Exploring Helmholtz Machine and Deep Belief Net in the Exponential Family Perspective

Conference Paper · July 2018

CITATION
1

READS
96

2 authors, including:

Yifeng Li
National Research Council Canada
64 PUBLICATIONS **460** CITATIONS

SEE PROFILE

Exploring Helmholtz Machine and Deep Belief Net in the Exponential Family Perspective

Yifeng Li [1] **Xiaodan Zhu** [2]

Abstract

Applications of directed deep generative models are still limited by the data types they can model. Here, we formulate the free energy function of exponential family restricted Boltzmann machine, extend the binary Helmholtz machine and deep belief network to the exponential family, and derive the corresponding wake-sleep learning algorithms. We demonstrate that appealing performance can be achieved by the generalized models.

1. Introduction

Generative models with latent variables concern joint distributions of observable measurements and hidden factors. While generative models, such as restricted Boltzmann machine (RBM) (Smolensky, 1986), are often employed to train deep discriminative models (Bengio et al., 2006), they also allow for addressing many more interesting problems. For example, sampling joint distribution $p(\boldsymbol{x}, \boldsymbol{h})$ to generate novel data could be useful for a wide variety of problems that may even reach some hard AI tasks, e.g., those involving creativity. A joint distribution also allows us to see association of variables as studied in causality. From the conditional $p(\boldsymbol{h}|\boldsymbol{x})$, states of hidden variables can be inferred, which are often used in dimensionality reduction, missing value estimation, prediction, and clustering. From $p(\boldsymbol{x}|\boldsymbol{h})$, effects of hidden factors can be evaluated.

Deep generative neural networks inherit this potential versatility and have the distinctive capacity of modeling phenomenons and systems of high complexity. For example, undirected generative models, e.g., RBM, have been extended to deep networks, e.g., deep Boltzmann machine (DBM) (Salakhutdinov & Hinton, 2009a) and multi-modal DBM (MDBM) (Srivastava & Salakhutdinov, 2014), to capture the complex nature of data and association of heterogeneous modalities. Also, directed deep generative models (DGMs), e.g., Helmholtz machine (HM) (Dayan et al., 1995) and deep belief net (DBN) (Hinton et al., 2006; Srivastava & Salakhutdinov, 2012), are used to model complicated (hierarchal or sequential) structures of hidden variables. These DGMs, however, face two major challenges: intractable inference [$p(\boldsymbol{h}|\boldsymbol{x})$] and limited forms of distributions. In undirected DGMs, hidden states could be approximated using variational approximation inference. In directed DGMs, hidden states may be obtained through recognition connections. Nevertheless, applications of existing DGMs are, to a large extent, restricted by the data types they can fit. For example, DBN is often applied to pretrain feedforward networks for classification problems when data are binary or can be scaled to the range of [0,1]. Generalizing DGMs to a larger pool of distributions in a united framework is of interest. Thanks to the exponential family RBM (exp-RBM) (Welling et al., 2005), the extension of binary DBM to the exponential family is straightforward. However, the generalization of binary wake-sleep algorithms (Hinton et al., 1995; 2006) for exponential family HM and DBN is nontrivial and remains a puzzle, thus is revisited here.

In summary, this paper addresses three issues: how to generically define deep generative models? how to derive the stochastic gradients in exponential family wake-sleep algorithms? and how to estimate their (lower bounds of) log-likelihoods? The major contributions of this work are threefold: (1) we discover that the free energy function of exp-RBM and the partition function of its base-rate model can be analytically computed, which enables the estimation of partition functions and (lower bounds of) likelihoods of DGMs; (2) inspired by the concepts of exp-RBMs, we formulate the exponential family HM and DBN (exp-HM and exp-DBN); and (3) we derive the update rules in the wake-sleep algorithms by virtue of an important property of exponential family distributions.

2. Exponential Family RBM

In addition to pretraining DGMs, the conditional distributions of RBM also play an important role in the modelling

[1]Digital Technologies Research Centre, National Research Council Canada, Ottawa, Ontario, Canada [2]Department of Electrical and Computer Engineering, Queen's University, Kingston, Ontario, Canada. Correspondence to: Yifeng Li <yifeng.li@nrc-cnrc.gc.ca>.

Presented at the ICML 2018 workshop on Theoretical Foundations and Applications of Deep Generative Models, Stockholm, Sweden, 2018.

of directed deep generative models. Prior to presenting our generalization, we briefly discuss key concepts of exponential family RBM, where we propose that (i) the free energy function can be formulated analytically and (ii) the log-partition function of the base-rate model can be computed analytically, enabling the estimation of partition function in exp-RBM using annealed importance sampling (AIS).

The exponential family includes univariate or multivariate distributions with the following natural parametric form:

$$p(\boldsymbol{x}) = h(\boldsymbol{x})e^{\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{s}(\boldsymbol{x})-A(\boldsymbol{\theta})}, \tag{1}$$

where $\boldsymbol{x}$ is either a univariate or multivariate random variable, $\boldsymbol{\theta}$ is a vector (or a scalar) of natural parameters, $\boldsymbol{s}(\boldsymbol{x})$ is a vector (or a scalar) of sufficient statistics, $A(\boldsymbol{\theta})$ is the log-partition function, and $h(\boldsymbol{x})$ is the base measure. It has several interesting properties. First, if Eq. (1) is rewritten to $p(\boldsymbol{x}) = \frac{1}{Z}e^{\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{s}(\boldsymbol{x})+\log h(\boldsymbol{x})}$, one will find that $\log Z = A(\boldsymbol{\theta})$, i.e. $A(\boldsymbol{\theta})$ normalizes the distribution:

$$A(\boldsymbol{\theta}) = \log\Big[\sum\!\!\!\!\!\!\int_{\boldsymbol{x}} h(\boldsymbol{x})e^{\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{s}(\boldsymbol{x})}\Big]. \tag{2}$$

This property about $A(\boldsymbol{\theta})$ is extremely useful in the estimation of the partition functions of exp-RBMs. Second, the first and second order derivatives of $A(\boldsymbol{\theta})$ equal to the mean and covariance of $\boldsymbol{s}(\boldsymbol{x})$, respectively:

$$\frac{\partial A(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathrm{E}_{p(\boldsymbol{x})}[\boldsymbol{s}(\boldsymbol{x})]; \tag{3}$$

$$\frac{\partial^2 A(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^{\mathrm{T}}} = \mathrm{Cov}_{p(\boldsymbol{x})}[\boldsymbol{s}(\boldsymbol{x})]. \tag{4}$$

$A(\boldsymbol{\theta})$ is convex, as the covariance matrix is positive semi-definite. Eq. (3) is extremely useful in deriving stochastic update rules for exp-HMs (see Section 3.2) and exp-DBNs. For the understanding of this article, we selectively provide the natural and standard forms of a range of univariate member of exponential family in Supplemental Table 1.

Exp-RBM (Welling et al., 2005) is generally formulated as

$$p(\boldsymbol{x}, \boldsymbol{h}) = \frac{1}{Z}e^{-E(\boldsymbol{x},\boldsymbol{h})}, \tag{5}$$

where $\boldsymbol{x} \in \mathbb{R}^M$ is the vector of visible random variables, $\boldsymbol{h} \in \mathbb{R}^K$ the vector of hidden random variables, $E(\boldsymbol{x}, \boldsymbol{h})$ the energy function, and $Z = \sum\!\!\!\!\!\!\int_{\boldsymbol{x}'} \sum\!\!\!\!\!\!\int_{\boldsymbol{h}'} e^{-E(\boldsymbol{x}',\boldsymbol{h}')}$ the partition function. The design of an exp-RBM requires three key components: (1) base distributions of $\boldsymbol{x}$ and $\boldsymbol{h}$, (2) energy function $E(\boldsymbol{x}, \boldsymbol{h})$, and (3) conditional distributions $p(\boldsymbol{x}|\boldsymbol{h})$ and $p(\boldsymbol{h}|\boldsymbol{x})$. They are elaborated below.

First, the base distributions for both $\boldsymbol{x}$ and $\boldsymbol{h}$ in the natural forms of exponential class can be defined as:

$$\begin{cases} p(\boldsymbol{x}) = \prod_{m=1}^{M} e^{\boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{s}_m+\log f_m(x_m)-A_m(\boldsymbol{a}_m)} & (6) \\ p(\boldsymbol{h}) = \prod_{k=1}^{K} e^{\boldsymbol{b}_k^{\mathrm{T}}\boldsymbol{t}_k+\log g_k(h_k)-B_k(\boldsymbol{b}_k)}. & (7) \end{cases}$$

Here, $\boldsymbol{a}_m$ and $\boldsymbol{b}_k$ are vectors/scalars of natural parameters for x_m and h_k, respectively; $\boldsymbol{s}_m$ and $\boldsymbol{t}_k$ are the corresponding vectors/scalars of sufficient statistics; $A_m(\boldsymbol{a}_m)$ and $B_k(\boldsymbol{b}_k)$ are the corresponding log-partition functions; $f_m(x_m)$ and $g_k(h_k)$ are the corresponding base measures. For example, if $p(\boldsymbol{x})$ is Gaussian with unknown mean and precision, then $\boldsymbol{a}_m = [a_m^{(1)}, a_m^{(2)}]^{\mathrm{T}} = [\mu_m\beta_m, -\frac{\beta_m}{2}]^{\mathrm{T}}$ and $\boldsymbol{s}_m = [s_m^{(1)}, s_m^{(2)}]^{\mathrm{T}} = [x_m, x_m^2]^{\mathrm{T}}$. If $p(\boldsymbol{h})$ is Bernoulli, then $b_k = \log\frac{p_k}{1-p_k}$ and $t_k = h_k$ are scalars.

Second, the energy function is defined via combining x_m- and h_k-related terms in Eq. (6) and (7) (including base measures), and enabling interactions between $\boldsymbol{s}_m$ and $\boldsymbol{t}_k$:

$$E(\boldsymbol{x}, \boldsymbol{h}) = -\sum_{m=1}^{M}\left(\boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{s}_m + \log f_m(x_m)\right) - \sum_{k=1}^{K}\left(\boldsymbol{b}_k^{\mathrm{T}}\boldsymbol{t}_k + \log g_k(h_k)\right) - \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{r=1}^{R}\sum_{u=1}^{U} s_{m,r}w_{m,k,r,u}t_{k,u}, \tag{8}$$

where R and U are the numbers of natural parameters for $p(x_m)$ and $p(h_k)$. The interaction strengths are represented by a tensor $\boldsymbol{W} \in \mathbb{R}^{M\times K\times R\times U}$. For Gaussian-Bernoulli RBMs (exp-RBMs with Gaussian visible and Bernoulli hidden variables), the weight tensor is of size $M \times K \times 2$, including $\boldsymbol{W}^{(1)}$ for interactions between $\boldsymbol{x}$ and $\boldsymbol{h}$, and $\boldsymbol{W}^{(2)}$ for $\boldsymbol{x}^{*2}$ and $\boldsymbol{h}$. To reduce model complexity, we may consider dropping off some interactions between the sufficient statistics. For instance, in Gaussian-Bernoulli RBMs, we may decide to disregard $\boldsymbol{W}^{(2)}$, reducing the weights in a tensor to a matrix of size $M \times K$. In fact, we only considered interactions between $\boldsymbol{x}$ and $\boldsymbol{h}$ in practice, reducing the interaction term to $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{h}$.

Third, from the energy function in Eq. (8) and the definition of exponential family in Eq. (1), $p(\boldsymbol{x}|\boldsymbol{h})$ can be derived as

$$p(\boldsymbol{x}|\boldsymbol{h}) = \frac{p(\boldsymbol{x}, \boldsymbol{h})}{p(\boldsymbol{h})} = e^{\sum_{m=1}^{M}\left((\hat{\boldsymbol{a}}_m)^{\mathrm{T}}\boldsymbol{s}_m+\log f_m(x_m)-A_m(\hat{\boldsymbol{a}}_m)\right)} = \prod_{m=1}^{M} p(x_m|\boldsymbol{h}, \boldsymbol{\eta}(\hat{\boldsymbol{a}}_m)), \tag{9}$$

where $\hat{\boldsymbol{a}}_m = \boldsymbol{a}_m + \sum_{k=1}^{K}\boldsymbol{W}_{m,k,:,:}\boldsymbol{t}_k$, and function $\boldsymbol{\eta}(\hat{\boldsymbol{a}}_m)$ maps the natural parameters in $\hat{\boldsymbol{a}}_m$ to the standard forms. As shown in Supplemental Table 1, if $p(x_m|\boldsymbol{h})$ is Gaussian with unknown mean and precision, we then have $\boldsymbol{\eta}(\hat{\boldsymbol{a}}_m) = [\hat{\mu}_m, \hat{\beta}_m]^{\mathrm{T}} = [-\frac{\hat{a}_m^{(1)}}{2\hat{a}_m^{(2)}}, -2\hat{a}_m^{(2)}]^{\mathrm{T}}$. Similarly, we can obtain

$$p(\boldsymbol{h}|\boldsymbol{x}) = \prod_{k=1}^{K} p(h_k|\boldsymbol{x}, \boldsymbol{\eta}(\hat{\boldsymbol{b}}_k)), \tag{10}$$

where $\hat{\boldsymbol{b}}_k = \boldsymbol{b}_k + \sum_{m=1}^{M}(\boldsymbol{W}_{m,k,:,:})^{\mathrm{T}}\boldsymbol{s}_m$ and $\boldsymbol{\eta}(\hat{\boldsymbol{b}}_k)$ maps the natural parameters in $\hat{\boldsymbol{b}}_k$ to the standard ones. For example, if $p(h_k|\boldsymbol{x})$ is Bernoulli, we have $\eta(\hat{b}_k) = \hat{p}_k = \sigma(\hat{b}_k)$. Importantly, the conditionals are decomposable and follow the same distributions as the bases, but use posterior parameters to reflect influence of their heterogeneous counterparts.

Stochastic gradient descent can be used to estimate exp-RBMs' parameters which can be represented by $\boldsymbol{\theta} = \{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{W}\}$ where $\boldsymbol{a} = \{\boldsymbol{a}^{(1)}, \cdots, \boldsymbol{a}^{(R)}\}$ and $\boldsymbol{b} = \{\boldsymbol{b}^{(1)}, \cdots, \boldsymbol{b}^{(U)}\}$ may respectively include several bias vectors, and $\boldsymbol{W} = \{\boldsymbol{W}^{(1,1)}, \cdots, \boldsymbol{W}^{(r,u)}, \cdots, \boldsymbol{W}^{(R,U)}\}$ (but only interactions between $\boldsymbol{x}$ and $\boldsymbol{h}$ are considered in practice). The likelihood $p(\boldsymbol{x})$ can obtained by marginalizing out $\boldsymbol{h}$:

$$p(\boldsymbol{x}) = \sum\!\!\!\!\!\!\int_{\boldsymbol{h}'} p(\boldsymbol{x}, \boldsymbol{h}') = \frac{1}{Z} e^{-F(\boldsymbol{x})}, \tag{11}$$

where $F(\boldsymbol{x}) = -\log \sum\!\!\!\!\!\!\int_{\boldsymbol{h}'} e^{-E(\boldsymbol{x}, \boldsymbol{h}')}$ is the free energy function. The model parameters are estimated by maximizing

$$\log p(\boldsymbol{x}) = -F(\boldsymbol{x}) - \log Z. \tag{12}$$

Given N samples, the gradient is computed as

$$\begin{aligned}\Delta_{\boldsymbol{\theta}} &= \frac{1}{N}\sum_{n=1}^{N}\Big(\frac{\partial F(\boldsymbol{x}_n)}{\partial \boldsymbol{\theta}} - \mathrm{E}_{p(\boldsymbol{x})}\Big[\frac{\partial F(\boldsymbol{x})}{\partial \boldsymbol{\theta}}\Big]\Big) \\ &= \frac{1}{N}\sum_{n=1}^{N}\Big(\mathrm{E}_{p(\boldsymbol{h}|\boldsymbol{x}_n)}\Big[\frac{\partial E(\boldsymbol{x}_n, \boldsymbol{h})}{\partial \boldsymbol{\theta}}\Big] - \mathrm{E}_{p(\boldsymbol{x}, \boldsymbol{h})}\Big[\frac{\partial E(\boldsymbol{x}, \boldsymbol{h})}{\partial \boldsymbol{\theta}}\Big]\Big),\end{aligned} \tag{13}$$

where $\frac{\partial E(\boldsymbol{x}, \boldsymbol{h})}{\partial \boldsymbol{\theta}}$ w.r.t. $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{W}$, respectively, are:

$$\frac{\partial E(\boldsymbol{x}, \boldsymbol{h})}{\partial a_m^{(r)}} = -s_m^{(r)}; \frac{\partial E(\boldsymbol{x}, \boldsymbol{h})}{\partial b_k^{(u)}} = -t_k^{(u)}; \frac{\partial E(\boldsymbol{x}, \boldsymbol{h})}{\partial w_{m,k}^{(r,u)}} = -s_m^{(r)} t_k^{(u)}. \tag{14}$$

The log-likelihood in Eq. (12) may serve as a measure of learning quality, where $\log Z$ must be estimated and we propose that $F(\boldsymbol{x})$ can be computed analytically:

$$F(\boldsymbol{x}) = -\zeta(\boldsymbol{x}) - \sum_{k=1}^{K} \log \sum\!\!\!\!\!\!\int_{h_k} e^{-\gamma_k(\boldsymbol{x}, h_k)} = -\zeta(\boldsymbol{x}) - \sum_{k=1}^{K} B_k(\hat{\boldsymbol{b}}_k), \tag{15}$$

where $\hat{\boldsymbol{b}}_k$ and $B_k(\cdot)$ are respectively the posterior parameter and log-partition function of a hidden variable distribution.

