, which is computationally infeasible due to the partition function in (3). In practice, the second terms are usually approximated by averaging over samples from $p(\boldsymbol{v};\boldsymbol{\theta})$, which in turn are drawn approximately using Markov-Chain Monte-Carlo (MCMC) methods such as Gibbs sampling. Please see one of the RBM tutorials mentioned in the abstract for details.
For notational brevity, let us introduce the following abbreviation:

$$\overline{h}_j(\boldsymbol{v},\boldsymbol{\theta}) \quad := \quad p(h_j = 1|\boldsymbol{v};\boldsymbol{\theta}) \tag{15}$$

This way, the gradients can be written as:

$$\frac{\partial}{\partial W_{ij}}\log p(\boldsymbol{v};\boldsymbol{\theta}) = v_i \cdot \overline{h}_j(\boldsymbol{v},\boldsymbol{\theta}) - \big\langle u_i \cdot \overline{h}_j(\boldsymbol{u},\boldsymbol{\theta})\big\rangle_{p(\boldsymbol{u};\boldsymbol{\theta})} \tag{16}$$

$$\frac{\partial}{\partial a_i}\log p(\boldsymbol{v};\boldsymbol{\theta}) = v_i - \langle u_i\rangle_{p(\boldsymbol{u};\boldsymbol{\theta})} \tag{17}$$

$$\frac{\partial}{\partial b_j}\log p(\boldsymbol{v};\boldsymbol{\theta}) = \overline{h}_j(\boldsymbol{v},\boldsymbol{\theta}) - \big\langle \overline{h}_j(\boldsymbol{u},\boldsymbol{\theta})\big\rangle_{p(\boldsymbol{u};\boldsymbol{\theta})} \tag{18}$$

In a practical implementation, one will usually draw samples $\boldsymbol{u}$, precompute all $\overline{h}_j(\boldsymbol{v},\boldsymbol{\theta})$ and $\overline{h}_j(\boldsymbol{u},\boldsymbol{\theta})$, and then use those to compute the gradients. Note that $\overline{h}_j(\boldsymbol{v},\boldsymbol{\theta})$ and $\overline{h}_j(\boldsymbol{u},\boldsymbol{\theta})$ denote the real-valued activation probabilties of hidden units, not binary samples.

3 Gaussian-Bernoulli RBM (Krizhevsky's variant)

Energy function [5, p. 12]:

$$E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{\theta}) = \sum_i \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{i,j} \frac{v_i}{\sigma_i} W_{ij} h_j - \sum_j h_j b_j \tag{19}$$

3.1 Conditional distribution of the visible units

We insert (19) into (5). We follow [5, p. 13], fixing one mistake and proceeding a lot slower. Again, some steps are marked with a number and explained below.

$$\begin{aligned}
p(\boldsymbol{v}|\boldsymbol{h};\boldsymbol{\theta}) &= \frac{\exp\big(-E(\boldsymbol{v},\boldsymbol{h},\boldsymbol{\theta})\big)}{\int_{\boldsymbol{u}} \exp\big(-E(\boldsymbol{u},\boldsymbol{h},\boldsymbol{\theta})\big) d\boldsymbol{u}} \\
&= \frac{\exp(-\sum_i \frac{(v_i-a_i)^2}{2\sigma_i^2} + \sum_{i,j} \frac{v_i}{\sigma_i} W_{ij} h_j + \sum_j h_j b_j)}{\int_{\boldsymbol{u}} \exp(-\sum_i \frac{(u_i-a_i)^2}{2\sigma_i^2} + \sum_{i,j} \frac{u_i}{\sigma_i} W_{ij} h_j + \sum_j h_j b_j) d\boldsymbol{u}} \\
&= \frac{\exp(-\sum_i \frac{(v_i-a_i)^2}{2\sigma_i^2} + \sum_{i,j} \frac{v_i}{\sigma_i} W_{ij} h_j) \exp(\sum_j h_j b_j)}{\int_{\boldsymbol{u}} \exp(-\sum_i \frac{(u_i-a_i)^2}{2\sigma_i^2} + \sum_{i,j} \frac{u_i}{\sigma_i} W_{ij} h_j) \exp(\sum_j h_j b_j) d\boldsymbol{u}} \\
&\overset{1}{=} \frac{\exp(-\sum_i \frac{(v_i-a_i)^2}{2\sigma_i^2} + \sum_i \frac{v_i}{\sigma_i} \sum_j W_{ij} h_j)}{\int_{\boldsymbol{u}} \exp(-\sum_i \frac{(u_i-a_i)^2}{2\sigma_i^2} + \sum_i \frac{u_i}{\sigma_i} \sum_j W_{ij} h_j) d\boldsymbol{u}} \\
&\overset{2}{=} \frac{\exp(-\sum_i \frac{(v_i-a_i)^2}{2\sigma_i^2} + \sum_i \frac{v_i}{\sigma_i} B_i)}{\int_{\boldsymbol{u}} \exp(-\sum_i \frac{(u_i-a_i)^2}{2\sigma_i^2} + \sum_i \frac{u_i}{\sigma_i} B_i) d\boldsymbol{u}} \qquad \text{with } B_i := \sum_j W_{ij} h_j \\
&= \frac{\prod_i \exp(-\frac{(v_i-a_i)^2}{2\sigma_i^2} + \frac{v_i}{\sigma_i} B_i)}{\int_{\boldsymbol{u}} \prod_i \exp(-\frac{(u_i-a_i)^2}{2\sigma_i^2} + \frac{u_i}{\sigma_i} B_i) d\boldsymbol{u}} \\
&\overset{3}{=} \frac{\cdots}{\prod_i \int_{u_i} \exp(-\frac{(u_i-a_i)^2}{2\sigma_i^2} + \frac{u_i}{\sigma_i} B_i) du_i} \\
&= \frac{\cdots}{\prod_i \int_{u_i} \exp(-\frac{u_i^2 - 2u_i a_i + a_i^2}{2\sigma_i^2} + \frac{u_i}{\sigma_i} B_i) du_i} \\
&= \frac{\cdots}{\prod_i \int_{u_i} \exp(-u_i^2 \frac{1}{2\sigma_i^2} + u_i(\frac{a_i}{\sigma_i^2} + \frac{B_i}{\sigma_i}) - \frac{a_i^2}{2\sigma_i^2}) du_i} \\
&\overset{4}{=} \frac{\cdots}{\prod_i \sqrt{\pi 2\sigma_i^2} \exp(\frac{2\sigma_i^2}{4}(\frac{a_i}{\sigma_i^2} + \frac{B_i}{\sigma_i})^2 - \frac{a_i^2}{2\sigma_i^2})} \\
&= \frac{\cdots}{\prod_i \sigma_i \sqrt{2\pi} \exp(\frac{1}{2}(\frac{a_i}{\sigma_i} + B_i)^2 - \frac{a_i^2}{2\sigma_i^2})} \\
&= \frac{\cdots}{\prod_i \sigma_i \sqrt{2\pi} \exp(\frac{1}{2}(\frac{a_i^2}{\sigma_i^2} + 2\frac{a_i}{\sigma_i} B_i + B_i^2) - \frac{a_i^2}{2\sigma_i^2})} \\
&= \frac{\prod_i \exp(-\frac{(v_i-a_i)^2}{2\sigma_i^2} + \frac{v_i}{\sigma_i} B_i)}{\prod_i \sigma_i \sqrt{2\pi} \exp(\frac{a_i}{\sigma_i} B_i + \frac{1}{2} B_i^2)} \\
&= \prod_i \frac{1}{\sigma_i \sqrt{2\pi}} \exp(-\frac{(v_i-a_i)^2}{2\sigma_i^2} + \frac{v_i}{\sigma_i} B_i - \frac{a_i}{\sigma_i} B_i - \frac{1}{2} B_i^2) \\
&= \prod_i \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_i^2}\Big((v_i-a_i)^2 - 2\sigma_i(v_i-a_i)B_i + \sigma_i^2 B_i^2\Big)\right) \\
&= \prod_i \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_i^2}\Big((v_i-a_i) - \sigma_i B_i\Big)^2\right) \\
&\overset{5}{=} \prod_i \mathcal{N}\Big(v_i; a_i + \sigma_i \sum_j W_{ij} h_j, \sigma_i^2\Big)
\end{aligned} \tag{20}$$