Using the free energy function, the AIS procedure for binary RBM (Salakhutdinov & Murray, 2008) can be generalized to estimate the log-partition function of exp-RBM, where the intermediate distribution for annealing can be defined as

$$p_t(\boldsymbol{x}, \boldsymbol{h}) = \frac{1}{Z_t} e^{(1-\beta_t)E_A(\boldsymbol{x}, \boldsymbol{h}^{(A)}) - \beta_t E_B(\boldsymbol{x}, \boldsymbol{h}^{(B)})}, \tag{16}$$

where β_t gradually changes from 0 to 1: $0 = \beta_0 < \cdots < \beta_t < \cdots < \beta_T = 1$, so that $p_0(\boldsymbol{x}, \boldsymbol{h})$ corresponds to the base-rate model A and $p_T(\boldsymbol{x}, \boldsymbol{h})$ corresponds to the focused model B. The intermediate marginals can be defined as:

$$p_t(\boldsymbol{x}) = \frac{1}{Z_t} e^{-F_t(\boldsymbol{x})} = \frac{1}{Z_t} p_t^*(\boldsymbol{x}), \tag{17}$$

where $p_t^*(\boldsymbol{x}) = e^{-F_t(\boldsymbol{x})}$ and the free energy function is

$$\begin{aligned}F_t(\boldsymbol{x}) = &-(1-\beta_t)\zeta^{(A)}(\boldsymbol{x}) - \sum_{k=1}^{K_A} B_k\big((1-\beta_t)\hat{\boldsymbol{b}}_k^{(A)}\big) \\ &- \beta_t \zeta^{(B)}(\boldsymbol{x}) - \sum_{k=1}^{K_B} B_k\big(\beta_t \hat{\boldsymbol{b}}_k^{(B)}\big).\end{aligned} \tag{18}$$

If model B is $\mathrm{RBM}_B(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{W})$, model A is $\mathrm{RBM}_A(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{0})$, we unveil that $\log Z_A$ can be analytically computed as

$$\log Z_A = \sum_{m=1}^{M} A_m(\boldsymbol{a}_m) + \sum_{k=1}^{K} B_k(\boldsymbol{b}_k). \tag{19}$$

where $A_m(\boldsymbol{a}_m)$ and $B_k(\boldsymbol{b}_k)$ are the log-partition functions of x_m and h_k, respectively.

The Markov transition $T(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ can be defined as

$$\begin{cases} p_t\big((x_t)_m|\boldsymbol{h}_{t-1}^{(A)}, \boldsymbol{h}_{t-1}^{(B)}\big) = p_{\text{vis}}\big(x|\boldsymbol{\eta}((1-\beta_t)\boldsymbol{a}_m + \beta_t\hat{\boldsymbol{a}}_m)\big) \\ p_t\big((h_t^{(A)})_k|\boldsymbol{x}_t\big) = p_{\text{hid}}\big(h|\boldsymbol{\eta}((1-\beta_t)\boldsymbol{b}_k)\big) \\ p_t\big((h_t^{(B)})_k|\boldsymbol{x}_t\big) = p_{\text{hid}}\big(h|\boldsymbol{\eta}(\beta_t\hat{\boldsymbol{b}}_k)\big). \end{cases}$$

Then, we can sample a sequence: $\{\boldsymbol{x}_t, \boldsymbol{h}_t^{(A)}, \boldsymbol{h}_t^{(B)}\}$ ($0 \le t \le T$-1). Hence, the importance weight is computed by

$$w_s = \frac{p_1^*(\boldsymbol{x}_0)}{p_0^*(\boldsymbol{x}_0)} \cdots \frac{p_T^*(\boldsymbol{x}_{T-1})}{p_{T-1}^*(\boldsymbol{x}_{T-1})} = \prod_{t=0}^{T-1} \frac{p_{t+1}^*(\boldsymbol{x}_t)}{p_t^*(\boldsymbol{x}_t)}. \tag{20}$$

Thus, the ratio of partition function can be estimated by $\frac{Z_B}{Z_A} \approx w_s$. Using a logarithm version to increase stability (Neal, 2001), we estimate $\log Z_B$ by

$$\log \hat{Z}_B = \log w_s + \log Z_A, \tag{21}$$

where $\log w_s = \sum_{t=0}^{T-1}\big(F_t(\boldsymbol{x}_t) - F_{t+1}(\boldsymbol{x}_t)\big)$ and $\log Z_A$ is computed using Eq. (19). These AIS components for exp-RBMs are summarized in Supplemental Table 3.

3. Exponential Family Helmholtz Machine

3.1. Generic Formulation of Exp-HM

A Helmholtz machine (HM) (Dayan et al., 1995) tackles the intractable inference problem $p(\boldsymbol{h}|\boldsymbol{x})$ raised in logistic belief nets (Neal, 1992) by introducing corresponding recognition connections to form an efficient approximate posterior distribution $q(\boldsymbol{h}|\boldsymbol{x})$. Thus, HM is a multilayer stochastic neural network with both top-down generative connections and bottom-up recognition connections. However, the original HM model only considers binary visible and hidden variables. Here, we extend it to the exponential family HM (exp-HM), so that a range of discrete and continuous data can be properly modelled.

An exp-HM model with L hidden layers can be generally formulated to a product of conditional distributions:

$$p(\boldsymbol{x}, \boldsymbol{h}^{(1)}, \cdots, \boldsymbol{h}^{(L)}) = p(\boldsymbol{x}|\boldsymbol{h}^{(1)}) \prod_{l=1}^{L-1} p(\boldsymbol{h}^{(l)}|\boldsymbol{h}^{(l+1)}) p(\boldsymbol{h}^{(L)}), \tag{22}$$

where the conditionals and $p(\boldsymbol{h}^{(L)})$ are driven by the generative connections and belong to the exponential family. The model parameters include generative parameters $\boldsymbol{\theta}^{(G)} = \{\boldsymbol{a}^{(G)}, \boldsymbol{b}^{(G,1)}, \cdots, \boldsymbol{b}^{(G,L)}, \boldsymbol{W}^{(G,1)}, \cdots, \boldsymbol{W}^{(G,L)}\}$

and recognition parameters $\boldsymbol{\theta}^{(R)} = \{\boldsymbol{b}^{(R,1)}, \cdots, \boldsymbol{b}^{(R,L)}, \boldsymbol{W}^{(R,1)}, \cdots, \boldsymbol{W}^{(R,L)}\}$, where $\boldsymbol{a}^{(G)}$ contains the generative bias parameters over the visible variables, $\boldsymbol{b}^{(G,l)}$ ($\boldsymbol{b}^{(R,l)}$) contains generative (recognition) bias parameters over the l-th hidden layer, and $\boldsymbol{W}^{(G,l)}$ ($\boldsymbol{W}^{(R,l)}$) is the generative (recognition) interaction weight matrix connecting the (l-1)- and l-th hidden layers. The recognition component does not need bias parameters on $\boldsymbol{x}$, as $\boldsymbol{x}$ is observed. By convention, we let size$\big(\boldsymbol{W}^{(R,l)}\big)$ = size$\big(\boldsymbol{W}^{(G,l)^{\mathrm{T}}}\big)$. Merely for the convenience of discussion, we allow $p(\boldsymbol{x}|\boldsymbol{h}^{(1)})$ to follow any suitable exponential family distribution, and restrict $p(\boldsymbol{h}^{(l)}|\boldsymbol{h}^{(l+1)})$ and $p(\boldsymbol{h}^{(L)})$ to be Bernoulli. Writing these distributions in natural forms, we have

$$p(x_m|\boldsymbol{h}^{(1)}) = e^{\hat{\boldsymbol{a}}_m^{(G)\mathrm{T}}\boldsymbol{s}_m(x_m)+\log f_m(x_m)-A_m(\hat{\boldsymbol{a}}_m^{(G)})} \tag{23}$$

$$p(h_k^{(l)}|\boldsymbol{h}^{(l+1)}) = e^{\hat{b}_k^{(G,l)}h_k^{(l)}+\log g_k^{(l)}(h_k^{(l)})-B_k^{(l)}(\hat{b}_k^{(G,l)})} \tag{24}$$

$$p(h_k^{(L)}) = e^{b_k^{(G,L)}h_k^{(L)}+\log g_k^{(L)}(h_k^{(L)})-B_k^{(L)}(b_k^{(G,L)})}, \tag{25}$$

where $0 \le l \le L-1$, $\hat{\boldsymbol{a}}_m^{(G)}$ hosts all the natural parameters of $p(x_m|\boldsymbol{h}^{(1)})$, and $\boldsymbol{s}_m(x_m)$ includes all corresponding natural forms of sufficient statistics for x_m. The posterior bias for the r-th sufficient statistic of x_m is computed as $\hat{a}_m^{(G,r)} = a_m^{(G,r)} + \boldsymbol{W}_{m,:}^{(G,1)}\boldsymbol{h}^{(1)}$, if its corresponding sufficient statistic $s_m^{(r)}$ is allowed to interact with $\boldsymbol{h}^{(1)}$; otherwise $\hat{a}_m^{(G,r)} = a_m^{(G,r)}$. Similarly, the posterior bias in $p(h_k^{(l)}|\boldsymbol{h}^{(l+1)})$ is computed as $\hat{b}_k^{(G,l)} = b_k^{(G,l)} + \boldsymbol{W}_{k,:}^{(G,l+1)}\boldsymbol{h}^{(l+1)}$.

The instances of conditional distribution $p(\boldsymbol{x}|\boldsymbol{h}^{(1)})$ for selected distributions in the exponential family are given in Supplemental Table 4. For example, if it follows a Poisson distribution, it is defined as

$$p(\boldsymbol{x}|\boldsymbol{h}^{(1)}) = \prod_{m=1}^{M} \mathcal{PO}\big(x_m|e^{\hat{a}_m^{(G)}}\big), \tag{26}$$

where $e^{\hat{a}_m^{(G)}}$ is the mean, and $\hat{a}_m^{(G)} = a_m^{(G)} + \boldsymbol{W}_{m,:}^{(G,1)}\boldsymbol{h}^{(1)}$. Conditional $p(\boldsymbol{h}^{(l)}|\boldsymbol{h}^{(l+1)})$ follow Bernoulli distributions:

$$p(\boldsymbol{h}^{(l)}|\boldsymbol{h}^{(l+1)}) = \prod_{k=1}^{K_l} \mathcal{BE}\big(h_k|\sigma(\hat{b}_k^{(G,l)})\big), 1 \le l \le L-1 \tag{27}$$

$$p(\boldsymbol{h}^{(L)}) = \prod_{k=1}^{K_L} \mathcal{BE}\big(h_k|\sigma(b_k^{(G,L)})\big), \tag{28}$$

where $\sigma(\hat{b}_k^{(G,l)})$ is the success rate.

In HM, the approximation $q(\boldsymbol{h}|\boldsymbol{x})$ is defined as

$$q(\boldsymbol{h}^{(1)}|\boldsymbol{x}) = \prod_{k=1}^{K_1} \mathcal{BE}\big(h_k|\sigma(\hat{b}_k^{(R,1)})\big) \tag{29}$$

$$q(\boldsymbol{h}^{(l)}|\boldsymbol{h}^{(l-1)}) = \prod_{k=1}^{K_l} \mathcal{BE}\big(h_k|\sigma(\hat{b}_k^{(R,l)})\big), 2 \le l \le L. \tag{30}$$

where $\hat{b}_k^{(R,1)} = b_k^{(R,1)} + \boldsymbol{W}_{k,:}^{(R,1)}\boldsymbol{x}$ and $\hat{b}_k^{(R,l)} = b_k^{(R,l)} + \boldsymbol{W}_{k,:}^{(R,l)}\boldsymbol{h}^{(l-1)}$ for $2 \le l \le L$.

3.2. Wake-Sleep Algorithm for Exp-HM

The wake-sleep algorithm (Hinton et al., 1995), originally proposed to learn binary HM, is generalized for exp-HM in this work. Generally speaking, the wake-sleep algorithm is in fact an expectation-maximization (EM) algorithm (Neal & Hinton, 1998) involving a wake phase and a sleep phase, and considers both generative and recognition capabilities. The wake phase is to improve the model to generate a sample close to the training distribution, and the sleep phase is to improve the capability of this model to recognize a fantasy. In the wake phase, values of hidden units are driven by the recognition parameters, and only the generative parameters are updated. In the sleep phase, values of hidden units and visible units are sampled using the generative (and maybe recognition) parameters, and only the recognition parameters are updated. Our wake-sleep algorithm for exp-HM is derived as follows. Note that $q(\boldsymbol{h}|\boldsymbol{x})$ [Eq. (29) and (30)] is factorizable and driven by the recognition connections, while $p(\boldsymbol{h}|\boldsymbol{x}) = \frac{p(\boldsymbol{x},\boldsymbol{h})}{p(\boldsymbol{x})}$ is nonfactorizable, intractable and driven by the generative model. As an objective, an upper bound of the negated log-likelihood is minimized in HM and exp-HM. According to the variational approximation theory, we know that

$$\log p(\boldsymbol{x}) = -j(\boldsymbol{x}) + \mathrm{KL}\big(q(\boldsymbol{h}|\boldsymbol{x})||p(\boldsymbol{h}|\boldsymbol{x})\big), \tag{31}$$

where $j(\boldsymbol{x}) = -l(\boldsymbol{x}) = -\mathrm{KL}\big(p(\boldsymbol{x},\boldsymbol{h})||q(\boldsymbol{h}|\boldsymbol{x})\big) = \sum\!\!\!\!\!\!\int_{\boldsymbol{h}} q(\boldsymbol{h}|\boldsymbol{x}) \log \frac{p(\boldsymbol{x},\boldsymbol{h})}{q(\boldsymbol{h}|\boldsymbol{x})}$ is the upper bound of the negated log-likelihood (equivalently $l(\boldsymbol{x})$ is the variational lower bound of the log-likelihood). The name comes from the fact that $\log p(x) \ge l(q)$, due to $\mathrm{KL}\big(q(\boldsymbol{h}|\boldsymbol{x})||p(\boldsymbol{h}|\boldsymbol{x})\big) \ge 0$. From this definition of $j(\boldsymbol{x})$, we have

$$\begin{aligned} j(\boldsymbol{x}) &= -\sum\!\!\!\!\!\!\int_{\boldsymbol{h}} q(\boldsymbol{h}|\boldsymbol{x}) \log \frac{p(\boldsymbol{x},\boldsymbol{h})}{q(\boldsymbol{h}|\boldsymbol{x})} \\ &= -\sum\!\!\!\!\!\!\int_{\boldsymbol{h}} q(\boldsymbol{h}|\boldsymbol{x}) \log p(\boldsymbol{x},\boldsymbol{h}) + \sum\!\!\!\!\!\!\int_{\boldsymbol{h}} q(\boldsymbol{h}|\boldsymbol{x}) \log q(\boldsymbol{h}|\boldsymbol{x}) \\ &= -\mathrm{E}_{q(\boldsymbol{h}|\boldsymbol{x})}[\log p(\boldsymbol{x},\boldsymbol{h})] - H\big(q(\boldsymbol{h}|\boldsymbol{x})\big). \end{aligned} \tag{32}$$

where $-\log p(\boldsymbol{x},\boldsymbol{h})$ is the description length of $p(\boldsymbol{x},\boldsymbol{h})$ and $H\big(q(\boldsymbol{h}|\boldsymbol{x})\big)$ is the functional entropy of the approximate distribution. When $q(\boldsymbol{h}|\boldsymbol{x})$ is decomposable, so is $H\big(q(\boldsymbol{h}|\boldsymbol{x})\big)$. Hence, computing $H\big(q(\boldsymbol{h}|\boldsymbol{x})\big)$ is technically easy. We can see that minimizing the upper bound $j(\boldsymbol{x})$ is equivalent to minimizing the expected description length $-\log p(\boldsymbol{x},\boldsymbol{h})$ (also called expected complete data negated log-likelihood) with respect to the approximate posterior distribution and maximizing the functional entropy of the approximate posterior distribution. In general, if $p(\boldsymbol{x},\boldsymbol{h}) = \frac{1}{Z'}e^{-E(\boldsymbol{x},\boldsymbol{h})}$ where $E(\boldsymbol{x},\boldsymbol{h})$ is the internal energy of the system, from a physical perspective, the cost $j(\boldsymbol{x})$ can also be written as

$$j(\boldsymbol{x}) = F(\boldsymbol{x}) = \mathrm{E}_{q(\boldsymbol{h}|\boldsymbol{x})}[E(\boldsymbol{x},\boldsymbol{h})] + \log Z' - H(q). \tag{33}$$

Thus, $j(\boldsymbol{x})$ is called Helmholtz free energy, denoted by $F(\boldsymbol{x})$ (hence the name of HM). For exp-HM, in the wake phase, $j(\boldsymbol{x})$ in Eq. (32) is the objective to be minimized.

In the wake phase, the first-order derivative of the objective $j(\boldsymbol{x})$ [Eq. (32)] w.r.t. the generative parameters is

$$\frac{\partial j(\boldsymbol{x})}{\partial \boldsymbol{\theta}^{(G)}} = -\mathrm{E}_{q(\boldsymbol{h}|\boldsymbol{x})}\Big[\frac{\partial \log p(\boldsymbol{x},\boldsymbol{h})}{\partial \boldsymbol{\theta}^{(G)}}\Big], \tag{34}$$

where $\boldsymbol{h}$ is approximated using $q(\boldsymbol{h}|\boldsymbol{x})$, and $\frac{\partial \log p(\boldsymbol{x},\boldsymbol{h})}{\partial \boldsymbol{\theta}^{(G)}}$ is crucial. We find that the first-order derivative of $\log p(\boldsymbol{x},\boldsymbol{h})$ w.r.t. $W^{(G,1)}_{m,k}$ can be computed as

$$\begin{aligned}\frac{\partial \log p(\boldsymbol{x},\boldsymbol{h})}{\partial W^{(G,1)}_{m,k}} &= \frac{\partial \log p(x_m|\boldsymbol{h}^{(1)})}{\partial W^{(G,1)}_{m,k}} \\ &= \frac{\partial \log p(x_m|\boldsymbol{h}^{(1)})}{\partial \hat{a}^{(G,r)}_m}\frac{\partial \hat{a}^{(G,r)}_m}{\partial W^{(G,1)}_{m,k}} \\ &= \big(s^{(r)}_m - \mathrm{E}_{p(x_m|\boldsymbol{h}^{(1)})}[s^{(r)}_m]\big)h^{(1)}_k \\ &= \big(x_m - \langle x_m\rangle\big)h^{(1)}_k, \end{aligned}\tag{35}$$

where in the second line we assume the r-th natural variable interacts with $\boldsymbol{h}^{(1)}$; in the third line we take advantage of the important property of exponential family from Eq. (3); and finally we only practically allow $\boldsymbol{x}$ to interact with $\boldsymbol{h}^{(1)}$. Eq. (35) tells us that the gradient to update $W^{(G,1)}$ equals to the correlation between the observed value of x_m and hidden state $h^{(1)}_k$ minus the correlation between the expected x_m and $h^{(1)}_k$. If $p(x_m|\boldsymbol{h}^{(1)})$ follows a Bernoulli distribution, then $\langle x_m\rangle = \sigma(\hat{a}^{(G)}_m)$, which is consistent with the logistic belief net (Neal, 1992) and the original HM (Hinton et al., 1995). If $p(x_m|\boldsymbol{h}^{(1)})$ is Gaussian, then $\langle x_m\rangle = -\frac{\hat{a}^{(G,1)}_m}{2a^{(G,1)}_m}$. Likewise, the first order derivative of $\log p(\boldsymbol{x},\boldsymbol{h})$ w.r.t. $a^{(G,r)}_m$ can be computed as

$$\begin{aligned}\frac{\partial \log p(\boldsymbol{x},\boldsymbol{h})}{\partial a^{(G,r)}_m} &= \frac{\partial \log p(x_m|\boldsymbol{h}^{(1)})}{\partial a^{(G,r)}_m} \\ &= \frac{\partial \log p(x_m|\boldsymbol{h}^{(1)})}{\partial \hat{a}^{(G,r)}_m}\frac{\partial \hat{a}^{(G,r)}_m}{\partial a^{(G,r)}_m} = s^{(r)}_m - \langle s^{(r)}_m\rangle. \end{aligned}\tag{36}$$

Similarly, the derivative of $\log p(\boldsymbol{x},\boldsymbol{h})$ w.r.t. $W^{(G,l)}_{k_1,k_2}$ $(2 \le l \le L)$ and $b^{(G,l)}_k$ $(1 \le l \le L)$ are respectively computed as

$$\frac{\partial \log p(\boldsymbol{x},\boldsymbol{h})}{\partial W^{(G,l)}_{k_1,k_2}} = \big(h^{(l-1)}_{k_1} - \langle h^{(l-1)}_{k_1}\rangle\big)h^{(l)}_{k_2},\ 2 \le l \le L \tag{37}$$

$$\frac{\partial \log p(\boldsymbol{x},\boldsymbol{h})}{\partial b^{(G,l)}_k} = h^{(l)}_k - \langle h^{(l)}_k\rangle, \quad 1 \le l \le L, \tag{38}$$

where $\langle h^{(l)}_k\rangle = \sigma(\hat{b}^{(G,l)}_k)$ for Bernoulli distributions. In short, we write the gradients in vector and matrix forms:

$$\Delta_{\boldsymbol{a}^{(G,r)}} = -(\boldsymbol{s}^{(r)} - \langle \boldsymbol{s}^{(r)}\rangle) \tag{39}$$

$$\Delta_{\boldsymbol{W}^{(G,1)}} = -(\boldsymbol{x} - \langle \boldsymbol{x}\rangle)\boldsymbol{h}^{(1)\mathrm{T}} \tag{40}$$

$$\Delta_{\boldsymbol{W}^{(G,l)}} = -(\boldsymbol{h}^{(l-1)} - \langle \boldsymbol{h}^{(l-1)}\rangle)\boldsymbol{h}^{(l)\mathrm{T}}, 2 \le l \le L \tag{41}$$

$$\Delta_{\boldsymbol{b}^{(G,l)}} = -(\boldsymbol{h}^{(l)} - \langle \boldsymbol{h}^{(l)}\rangle), \quad 1 \le l \le L, \tag{42}$$

where $\boldsymbol{x}$ is an actual training sample, and $\boldsymbol{h} = \{\boldsymbol{h}^{(1)}, \cdots, \boldsymbol{h}^{(L)}\}$ are sampled from $q(\boldsymbol{h}|\boldsymbol{x})$ using the recognition parameters. With these gradients, the generative parameters are updated using rule: $\boldsymbol{\theta}^{(G)} = \boldsymbol{\theta}^{(G)} - \epsilon\Delta_{\theta^{(G)}}$.

In the sleep phase, starting from the top layer, a fantasy can be either (i) unbiasedly sampled from the generative distribution $p(\boldsymbol{h}^{(L)})$ or (ii) initialized using the states of $\boldsymbol{h}^{(L)}$ obtained in the wake phase. Values of downstream units are sampled by the generative component. We thus generate a total fantasy, using which the recognition parameters are updated accordingly. The second method (used in our implementation) connects both objectives and is akin to the contrastive divergence (CD) algorithm (Hinton, 2002; Hinton et al., 2006). To remove sampling bias, Gibbs sampling with multiple alternating top-down and bottom-up passes could be adopted, similar to (persistent) CD-k algorithms (Tieleman, 2008). Since the recognition and generative components are structurally opposite, to derive update rules for the recognition parameters, we just need to swap their roles in the objective function [Eq. (32)], resulting in the corresponding gradients:

$$\Delta_{\boldsymbol{b}^{(R,l)}} = -(\boldsymbol{h}^{(l)} - \langle \boldsymbol{h}^{(l)}\rangle), \quad 1 \le l \le L \tag{43}$$

$$\Delta_{\boldsymbol{W}^{(R,1)}} = -(\boldsymbol{h}^{(1)} - \langle \boldsymbol{h}^{(1)}\rangle)\boldsymbol{x}^{\mathrm{T}} \tag{44}$$

$$\Delta_{\boldsymbol{W}^{(R,l)}} = -(\boldsymbol{h}^{(l)} - \langle \boldsymbol{h}^{(l)}\rangle)\boldsymbol{h}^{(l-1)\mathrm{T}}, \quad 2 \le l \le L, \tag{45}$$

where $\{\boldsymbol{h}^{(L)}, \cdots, \boldsymbol{h}^{(1)}, \boldsymbol{x}\} \sim p(\boldsymbol{x},\boldsymbol{h})$; $\langle h^{(l)}_k\rangle = \sigma(\hat{b}^{(R,l)}_k)$.

In the algorithm, the wake phase and sleep phase alternatingly iterate for a pre-specified steps or until the objective $j(\boldsymbol{x})$ [Eq. (32)] does not reduce dramatically. The gradients for selected instances of exp-HMs are given in Supplemental Table 4. The wake-sleep algorithm with a random initialization may not perform well for deep exp-HMs. Stacked exp-RBMs can be first applied to provide a warm start.

4. Exponential Family DBN

When generating fantasies, one critical issue with HM models is that a point sampled from $p(\boldsymbol{h}^{(L)})$ may not represent a meaningful object, hence may eventually generate an abnormal $\boldsymbol{x}$. To address this issue, DBN adds an RBM on top of an HM, so that a point generated using Gibbs sampling from the RBM is from the domain of interest (Hinton et al., 2006). However, the original DBN was designed for binary variables, limiting its applications in many fields. To overcome it, we generalize it to have visible variables follow any distributions from the exponential family. Although hidden variables can follow any exponential family distributions too, for convenience of derivation, we only discuss Bernoulli hidden variables below. Such an exponential family DBN (exp-DBN) with L hidden layers is defined as

$$\begin{aligned}p(\boldsymbol{x},\boldsymbol{h}) =& p(\boldsymbol{x}|\boldsymbol{h}^{(1)})p(\boldsymbol{h}^{(1)}|\boldsymbol{h}^{(2)})\cdots \\ & p(\boldsymbol{h}^{(L-2)}|\boldsymbol{h}^{(L-1)})p(\boldsymbol{h}^{(L-1)},\boldsymbol{h}^{(L)}), \end{aligned}\tag{46}$$

where we denote $\boldsymbol{h} = \{\boldsymbol{h}^{(1)}, \cdots, \boldsymbol{h}^{(L)}\}$ and define

$$p(\boldsymbol{x}|\boldsymbol{h}^{(1)}) = \prod_{m=1}^{M} p\big(x_m|\boldsymbol{\eta}(\hat{\boldsymbol{a}}_m^{(G)})\big) \tag{47}$$

$$p(\boldsymbol{h}^{(l)}|\boldsymbol{h}^{(l+1)}) = \prod_{k=1}^{K_l} \mathcal{BE}\big(h_k|\sigma(\hat{b}_k^{(G,l)})\big),\ 1 \le l \le L-2 \tag{48}$$

$$p(\boldsymbol{h}^{(L-1)}, \boldsymbol{h}^{(L)}) = \frac{1}{Z} e^{-E(\boldsymbol{h}^{(L-1)}, \boldsymbol{h}^{(L)})}, \tag{49}$$

where $\hat{\boldsymbol{a}}_m^{(G)}$ and $\hat{b}_k^{(G,l)}$ are defined as in exp-HM and $E(\boldsymbol{h}^{(L-1)}, \boldsymbol{h}^{(L)}) = -{\boldsymbol{b}^{(G,L-1)}}^{\mathrm{T}}\boldsymbol{h}^{(L-1)} - {\boldsymbol{b}^{(G,L)}}^{\mathrm{T}}\boldsymbol{h}^{(L)} - {\boldsymbol{h}^{(G,L-1)}}^{\mathrm{T}}\boldsymbol{W}^{(G,L)}\boldsymbol{h}^{(G,L)}$. The undirected component is formulated by a joint distribution, while the directed component are formulated by conditional distributions. Eq. (47) formulates the exponential family conditional distributions of visible variables. For example, a Bernoulli visible unit has $p(x_m|\boldsymbol{h}^{(1)}) = \mathcal{BE}(x_m|\sigma(\hat{a}_m^{(G)}))$, where $\hat{a}_m^{(G)} = a_m^{(G)} + \boldsymbol{W}_{m,:}^{(G,1)}\boldsymbol{h}^{(1)}$. Other distributions are given in Supplemental Table 4. Same as in exp-HMs, the conditional over a hidden layer follows a Bernoulli distribution which is parameterized by its prior bias and interactions with its upper layer. The top two layers form a Bernoulli-Bernoulli RBM as associative memory. The inference in DBN is also intractable. Same as in exp-HMs, recognition connections define a factorizable approximate distribution:

$$q(\boldsymbol{h}^{(1\cdots L-1)}|\boldsymbol{x}) = q(\boldsymbol{h}^{(1)}|\boldsymbol{x}) \cdots q(\boldsymbol{h}^{(L-1)}|\boldsymbol{h}^{(L-2)}), \tag{50}$$

where $\boldsymbol{h}^{(1\cdots L-1)} = \{\boldsymbol{h}^{(1)}, \cdots, \boldsymbol{h}^{(L-1)}\}$. In summary, the model parameters in exp-DBN include generative parameters $\boldsymbol{\theta}^{(G)}=\{\boldsymbol{a}^{(G)}, \boldsymbol{b}^{(G,1)}, \cdots, \boldsymbol{b}^{(G,L-1)}, \boldsymbol{b}^{(G,L)}, \boldsymbol{W}^{(G,1)}, \cdots, \boldsymbol{W}^{(G,L-1)}, \boldsymbol{W}^{(G,L)}\}$ and recognition parameters $\boldsymbol{\theta}^{(R)}=\{\boldsymbol{b}^{(R,1)}, \cdots, \boldsymbol{b}^{(R,L-1)}, \boldsymbol{W}^{(R,1)}, \cdots, \boldsymbol{W}^{(R,L-1)}\}$.

4.1. Model Learning for Exp-DBN

In the original DBN (Hinton et al., 2006), the model parameters are pretrained using binary RBMs (with $\boldsymbol{\theta}^{(R)}$ tied with $\boldsymbol{\theta}^{(G)}$) and then fine-tuned using a wake-sleep (up-down) algorithm (with $\boldsymbol{\theta}^{(R)}$ and $\boldsymbol{\theta}^{(G)}$ untied). We generalize this procedure to exp-DBN, such that its model parameters can be pretrained using exp-RBMs and then fine-tuned using an exponential family wake-sleep algorithm. In the wake phase of this extended algorithm, $\boldsymbol{h}^{(1\cdots L-1)}$ are first sampled using the corresponding recognition parameters in the exp-HM component. Then, using $\boldsymbol{h}^{(L-1)}$ as input, (persistent) CD-k sampling (Tieleman, 2008) can be employed in the top exp-RBM and its parameters are updated using the rules for exp-RBM. After that, generative parameters below the top exp-RBM are updated using the same rules as in exp-HMs. In the sleep phase, $\boldsymbol{h}^{(L-1)}$ sampled during the wake phase is used to generate states of the subsequent layers driven by generative parameters, and the recognition parameters are updated as in exp-HMs. The gradients to update the model parameters for various distributions are given in Supplemental Table 4.

4.2. Variational Lower Bound of Exp-DBN

As indicated in Eq. (49), the partition function only appears at the top exp-RBM. Inspired by (Salakhutdinov & Murray, 2008), to obtain a variational lower bound of the log-likelihood, $\boldsymbol{h}^{(L)}$ thus is summed out, leading to

$$p(\boldsymbol{x}, \boldsymbol{h}^{(1\cdots L-1)}) = p(\boldsymbol{x}|\boldsymbol{h}^{(1)}) \cdots p(\boldsymbol{h}^{(L-2)}|\boldsymbol{h}^{(L-1)}) \times \frac{1}{Z} e^{-F(\boldsymbol{h}^{(L-1)})} = \frac{1}{Z} p^*(\boldsymbol{x}, \boldsymbol{h}^{(1\cdots L-1)}). \tag{51}$$

The variational lower bound is hence computed as

$$l(\boldsymbol{x}) = \mathrm{E}_{q(\boldsymbol{h}^{(1\cdots L-1)}|\boldsymbol{x})}[\log p^*(\boldsymbol{x}, \boldsymbol{h}^{(1\cdots L-1)})] - \log Z + H\big(q(\boldsymbol{h}^{(1\cdots L-1)}|\boldsymbol{x})\big). \tag{52}$$

The first term can be estimated by Monte Carlo approximation by sampling from $q(\boldsymbol{h}^{(1\cdots L-1)}|\boldsymbol{x})$ with $\boldsymbol{x}$ clamped. The second term $\log Z$ can be estimated using the generalized AIS procedure (Section 2) to the top exp-RBM.

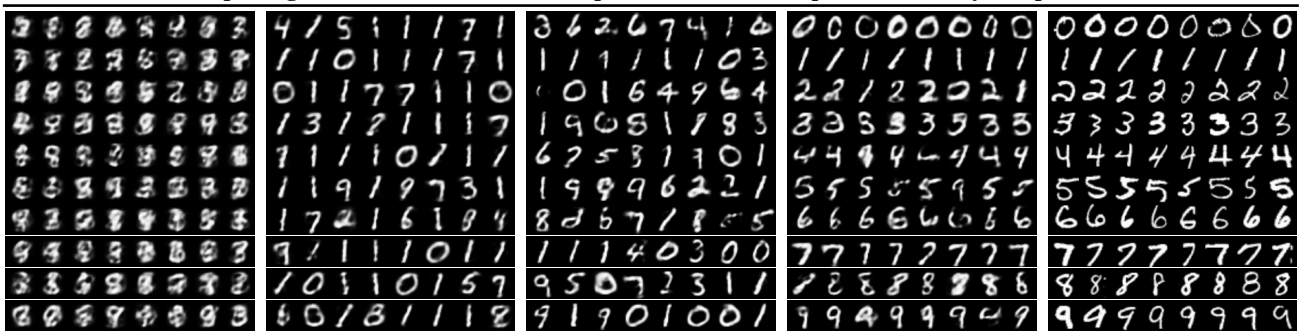
4.3. Applications of Exp-DBN

Certainly, exp-DBN is applicable to pretrain deep discriminative models. More importantly, exp-DBN can be further extended to multi-modal exp-DBNs (see Supplemental Figure 1 for an example) which can be potentially applied in multi-modal learning (Ramachandram & Taylor, 2017; Li et al., 2018) tasks, such as data fusion, integrative classification and clustering, multi-label learning, transfer learning, missing value estimation, information retrieval, machine translation, autonomous navigation, and Internet of things. Our future work will focus on these applications. In our experiments, we designed a two-modal exp-DBN to showcase some potentials of our exponential family generalization.