Explanation of steps marked with a number:

1. The $\exp(\sum_j h_j b_j)$ term in the denominator is constant w.r.t. $\boldsymbol{u}$, thus can be dragged out of the integral as a factor and cancels with the same factor in the nominator.
2. Just to make notation easier, we substitute $B_i := \sum_j W_{ij} h_j$.
3. Note that $\int_{\boldsymbol{u}} \prod_i f(u_i) d\boldsymbol{u} = \int_{u_1} \int_{u_2} \dots \int_{u_n} f(u_1) f(u_2) \dots f(u_n) du_1 du_2 \dots du_n$, which can be freely rearranged to $\int_{u_1} f(u_1) du_1 \int_{u_2} f(u_2) du_2 \dots \int_{u_n} f(u_n) du_n = \prod_i \int_{u_i} f(u_i) du_i$ because any of the integrals is just a constant factor w.r.t. the other integration variables.
4. We solve the Gaussian function integral $\int_x \exp(-x^2 a + xb + c) dx = \sqrt{\frac{\pi}{a}} \exp(\frac{b^2}{4a} + c)$.
5. For your convenience, the Normal distribution is $\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$.

3.2 Conditional distribution of the hidden units

We start from an intermediate step of the derivation in Section 2.2 (the one between the steps marked with 3 and 4). Again, some steps are marked with a number and explained below.

$$
\begin{aligned}
p(h_k = 1|\boldsymbol{v}) &= \frac{1}{1 + \frac{\sum_{\{\boldsymbol{g}|g_k=0\}} \exp\left(-E(\boldsymbol{v},\boldsymbol{g},\boldsymbol{\theta})\right)}{\sum_{\{\boldsymbol{g}|g_k=1\}} \exp\left(-E(\boldsymbol{v},\boldsymbol{g},\boldsymbol{\theta})\right)}} \\
&\overset{1}{=} \frac{1}{1 + \frac{\sum_{\{\boldsymbol{g}|g_k=0\}} \exp(\sum_{i,j} \frac{v_i}{\sigma_i} W_{ij} g_j - \sum_i \frac{(v_i - a_i)^2}{\sigma_i^2} + \sum_j g_j b_j)}{\sum_{\{\boldsymbol{g}|g_k=1\}} \exp(\sum_{i,j} \frac{v_i}{\sigma_i} W_{ij} g_j - \sum_i \frac{(v_i - a_i)^2}{\sigma_i^2} + \sum_j g_j b_j)}} \\
&\overset{2}{=} \frac{1}{1 + \frac{\exp(\sum_i \frac{v_i}{\sigma_i} W_{ik} \cdot 0 + 0 \cdot b_k)}{\exp(\sum_i \frac{v_i}{\sigma_i} W_{ik} \cdot 1 + 1 \cdot b_k)}} \\
&= \frac{1}{1 + \frac{1}{\exp(\sum_i \frac{v_i}{\sigma_i} W_{ik} + b_k)}} \\
&= \frac{1}{1 + \exp(-(b_k + \sum_i \frac{v_i}{\sigma_i} W_{ik}))} \\
&= \sigma\left(b_k + \sum_i \frac{v_i}{\sigma_i} W_{ik}\right)
\end{aligned} \tag{21}
$$

Explanation of steps marked with a number:

1. We substitute the energy function (19).
2. We substitute the known values for $g_j = g_k$ and cancel all terms not depending on g_k (cf. steps 5 and 6 in Section 2.2).

3.3 Log likelihood gradient

As in Section 2.3, we compute $G_\theta(\boldsymbol{v}, \boldsymbol{\theta})$ for the model parameters θ we are interested in: a connection weight W_{ij}, visible bias a_i, hidden bias b_j and visible standard deviation σ_i. Again, some steps are marked with a number and explained in the end of the section.
We start with the connection weight:

$$
\begin{aligned}
G_{W_{ij}}(\boldsymbol{v}, \boldsymbol{\theta}) &= \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v}; \boldsymbol{\theta}) \cdot \frac{\partial}{\partial W_{ij}} \left(-E(\boldsymbol{v}, \boldsymbol{g}, \boldsymbol{\theta})\right) \\
&\overset{1}{=} \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v}; \boldsymbol{\theta}) \cdot \frac{v_i}{\sigma_i} g_j \\
&\overset{2}{=} \frac{v_i}{\sigma_i} \cdot \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v}; \boldsymbol{\theta}) g_j
\end{aligned}
$$

$$\stackrel{3}{=} \frac{v_i}{\sigma_i} \cdot p(h_j = 1|\boldsymbol{v};\boldsymbol{\theta})$$
$$\stackrel{4}{=} \frac{v_i}{\sigma_i} \cdot \overline{h}_j(\boldsymbol{v},\boldsymbol{\theta})$$

We continue with the gradient for the visible bias:

$$\begin{aligned} G_{a_i}(\boldsymbol{v},\boldsymbol{\theta}) &= \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v};\boldsymbol{\theta}) \cdot \frac{\partial}{\partial a_i}\Big(-E(\boldsymbol{v},\boldsymbol{g},\boldsymbol{\theta})\Big) \\ &\stackrel{1}{=} \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v};\boldsymbol{\theta}) \cdot \frac{v_i - a_i}{\sigma_i^2} \\ &\stackrel{5}{=} \frac{v_i - a_i}{\sigma_i^2} \cdot \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v};\boldsymbol{\theta}) \\ &\stackrel{6}{=} \frac{v_i - a_i}{\sigma_i^2} \end{aligned}$$

We proceed with the gradient for the hidden bias:

$$\begin{aligned} G_{b_j}(\boldsymbol{v},\boldsymbol{\theta}) &= \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v};\boldsymbol{\theta}) \cdot \frac{\partial}{\partial b_j}\Big(-E(\boldsymbol{v},\boldsymbol{g},\boldsymbol{\theta})\Big) \\ &\stackrel{1}{=} \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v};\boldsymbol{\theta}) \cdot g_j \\ &\stackrel{3,4}{=} \overline{h}_j(\boldsymbol{v},\boldsymbol{\theta}) \end{aligned}$$

Finally, we compute the gradient for the visible standard derivation:

$$\begin{aligned} G_{\sigma_i}(\boldsymbol{v},\boldsymbol{\theta}) &= \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v};\boldsymbol{\theta}) \cdot \frac{\partial}{\partial \sigma_i}\Big(-E(\boldsymbol{v},\boldsymbol{g},\boldsymbol{\theta})\Big) \\ &\stackrel{1}{=} \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v};\boldsymbol{\theta}) \cdot \left(\frac{(v_i - a_i)^2}{\sigma_i^3} - \frac{v_i}{\sigma_i^2} \sum_j W_{ij} g_j\right) \\ &= \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v};\boldsymbol{\theta}) \frac{(v_i - a_i)^2}{\sigma_i^3} - \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v};\boldsymbol{\theta}) \frac{v_i}{\sigma_i^2} \sum_j W_{ij} g_j \\ &\stackrel{7}{=} \frac{(v_i - a_i)^2}{\sigma_i^3} \cdot \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v};\boldsymbol{\theta}) - \frac{v_i}{\sigma_i^2} \sum_j W_{ij} \cdot \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v};\boldsymbol{\theta}) g_j \\ &\stackrel{6,3}{=} \frac{(v_i - a_i)^2}{\sigma_i^3} - \frac{v_i}{\sigma_i^2} \sum_j W_{ij} \cdot p(h_j = 1|\boldsymbol{v};\boldsymbol{\theta}) \\ &\stackrel{4}{=} \frac{(v_i - a_i)^2}{\sigma_i^3} - \frac{v_i}{\sigma_i^2} \sum_j W_{ij} \cdot \overline{h}_j(\boldsymbol{v},\boldsymbol{\theta}) \end{aligned}$$