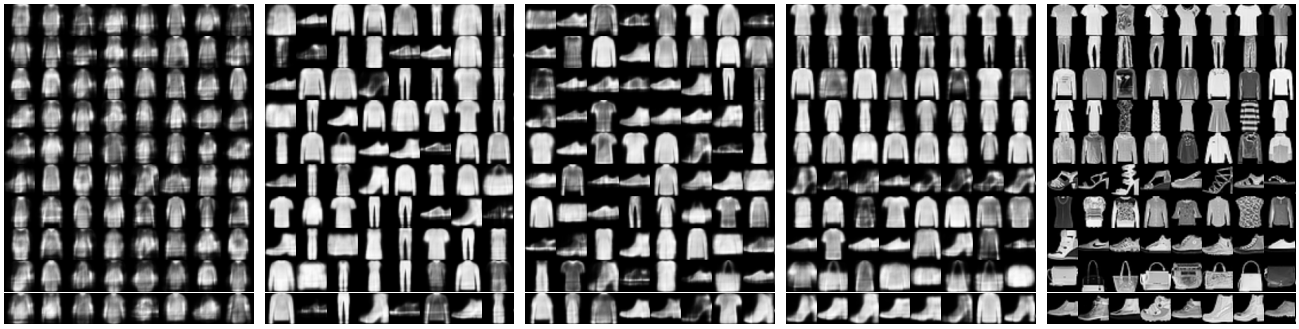
5. Experiments

5.1. Modelling Image Data

To demonstrate the potential of these exponential family models for image generation, we first designed and learned exp-HM, exp-DBN, and two-modal exp-DBN models on the MNIST (LeCun et al., 1998) and Fashion-MNIST (Xiao et al., 2017) data sets respectively. Each data set contains 60,000 training and 10,000 test images of size 28×28. Bernoulli and Multinoulli distributions were assumed respectively for the image (pixel intensities in range [0,1]) and class (categories from 0 to 9) modalities in the two-modal exp-DBN. All models were pretrained using exp-RBMs and fine-tuned using our generalized wake-sleep algorithms. The structures of these models and the generated images are



(a) Exp-HM (Ancestral) $784 \Leftarrow 500 \Leftarrow 500 \Leftarrow 500$ (b) Exp-HM (Gibbs) $784 \Leftarrow 500 \Leftarrow 500 \Leftarrow 500$ (c) Exp-DBN $784 \Leftarrow 500 \Leftarrow 500 \Leftrightarrow 500$ (d) 2-modal exp-DBN $784 \Leftarrow 500 \Leftarrow 500 \Leftrightarrow 1000 \Leftrightarrow 10$ (e) Actual



(f) Exp-HM (Ancestral) $784 \Leftarrow 500 \Leftarrow 500 \Leftarrow 1000$ (g) Exp-HM (Gibbs) $784 \Leftarrow 500 \Leftarrow 500 \Leftarrow 1000$ (h) Exp-DBN $784 \Leftarrow 500 \Leftarrow 500 \Leftrightarrow 1000$ (i) 2-modal exp-DBN $784 \Leftarrow 500 \Leftarrow 500 \Leftrightarrow 1000 \Leftrightarrow 10$ (j) Actual

Figure 1. Actual and generated images on MNIST (top) and Fashion-MNIST (bottom) data sets.

given in Figure 1. First, from Figures 1a and 1f, we see that exp-HM did not generate good images using a traditional ancestral sampling, indicating a hidden state sampled from the generative distribution of the top hidden layer is unlikely from the domain of interest when its dimensionality is large. Second, we uncover that exp-HM's capacity of generating fantasies can be improved by a Gibbs sampling using multiple alternating top-down and bottom-up passes (see Figures 1b and 1g). This discovery was not reported before. Third, Figures 1c and 1h corroborate the early finding in (Hinton et al., 2006) that, by adding an undirected network on top of an exp-HM, exp-DBN could generate images of good quality, the majority of which look similar to the actual examples in Figures 1e and 1j. Fourth, as show in Figures 1d and 1i, by fixing the class modality, the two-modal exp-DBN could take attention to generating images of specified classes.

5.2. Modelling Text Data

To further show the importance of proper distribution in modelling, we applied a two-modal exp-DBN on the 20 Newsgroups data (Lang, 1995). The 11,269 training and 7,505 test documents from 20 topics were transformed to word-count profiles. Stop-words and words appear in less than 60 training samples were discarded, leaving 4,031 words for modelling. The network has one modality for word counts with a hidden layer of 500 Bernoulli units, one modality for class labels without any hidden layer, and a joint hidden layer with 520 Bernoulli units. We used Bernoulli, Poisson, and Multinomial distributions, separately, for word counts, and Multinoulli distribution for class labels. Each profile was normalized to have 1,000 word counts for Poisson and Multinomial models, but binarized for Bernoulli models. For comparison, we extended exp-RBM to a two-modal exp-DBM with same network architecture as the two-modal exp-DBN. Both types of generative models were pretrained using exp-RBMs. The two-modal exp-DBNs were fine-tuned by our generalized wake-sleep algorithms, while the two-modal exp-DBMs were fine-tuned by a modified algorithm originally proposed in (Salakhutdinov & Hinton, 2009a). A multilayer perceptron with same network skeleton but RELU units was also applied as a discriminative model on the tf-idf normalized data.

The log-partition functions and variational lower bounds of these generative models were estimated and shown in Table 1. We see that log-partition functions can be estimated stably with tiny variances in the exponential family framework, and the log-likelihoods of Poisson and Multinomial models are roughly similar (note: we cannot compare their log-likelihoods with these of the Bernoulli models, because they were experimented on different formats of the data). After training, fantasies were drawn using Gibbs sampling from the two-modal exp-DBN with Poisson distribution. The top 15 words and class label of some cherry-picked generations are given in Table 2, which implies that the model learned useful information and generated meaningful samples. In these generative models, class labels of test samples can

be treated as missing data and inferred using a mean-field method. Table 1 also shows the test accuracies after pre-training and after fine-tuning, respectively. First, we find that the Poisson models outperform Multinomoal models which are named replicated softmax models in (Salakhutdinov & Hinton, 2009b). The models based on conventionally used Bernoulli distributions obtained non-competitive results. Second, the two-modal exp-DBNs achieved better accuracies than their two-modal exp-DBM rivals. Moreover, although discriminative models often claim victories over generative models in classification tasks, we find that with appropriate distributions, the directed and undirected generative models under investigation may perform better than them which are less advantageous in handling data types. Of course, classification performance can be further improved if sequential dependency is properly modelled.

Table 1. Log-partition functions, variational lower bounds, and prediction accuracies on 20 Newsgroups dataset. BE: Bernoulli, MU: Multinoulli, PO: Poisson, MN: Multinomial.

	Vis. Type	Data	log Z (STD)	$l(\boldsymbol{x})$	Acc. pretrain	Acc. fine-tune
MDBN	BE+MU	Train	3105.74($\pm$0.35)	-2691.43	35.03%	38.81%
		Test	3106.05($\pm$0.46)	-2694.58		
	PO+MU	Train	1469.18($\pm$0.44)	-3976.12	**68.46%**	**69.30%**
		Test	1468.72($\pm$0.39)	-4273.88		
	MN+MU	Train	1617.29($\pm$0.64)	-3897.86	66.80 %	68.27%
		Test	1616.88($\pm$0.63)	-4216.70		
MDBM	BE+MU	Train	3028.27($\pm$2.61)	-1754.63	27.24%	40.58%
		Test	3028.28($\pm$1.43)	-1757.03		
	PO+MU	Train	2261.53($\pm$1.76)	-3772.34	67.31%	66.84%
		Test	2260.72($\pm$1.76)	-4043.79		
	MN+MU	Train	9761.72($\pm$0.96)	-3873.40	63.94 %	63.96%
		Test	9762.22($\pm$0.76)	-4108.18		
Multilayer perceptron with RELU units					58.53%	

Table 2. Top 15 words and classes of samples generated by the Poisson-Multinoulli two-modal exp-DBN.

$\boldsymbol{x}$	y
government system writes key clipper encryption keys don nsa article escrow secure chip people time	sci.crypt
article writes people killed gun fire apr control country children weapons government time guns police	talk.politics.guns
sale offer university email includes time condition windows call contact shipping interested sell computer original	misc.forsale
division games toronto boston hockey st team cup pittsburgh series detroit play chicago game montreal	rec.sport.hockey
card bus motherboard ram controller board isa friend machine pc slot scsi work dx cards	comp.sys.mac.hardware

6. Conclusion and Discussion

We discuss RBM's free energy function, HM and DBN from the exponential family perspective. We derive the corresponding wake-sleep algorithms in a general framework. Our experiments show that exp-DBN and exp-HM with Gibbs sampling are superior to exp-HM with ancestral sampling in generating samples. With appropriate distributions, the two-modal exp-DBN outperformed the two-modal exp-DBM and a discriminative model on a text data. We clarify that, after finishing this work, we identified that a model, named deep exponential family (DEF), is independently explored in (Ranganath et al., 2015). DEF and our work share similar motivation, but complement each other. The differences are explained below. (1) DEF was a result of hierarchical chaining of exponential family distributions, while our work was inspired by existing and our new findings of exp-RBMs. In the DEF framework, only means of nodes within current layer are affected by their parents. But in our framework, any parameters of nodes in current layer can be modified by their parents in theory. (2) We generalize the wake-sleep algorithm for directed DGM, which is not trivial because strict derivation of generic update rules requires wise application of several properties of exp-RBM and exponential family distributions. DEF resorts to an algorithm similar to neural variational inference and learning (NVIL) (Mnih & Gregor, 2014), a REINFORCE algorithm. (3) We find that the free energy function of exp-RBM is a linear combination of log-partition functions of hidden variable, and the log-partition function of a base-rate exp-RBM can be beautifully computed as a summation of log-partition functions of individual variables. This discovery enables us to easily estimate log-partition functions of exp-RBM, exp-DBN and exp-DBM using AIS, and thus their (lower bounds of) likelihoods. This is not addressed in DEF.

Current DGMs with inference networks (including ours) are partially united under the umbrella of variational inference, using either separate or joint objectives for generative and recognition components. Specific learning and inference methods, such as wake-sleep algorithms (Bornschein & Bengio, 2015), NVIL based algorithms (Mnih & Gregor, 2014; Ranganath et al., 2015) and variational autoencoder (VAE) based algorithms (Rezende et al., 2014; Kingma & Welling, 2014), mainly differ in how the inference component is trained. However, these methods could be transformable under mild conditions. For example, if we apply mean-field approximation to alternating hidden layers in exp-HM, we would obtain a VAE-like model which uses deterministic networks to regress parameters of stochastic layers. Moreover, level of novelty and quality within generated samples vary among models, depending on how the generative and inference networks coordinate (Burda et al., 2016).

Our investigation indicates that ancestral sampling is not always effective. Association and (or) attention may be needed in both directed and undirected generative networks for novelty generation. While celebrating limited successes in current DGMs [including generative adversarial network (GAN) (Goodfellow et al., 2014)], methodological engineering may only partially solve the puzzle, we thus need to think beyond this niche. From the cognitive science perspective, innovation comes from interaction and combination of cognits, units of any knowledge or concept representation in the cerebral cortex (Fuster, 2003). Many existing DGMs require justification in cognitive neuroscience. Certainly, inspiration from hierarchies of perceptual, associative and executive networks would create a new generation of DGMs which might enhance the statistical foundation and drive us closer towards the long-dreamed strong AI.

References

Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. Greedy layer-wise training of deep networks. In *Advances in Neural Information Processing Systems*, pp. 153–160, 2006.

Bornschein, J. and Bengio, Y. Reweighted wake-sleep. In *International Conference on Learning Representations*, 2015.

Burda, Y., Grosse, R., and Salakhutdinov, R. Importance weighted autoencoders. In *International Conference on Learning Representations*, 2016.

Dayan, P., Hinton, G.E., Neal, R., and Zemel, R.S. The Helmholtz machine. *Neural Computation*, 7:1022–1037, 1995.

Fuster, J.M. *Cortex and Mind*. Oxford University Press, 2003.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Hinton, G. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

Hinton, G., Dayan, P., Frey, B., and Neal, R. The wake-sleep algorithm for unsupervised neural networks. *Science*, 268: 1558–1161, 1995.

Hinton, G.E., Osindero, S., and Teh, Y. A fast learning algorithm for deep belief nets. *Neural Computation*, 18: 1527–1554, 2006.

Kingma, D.P. and Welling, M. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Lang, Ken. Newsweeder: Learning to filter netnews. In *International Conference on Machine Learning*, pp. 331–339, 1995. http://qwone.com/ jason/20Newsgroups.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Li, Y., Wu, F.X., and Ngom, A. A review on machine learning principles for multi-view biological data integration. *Briefings in Bioinformatics*, 29(2):325–340, 2018.

Mnih, A. and Gregor, K. Neural variational inference and learning in belief networks. In *International Conference on Machine Learning*, pp. II–1791–II–1799, 2014.

Neal, R. M. and Hinton, G.E. A view of the EM algorithm that justifies incremental, sparse, and other variants. In Jordan, M.I. (ed.), *Learning in Graphical Models*, Adaptive Computation and Machine Learning series, chapter 11, pp. 355–368. MIT, Cambridge, MA, 1998.

Neal, R.M. Connectionist learning of belief networks. *Artificial Intelligence*, 56:71–113, 1992.

Neal, R.M. Annealed importance sampling. *Statistics and Computing*, 11:125–139, 2001.

Ramachandram, D. and Taylor, G.W. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017.

Ranganath, R., Tang, L., Charlin, L., and Blei, D.M. Deep exponential families. In *International Conference on Artificial Intelligence and Statistics*, pp. 762–771, 2015.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pp. II–1278–II–1286, 2014.

Salakhutdinov, R. and Hinton, G. Deep Boltzmann machine. In *International Conference on Artificial Intelligence and Statistics*, pp. 448–455, 2009a.

Salakhutdinov, R. and Hinton, G. Replicated softmax: An undirected topic model. In *Advances in Neural Information Processing Systems*, pp. 1607–1614, 2009b.

Salakhutdinov, R. and Murray, I. On the quantitative analysis of deep belief networks. In *International Conference on Machine Learning*, pp. 872–879, 2008.

Smolensky, P. Information processing in dynamical systems: Foundations of harmony theory. In Rumelhart, D.E. and McLelland, J.L. (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundation*, chapter 6, pp. 194–281. MIT, Cambridge, MA, 1986.

Srivastava, N. and Salakhutdinov, R. Learning representations for multimodal data with deep belief nets. In *International Conference on Machine Learning Workshop on Representation Learning*, 2012.

Srivastava, N. and Salakhutdinov, R. Multimodal learning with deep Boltzmann machines. *Journal of Machine Learning Research*, 15:2949–2980, 2014.

Tieleman, T. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *International Conference on Machine Learning*, pp. 1064–1071, 2008.

Welling, M., Rosen-zvi, M., and Hinton, G. Exponential family harmoniums with an application to information retrieval. In *Advances in Neural Information Processing Systems*, pp. 1481–1488, 2005.

Xiao, Han, Rasul, Kashif, and Vollgraf, Roland. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms, 2017.

Supplemental Figures and Tables for "Exploring Helmholtz Machine and Deep Belief Net in the Exponential Family Perspective"

Yifeng Li*[1] and Xiaodan Zhu†[2]

[1]Digital Technologies Research Centre, National Research Council Canada, Ottawa, Ontario, K1A 0R6, Canada

[2]Department of Electrical and Computer Engineering, Queen's University, Kingston, Ontario, K7L 3N6 Canada

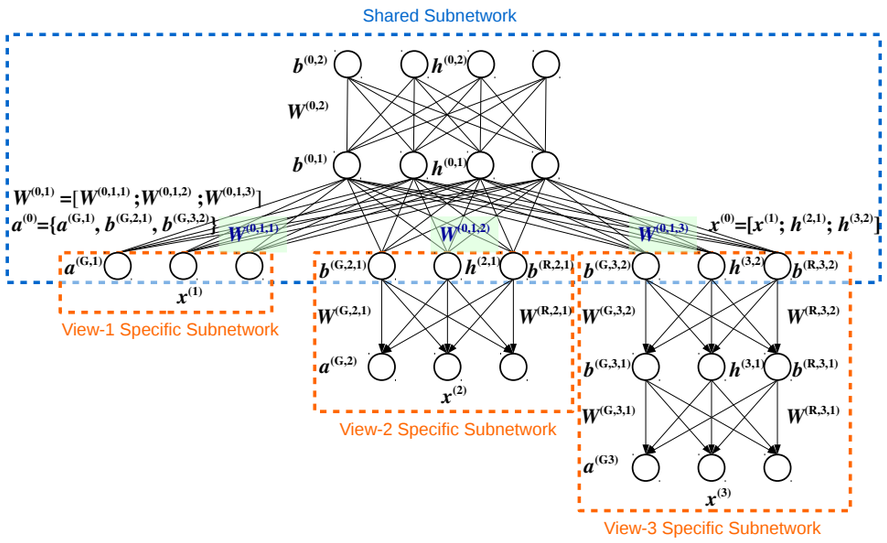


Figure 1: A schematic example of exponential family multi-modal deep belief net (exp-MDBN). It integrates three modalities/views whose visible layers can respectively follow any types of distributions from the exponential family. Modality 1 has a trivial structure without any hidden layers, and may be used for class labels. The subnetworks of modalities 2 and 3 are exp-HMs. The shared/joint subnetwork is a DBM, functioning as a deep associative memory.