Explanation of steps marked with a number:

1. We substitute the energy function (19) and differentiate.
2. The factor $\frac{v_i}{\sigma_i}$ is the same for every $\boldsymbol{g}$, so we can pull it out of the sum.
3. We substitute (11) from Section 2.3.
4. We substitute the abbreviation (15).
5. The factor $\frac{v_i - a_i}{\sigma_i^2}$ is the same for every $\boldsymbol{g}$, so we can pull it out of the sum.
6. The sum over a probability distribution evaluates to 1 and can be dropped.
7. We pull out of the sums over $\boldsymbol{g}$ all factors that are independent of $\boldsymbol{g}$.

Inserting our results for $G_{W_{ij}}(\boldsymbol{v},\boldsymbol{\theta})$, $G_{a_i}(\boldsymbol{v},\boldsymbol{\theta})$, $G_{b_j}(\boldsymbol{v},\boldsymbol{\theta})$, $G_{\sigma_i}(\boldsymbol{v},\boldsymbol{\theta})$ into (7), we obtain:

$$\frac{\partial}{\partial W_{ij}} \log p(\boldsymbol{v};\boldsymbol{\theta}) = \frac{v_i}{\sigma_i} \cdot \overline{h}_j(\boldsymbol{v},\boldsymbol{\theta}) - \left\langle \frac{u_i}{\sigma_i} \cdot \overline{h}_j(\boldsymbol{u},\boldsymbol{\theta}) \right\rangle_{p(\boldsymbol{u};\boldsymbol{\theta})} \tag{22}$$

$$\begin{aligned}
\frac{\partial}{\partial a_i} \log p(\boldsymbol{v};\boldsymbol{\theta}) &= \frac{v_i - a_i}{\sigma_i^2} - \left\langle \frac{u_i - a_i}{\sigma_i^2} \right\rangle_{p(\boldsymbol{u};\boldsymbol{\theta})} \\
&= \frac{v_i}{\sigma_i^2} - \frac{a_i}{\sigma_i^2} - \left\langle \frac{u_i}{\sigma_i^2} - \frac{a_i}{\sigma_i^2} \right\rangle_{p(\boldsymbol{u};\boldsymbol{\theta})} \\
&= \frac{v_i}{\sigma_i^2} - \frac{a_i}{\sigma_i^2} + \frac{a_i}{\sigma_i^2} - \left\langle \frac{u_i}{\sigma_i^2} \right\rangle_{p(\boldsymbol{u};\boldsymbol{\theta})} \\
&= \frac{v_i}{\sigma_i^2} - \left\langle \frac{u_i}{\sigma_i^2} \right\rangle_{p(\boldsymbol{u};\boldsymbol{\theta})}
\end{aligned} \tag{23}$$

$$\frac{\partial}{\partial b_j} \log p(\boldsymbol{v};\boldsymbol{\theta}) = \overline{h}_j(\boldsymbol{v},\boldsymbol{\theta}) - \left\langle \overline{h}_j(\boldsymbol{u},\boldsymbol{\theta}) \right\rangle_{p(\boldsymbol{u};\boldsymbol{\theta})} \tag{24}$$

$$\begin{aligned}
\frac{\partial}{\partial \sigma_i} \log p(\boldsymbol{v};\boldsymbol{\theta}) &= \frac{(v_i - a_i)^2}{\sigma_i^3} - \frac{v_i}{\sigma_i^2} \sum_j W_{ij} \cdot \overline{h}_j(\boldsymbol{v},\boldsymbol{\theta}) \\
&\quad - \left\langle \frac{(u_i - a_i)^2}{\sigma_i^3} - \frac{u_i}{\sigma_i^2} \sum_j W_{ij} \cdot \overline{h}_j(\boldsymbol{u},\boldsymbol{\theta}) \right\rangle_{p(\boldsymbol{u};\boldsymbol{\theta})}
\end{aligned} \tag{25}$$

4 Gaussian-Bernoulli RBM (Cho's variant)

Energy function [1, p. 2]:

$$E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{\theta}) = \sum_i \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{i,j} \frac{v_i}{\sigma_i^2} W_{ij} h_j - \sum_j h_j b_j \tag{26}$$

4.1 Conditional distribution of the visible units

Note that we can obtain the energy function (26) from (19) by simply replacing W_{ij} with $\frac{1}{\sigma_i} W_{ij}$. This substitution carries through the derivation (Section 3.1) and results in:

$$p(\boldsymbol{v}|\boldsymbol{h}; \boldsymbol{\theta}) = \prod_i \mathcal{N}\Big(v_i; a_i + \sum_j W_{ij} h_j, \sigma_i^2\Big) \tag{27}$$

4.2 Conditional distribution of the hidden units

Similarly, we can substitute W_{ij} with $\frac{1}{\sigma_i} W_{ij}$ in the derivation of Section 3.2 to arrive at:

$$p(h_k = 1|\boldsymbol{v}) = \sigma\Big(b_k + \sum_i \frac{v_i}{\sigma_i^2} W_{ik}\Big) \tag{28}$$

4.3 Log likelihood gradient

The gradient derivations are very similar to Section 3.3. For the biases a_i and b_j, nothing changes. The connection weight W_{ij} and visible standard deviation σ_i gradients are affected by the slightly different $\frac{\partial}{\partial W_{ij}}\big(-E(\boldsymbol{v}, \boldsymbol{g}, \boldsymbol{\theta})\big)$ and $\frac{\partial}{\partial \sigma_i}\big(-E(\boldsymbol{v}, \boldsymbol{g}, \boldsymbol{\theta})\big)$, respectively.
In summary, we obtain:

$$\frac{\partial}{\partial W_{ij}} \log p(\boldsymbol{v}; \boldsymbol{\theta}) = \frac{v_i}{\sigma_i^2} \cdot \overline{h}_j(\boldsymbol{v}, \boldsymbol{\theta}) - \left\langle \frac{u_i}{\sigma_i^2} \cdot \overline{h}_j(\boldsymbol{u}, \boldsymbol{\theta}) \right\rangle_{p(\boldsymbol{u}; \boldsymbol{\theta})} \tag{29}$$

$$\frac{\partial}{\partial a_i} \log p(\boldsymbol{v}; \boldsymbol{\theta}) = \frac{v_i}{\sigma_i^2} - \left\langle \frac{u_i}{\sigma_i^2} \right\rangle_{p(\boldsymbol{u}; \boldsymbol{\theta})} \tag{30}$$

$$\frac{\partial}{\partial b_j} \log p(\boldsymbol{v}; \boldsymbol{\theta}) = \overline{h}_j(\boldsymbol{v}, \boldsymbol{\theta}) - \big\langle \overline{h}_j(\boldsymbol{u}, \boldsymbol{\theta}) \big\rangle_{p(\boldsymbol{u}; \boldsymbol{\theta})} \tag{31}$$

$$\begin{aligned} \frac{\partial}{\partial \sigma_i} \log p(\boldsymbol{v}; \boldsymbol{\theta}) = {} & \frac{(v_i - a_i)^2}{\sigma_i^3} - \frac{v_i}{2\sigma_i^3} \sum_j W_{ij} \cdot \overline{h}_j(\boldsymbol{v}, \boldsymbol{\theta}) \\ & - \left\langle \frac{(u_i - a_i)^2}{\sigma_i^3} - \frac{u_i}{2\sigma_i^3} \sum_j W_{ij} \cdot \overline{h}_j(\boldsymbol{u}, \boldsymbol{\theta}) \right\rangle_{p(\boldsymbol{u}; \boldsymbol{\theta})} \end{aligned} \tag{32}$$