*Corresponding author. E-mail address: yifeng.li@nrc-cnrc.gc.ca yifeng.li.cn@gmail.com

†Email address: xiaodan.zhu@queensu.ca

Table 1: Selected members of the exponential family.

Distribution	Standard Form	Natural Form	$\boldsymbol{\theta}(\boldsymbol{\eta})$	$\boldsymbol{\eta}(\boldsymbol{\theta})$	$\boldsymbol{s}(x)$	$h(x)$	$A(\boldsymbol{\eta})$	$A(\boldsymbol{\theta})$	Sufficient Statistics
Gaussian	$p(x\|\mu,\lambda^{-1}) = \frac{\sqrt{\lambda}}{\sqrt{2\pi}} e^{-\frac{\lambda}{2}(x-\mu)^2}$, $\lambda > 0$	$p(x\|\boldsymbol{\theta}) = e^{\theta_1 s_1 + \theta_2 s_2 - A(\boldsymbol{\theta})}$, $\theta_2 < 0$	$\begin{bmatrix} \mu\lambda \\ -\frac{\lambda}{2} \end{bmatrix}$	$\begin{bmatrix} -\frac{\theta_1}{2\theta_2} \\ -2\theta_2 \end{bmatrix}$	$\begin{bmatrix} x \\ x^2 \end{bmatrix}$	1	$\frac{1}{2}\log\frac{2\pi}{\lambda} + \frac{\mu^2\lambda}{2}$	$\frac{1}{2}\log\frac{\pi}{-\theta_2} - \frac{\theta_1^2}{4\theta_2}$	$\mathrm{E}[x] = \mu = -\frac{\theta_1}{2\theta_2}$ $\mathrm{Var}[x] = \frac{1}{\lambda} = -\frac{1}{2\theta_2}$ $\mathrm{E}[x^2] = \mu^2 + \frac{1}{\lambda} = \frac{\theta_1^2}{4\theta_2^2} + \frac{1}{-2\theta_2}$
Gaussian (fix precision λ)	$p(x\|\mu,\lambda^{-1}) = \frac{\sqrt{\lambda}}{\sqrt{2\pi}} e^{-\frac{\lambda}{2}(x-\mu)^2}$, $\lambda > 0$	$p(x\|\theta) = (\frac{\sqrt{\lambda}}{\sqrt{2\pi}} e^{-\frac{\lambda}{2}x^2}) e^{\theta s - A(\boldsymbol{\theta})}$	μ	θ	λx	$\frac{\sqrt{\lambda}}{\sqrt{2\pi}} e^{-\frac{\lambda}{2}x^2}$	$\frac{\mu^2\lambda}{2}$	$\frac{\theta^2\lambda}{2}$	$\mathrm{E}[x] = \mu = \theta$ $\mathrm{Var}[x] = \frac{1}{\lambda}$ $\mathrm{E}[x^2] = \mu^2 + \frac{1}{\lambda} = \theta^2 + \frac{1}{\lambda}$
Gaussian (fix precision λ)	$p(x\|\mu,\lambda^{-1}) = \frac{\sqrt{\lambda}}{\sqrt{2\pi}} e^{-\frac{\lambda}{2}(x-\mu)^2}$, $\lambda > 0$	$p(x\|\theta) = (\frac{\sqrt{\lambda}}{\sqrt{2\pi}} e^{-\frac{\lambda}{2}x^2}) e^{\theta s - A(\boldsymbol{\theta})}$	$\mu\lambda$	$\frac{\theta}{\lambda}$	x	$\frac{\sqrt{\lambda}}{\sqrt{2\pi}} e^{-\frac{\lambda}{2}x^2}$	$\frac{\mu^2\lambda}{2}$	$\frac{\theta^2}{2\lambda}$	$\mathrm{E}[x] = \mu = \frac{\theta}{\lambda}$ $\mathrm{Var}[x] = \frac{1}{\lambda}$ $\mathrm{E}[x^2] = \mu^2 + \frac{1}{\lambda} = \frac{\theta^2}{\lambda^2} + \frac{1}{\lambda}$
Poisson	$p(x\|\lambda) = \frac{e^{-\lambda}}{x!}\lambda^x$, $\lambda > 0$, $x \geq 0$	$p(x\|\theta) = \frac{1}{x!} e^{\theta s - A(\theta)}$	$\log\lambda$	e^θ	x	$\frac{1}{x!}$	λ	e^θ	$\mathrm{E}[x] = \lambda = e^\theta$ $\mathrm{Var}[x] = \lambda = e^\theta$
Bernoulli	$p(x\|p) = p^x(1-p)^{1-x}$, $p \in (0,1)$, $x \in \{0,1\}$	$p(x\|\theta) = e^{\theta s - A(\theta)}$	$\log\frac{p}{1-p}$	$\frac{e^\theta}{1+e^\theta} = \sigma(\theta)$	x	1	$-\log(1-p)$	$-\log(1-\sigma(\theta)) = \log(1+e^\theta)$	$\mathrm{E}[x] = p = \sigma(\theta)$ $\mathrm{Var}[x] = p(1-p) = \sigma(\theta)(1-\sigma(\theta))$
Binomial (fix number of trials n)	$p(x\|n,p) = \binom{n}{x} p^x(1-p)^{n-x}$, $p \in (0,1)$, $x \geq 0$	$p(x\|\theta) = \frac{n!}{x!(n-x)!} e^{\theta s - A(\theta)}$	$\log\frac{p}{1-p}$	$\frac{e^\theta}{1+e^\theta} = \sigma(\theta)$	x	$\frac{n!}{x!(n-x)!}$	$-n\log(1-p)$	$-n\log(1-\sigma(\theta))$	$\mathrm{E}[x] = np = n\sigma(\theta)$ $\mathrm{Var}[x] = np(1-p) = n\sigma(\theta)(1-\sigma(\theta))$
Negative Binomial (fix number of successes k, success rate p)	$p(x\|k,p) = \binom{x+k-1}{k-1} p^k(1-p)^x$, $p \in (0,1)$, $x \geq 0$	$p(x\|\theta) = \frac{(x+k-1)!}{(k-1)!x!} e^{\theta s - A(\theta)}$, $\theta < 0$	$\log(1-p)$	$1 - e^\theta$	x	$\frac{(x+k-1)!}{(k-1)!x!}$	$-k\log p$	$-k\log(1-e^\theta)$	$\mathrm{E}[x] = k\frac{1-p}{p} = k\frac{e^\theta}{1-e^\theta}$ $\mathrm{Var}[x] = k\frac{1-p}{p^2} = k\frac{e^\theta}{(1-e^\theta)^2}$
Multinoulli	$p(x_1,\cdots,x_M\|p_1,\cdots,p_M) = p_1^{x_1}\cdots p_M^{x_M}$, $p_m \geq 0$, $\sum_{m=1}^M p_m = 1$, $x_m \in \{0,1\}$, $\sum_{m=1}^M x_m = 1$	$p(x\|\boldsymbol{\theta}) = \sum_{m=1}^M e^{\theta_m} e^{\theta_1 s_m + \cdots + \theta_M s_M - \log C}$, where $C = \sum_{m=1}^M e^{\theta_m}$	$\begin{bmatrix} \log p_1 + \log C \\ \vdots \\ \log p_M + \log C \end{bmatrix}$	$\begin{bmatrix} \frac{e^{\theta_1}}{C} \\ \vdots \\ \frac{e^{\theta_M}}{C} \end{bmatrix}$	$\begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix}$	1	$\log C$	$\log C$	$\mathrm{E}[x_m = 1] = p_m = \frac{e^{\theta_m}}{C}$ $\mathrm{Var}[x_m = 1] = p_m(1-p_m) = \frac{e^{\theta_m}}{C - \theta_m}$
Multinomial (fix number of trials n)	$p(x_1,\cdots,x_M\|n,p_1,\cdots,p_M) = \frac{n!}{\prod_{m=1}^M x_m!} p_1^{x_1}\cdots p_M^{x_M}$, $p_m \geq 0$, $\sum_{m=1}^M p_m = 1$, $x_m \in \{0,1,\cdots,n\}$, $\sum_{m=1}^M x_m = n$	$p(x\|\theta) = \frac{n!}{\prod_{m=1}^M x_m!} e^{\theta_1 s_m + \cdots + \theta_M s_M - n\log C}$, where $C = \sum_{m=1}^M e^{\theta_m}$	$\begin{bmatrix} \log p_1 + \log C \\ \vdots \\ \log p_M + \log C \end{bmatrix}$	$\begin{bmatrix} \frac{e^{\theta_1}}{C} \\ \vdots \\ \frac{e^{\theta_M}}{C} \end{bmatrix}$	$\begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix}$	$\frac{n!}{\prod_{m=1}^M x_m!}$	$n\log C$	$n\log C$	$\mathrm{E}[x_m] = np_m = n\frac{e^{\theta_m}}{C}$ $\mathrm{Var}[x] = np_m(1-p_m) = n\frac{e^\theta}{C - e^\theta}$

Table 2: Selected instances of exp-RBM.