To make learning of the visible standard deviations (or variances) more stable, Cho et al. [1, pp. 2–3] propose to reparameterize the energy function by substituting σ_i^2 with $\exp(z_i)$ and learn z_i, the logarithms of the variances. The reparameterized energy function becomes:

$$E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{\theta}) = \sum_i \frac{(v_i - a_i)^2}{2\exp(z_i)} - \sum_{i,j} \frac{v_i}{\exp(z_i)} W_{ij} h_j - \sum_j h_j b_j \tag{33}$$

The corresponding gradient becomes:

$$\begin{aligned} G_{z_i}(\boldsymbol{v}, \boldsymbol{\theta}) &= \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v}; \boldsymbol{\theta}) \cdot \frac{\partial}{\partial z_i}\Big(-E(\boldsymbol{v}, \boldsymbol{g}, \boldsymbol{\theta})\Big) \\ &= \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v}; \boldsymbol{\theta}) \cdot \left(\frac{(v_i - a_i)^2}{2\exp(z_i)} - \frac{v_i}{\exp(z_i)} \sum_j W_{ij} g_j \right) \\ &= \frac{(v_i - a_i)^2}{2\exp(z_i)} - \frac{v_i}{\exp(z_i)} \sum_j W_{ij} \cdot \overline{h}_j(\boldsymbol{v}, \boldsymbol{\theta}) \end{aligned}$$

$$\frac{\partial}{\partial z_i} \log p(\boldsymbol{v}; \boldsymbol{\theta}) = \frac{(v_i - a_i)^2}{2\exp(z_i)} - \frac{v_i}{\exp(z_i)} \sum_j W_{ij} \cdot \overline{h}_j(\boldsymbol{v}, \boldsymbol{\theta}) \\ - \left\langle \frac{(u_i - a_i)^2}{2\exp(z_i)} - \frac{u_i}{\exp(z_i)} \sum_j W_{ij} \cdot \overline{h}_j(\boldsymbol{u}, \boldsymbol{\theta}) \right\rangle_{p(\boldsymbol{u}; \boldsymbol{\theta})} \tag{34}$$

5 Gaussian-Bernoulli RBM (Goodfellow's variant)

Energy function [4, p. 8]:

$$E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{\theta}) = \sum_i \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{i,j} \frac{v_i - a_i}{\sigma_i^2} W_{ij} h_j - \sum_j h_j b_j \tag{35}$$

5.1 Conditional distribution of the visible units

Note that the energy function can be rewritten as:

$$E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{\theta}) = \sum_i \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{i,j} \frac{v_i}{\sigma_i^2} W_{ij} h_j + \sum_{i,j} \frac{a_i}{\sigma_i^2} W_{ij} h_j - \sum_j h_j b_j \tag{36}$$

Compared to (26) this just adds a term that is independent of $\boldsymbol{v}$. In the derivation of $p(\boldsymbol{v}|\boldsymbol{h};\boldsymbol{\theta})$ this term cancels just like the $\sum_j h_j b_j$ term (cf. Section 3.1), so the resulting conditional distribution is identical to Cho's variant:

$$p(\boldsymbol{v}|\boldsymbol{h};\boldsymbol{\theta}) = \prod_i \mathcal{N}\Big(v_i; a_i + \sum_j W_{ij} h_j, \sigma_i^2\Big) \tag{37}$$

5.2 Conditional distribution of the hidden units

The derivation is very similar to the one in Section 3.2:

$$\begin{aligned}
p(h_k = 1|\boldsymbol{v}) &= \frac{1}{1 + \frac{\sum_{\{\boldsymbol{g}|g_k=0\}} \exp\Big(-E(\boldsymbol{v},\boldsymbol{g},\boldsymbol{\theta})\Big)}{\sum_{\{\boldsymbol{g}|g_k=1\}} \exp\Big(-E(\boldsymbol{v},\boldsymbol{g},\boldsymbol{\theta})\Big)}} \\
&\overset{1}{=} \frac{1}{1 + \frac{\sum_{\{\boldsymbol{g}|g_k=0\}} \exp\left(\sum_{i,j} \frac{v_i - a_i}{\sigma_i^2} W_{ij} g_j - \sum_i \frac{(v_i-a_i)^2}{\sigma_i^2} + \sum_j g_j b_j\right)}{\sum_{\{\boldsymbol{g}|g_k=1\}} \exp\left(\sum_{i,j} \frac{v_i - a_i}{\sigma_i^2} W_{ij} g_j - \sum_i \frac{(v_i-a_i)^2}{\sigma_i^2} + \sum_j g_j b_j\right)}} \\
&\overset{2}{=} \frac{1}{1 + \frac{\exp\left(\sum_i \frac{v_i - a_i}{\sigma_i^2} W_{ik} \cdot 0 + 0 \cdot b_k\right)}{\exp\left(\sum_i \frac{v_i - a_i}{\sigma_i^2} W_{ik} \cdot 1 + 1 \cdot b_k\right)}} \\
&= \frac{1}{1 + \frac{1}{\exp\left(\sum_i \frac{v_i - a_i}{\sigma_i^2} W_{ik} + b_k\right)}} \\
&= \sigma\Big(b_k + \sum_i \frac{v_i - a_i}{\sigma_i^2} W_{ik}\Big)
\end{aligned} \tag{38}$$

Explanation of steps marked with a number:

1. We substitute the energy function (35).
2. We substitute the known values for $g_j = g_k$ and cancel all terms not depending on g_k.

5.3 Log likelihood gradient

Again, the gradient derivations are similar to Section 3.3. For the connection weight W_{ij}, hidden bias b_j and visible standard deviation σ_i, the gradients only change a little. For the visible bias, we will take a more detailed look at the derivation:

$$\begin{aligned}
G_{a_i}(\boldsymbol{v}, \boldsymbol{\theta}) &= \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v};\boldsymbol{\theta}) \cdot \frac{\partial}{\partial a_i}\Big(-E(\boldsymbol{v},\boldsymbol{g},\boldsymbol{\theta})\Big) \\
&\overset{1}{=} \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v};\boldsymbol{\theta}) \cdot \left(\frac{v_i - a_i}{\sigma_i^2} - \frac{1}{\sigma_i^2} \sum_j W_{ij} g_j\right)
\end{aligned}$$

$$
\begin{aligned}
&\overset{2}{=} \frac{v_i - a_i}{\sigma_i^2} \cdot \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v};\boldsymbol{\theta}) - \frac{1}{\sigma_i^2} \sum_j W_{ij} \cdot \sum_{\boldsymbol{g}} p(\boldsymbol{g}|\boldsymbol{v};\boldsymbol{\theta}) g_j \\
&\overset{3,4}{=} \frac{v_i - a_i}{\sigma_i^2} - \frac{1}{\sigma_i^2} \sum_j W_{ij} \cdot p(h_j = 1|\boldsymbol{v};\boldsymbol{\theta}) \\
&\overset{5}{=} \frac{v_i - a_i}{\sigma_i^2} - \frac{1}{\sigma_i^2} \sum_j W_{ij} \cdot \overline{h}_j(\boldsymbol{v},\boldsymbol{\theta})
\end{aligned}
$$

Explanation of steps marked with a number:

1. We substitute the energy function (35) and differentiate.
2. We pull out of the sums over $\boldsymbol{g}$ all factors that are independent of $\boldsymbol{g}$.
3. The sum over a probability distribution evaluates to 1 and can be dropped.
4. We substitute (11) from Section 2.3.
5. We substitute the abbreviation (15).