Model	Energy Function	Conditional	Gradient
Bernoulli-Bernoulli	$E(\boldsymbol{x},\boldsymbol{h}) = -\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{h} - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{h}$, where $\boldsymbol{a} = [\log\frac{p_1}{1-p_1}, \cdots, \log\frac{p_M}{1-p_M}]^{\mathrm{T}}$, $\boldsymbol{b} = [\log\frac{p_1}{1-p_1}, \cdots, \log\frac{p_K}{1-p_K}]$	$p(\boldsymbol{x}\vert\boldsymbol{h}) = \prod_{m=1}^{M}\mathcal{BE}(x_m\vert\sigma(\hat{a}_m)),\ \hat{a}_m = a_m + \boldsymbol{W}_{m,:}\boldsymbol{h}$ $p(\boldsymbol{h}\vert\boldsymbol{x}) = \prod_{k=1}^{K}\mathcal{BE}(h_k\vert\sigma(\hat{b}_k)),\ \hat{b}_k = b_k + (\boldsymbol{W}_{:,k})^{\mathrm{T}}\boldsymbol{x}$	$\Delta_{\boldsymbol{a}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{x}_n\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{x}_s\right)$ $\Delta_{\boldsymbol{b}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{h}_n\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{h}_s\right)$ $\Delta_{\boldsymbol{W}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{x}_n\boldsymbol{h}_n^{\mathrm{T}}\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{x}_s\boldsymbol{h}_s^{\mathrm{T}}\right)$
Gaussian-Bernoulli	$E(\boldsymbol{x},\boldsymbol{h}) = -\boldsymbol{a}^{(1)\mathrm{T}}\boldsymbol{x} - \boldsymbol{a}^{(2)\mathrm{T}}\boldsymbol{x}^{*\boldsymbol{2}} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{h} - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{h}$, where $\boldsymbol{a}^{(1)} = [\mu_1\lambda_1, \cdots, \mu_M\lambda_M]^{\mathrm{T}}$, $\boldsymbol{a}^{(2)} = [\frac{-\lambda_1}{2}, \cdots, \frac{-\lambda_M}{2}]^{\mathrm{T}}$, $\boldsymbol{a}^{(2)} < \boldsymbol{0}$, $\boldsymbol{b} = [\log\frac{p_1}{1-p_1}, \cdots, \log\frac{p_K}{1-p_K}]$	$p(\boldsymbol{x}\vert\boldsymbol{h}) = \prod_{m=1}^{M}\mathcal{GS}(x_m\vert -\frac{\hat{a}_m^{(1)}}{2a_m^{(2)}}, (-2a_m^{(2)})^{-1}),\ \hat{a}_m^{(1)} = a_m^{(1)} + \boldsymbol{W}_{m,:}\boldsymbol{h}$ $p(\boldsymbol{h}\vert\boldsymbol{x}) = \prod_{k=1}^{K}\mathcal{BE}(h_k\vert\sigma(\hat{b}_k)),\ \hat{b}_k = b_k + (\boldsymbol{W}_{:,k})^{\mathrm{T}}\boldsymbol{x}$	$\Delta_{\boldsymbol{a}^{(1)}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{x}_n\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{x}_s\right)$ $\Delta_{\boldsymbol{a}^{(2)}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{x}_n^{*2}\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{x}_s^{*2}\right)$ $\Delta_{\boldsymbol{b}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{h}_n\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{h}_s\right)$ $\Delta_{\boldsymbol{W}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{x}_n\boldsymbol{h}_n^{\mathrm{T}}\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{x}_s\boldsymbol{h}_s^{\mathrm{T}}\right)$
Gaussian-Bernoulli fix precision $\boldsymbol{\lambda}$, model 1	$E(\boldsymbol{x},\boldsymbol{h}) = -\boldsymbol{a}^{\mathrm{T}}(\boldsymbol{\lambda}*\boldsymbol{x}) - \sum_{m=1}^{M}(\log\frac{\sqrt{\lambda_m}}{\sqrt{2\pi}} - \frac{\lambda_m}{2}x_m^2) - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{h} - (\boldsymbol{\lambda}*\boldsymbol{x})^{\mathrm{T}}\boldsymbol{W}\boldsymbol{h}$, where $\boldsymbol{a} = [\mu_1, \cdots, \mu_M]^{\mathrm{T}}$, $\boldsymbol{b} = [\log\frac{p_1}{1-p_1}, \cdots, \log\frac{p_K}{1-p_K}]$	$p(\boldsymbol{x}\vert\boldsymbol{h}) = \prod_{m=1}^{M}\mathcal{GS}(x_m\vert\hat{a}_m, \lambda_m),\ \hat{a}_m = a_m + \boldsymbol{W}_{m,:}\boldsymbol{h}$ $p(\boldsymbol{h}\vert\boldsymbol{x}) = \prod_{k=1}^{K}\mathcal{BE}(h_k\vert\sigma(\hat{b}_k)),\ \hat{b}_k = b_k + (\boldsymbol{W}_{:,k})^{\mathrm{T}}(\boldsymbol{\lambda}*\boldsymbol{x})$	$\Delta_{\boldsymbol{a}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{\lambda}*\boldsymbol{x}_n\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{\lambda}*\boldsymbol{x}_s\right)$ $\Delta_{\boldsymbol{b}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{h}_n\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{h}_s\right)$ $\Delta_{\boldsymbol{W}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-(\boldsymbol{\lambda}*\boldsymbol{x}_n)\boldsymbol{h}_n^{\mathrm{T}}\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-(\boldsymbol{\lambda}*\boldsymbol{x}_s)\boldsymbol{h}_s^{\mathrm{T}}\right)$
Gaussian-Bernoulli fix precision $\boldsymbol{\lambda}$, model 2	$E(\boldsymbol{x},\boldsymbol{h}) = -\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} - \sum_{m=1}^{M}(\log\frac{\sqrt{\lambda_m}}{\sqrt{2\pi}} - \frac{\lambda_m}{2}x_m^2) - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{h} - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{h}$, where $\boldsymbol{a} = [\mu_1\lambda_1, \cdots, \mu_M\lambda_M]^{\mathrm{T}}$, $\boldsymbol{b} = [\log\frac{p_1}{1-p_1}, \cdots, \log\frac{p_K}{1-p_K}]$	$p(\boldsymbol{x}\vert\boldsymbol{h}) = \prod_{m=1}^{M}\mathcal{GS}(x_m\vert\frac{\hat{a}_m}{\lambda_m}, \lambda_m),\ \hat{a}_m = a_m + \boldsymbol{W}_{m,:}\boldsymbol{h}$ $p(\boldsymbol{h}\vert\boldsymbol{x}) = \prod_{k=1}^{K}\mathcal{BE}(h_k\vert\sigma(\hat{b}_k)),\ \hat{b}_k = b_k + (\boldsymbol{W}_{:,k})^{\mathrm{T}}\boldsymbol{x}$	$\Delta_{\boldsymbol{a}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{x}_n\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{x}_s\right)$ $\Delta_{\boldsymbol{b}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{h}_n\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{h}_s\right)$ $\Delta_{\boldsymbol{W}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{x}_n\boldsymbol{h}_n^{\mathrm{T}}\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{x}_s\boldsymbol{h}_s^{\mathrm{T}}\right)$
Poisson-Bernoulli	$E(\boldsymbol{x},\boldsymbol{h}) = -\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} + \sum_{m=1}^{M}\log(x_m!) - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{h} - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{h}$, where $\boldsymbol{a} = [\log\lambda_1, \cdots, \log\lambda_M]^{\mathrm{T}}$, $\boldsymbol{b}_k = [\log\frac{p_1}{1-p_1}, \cdots, \log\frac{p_K}{1-p_K}]^{\mathrm{T}}$	$p(\boldsymbol{x}\vert\boldsymbol{h}) = \prod_{m=1}^{M}\mathcal{PO}(x_m\vert e^{\hat{a}_m}),\ \hat{a}_m = a_m + \boldsymbol{W}_{m,:}\boldsymbol{h}$ $p(\boldsymbol{h}\vert\boldsymbol{x}) = \prod_{k=1}^{K}\mathcal{BE}(h_k\vert\sigma(\hat{b}_k)),\ \hat{b}_k = b_k + (\boldsymbol{W}_{:,k})^{\mathrm{T}}\boldsymbol{x}$	$\Delta_{\boldsymbol{a}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{x}_n\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{x}_s\right)$ $\Delta_{\boldsymbol{b}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{h}_n\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{h}_s\right)$ $\Delta_{\boldsymbol{W}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{x}_n\boldsymbol{h}_n^{\mathrm{T}}\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{x}_s\boldsymbol{h}_s^{\mathrm{T}}\right)$
Poisson-Binomial	$E(\boldsymbol{x},\boldsymbol{h}) = -\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} + \sum_{m=1}^{M}\log(x_m!) - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{h} - \sum_{k=1}^{K}\log\frac{n!}{h_k!(n-h_k!)} - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{h}$, where $\boldsymbol{a} = [\log\lambda_1, \cdots, \log\lambda_M]^{\mathrm{T}}$, $\boldsymbol{b} = [\log\frac{p_1}{1-p_1}, \cdots, \log\frac{p_K}{1-p_K}]^{\mathrm{T}}$	$p(\boldsymbol{x}\vert\boldsymbol{h}) = \prod_{m=1}^{M}\mathcal{PO}(x_m\vert e^{\hat{a}_m}),\ \hat{a}_m = a_m + \boldsymbol{W}_{m,:}\boldsymbol{h}$ $p(\boldsymbol{h}\vert\boldsymbol{x}) = \prod_{k=1}^{K}\mathcal{BN}(n_k, \sigma(\hat{b}_k)),\ \hat{b}_k = b_k + (\boldsymbol{W}_{:,k})^{\mathrm{T}}\boldsymbol{x}$	$\Delta_{\boldsymbol{a}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{x}_n\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{x}_s\right)$ $\Delta_{\boldsymbol{b}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{h}_n\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{h}_s\right)$ $\Delta_{\boldsymbol{W}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{x}_n\boldsymbol{h}_n^{\mathrm{T}}\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{x}_s\boldsymbol{h}_s^{\mathrm{T}}\right)$
Negative-Binomial-Bernoulli	$E(\boldsymbol{x},\boldsymbol{h}) = -\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} - \sum_{m=1}^{M}\log\frac{(x_m+s_m-1)!}{(s_m-1)!x_m!} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{h} - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{h}$, where $\boldsymbol{a} = [\log p_1, \cdots, \log p_M]^{\mathrm{T}}$, $\boldsymbol{a} < \boldsymbol{0}$ $\boldsymbol{b} = [\log\frac{p_1}{1-p_1}, \cdots, \log\frac{p_K}{1-p_K}]^{\mathrm{T}}$	$p(\boldsymbol{x}\vert\boldsymbol{h}) = \prod_{m=1}^{M}\mathcal{NB}(x_m\vert s_m, 1-e^{\hat{a}_m}),\ \hat{a}_m = a_m + \boldsymbol{W}_{m,:}\boldsymbol{h} < 0$ $p(\boldsymbol{h}\vert\boldsymbol{x}) = \prod_{k=1}^{K}\mathcal{BE}(\sigma(\hat{b}_k)),\ \hat{b}_k = b_k + (\boldsymbol{W}_{:,k})^{\mathrm{T}}\boldsymbol{x}$	$\Delta_{\boldsymbol{a}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{x}_n\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{x}_s\right)$ $\Delta_{\boldsymbol{b}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{h}_n\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{h}_s\right)$ $\Delta_{\boldsymbol{W}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{x}_n\boldsymbol{h}_n^{\mathrm{T}}\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{x}_s\boldsymbol{h}_s^{\mathrm{T}}\right)$
Multinoulli-Bernoulli	$E(\boldsymbol{x},\boldsymbol{h}) = -\sum_{m=1}^{M}\boldsymbol{a}^{(m)\mathrm{T}}\boldsymbol{x}^{(m)} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{h} - \sum_{m=1}^{M}(\boldsymbol{x}^{(m)})^{\mathrm{T}}\boldsymbol{W}^{(m)}\boldsymbol{h}$, where $\boldsymbol{a}^{(m)} = [\log p_1^{(m)} + \log C^{(m)}, \cdots, \log p_{C_m}^{(m)} + \log C^{(m)}]^{\mathrm{T}}$, $C_m = \mathrm{size}(\boldsymbol{x}^{(m)})$, $C^{(m)} = \sum_{c'=1}^{C_m}\exp(a_{c'}^{(m)})$, $\boldsymbol{b} = [\log\frac{p_1}{1-p_1}, \cdots, \log\frac{p_K}{1-p_K}]^{\mathrm{T}}$, $\boldsymbol{W}^{(m)} \in \mathbb{R}^{C_m\times K}$	$p(\boldsymbol{x}^{(m)}\vert\boldsymbol{h}) = \mathcal{MU}(\boldsymbol{x}^{(m)}\vert\hat{\boldsymbol{p}}^{(m)}),\ \hat{a}_c^{(m)} = a_c^{(m)} + \boldsymbol{W}_{c,:}^{(m)}\boldsymbol{h},\ \hat{\boldsymbol{p}}^{(m)} = [\frac{\exp(\hat{a}_1^{(m)})}{\sum_{c'=1}^{C_m}\exp(\hat{a}_{c'}^{(m)})}, \cdots, \frac{\exp(\hat{a}_{C_m}^{(m)})}{\sum_{c'=1}^{C_m}\exp(\hat{a}_{c'}^{(m)})}]^{\mathrm{T}}$ $p(\boldsymbol{h}\vert\boldsymbol{x}) = \prod_{k=1}^{K}\mathcal{BE}(\sigma(\hat{b}_k)),\ \hat{b}_k = b_k + \sum_{m=1}^{M}(\boldsymbol{W}_{:,k}^{(m)})^{\mathrm{T}}\boldsymbol{x}$	$\Delta_{\boldsymbol{a}^{(m)}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{x}_n^{(m)}\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{x}_s^{(m)}\right)$ $\Delta_{\boldsymbol{b}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{h}_n\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{h}_s\right)$ $\Delta_{\boldsymbol{W}^{(m)}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{x}_n^{(m)}\boldsymbol{h}_n^{\mathrm{T}}\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{x}_s^{(m)}\boldsymbol{h}_s^{\mathrm{T}}\right)$