In summary, we obtain:

$$
\frac{\partial}{\partial W_{ij}} \log p(\boldsymbol{v};\boldsymbol{\theta}) = \frac{v_i - a_i}{\sigma_i^2} \cdot \overline{h}_j(\boldsymbol{v},\boldsymbol{\theta}) - \left\langle \frac{u_i - a_i}{\sigma_i^2} \cdot \overline{h}_j(\boldsymbol{u},\boldsymbol{\theta}) \right\rangle_{p(\boldsymbol{u};\boldsymbol{\theta})} \tag{39}
$$

$$
\begin{aligned}
\frac{\partial}{\partial a_i} \log p(\boldsymbol{v};\boldsymbol{\theta}) &= \frac{v_i - a_i}{\sigma_i^2} - \frac{1}{\sigma_i^2} \sum_j W_{ij} \cdot \overline{h}_j(\boldsymbol{v},\boldsymbol{\theta}) \\
&\quad - \left\langle \frac{u_i - a_i}{\sigma_i^2} - \frac{1}{\sigma_i^2} \sum_j W_{ij} \cdot \overline{h}_j(\boldsymbol{u},\boldsymbol{\theta}) \right\rangle_{p(\boldsymbol{u};\boldsymbol{\theta})} \\
&= \frac{v_i}{\sigma_i^2} - \frac{1}{\sigma_i^2} \sum_j W_{ij} \cdot \overline{h}_j(\boldsymbol{v},\boldsymbol{\theta}) \\
&\quad - \left\langle \frac{u_i}{\sigma_i^2} - \frac{1}{\sigma_i^2} \sum_j W_{ij} \cdot \overline{h}_j(\boldsymbol{u},\boldsymbol{\theta}) \right\rangle_{p(\boldsymbol{u};\boldsymbol{\theta})}
\end{aligned} \tag{40}
$$

$$
\frac{\partial}{\partial b_j} \log p(\boldsymbol{v};\boldsymbol{\theta}) = \overline{h}_j(\boldsymbol{v},\boldsymbol{\theta}) - \left\langle \overline{h}_j(\boldsymbol{u},\boldsymbol{\theta}) \right\rangle_{p(\boldsymbol{u};\boldsymbol{\theta})} \tag{41}
$$

$$
\begin{aligned}
\frac{\partial}{\partial \sigma_i} \log p(\boldsymbol{v};\boldsymbol{\theta}) &= \frac{(v_i - a_i)^2}{\sigma_i^3} - \frac{v_i - a_i}{2\sigma_i^3} \sum_j W_{ij} \cdot \overline{h}_j(\boldsymbol{v},\boldsymbol{\theta}) \\
&\quad - \left\langle \frac{(u_i - a_i)^2}{\sigma_i^3} - \frac{u_i - a_i}{2\sigma_i^3} \sum_j W_{ij} \cdot \overline{h}_j(\boldsymbol{u},\boldsymbol{\theta}) \right\rangle_{p(\boldsymbol{u};\boldsymbol{\theta})}
\end{aligned} \tag{42}
$$

As in Section 4.3, we can reparameterize σ_i^2 as $\exp(z_i)$ and learn z_i instead:

$$
\begin{aligned}
\frac{\partial}{\partial z_i} \log p(\boldsymbol{v};\boldsymbol{\theta}) &= \frac{(v_i - a_i)^2}{2\exp(z_i)} - \frac{v_i - a_i}{\exp(z_i)} \sum_j W_{ij} \cdot \overline{h}_j(\boldsymbol{v},\boldsymbol{\theta}) \\
&\quad - \left\langle \frac{(u_i - a_i)^2}{2\exp(z_i)} - \frac{u_i - a_i}{\exp(z_i)} \sum_j W_{ij} \cdot \overline{h}_j(\boldsymbol{u},\boldsymbol{\theta}) \right\rangle_{p(\boldsymbol{u};\boldsymbol{\theta})}
\end{aligned} \tag{43}
$$

An interesting aspect of this GRBM variant is that the visible bias is subtracted in all occurrences of the visible unit states, similar to the *centering trick* [6] discussed in the next section. We will see that this formulation is ill-suited for centering the hidden units, though.

6 Centering Trick

The *centering trick* [6] is a temporary reparameterization of an RBM's energy function that helps learning without changing the modelled probability distribution. It introduces a scalar offset for each visible and hidden unit that is updated to subtract its mean activation during training, and (optionally) reset to zero after training to obtain a standard uncentered RBM. In order to keep the probability distribution unchanged and not create a moving target, any changes in the offsets have to be countered by updates to other model parameters. In this section, we will show how to derive these reparameterization rules for the RBM variants considered before.

6.1 Bernoulli-Bernoulli RBM

We begin by introducing offset vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ for the visible and hidden units, respectively, resulting in the following energy function:

$$E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{\theta}) = -\sum_i (v_i - \alpha_i) a_i - \sum_{i,j} (v_i - \alpha_i) W_{ij} (h_j - \beta_j) - \sum_j (h_j - \beta_j) b_j \tag{44}$$

$$= -(\boldsymbol{v} - \boldsymbol{\alpha})^T \boldsymbol{a} - (\boldsymbol{v} - \boldsymbol{\alpha})^T \boldsymbol{W} (\boldsymbol{h} - \boldsymbol{\beta}) - (\boldsymbol{h} - \boldsymbol{\beta})^T \boldsymbol{b} \tag{45}$$

The substitution of v_i by $(v_i - \alpha_i)$ and of h_j by $(h_j - \beta_j)$ simply carries through the derivations of the conditional probabilities (9) and (10) and the log likelihood gradients in Section 2.3.
When recentering the hidden units, we set $\boldsymbol{\beta}$ to a new value $\boldsymbol{\beta'}$ (during training, this will be the mean activation of hidden units, usually computed as a running average as per line 20 of Fig. 3 in [6], and after training this may be zero to eliminate the offset terms). We want to ensure that the modelled probability distribution does not change, so we need to choose the remaining model parameters in $\boldsymbol{\theta'}$ to fulfill

$$p(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta}) = p(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta'}). \tag{46}$$

One way to satisfy this constraint is by requiring the enery function to only change by a constant term (which will drop out as a common factor in the nominator and denominator of (2)), as in [6, Sect. 3.1]:

$$E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{\theta}) = E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{\theta'}) + \text{const} \tag{47}$$

Thus, we insert (45) into (47) and solve for $\boldsymbol{a'}, \boldsymbol{b'}$. We set $\boldsymbol{\alpha'} = \boldsymbol{\alpha}$ because we need to be able to choose the visible unit offsets independently of the hidden unit offsets. We set $\boldsymbol{W'} = \boldsymbol{W}$ because we cannot use it to counter the offsets.