Multinomial-Bernoulli	$E(\boldsymbol{x},\boldsymbol{h}) = -\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x} - \log\frac{n!}{\prod_{m=1}^{M}x_m!} - \boldsymbol{b}^{\mathrm{T}}\boldsymbol{h} - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{h}$, where $\boldsymbol{a} = [\log p_1 + \log C, \cdots, \log p_M + \log C]^{\mathrm{T}}$, $\boldsymbol{b} = [\log\frac{p_1}{1-p_1}, \cdots, \log\frac{p_K}{1-p_K}]^{\mathrm{T}}$	$p(\boldsymbol{x}\vert\boldsymbol{h}) = \mathcal{MN}(\boldsymbol{x}\vert n, \hat{\boldsymbol{p}}),\ \hat{a}_m = a_m + \boldsymbol{W}_{m,:}\boldsymbol{h},\ \hat{p}_m = \frac{e^{\hat{a}_m}}{\sum_{m'=1}^{M}e^{\hat{a}_{m'}}}$ $p(\boldsymbol{h}\vert\boldsymbol{x}) = \prod_{k=1}^{K}\mathcal{BE}(\sigma(\hat{b}_k)),\ \hat{b}_k = b_k + (\boldsymbol{W}_{:,k})^{\mathrm{T}}\boldsymbol{x}$	$\Delta_{\boldsymbol{a}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{x}_n\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{x}_s\right)$ $\Delta_{\boldsymbol{b}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{h}_n\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{h}_s\right)$ $\Delta_{\boldsymbol{W}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(-\boldsymbol{x}_n\boldsymbol{h}_n^{\mathrm{T}}\right) - \frac{1}{S}\sum_{s=1}^{S}\left(-\boldsymbol{x}_s\boldsymbol{h}_s^{\mathrm{T}}\right)$

Table 3: Estimatation of log-partition functions of exp-RBMs.

Model	$F(\boldsymbol{x})$	$F_t(\boldsymbol{x})$ in $p_t(\boldsymbol{x})$	$T(\boldsymbol{x}_t \vert \boldsymbol{x}_{t-1})$	$\log Z_A$
Bernoulli-Bernoulli	$-\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}$ $-\sum_{k=1}^{K}\log(1+e^{b_k+\boldsymbol{W}^{\mathrm{T}}{}_{k,:}\boldsymbol{x}})$	$-\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}$ $-\sum_{k=1}^{K}\left(\log(1+e^{(1-\beta_t)b_k})+\log(1+e^{\beta_t(b_k+\boldsymbol{W}^{\mathrm{T}}{}_{k,:}\boldsymbol{x})})\right)$	$p_t((x_t)_m \vert \boldsymbol{h}_{t-1}^{(A)}, \boldsymbol{h}_{t-1}^{(B)}) = \mathcal{BE}(x_m \vert \sigma((1-\beta_t)a_m+\beta_t(a_m+\boldsymbol{W}_{m,:}\boldsymbol{h}_{t-1}^{(B)})))$ $p_t((h_t^{(A)})_k \vert \boldsymbol{x}_t) = \mathcal{BE}((h_t^{(A)})_k \vert \sigma((1-\beta_t)b_k))$ $p_t((h_t^{(B)})_k \vert \boldsymbol{x}_t) = \mathcal{BE}((h_t^{(B)})_k \vert \sigma(\beta_t(b_k+\boldsymbol{W}^{\mathrm{T}}{}_{k,:}\boldsymbol{x}_t)))$	$\sum_{m=1}^{M}\log(1+e^{a_m})+\sum_{k=1}^{K}\log(1+e^{b_k})$
Gaussian-Bernoulli	$-\boldsymbol{a}^{(1)\mathrm{T}}\boldsymbol{x}-\boldsymbol{a}^{(2)\mathrm{T}}\boldsymbol{x}^{*2}$ $-\sum_{k=1}^{K}\log(1+e^{b_k+\boldsymbol{W}^{\mathrm{T}}{}_{k,:}\boldsymbol{x}})$	$-\boldsymbol{a}^{(1)\mathrm{T}}\boldsymbol{x}-\boldsymbol{a}^{(2)\mathrm{T}}\boldsymbol{x}^{*2}$ $-\sum_{k=1}^{K}\left(\log(1+e^{(1-\beta_t)b_k})+\log(1+e^{\beta_t(b_k+\boldsymbol{W}^{\mathrm{T}}{}_{k,:}\boldsymbol{x})})\right)$	$p_t((x_t)_m \vert \boldsymbol{h}_{t-1}^{(A)}, \boldsymbol{h}_{t-1}^{(B)}) = \mathcal{GS}\Big((x_t)_m \vert -\frac{(1-\beta_t)a_m^{(1)}+\beta_t(a_m^{(1)}+\boldsymbol{W}_{m,:}\boldsymbol{h}_{t-1}^{(B)})}{a_m^{(2)}}, (-2a_m^{(2)})^{-1}\Big)$ $p_t((h_t^{(A)})_k \vert \boldsymbol{x}_t) = \mathcal{BE}((h_t^{(A)})_k \vert \sigma((1-\beta_t)b_k))$ $p_t((h_t^{(B)})_k \vert \boldsymbol{x}_t) = \mathcal{BE}((h_t^{(B)})_k \vert \sigma(\beta_t(b_k+\boldsymbol{W}^{\mathrm{T}}{}_{k,:}\boldsymbol{x}_t)))$	$\sum_{m=1}^{M}(\frac{1}{2}\log\frac{\pi}{-a_m^{(2)}}-\frac{a_m^{(1)2}}{4a_m^{(2)}})+\sum_{k=1}^{K}\log(1+e^{b_k})$
Gaussian-Bernoulli fix precision, model 1	$-\boldsymbol{a}^{\mathrm{T}}(\boldsymbol{\lambda}*\boldsymbol{x})-\sum_{m=1}^{M}\left(\log\frac{\sqrt{\lambda_m}}{\sqrt{2\pi}}-\frac{\lambda_m}{2}x_m^2\right)$ $-\sum_{k=1}^{K}\log(1+e^{b_k+\boldsymbol{W}^{\mathrm{T}}{}_{k,:}(\boldsymbol{\lambda}*\boldsymbol{x})})$	$-\boldsymbol{a}^{\mathrm{T}}(\boldsymbol{\lambda}*\boldsymbol{x})-\sum_{m=1}^{M}\left(\log\frac{\sqrt{\lambda_m}}{\sqrt{2\pi}}-\frac{\lambda_m}{2}x_m^2\right)$ $-\sum_{k=1}^{K}\left(\log(1+e^{(1-\beta_t)b_k})+\log(1+e^{\beta_t(b_k+\boldsymbol{W}^{\mathrm{T}}{}_{k,:}(\boldsymbol{\lambda}*\boldsymbol{x})})\right)$	$p_t((x_t)_m \vert \boldsymbol{h}_{t-1}^{(A)}, \boldsymbol{h}_{t-1}^{(B)}) = \mathcal{GS}\Big((x_t)_m \vert (1-\beta_t)a_m+\beta_t(a_m+\boldsymbol{W}_{m,:}\boldsymbol{h}_{t-1}^{(B)}), \lambda_m\Big)$ $p_t((h_t^{(A)})_k \vert \boldsymbol{x}_t) = \mathcal{BE}((h_t^{(A)})_k \vert \sigma((1-\beta_t)b_k))$ $p_t((h_t^{(B)})_k \vert \boldsymbol{x}_t) = \mathcal{BE}((h_t^{(B)})_k \vert \sigma(\beta_t(b_k+\boldsymbol{W}^{\mathrm{T}}{}_{k,:}(\boldsymbol{\lambda}*\boldsymbol{x}_t))))$	$\sum_{m=1}^{M}(\frac{\lambda_m}{2}a_m^2)+\sum_{k=1}^{K}\log(1+e^{b_k})$
Gaussian-Bernoulli fix precision, model 2	$-\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}-\sum_{m=1}^{M}\left(\log\frac{\sqrt{\lambda_m}}{\sqrt{2\pi}}-\frac{\lambda_m}{2}x_m^2\right)$ $-\sum_{k=1}^{K}\log(1+e^{b_k+\boldsymbol{W}^{\mathrm{T}}{}_{k,:}\boldsymbol{x}})$	$-\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}-\sum_{m=1}^{M}\left(\log\frac{\sqrt{\lambda_m}}{\sqrt{2\pi}}-\frac{\lambda_m}{2}x_m^2\right)$ $-\sum_{k=1}^{K}\left(\log(1+e^{(1-\beta_t)b_k})+\log(1+e^{\beta_t(b_k+\boldsymbol{W}^{\mathrm{T}}{}_{k,:}\boldsymbol{x})})\right)$	$p_t((x_t)_m \vert \boldsymbol{h}_{t-1}^{(A)}, \boldsymbol{h}_{t-1}^{(B)}) = \mathcal{GS}\Big((x_t)_m \vert \frac{(1-\beta_t)a_m+\beta_t(a_m+\boldsymbol{W}_{m,:}\boldsymbol{h}_{t-1}^{(B)})}{\lambda}, \lambda_m\Big)$ $p_t((h_t^{(A)})_k \vert \boldsymbol{x}_t) = \mathcal{BE}((h_t^{(A)})_k \vert \sigma((1-\beta_t)b_k))$ $p_t((h_t^{(B)})_k \vert \boldsymbol{x}_t) = \mathcal{BE}((h_t^{(B)})_k \vert \sigma(\beta_t(b_k+\boldsymbol{W}^{\mathrm{T}}{}_{k,:}\boldsymbol{x}_t)))$	$\sum_{m=1}^{M}(\frac{a_m^2}{2\lambda_m})+\sum_{k=1}^{K}\log(1+e^{b_k})$
Poisson-Bernoulli	$-\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}+\sum_{m=1}^{M}\log(x_m!)$ $-\sum_{k=1}^{K}\log(1+e^{b_k+\boldsymbol{W}^{\mathrm{T}}{}_{k,:}\boldsymbol{x}})$	$-\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}+\sum_{m=1}^{M}\log(x_m!)$ $-\sum_{k=1}^{K}\left(\log(1+e^{(1-\beta_t)b_k})+\log(1+e^{\beta_t(b_k+\boldsymbol{W}^{\mathrm{T}}{}_{k,:}\boldsymbol{x})})\right)$	$p_t((x_t)_m \vert \boldsymbol{h}_{t-1}^{(A)}, \boldsymbol{h}_{t-1}^{(B)}) = \mathcal{PO}((x_t)_m \vert e^{(1-\beta_t)a_m+\beta_t(a_m+\boldsymbol{W}_{m,:}\boldsymbol{h}_{t-1})})$ $p_t((h_t^{(A)})_k \vert \boldsymbol{x}_t) = \mathcal{BE}((h_t^{(A)})_k \vert \sigma((1-\beta_t)b_k))$ $p_t((h_t^{(B)})_k \vert \boldsymbol{x}_t) = \mathcal{BE}((h_t^{(B)})_k \vert \sigma(\beta_t(b_k+\boldsymbol{W}^{\mathrm{T}}{}_{k,:}\boldsymbol{x}_t)))$	$\sum_{m=1}^{M}e^{a_m}+\sum_{k=1}^{K}\log(1+e^{b_k})$
Poisson-Binomial	$-\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}+\sum_{m=1}^{M}\log(x_m!)$ $-\sum_{k=1}^{K}n_k\log(1+e^{b_k+\boldsymbol{W}^{\mathrm{T}}{}_{k,:}\boldsymbol{x}})$	$-\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}+\sum_{m=1}^{M}\log(x_m!)$ $-\sum_{k=1}^{K}\left(n_k\log(1+e^{(1-\beta_t)b_k})+n_k\log(1+e^{\beta_t(b_k+\boldsymbol{W}^{\mathrm{T}}{}_{k,:}\boldsymbol{x})})\right)$	$p_t((x_t)_m \vert \boldsymbol{h}_{t-1}^{(A)}, \boldsymbol{h}_{t-1}^{(B)}) = \mathcal{PO}((x_t)_m \vert e^{(1-\beta_t)a_m+\beta_t(a_m+\boldsymbol{W}_{m,:}\boldsymbol{h}_{t-1})})$ $p_t((h_t^{(A)})_k \vert \boldsymbol{x}_t) = \mathcal{BN}((h_t^{(A)})_k \vert n_k, \sigma((1-\beta_t)b_k))$ $p_t((h_t^{(B)})_k \vert \boldsymbol{x}_t) = \mathcal{BN}((h_t^{(B)})_k \vert n_k, \sigma(\beta_t(b_k+\boldsymbol{W}^{\mathrm{T}}{}_{k,:}\boldsymbol{x}_t)))$	$\sum_{m=1}^{M}e^{a_m}+\sum_{k=1}^{K}n_k\log(1+e^{b_k})$
Negative-Binomial-Bernoulli	$-\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}-\sum_{m=1}^{M}\log\frac{(x_m+s_m-1)!)}{(s_m-1)!x_m!}$ $-\sum_{k=1}^{K}\log(1+e^{b_k+\boldsymbol{W}^{\mathrm{T}}{}_{k,:}\boldsymbol{x}})$	$-\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}-\sum_{m=1}^{M}\log\frac{(x_m+s_m-1)!)}{(s_m-1)!x_m!}$ $-\sum_{k=1}^{K}\left(\log(1+e^{(1-\beta_t)b_k})+\log(1+e^{\beta_t(b_k+\boldsymbol{W}^{\mathrm{T}}{}_{k,:}\boldsymbol{x})})\right)$	$p_t((x_t)_m \vert \boldsymbol{h}_{t-1}^{(A)}, \boldsymbol{h}_{t-1}^{(B)}) = \mathcal{NB}((x_t)_m \vert s_m, 1-e^{(1-\beta_t)a_m+\beta_t(a_m+\boldsymbol{W}_{m,:}\boldsymbol{h}_{t-1})})$ $p_t((h_t^{(A)})_k \vert \boldsymbol{x}_t) = \mathcal{BE}((h_t^{(A)})_k \vert \sigma((1-\beta_t)b_k))$ $p_t((h_t^{(B)})_k \vert \boldsymbol{x}_t) = \mathcal{BE}((h_t^{(B)})_k \vert \sigma(\beta_t(b_k+\boldsymbol{W}^{\mathrm{T}}{}_{k,:}\boldsymbol{x}_t)))$	$\sum_{m=1}^{M}-s_m\log(1-e^{a_m})+\sum_{k=1}^{K}\log(1+e^{b_k})$
Multinoulli-Bernoulli	$-\sum_{m=1}^{M}\boldsymbol{a}^{(m)\mathrm{T}}\boldsymbol{x}^{(m)}$ $-\sum_{k=1}^{K}\log(1+e^{b_k+\sum_{m=1}^{M}(\boldsymbol{W}^{(m)\mathrm{T}})_{k,:}\boldsymbol{x}^{(m)}})$	$-\sum_{m=1}^{M}\boldsymbol{a}^{(m)\mathrm{T}}\boldsymbol{x}^{(m)}$ $-\sum_{k=1}^{K}\Big(\log(1+e^{(1-\beta_t)b_k})$ $+\log(1+e^{\beta_t(b_k+\sum_{m=1}^{M}(\boldsymbol{W}^{(m)\mathrm{T}})_{k,:}\boldsymbol{x}^{(m)})})\Big)$	$p_t(\boldsymbol{x}_t^{(m)} \vert \boldsymbol{h}_{t-1}^{(A)}, \boldsymbol{h}_{t-1}^{(B)}) = \mathcal{MU}\Big(\boldsymbol{x}_t^{(m)} \vert \big[\frac{e^{(1-\beta_t)a_c^{(m)}+\beta_t(a_c^{(m)}+\boldsymbol{W}_{c,:}^{(m)}\boldsymbol{h}_{t-1})}}{\sum_{c'=1}^{C_m}e^{(1-\beta_t)a_{c'}^{(m)}+\beta_t(a_{c'}^{(m)}+\boldsymbol{W}_{c',:}^{(m)}\boldsymbol{h}_{t-1})}}\big]\Big)$ $p_t((h_t^{(A)})_k \vert \boldsymbol{x}_t) = \mathcal{BE}((h_t^{(A)})_k \vert \sigma((1-\beta_t)b_k))$ $p_t((h_t^{(B)})_k \vert \boldsymbol{x}_t) = \mathcal{BE}((h_t^{(B)})_k \vert \sigma(\beta_t(b_k+\sum_{m=1}^{M}(\boldsymbol{W}^{(m)\mathrm{T}})_{k,:}\boldsymbol{x}_t^{(m)})))$	$\sum_{m=1}^{M}\log\sum_{c'=1}^{C_m}e^{a_{c'}^{(m)}}+\sum_{k=1}^{K}\log(1+e^{b_k})$
Multinomial-Bernoulli	$-\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}-\log\frac{n!}{\prod_{m=1}^{M}x_m!}$ $-\sum_{k=1}^{K}\log(1+e^{b_k+\boldsymbol{W}^{\mathrm{T}}{}_{k,:}\boldsymbol{x}})$	$-\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}-\log\frac{n!}{\prod_{m=1}^{M}x_m!}$ $-\sum_{k=1}^{K}\left(\log(1+e^{(1-\beta_t)b_k})+\log(1+e^{\beta_t(b_k+\boldsymbol{W}^{\mathrm{T}}{}_{k,:}\boldsymbol{x})})\right)$	$p_t(\boldsymbol{x}_t \vert \boldsymbol{h}_{t-1}^{(A)}, \boldsymbol{h}_{t-1}^{(B)}) = \mathcal{MN}\Big(\boldsymbol{x}_t \vert n, \big[\frac{e^{(1-\beta_t)a_m+\beta_t(a_m+\boldsymbol{W}_{m,:}\boldsymbol{h}_{t-1})}}{\sum_{m'=1}^{M}e^{(1-\beta_t)a_{m'}+\beta_t(a_{m'}+\boldsymbol{W}_{m',:}\boldsymbol{h}_{t-1})}}\big]\Big)$ $p_t((h_t^{(A)})_k \vert \boldsymbol{x}_t) = \mathcal{BE}((h_t^{(A)})_k \vert \sigma((1-\beta_t)b_k))$ $p_t((h_t^{(B)})_k \vert \boldsymbol{x}_t) = \mathcal{BE}((h_t^{(B)})_k \vert \sigma(\beta_t(b_k+\boldsymbol{W}^{\mathrm{T}}{}_{k,:}\boldsymbol{x}_t)))$	$n\log\sum_{m=1}^{M}e^{a_m}+\sum_{k=1}^{K}\log(1+e^{b_k})$