$$\begin{aligned}
E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{\theta}) &= E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{\theta'}) + \text{const} \\
-(\boldsymbol{v} - \boldsymbol{\alpha})^T \boldsymbol{a} - (\boldsymbol{v} - \boldsymbol{\alpha})^T \boldsymbol{W} (\boldsymbol{h} - \boldsymbol{\beta}) & \\
-(\boldsymbol{h} - \boldsymbol{\beta})^T \boldsymbol{b} &= -(\boldsymbol{v} - \boldsymbol{\alpha})^T \boldsymbol{a'} - (\boldsymbol{v} - \boldsymbol{\alpha})^T \boldsymbol{W} (\boldsymbol{h} - \boldsymbol{\beta'}) \\
& \quad -(\boldsymbol{h} - \boldsymbol{\beta'})^T \boldsymbol{b'} + \text{const} \\
(\boldsymbol{v} - \boldsymbol{\alpha})^T (\boldsymbol{a'} - \boldsymbol{a}) & \overset{1}{=} (\boldsymbol{v} - \boldsymbol{\alpha})^T \boldsymbol{W} (\boldsymbol{h} - \boldsymbol{\beta} - \boldsymbol{h} + \boldsymbol{\beta'}) \\
-(\boldsymbol{h} - \boldsymbol{\beta})^T \boldsymbol{b} & \quad -(\boldsymbol{h} - \boldsymbol{\beta'})^T \boldsymbol{b'} + \text{const} \\
\boldsymbol{v}^T (\boldsymbol{a'} - \boldsymbol{a}) - \boldsymbol{\alpha}^T (\boldsymbol{a'} - \boldsymbol{a}) & = \boldsymbol{v}^T \boldsymbol{W} (\boldsymbol{\beta'} - \boldsymbol{\beta}) - \boldsymbol{\alpha}^T \boldsymbol{W} (\boldsymbol{\beta'} - \boldsymbol{\beta}) \\
-\boldsymbol{h}^T \boldsymbol{b} + \boldsymbol{\beta}^T \boldsymbol{b} & \quad -\boldsymbol{h}^T \boldsymbol{b'} + \boldsymbol{\beta'}^T \boldsymbol{b'} + \text{const} \\
\boldsymbol{v}^T (\boldsymbol{a'} - \boldsymbol{a}) & \overset{2}{=} \boldsymbol{v}^T \boldsymbol{W} (\boldsymbol{\beta'} - \boldsymbol{\beta}) \\
-\boldsymbol{h}^T \boldsymbol{b} & \quad -\boldsymbol{h}^T \boldsymbol{b'} + \text{const} \\
\boldsymbol{a'} - \boldsymbol{a} &= \boldsymbol{W} (\boldsymbol{\beta'} - \boldsymbol{\beta}) \\
\boldsymbol{b'} &= \boldsymbol{b}
\end{aligned}$$

Explanation of steps marked with a number:

1. We add $(\boldsymbol{v} - \boldsymbol{\alpha})^T \boldsymbol{a'}$ and $(\boldsymbol{v} - \boldsymbol{\alpha})^T \boldsymbol{W} (\boldsymbol{h} - \boldsymbol{\beta})$ to both sides of the equation.
2. We absorb any terms not involving $\boldsymbol{v}$ or $\boldsymbol{h}$ in the constant term.

So we find that when we change the hidden unit offsets from $\boldsymbol{\beta}$ to $\boldsymbol{\beta'}$, we need to update the visible biases to $\boldsymbol{a'} = \boldsymbol{a} + \boldsymbol{W} (\boldsymbol{\beta'} - \boldsymbol{\beta})$ and leave the hidden unit biases unchanged – i.e., we use the visible unit biases to exactly compensate the change in top-down input caused by changing the hidden unit offsets.

An alternative way to satisfy (46) is by requiring the following set of equations to hold:

$$p(\boldsymbol{v}|\boldsymbol{h};\boldsymbol{\theta'}) = p(\boldsymbol{v}|\boldsymbol{h};\boldsymbol{\theta}) \tag{48a}$$
$$p(\boldsymbol{h}|\boldsymbol{v};\boldsymbol{\theta'}) = p(\boldsymbol{h}|\boldsymbol{v};\boldsymbol{\theta}) \tag{48b}$$

We insert the centered version of (9) into (48a) and the centered version of (10) into (48b) and solve the system of equations for $\boldsymbol{a}, \boldsymbol{b}$, again setting $\boldsymbol{W'} = \boldsymbol{W}$ and $\boldsymbol{\alpha'} = \boldsymbol{\alpha}$ beforehand.

$$\begin{cases} \forall_k \quad \sigma\Big(a'_k + \sum_j W_{kj}(h_j - \beta'_j)\Big) = \sigma\Big(a_k + \sum_j W_{kj}(h_j - \beta_j)\Big) \\ \forall_k \quad \sigma\Big(b'_k + \sum_i (v_i - \alpha_i)W_{ik}\Big) = \sigma\Big(b_k + \sum_i (v_i - \alpha_i)W_{ik}\Big) \end{cases}$$

$$\begin{cases} \forall_k \quad a'_k + \sum_j W_{kj}(h_j - \beta'_j) = a_k + \sum_j W_{kj}(h_j - \beta_j) \\ \forall_k \quad\quad\quad\quad\quad\quad\quad\quad b'_k = b_k \end{cases}$$

$$\begin{cases} \forall_k \quad a'_k = a_k + \sum_j W_{kj}(h_j - \beta_j - h_j + \beta'_j) \\ \qquad\quad\quad = a_k + \sum_j W_{kj}(\beta'_j - \beta_j) \\ \forall_k \quad b'_k = b_k \end{cases}$$

Not surprisingly, we obtain the same solution $\boldsymbol{a'} = \boldsymbol{a} + \boldsymbol{W}(\boldsymbol{\beta'} - \boldsymbol{\beta})$ and $\boldsymbol{b'} = \boldsymbol{b}$. However, for the Gaussian-Binary RBMs, this way of deriving the reparameterization rule results in far less writing, so we will use this for the remaining derivations.
When recentering the visible units, we need to set $\boldsymbol{b'} = \boldsymbol{b} + (\boldsymbol{\alpha'} - \boldsymbol{\alpha})^T\boldsymbol{W}$ and $\boldsymbol{a'} = \boldsymbol{a}$. The derivation is symmetric to the ones above. Note that when the offsets are only learned from data-dependent states as in [6] and not from model samples, $\boldsymbol{\alpha}$ can be initialized to the mean of training data and left untouched during training. For an RBM (as opposed to a Deep Boltzmann Machine), this also avoids the need of reparameterizing the hidden unit biases $\boldsymbol{b}$.

6.2 Gaussian-Bernoulli RBMs

We start with the standard Gaussian-Bernoulli RBM as used by Krizhevsky [5] and discussed in Section 3. As before, we introduce offset vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ for the visible and hidden units, respectively, obtaining the following energy function:

$$E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{\theta}) = \sum_i \frac{(v_i - a_i - \alpha_i)^2}{2\sigma_i^2} - \sum_{i,j} \frac{v_i - \alpha_i}{\sigma_i} W_{ij}(h_j - \beta_j) - \sum_j (h_j - \beta_j)b_j \tag{49}$$
$$= \frac{1}{2}(\boldsymbol{v} - \boldsymbol{a} - \boldsymbol{\alpha})^T\boldsymbol{\Lambda}(\boldsymbol{v} - \boldsymbol{a} - \boldsymbol{\alpha})^T - (\boldsymbol{v} - \boldsymbol{\alpha})^T\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{W}(\boldsymbol{h} - \boldsymbol{\beta}) - (\boldsymbol{h} - \boldsymbol{\beta})^T\boldsymbol{b}, \tag{50}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix of precisions, or inverse variances, σ_i^{-2}. The substitutions of v_i by $(v_i - \alpha_i)$ and of h_j by $(h_j - \beta_j)$ carry through the derivations in Sections 3–5 except for a minor change in $p(\boldsymbol{v}|\boldsymbol{h})$. To derive the reparameterization rule for recentering the hidden units, we insert the centered version of (20) into (48a) and the centered version of (21) into (48b) and solve the system of equations for $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\Lambda}$, setting $\boldsymbol{W'} = \boldsymbol{W}$ and $\boldsymbol{\alpha'} = \boldsymbol{\alpha}$ beforehand:

$$\begin{cases} \prod_i \mathcal{N}\Big(v_i; a'_i + \alpha_i + \sigma'_i \sum_j W_{ij}(h_j - \beta'_j), \sigma'^2_i\Big) = \prod_i \mathcal{N}\Big(v_i; a_i + \alpha_i + \sigma_i \sum_j W_{ij}(h_j - \beta_j), \sigma_i^2\Big) \\ \forall_k \quad\quad\quad\quad\quad\quad\quad\quad \sigma\Big(b'_k + \sum_i \frac{v_i - \alpha_i}{\sigma'_i} W_{ik}\Big) = \sigma\Big(b_k + \sum_i \frac{v_i - \alpha_i}{\sigma_i} W_{ik}\Big) \end{cases}$$

$$\begin{cases} \forall_i \quad\quad\quad\quad\quad\quad\quad\quad \sigma'_i = \sigma_i \\ \forall_i \quad a'_i + \alpha_i + \sigma_i \sum_j W_{ij}(h_j - \beta'_j) = a_i + \alpha_i + \sigma_i \sum_j W_{ij}(h_j - \beta_j) \\ \forall_k \quad\quad\quad\quad\quad\quad\quad\quad b'_k + \sum_i \frac{v_i - \alpha_i}{\sigma_i} W_{ik} = b_k + \sum_i \frac{v_i - \alpha_i}{\sigma_i} W_{ik} \end{cases}$$

$$\begin{cases} \forall_i & \sigma'_i = \sigma_i \\ \forall_i & a'_i = a_i + \sigma_i \sum_j W_{ij}(h_j - \beta_j - h_j + \beta'_j) \\ & \quad = a_i + \sigma_i \sum_j W_{ij}(\beta'_j - \beta_j) \\ \forall_k & b'_k = b_k \end{cases}$$

We see that the hidden biases and the visible standard deviations stay unmodified, and we need to set $\boldsymbol{a'} = \boldsymbol{a} + \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{W}(\boldsymbol{\beta'} - \boldsymbol{\beta})$ to counter the change of hidden offsets.
For recentering the visible units, we solve the same system of equations for $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\Lambda}$, this time setting $\boldsymbol{W'} = \boldsymbol{W}$ and $\boldsymbol{\beta'} = \boldsymbol{\beta}$ beforehand:

$$\begin{cases} \prod_i \mathcal{N}\Big(v_i; a'_i + \alpha'_i + \sigma'_i \sum_j W_{ij}(h_j - \beta_j), \sigma'^2_i\Big) = \prod_i \mathcal{N}\Big(v_i; a_i + \alpha_i + \sigma_i \sum_j W_{ij}(h_j - \beta_j), \sigma_i^2\Big) \\ \forall_k \quad \sigma\Big(b'_k + \sum_i \frac{v_i - \alpha'_i}{\sigma'_i} W_{ik}\Big) = \sigma\Big(b_k + \sum_i \frac{v_i - \alpha_i}{\sigma_i} W_{ik}\Big) \end{cases}$$

$$\begin{cases} \forall_i & \sigma'_i = \sigma_i \\ \forall_i & a'_i + \alpha'_i + \sigma_i \sum_j W_{ij}(h_j - \beta_j) = a_i + \alpha_i + \sigma_i \sum_j W_{ij}(h_j - \beta_j) \\ \forall_k & b'_k + \sum_i \frac{v_i - \alpha'_i}{\sigma_i} W_{ik} = b_k + \sum_i \frac{v_i - \alpha_i}{\sigma_i} W_{ik} \end{cases}$$

$$\begin{cases} \forall_i & \sigma'_i = \sigma_i \\ \forall_i & a'_i = a_i + \alpha_i - \alpha'_i \\ \forall_k & b'_k = b_k + \sum_i \frac{v_i - \alpha_i - v_i + \alpha'_i}{\sigma_i} W_{ik} \\ & \quad = b_k + \sum_i \frac{v_i + \alpha'_i - \alpha_i}{\sigma_i} W_{ik} \end{cases}$$

Here, both the visible and hidden unit biases have to be adapted. When changing the visible unit offsets from $\boldsymbol{\alpha}$ to $\boldsymbol{\alpha'}$, we have to counter that by setting $\boldsymbol{a'} = \boldsymbol{a} + \boldsymbol{\alpha} - \boldsymbol{\alpha'}$ and $\boldsymbol{b'} = \boldsymbol{b} + (\boldsymbol{\alpha'} - \boldsymbol{\alpha})^T \boldsymbol{\Lambda}^{\frac{1}{2}} \boldsymbol{W}$.

For Cho's variant of the Gaussian-Bernoulli RBM discussed in Section 4, nothing much changes. As noted in Section 4.1, its energy function can be derived from Krizhevsky's variant by simply replacing W_{ij} with $\frac{1}{\sigma_i} W_{ij}$ (or $\boldsymbol{W}$ with $\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{W}$), resulting in:

$$E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{\theta}) = \sum_i \frac{(v_i - a_i - \alpha_i)^2}{2\sigma_i^2} - \sum_{i,j} \frac{v_i - \alpha_i}{\sigma_i^2} W_{ij}(h_j - \beta_j) - \sum_j (h_j - \beta_j) b_j \tag{51}$$

$$= \frac{1}{2}(\boldsymbol{v} - \boldsymbol{a} - \boldsymbol{\alpha})^T \boldsymbol{\Lambda} (\boldsymbol{v} - \boldsymbol{a} - \boldsymbol{\alpha})^T - (\boldsymbol{v} - \boldsymbol{\alpha})^T \boldsymbol{\Lambda} \boldsymbol{W} (\boldsymbol{h} - \boldsymbol{\beta}) - (\boldsymbol{h} - \boldsymbol{\beta})^T \boldsymbol{b} \tag{52}$$

This substitution carries through the derivation of the reparameterization rule: When updating the visible and hidden offsets from $\boldsymbol{\alpha}$ to $\boldsymbol{\alpha'}$ and $\boldsymbol{\beta}$ to $\boldsymbol{\beta'}$, the biases need to be set according to $\boldsymbol{a'} = \boldsymbol{a} + \boldsymbol{\alpha} - \boldsymbol{\alpha'} + \boldsymbol{W}(\boldsymbol{\beta'} - \boldsymbol{\beta})$ and $\boldsymbol{b'} = \boldsymbol{b} + (\boldsymbol{\alpha'} - \boldsymbol{\alpha})^T \boldsymbol{\Lambda} \boldsymbol{W}$.

For Goodfellow's variant discussed in Section 5, the centered energy function would be:

$$E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{\theta}) = \sum_i \frac{(v_i - a_i - \alpha_i)^2}{2\sigma_i^2} - \sum_{i,j} \frac{v_i - a_i - \alpha_i}{\sigma_i^2} W_{ij}(h_j - \beta_j) - \sum_j (h_j - \beta_j) b_j \tag{53}$$

$$= \frac{1}{2}(\boldsymbol{v} - \boldsymbol{a} - \boldsymbol{\alpha})^T \boldsymbol{\Lambda} (\boldsymbol{v} - \boldsymbol{a} - \boldsymbol{\alpha})^T - (\boldsymbol{v} - \boldsymbol{a} - \boldsymbol{\alpha})^T \boldsymbol{\Lambda} \boldsymbol{W} (\boldsymbol{h} - \boldsymbol{\beta}) - (\boldsymbol{h} - \boldsymbol{\beta})^T \boldsymbol{b} \tag{54}$$

Note that the visible bias $\boldsymbol{a}$ and offset $\boldsymbol{\alpha}$ assume the same role, so they could be merged. We will leave them separated for clarity. As before, we derive the reparameterization rule for updating the hidden

offsets by inserting the centered versions of (37) and (38) into (48), setting $\boldsymbol{W'} = \boldsymbol{W}$, $\boldsymbol{\alpha'} = \boldsymbol{\alpha}$ and solving for $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\Lambda}$:

$$\begin{cases} \displaystyle\prod_i \mathcal{N}\Big(v_i; a'_i + \alpha_i + \sum_j W_{ij}(h_j - \beta'_j), \sigma'^2_i\Big) = \prod_i \mathcal{N}\Big(v_i; a_i + \alpha_i + \sum_j W_{ij}(h_j - \beta_j), \sigma_i^2\Big) \\ \forall_k \quad \sigma\Big(b'_k + \displaystyle\sum_i \frac{v_i - a'_i - \alpha_i}{\sigma'^2_i} W_{ik}\Big) = \sigma\Big(b_k + \sum_i \frac{v_i - a_i - \alpha_i}{\sigma_i^2} W_{ik}\Big) \end{cases}$$

$$\begin{cases} \forall_i \quad \sigma'_i = \sigma_i \\ \forall_i \quad a'_i + \alpha_i + \displaystyle\sum_j W_{ij}(h_j - \beta'_j) = a_i + \alpha_i + \sum_j W_{ij}(h_j - \beta_j) \\ \forall_k \quad b'_k + \displaystyle\sum_i \frac{v_i - a'_i - \alpha_i}{\sigma_i^2} W_{ik} = b_k + \sum_i \frac{v_i - a_i - \alpha_i}{\sigma_i^2} W_{ik} \end{cases}$$

$$\begin{cases} \forall_i \quad \sigma'_i = \sigma_i \\ \forall_i \quad a'_i = a_i + \displaystyle\sum_j W_{ij}(\beta'_j - \beta_j) \\ \forall_k \quad b'_k = b_k + \displaystyle\sum_i \frac{a'_i - a_i}{\sigma_i^2} W_{ik} \\ \qquad\quad = b_k + \displaystyle\sum_i \frac{\sum_j W_{ij}(\beta'_j - \beta_j)}{\sigma_i^2} W_{ik} \end{cases}$$

That is, when changing the hidden offsets from $\boldsymbol{\beta}$ to $\boldsymbol{\beta'}$, we need to set $\boldsymbol{a'} = \boldsymbol{a} + \boldsymbol{W}(\boldsymbol{\beta'} - \boldsymbol{\beta})$ and $\boldsymbol{b'} = \boldsymbol{b} + (\boldsymbol{a} - \boldsymbol{a'})^T \boldsymbol{\Lambda} \boldsymbol{W} = \boldsymbol{b} + (\boldsymbol{W}(\boldsymbol{\beta'} - \boldsymbol{\beta}))^T \boldsymbol{\Lambda} \boldsymbol{W} = \boldsymbol{b} + (\boldsymbol{\beta'} - \boldsymbol{\beta})^T \boldsymbol{W}^T \boldsymbol{\Lambda} \boldsymbol{W}$.
For recentering the visible units, we solve the same system of equations for $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\Lambda}$, setting $\boldsymbol{W'} = \boldsymbol{W}$ and $\boldsymbol{\beta'} = \boldsymbol{\beta}$ beforehand:

$$\begin{cases} \displaystyle\prod_i \mathcal{N}\Big(v_i; a'_i + \alpha'_i + \sum_j W_{ij}(h_j - \beta_j), \sigma'^2_i\Big) = \prod_i \mathcal{N}\Big(v_i; a_i + \alpha_i + \sum_j W_{ij}(h_j - \beta_j), \sigma_i^2\Big) \\ \forall_k \quad \sigma\Big(b'_k + \displaystyle\sum_i \frac{v_i - a'_i - \alpha'_i}{\sigma'^2_i} W_{ik}\Big) = \sigma\Big(b_k + \sum_i \frac{v_i - a_i - \alpha_i}{\sigma_i^2} W_{ik}\Big) \end{cases}$$

$$\begin{cases} \forall_i \quad \sigma'_i = \sigma_i \\ \forall_i \quad a'_i + \alpha'_i + \displaystyle\sum_j W_{ij}(h_j - \beta_j) = a_i + \alpha_i + \sum_j W_{ij}(h_j - \beta_j) \\ \forall_k \quad b'_k + \displaystyle\sum_i \frac{v_i - a'_i - \alpha'_i}{\sigma_i^2} W_{ik} = b_k + \sum_i \frac{v_i - a_i - \alpha_i}{\sigma_i^2} W_{ik} \end{cases}$$

$$\begin{cases} \forall_i \quad \sigma'_i = \sigma_i \\ \forall_i \quad a'_i = a_i + \alpha_i - \alpha'_i \\ \forall_k \quad b'_k = b_k + \displaystyle\sum_i \frac{a_i - a'_i + \alpha_i - \alpha'_i}{\sigma_i^2} W_{ik} \\ \qquad\quad = b_k + \displaystyle\sum_i \frac{a_i - a_i - \alpha_i + \alpha'_i + \alpha_i - \alpha'_i}{\sigma_i^2} W_{ik} \\ \qquad\quad = b_k \end{cases}$$

That is, when changing the visible offsets from $\boldsymbol{\alpha}$ to $\boldsymbol{\alpha'}$, we need to set $\boldsymbol{a'} = \boldsymbol{\alpha} - \boldsymbol{\alpha'}$ and $\boldsymbol{b'} = \boldsymbol{b}$. This is of course a much simpler reparameterization rule than for the other two Gaussian-Bernoulli RBM variants, but given that the visible offsets $\boldsymbol{\alpha}$ are typically set in advance and not changed during training, this is only a virtual advantage and does not outweight the disadvantage of the much more involved reparameterization rule for the hidden offsets derived before.

Acknowledgements

Thanks to Alex Krizhevsky and Asja Fischer et al. for publishing their respective derivations [5, 3], and thanks to Karen Ullrich for helping to get the derivation in 3.1 straight.

References

[1] KyungHyun Cho, Alexander Ilin, and Tapani Raiko. Improved learning of gaussian-bernoulli restricted boltzmann machines. In *Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN)*, pages 10–17, Espoo, Finland, 2011.

[2] Deep learning tutorials: Restricted Boltzmann Machines (RBM). `http://deeplearning.net/tutorial/rbm.html`, January 2014.

[3] Asja Fischer and Christian Igel. An introduction to restricted boltzmann machines. In *Proceedings of the 17th Iberoamerican Congress on Pattern Recognition (CIARP)*, Buenos Aires, Argentinia, 2012.

[4] Ian Goodfellow, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Multi-prediction deep boltzmann machines. In *Advances in Neural Information Processing Systems (NIPS) 26*, pages 548–556. 2013.

[5] Alex Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, Dept. of Comp. Science, Univ. of Toronto, 2009.

[6] Grégoire Montavon and Klaus-Robert Müller. Deep boltzmann machines and the centering trick. In Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade*, volume 7700 of *Lecture Notes in Computer Science*, pages 621–637. Springer Berlin Heidelberg, 2012.

[7] Jan Schlüter. Unsupervised Audio Feature Extraction for Music Similarity Estimation. Master's thesis, Technische Universität München, Munich, Germany, October 2011.