Table 4: Selected instances of exp-HM and exp-DBN.

Model	$p(\boldsymbol{x}\|\boldsymbol{h}^{(1)})$	Gradient for Exp-HMs	Gradient for Exp-DBNs
Bernoulli-Bernoulli	$\prod_{m=1}^{M}\mathcal{BE}(x_m\|\sigma(\hat{a}_m))$, $\hat{a}_m^{(G)} = a_m^{(G)} + \boldsymbol{W}_{m,:}^{(G,1)}\boldsymbol{h}^{(1)}$	$\Delta_{\boldsymbol{a}^{(G)}} \approx -(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)$, $\langle\boldsymbol{x}\rangle = \sigma(\hat{\boldsymbol{a}}^{(G)})$, $\hat{\boldsymbol{a}}^{(G)} = \boldsymbol{a}^{(G)} + \boldsymbol{W}^{(G,1)}\boldsymbol{h}^{(1)}$ $\Delta_{\boldsymbol{b}^{(G,l)}} \approx -(\boldsymbol{h}^{(l)} - \langle\boldsymbol{h}^{(l)}\rangle)$, $\langle\boldsymbol{h}^{(l)}\rangle = \sigma(\hat{\boldsymbol{b}}^{(G,l)})$, $\hat{\boldsymbol{b}}^{(G,l)} = \boldsymbol{b}_k^{(G,l)} + \boldsymbol{W}^{(G,l+1)}\boldsymbol{h}^{(l+1)}$, $1 \le l \le L-1$ $\Delta_{\boldsymbol{b}^{(G,L)}} \approx -(\boldsymbol{h}^{(L)} - \langle\boldsymbol{h}^{(L)}\rangle)$, $\langle\boldsymbol{h}^{(L)}\rangle = \sigma(\boldsymbol{b}^{(G,L)})$ $\Delta_{\boldsymbol{W}^{(G,1)}} \approx -(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)\boldsymbol{h}^{(1)\mathrm{T}}$ $\Delta_{\boldsymbol{W}^{(G,l)}} \approx -(\boldsymbol{h}^{(l-1)} - \langle\boldsymbol{h}^{(l-1)}\rangle)\boldsymbol{h}^{(l)\mathrm{T}}$, $2 \le l \le L$ $\Delta_{\boldsymbol{b}^{(R,1)}} \approx -(\boldsymbol{h}^{(1)} - \langle\boldsymbol{h}^{(1)}\rangle)$, $\langle\boldsymbol{h}^{(1)}\rangle = \sigma(\hat{\boldsymbol{b}}^{(R,1)})$, $\hat{\boldsymbol{b}}^{(R,1)} = \boldsymbol{b}^{(R,1)} + \boldsymbol{W}^{(R,1)}\boldsymbol{x}$ $\Delta_{\boldsymbol{b}^{(R,l)}} \approx -(\boldsymbol{h}^{(l)} - \langle\boldsymbol{h}^{(l)}\rangle)$, $\langle\boldsymbol{h}^{(l)}\rangle = \sigma(\hat{\boldsymbol{b}}^{(R,l)})$, $\hat{\boldsymbol{b}}^{(R,l)} = \boldsymbol{b}^{(R,l)} + \boldsymbol{W}^{(R,l)}\boldsymbol{h}^{(l-1)}$, $2 \le l \le L$ $\Delta_{\boldsymbol{W}^{(R,1)}} \approx -(\boldsymbol{h}^{(1)} - \langle\boldsymbol{h}^{(1)}\rangle)\boldsymbol{x}^{\mathrm{T}}$ $\Delta_{\boldsymbol{W}^{(R,l)}} \approx -(\boldsymbol{h}^{(l)} - \langle\boldsymbol{h}^{(l)}\rangle)\boldsymbol{h}^{(l-1)\mathrm{T}}$, $2 \le l \le L$	$\Delta_{\boldsymbol{a}^{(G)}} \approx -(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)$, $\langle\boldsymbol{x}\rangle = \sigma(\hat{\boldsymbol{a}}^{(G)})$, $\hat{\boldsymbol{a}}^{(G)} = \boldsymbol{a}^{(G)} + \boldsymbol{W}^{(G,1)}\boldsymbol{h}^{(1)}$ $\Delta_{\boldsymbol{b}^{(G,l)}} \approx -(\boldsymbol{h}^{(l)} - \langle\boldsymbol{h}^{(l)}\rangle)$, $\langle\boldsymbol{h}^{(l)}\rangle = \sigma(\hat{\boldsymbol{b}}^{(G,l)})$, $\hat{\boldsymbol{b}}^{(G,l)} = \boldsymbol{b}_k^{(G,l)} + \boldsymbol{W}^{(G,l+1)}\boldsymbol{h}^{(l+1)}$, $1 \le l \le L-2$ $\Delta_{\boldsymbol{W}^{(G,1)}} \approx -(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)\boldsymbol{h}^{(1)\mathrm{T}}$ $\Delta_{\boldsymbol{W}^{(G,l)}} \approx -(\boldsymbol{h}^{(l-1)} - \langle\boldsymbol{h}^{(l-1)}\rangle)\boldsymbol{h}^{(l)\mathrm{T}}$, $2 \le l \le L-1$ $\Delta_{\boldsymbol{b}^{(G,L-1)}} \approx -(\boldsymbol{h}^{(G,L-1)} - \tilde{\boldsymbol{h}}^{(G,L-1)})$, $\{\tilde{\boldsymbol{h}}^{(G,L-1)}, \tilde{\boldsymbol{h}}^{(G,L)}\} \sim p(\boldsymbol{h}^{(G,L-1)}, \boldsymbol{h}^{(G,L)})$ $\Delta_{\boldsymbol{b}^{(G,L)}} \approx -(\boldsymbol{h}^{(G,L)} - \tilde{\boldsymbol{h}}^{(G,L)})$ $\Delta_{\boldsymbol{W}^{(G,L)}} \approx -(\boldsymbol{h}^{(L-1)}\boldsymbol{h}^{(L)\mathrm{T}} - \tilde{\boldsymbol{h}}^{(L-1)}\tilde{\boldsymbol{h}}^{(L)\mathrm{T}})$ $\Delta_{\boldsymbol{b}^{(R,1)}} \approx -(\boldsymbol{h}^{(1)} - \langle\boldsymbol{h}^{(1)}\rangle)$, $\langle\boldsymbol{h}^{(1)}\rangle = \sigma(\hat{\boldsymbol{b}}^{(R,1)})$, $\hat{\boldsymbol{b}}^{(R,1)} = \boldsymbol{b}^{(R,1)} + \boldsymbol{W}^{(R,1)}\boldsymbol{x}$ $\Delta_{\boldsymbol{b}^{(R,l)}} \approx -(\boldsymbol{h}^{(l)} - \langle\boldsymbol{h}^{(l)}\rangle)$, $\langle\boldsymbol{h}^{(l)}\rangle = \sigma(\hat{\boldsymbol{b}}^{(R,l)})$, $\hat{\boldsymbol{b}}^{(R,l)} = \boldsymbol{b}^{(R,l)} + \boldsymbol{W}^{(R,l)}\boldsymbol{h}^{(l-1)}$, $2 \le l \le L-1$ $\Delta_{\boldsymbol{W}^{(R,1)}} \approx -(\boldsymbol{h}^{(1)} - \langle\boldsymbol{h}^{(1)}\rangle)\boldsymbol{x}^{\mathrm{T}}$ $\Delta_{\boldsymbol{W}^{(R,l)}} \approx -(\boldsymbol{h}^{(l)} - \langle\boldsymbol{h}^{(l)}\rangle)\boldsymbol{h}^{(l-1)\mathrm{T}}$, $2 \le l \le L-1$
Gaussian-Bernoulli	$\prod_{m=1}^{M}\mathcal{GS}(x_m\| -\frac{\hat{a}_m^{(G,1)}}{2a_m^{(G,2)}}, (-2a_m^{(2)})^{-1})$, $\hat{a}_m^{(G,1)} = a_m^{(G,1)} + \boldsymbol{W}_{m,:}^{(G,1)}\boldsymbol{h}^{(1)}$	$\Delta_{\boldsymbol{a}^{(G,1)}} \approx -(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)$, $\langle\boldsymbol{x}\rangle = -\frac{\hat{\boldsymbol{a}}^{(G,1)}}{2\boldsymbol{a}^{(G,2)}}$, $\hat{\boldsymbol{a}}^{(G,1)} = \boldsymbol{a}^{(G,1)} + \boldsymbol{W}^{(G,1)}\boldsymbol{h}^{(1)}$ $\Delta_{\boldsymbol{a}^{(G,2)}} \approx -(\boldsymbol{x}^{*2} - \langle\boldsymbol{x}^{*2}\rangle)$, $\langle\boldsymbol{x}^{*2}\rangle = \frac{(\hat{\boldsymbol{a}}^{(G,1)})^2}{4(\boldsymbol{a}^{(G,2)})^2} - \frac{1}{2\boldsymbol{a}^{(G,2)}}$ $\Delta_{\boldsymbol{W}^{(G,1)}} \approx -(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)\boldsymbol{h}^{(1)\mathrm{T}}$ $\Delta_{\boldsymbol{b}^{(R,1)}} \approx -(\boldsymbol{h}^{(1)} - \langle\boldsymbol{h}^{(1)}\rangle)$, $\langle\boldsymbol{h}^{(1)}\rangle = \sigma(\hat{\boldsymbol{b}}^{(R,1)})$, $\hat{\boldsymbol{b}}^{(R,1)} = \boldsymbol{b}^{(R,1)} + \boldsymbol{W}^{(R,1)}\boldsymbol{x}$ $\Delta_{\boldsymbol{W}^{(R,1)}} \approx -(\boldsymbol{h}^{(1)} - \langle\boldsymbol{h}^{(1)}\rangle)\boldsymbol{x}^{\mathrm{T}}$ the rest is the same as in Bernoulli-Bernoulli HM	$\Delta_{\boldsymbol{a}^{(G,1)}} \approx -(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)$, $\langle\boldsymbol{x}\rangle = -\frac{\hat{\boldsymbol{a}}^{(G,1)}}{2\boldsymbol{a}^{(G,2)}}$, $\hat{\boldsymbol{a}}^{(G,1)} = \boldsymbol{a}^{(G,1)} + \boldsymbol{W}^{(G,1)}\boldsymbol{h}^{(1)}$ $\Delta_{\boldsymbol{a}^{(G,2)}} \approx -(\boldsymbol{x}^{*2} - \langle\boldsymbol{x}^{*2}\rangle)$, $\langle\boldsymbol{x}^{*2}\rangle = \frac{(\hat{\boldsymbol{a}}^{(G,1)})^2}{4(\boldsymbol{a}^{(G,2)})^2} - \frac{1}{2\boldsymbol{a}^{(G,2)}}$ $\Delta_{\boldsymbol{W}^{(G,1)}} \approx -(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)\boldsymbol{h}^{(1)\mathrm{T}}$ $\Delta_{\boldsymbol{b}^{(R,1)}} \approx -(\boldsymbol{h}^{(1)} - \langle\boldsymbol{h}^{(1)}\rangle)$, $\langle\boldsymbol{h}^{(1)}\rangle = \sigma(\hat{\boldsymbol{b}}^{(R,1)})$, $\hat{\boldsymbol{b}}^{(R,1)} = \boldsymbol{b}^{(R,1)} + \boldsymbol{W}^{(R,1)}\boldsymbol{x}$ $\Delta_{\boldsymbol{W}^{(R,1)}} \approx -(\boldsymbol{h}^{(1)} - \langle\boldsymbol{h}^{(1)}\rangle)\boldsymbol{x}^{\mathrm{T}}$ the rest is the same as in Bernoulli-Bernoulli DBN
Poisson-Bernoulli	$\prod_{m=1}^{M}\mathcal{PO}(x_m\|e^{\hat{a}_m^{(G)}})$, $\hat{a}_m^{G} = a_m^{(G)} + \boldsymbol{W}_{m,:}^{(G,1)}\boldsymbol{h}^{(1)}$	$\Delta_{\boldsymbol{a}^{(G)}} \approx -(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)$, $\langle\boldsymbol{x}\rangle = e^{\hat{\boldsymbol{a}}^{(G)}}$, $\hat{\boldsymbol{a}}^{(G)} = \boldsymbol{a}^{(G)} + \boldsymbol{W}^{(G,1)}\boldsymbol{h}^{(1)}$ $\Delta_{\boldsymbol{W}^{(G,1)}} \approx -(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)\boldsymbol{h}^{(1)\mathrm{T}}$ $\Delta_{\boldsymbol{b}^{(R,1)}} \approx -(\boldsymbol{h}^{(1)} - \langle\boldsymbol{h}^{(1)}\rangle)$, $\langle\boldsymbol{h}^{(1)}\rangle = \sigma(\hat{\boldsymbol{b}}^{(R,1)})$, $\hat{\boldsymbol{b}}^{(R,1)} = \boldsymbol{b}^{(R,1)} + \boldsymbol{W}^{(R,1)}\boldsymbol{x}$ $\Delta_{\boldsymbol{W}^{(R,1)}} \approx -(\boldsymbol{h}^{(1)} - \langle\boldsymbol{h}^{(1)}\rangle)\boldsymbol{x}^{\mathrm{T}}$ the rest is the same as in Bernoulli-Bernoulli HM	$\Delta_{\boldsymbol{a}^{(G)}} \approx -(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)$, $\langle\boldsymbol{x}\rangle = e^{\hat{\boldsymbol{a}}^{(G)}}$, $\hat{\boldsymbol{a}}^{(G)} = \boldsymbol{a}^{(G)} + \boldsymbol{W}^{(G,1)}\boldsymbol{h}^{(1)}$ $\Delta_{\boldsymbol{W}^{(G,1)}} \approx -(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)\boldsymbol{h}^{(1)\mathrm{T}}$ $\Delta_{\boldsymbol{b}^{(R,1)}} \approx -(\boldsymbol{h}^{(1)} - \langle\boldsymbol{h}^{(1)}\rangle)$, $\langle\boldsymbol{h}^{(1)}\rangle = \sigma(\hat{\boldsymbol{b}}^{(R,1)})$, $\hat{\boldsymbol{b}}^{(R,1)} = \boldsymbol{b}^{(R,1)} + \boldsymbol{W}^{(R,1)}\boldsymbol{x}$ $\Delta_{\boldsymbol{W}^{(R,1)}} \approx -(\boldsymbol{h}^{(1)} - \langle\boldsymbol{h}^{(1)}\rangle)\boldsymbol{x}^{\mathrm{T}}$ the rest is the same as in Bernoulli-Bernoulli DBN
Negative-Binomial-Bernoulli	$\prod_{m=1}^{M}\mathcal{NB}(x_m\|s_m, 1-e^{\hat{a}_m^{(G)}})$, $\hat{a}_m^{(G)} = a_m^{(G)} + \boldsymbol{W}_{m,:}^{(G,1)}\boldsymbol{h}^{(1)} < 0$	$\Delta_{\boldsymbol{a}^{(G)}} \approx -(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)$, $\langle\boldsymbol{x}\rangle = s_m\frac{e^{\hat{\boldsymbol{a}}^{(G)}}}{1-e^{\hat{\boldsymbol{a}}^{(G)}}}$, $\hat{\boldsymbol{a}}^{(G)} = \boldsymbol{a}^{(G)} + \boldsymbol{W}^{(G,1)}\boldsymbol{h}^{(1)}$ $\Delta_{\boldsymbol{W}^{(G,1)}} \approx -(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)\boldsymbol{h}^{(1)\mathrm{T}}$ $\Delta_{\boldsymbol{b}^{(R,1)}} \approx -(\boldsymbol{h}^{(1)} - \langle\boldsymbol{h}^{(1)}\rangle)$, $\langle\boldsymbol{h}^{(1)}\rangle = \sigma(\hat{\boldsymbol{b}}^{(R,1)})$, $\hat{\boldsymbol{b}}^{(R,1)} = \boldsymbol{b}^{(R,1)} + \boldsymbol{W}^{(R,1)}\boldsymbol{x}$ $\Delta_{\boldsymbol{W}^{(R,1)}} \approx -(\boldsymbol{h}^{(1)} - \langle\boldsymbol{h}^{(1)}\rangle)\boldsymbol{x}^{\mathrm{T}}$ the rest is the same as in Bernoulli-Bernoulli HM	$\Delta_{\boldsymbol{a}^{(G)}} \approx -(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)$, $\langle\boldsymbol{x}\rangle = s_m\frac{e^{\hat{\boldsymbol{a}}^{(G)}}}{1-e^{\hat{\boldsymbol{a}}^{(G)}}}$, $\hat{\boldsymbol{a}}^{(G)} = \boldsymbol{a}^{(G)} + \boldsymbol{W}^{(G,1)}\boldsymbol{h}^{(1)}$ $\Delta_{\boldsymbol{W}^{(G,1)}} \approx -(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)\boldsymbol{h}^{(1)\mathrm{T}}$ $\Delta_{\boldsymbol{b}^{(R,1)}} \approx -(\boldsymbol{h}^{(1)} - \langle\boldsymbol{h}^{(1)}\rangle)$, $\langle\boldsymbol{h}^{(1)}\rangle = \sigma(\hat{\boldsymbol{b}}^{(R,1)})$, $\hat{\boldsymbol{b}}^{(R,1)} = \boldsymbol{b}^{(R,1)} + \boldsymbol{W}^{(R,1)}\boldsymbol{x}$ $\Delta_{\boldsymbol{W}^{(R,1)}} \approx -(\boldsymbol{h}^{(1)} - \langle\boldsymbol{h}^{(1)}\rangle)\boldsymbol{x}^{\mathrm{T}}$ the rest is the same as in Bernoulli-Bernoulli DBN
Multinoulli-Bernoulli	$p(\boldsymbol{x}^{(m)}\|\boldsymbol{h}^{(1)}) = \mathcal{MU}(\boldsymbol{x}^{(m)}\|\hat{\boldsymbol{p}}^{(m)})$, $\hat{a}_c^{(G,m)} = a_c^{(G,m)} + \boldsymbol{W}_{c,:}^{(G,m)}\boldsymbol{h}^{(1)}$, $\hat{\boldsymbol{p}}^{(m)} = [\frac{\exp(\hat{a}_1^{(G,m)})}{\sum_{c'=1}^{C_m}\exp(\hat{a}_{c'}^{(G,m)})}, \cdots, \frac{\exp(\hat{a}_{C_m}^{(G,m)})}{\sum_{c'=1}^{C_m}\exp(\hat{a}_{c'}^{(G,m)})}]^{\mathrm{T}}$	$\Delta_{\boldsymbol{a}^{(G,m)}} \approx -(\boldsymbol{x}^{(m)} - \langle\boldsymbol{x}^{(m)}\rangle)$, $\langle\boldsymbol{x}^{(m)}\rangle = \hat{\boldsymbol{p}}^{(G,m)})$, $\hat{\boldsymbol{a}}^{(G,m)} = \boldsymbol{a}^{(G,m)} + \boldsymbol{W}^{(G,1,m)}\boldsymbol{h}^{(1)}$, $\hat{\boldsymbol{p}}^{(m)} = [\frac{\exp(\hat{a}_1^{(G,m)})}{\sum_{c'=1}^{C_m}\exp(\hat{a}_{c'}^{(G,m)})}, \cdots, \frac{\exp(\hat{a}_{C_m}^{(G,m)})}{\sum_{c'=1}^{C_m}\exp(\hat{a}_{c'}^{(G,m)})}]^{\mathrm{T}}$ $\Delta_{\boldsymbol{W}^{(G,1,m)}} \approx -(\boldsymbol{x}^{(m)} - \langle\boldsymbol{x}^{(m)}\rangle)\boldsymbol{h}^{(1)\mathrm{T}}$ $\Delta_{\boldsymbol{b}^{(R,1)}} \approx -(\boldsymbol{h}^{(1)} - \langle\boldsymbol{h}^{(1)}\rangle)$, $\langle\boldsymbol{h}^{(1)}\rangle = \sigma(\hat{\boldsymbol{b}}^{(R,1)})$, $\hat{\boldsymbol{b}}^{(R,1)} = \boldsymbol{b}^{(R,1)} + \sum_{m'=1}^{M}\boldsymbol{W}^{(R,1,m')}\boldsymbol{x}^{(m')}$ $\Delta_{\boldsymbol{W}^{(R,1,m)}} \approx -(\boldsymbol{h}^{(1)} - \langle\boldsymbol{h}^{(1)}\rangle)\boldsymbol{x}^{(m)\mathrm{T}}$ the rest is the same as in Bernoulli-Bernoulli HM	$\Delta_{\boldsymbol{a}^{(G,m)}} \approx -(\boldsymbol{x}^{(m)} - \langle\boldsymbol{x}^{(m)}\rangle)$, $\langle\boldsymbol{x}^{(m)}\rangle = \hat{\boldsymbol{p}}^{(G,m)})$, $\hat{\boldsymbol{a}}^{(G,m)} = \boldsymbol{a}^{(G,m)} + \boldsymbol{W}^{(G,1,m)}\boldsymbol{h}^{(1)}$, $\hat{\boldsymbol{p}}^{(m)} = [\frac{\exp(\hat{a}_1^{(G,m)})}{\sum_{c'=1}^{C_m}\exp(\hat{a}_{c'}^{(G,m)})}, \cdots, \frac{\exp(\hat{a}_{C_m}^{(G,m)})}{\sum_{c'=1}^{C_m}\exp(\hat{a}_{c'}^{(G,m)})}]^{\mathrm{T}}$ $\Delta_{\boldsymbol{W}^{(G,1,m)}} \approx -(\boldsymbol{x}^{(m)} - \langle\boldsymbol{x}^{(m)}\rangle)\boldsymbol{h}^{(1)\mathrm{T}}$ $\Delta_{\boldsymbol{b}^{(R,1)}} \approx -(\boldsymbol{h}^{(1)} - \langle\boldsymbol{h}^{(1)}\rangle)$, $\langle\boldsymbol{h}^{(1)}\rangle = \sigma(\hat{\boldsymbol{b}}^{(R,1)})$, $\hat{\boldsymbol{b}}^{(R,1)} = \boldsymbol{b}^{(R,1)} + \sum_{m'=1}^{M}\boldsymbol{W}^{(R,1,m')}\boldsymbol{x}^{(m')}$ $\Delta_{\boldsymbol{W}^{(R,1,m)}} \approx -(\boldsymbol{h}^{(1)} - \langle\boldsymbol{h}^{(1)}\rangle)\boldsymbol{x}^{(m)\mathrm{T}}$ the rest is the same as in Bernoulli-Bernoulli DBN
Multinomial-Bernoulli	$\mathcal{MN}(\boldsymbol{x}\|n, \hat{\boldsymbol{p}})$, $\hat{a}_m^{(G)} = a_m + \boldsymbol{W}_{m,:}^{(G)}\boldsymbol{h}^{(1)}$, $\hat{p}_m = \frac{\exp(\hat{a}_m^{(G)})}{\sum_{m'=1}^{M}\exp(\hat{a}_{m'}^{(G)})}$	$\Delta_{\boldsymbol{a}^{(G)}} \approx -(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)$, $\langle\boldsymbol{x}\rangle = n\hat{\boldsymbol{p}}$, $\hat{\boldsymbol{a}}^{(G)} = \boldsymbol{a}^{(G)} + \boldsymbol{W}^{(G,1)}\boldsymbol{h}^{(1)}$, $\hat{\boldsymbol{p}} = [\frac{\exp(\hat{a}_1^{(G)})}{\sum_{m'=1}^{M}\exp(\hat{a}_{m'}^{(G)})}, \cdots, \frac{\exp(\hat{a}_M^{(G)})}{\sum_{m'=1}^{M}\exp(\hat{a}_{m'}^{(G)})}]^{\mathrm{T}}$ $\Delta_{\boldsymbol{W}^{(G,1)}} \approx -(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)\boldsymbol{h}^{(1)\mathrm{T}}$ $\Delta_{\boldsymbol{b}^{(R,1)}} \approx -(\boldsymbol{h}^{(1)} - \langle\boldsymbol{h}^{(1)}\rangle)$, $\langle\boldsymbol{h}^{(1)}\rangle = \sigma(\hat{\boldsymbol{b}}^{(R,1)})$, $\hat{\boldsymbol{b}}^{(R,1)} = \boldsymbol{b}^{(R,1)} + \boldsymbol{W}^{(R,1)}\boldsymbol{x}$ $\Delta_{\boldsymbol{W}^{(R,1)}} \approx -(\boldsymbol{h}^{(1)} - \langle\boldsymbol{h}^{(1)}\rangle)\boldsymbol{x}^{\mathrm{T}}$ the rest is the same as in Bernoulli-Bernoulli HM	$\Delta_{\boldsymbol{a}^{(G)}} \approx -(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)$, $\langle\boldsymbol{x}\rangle = n\hat{\boldsymbol{p}}$, $\hat{\boldsymbol{a}}^{(G)} = \boldsymbol{a}^{(G)} + \boldsymbol{W}^{(G,1)}\boldsymbol{h}^{(1)}$, $\hat{\boldsymbol{p}} = [\frac{\exp(\hat{a}_1^{(G)})}{\sum_{m'=1}^{M}\exp(\hat{a}_{m'}^{(G)})}, \cdots, \frac{\exp(\hat{a}_M^{(G)})}{\sum_{m'=1}^{M}\exp(\hat{a}_{m'}^{(G)})}]^{\mathrm{T}}$ $\Delta_{\boldsymbol{W}^{(G,1)}} \approx -(\boldsymbol{x} - \langle\boldsymbol{x}\rangle)\boldsymbol{h}^{(1)\mathrm{T}}$ $\Delta_{\boldsymbol{b}^{(R,1)}} \approx -(\boldsymbol{h}^{(1)} - \langle\boldsymbol{h}^{(1)}\rangle)$, $\langle\boldsymbol{h}^{(1)}\rangle = \sigma(\hat{\boldsymbol{b}}^{(R,1)})$, $\hat{\boldsymbol{b}}^{(R,1)} = \boldsymbol{b}^{(R,1)} + \boldsymbol{W}^{(R,1)}\boldsymbol{x}$ $\Delta_{\boldsymbol{W}^{(R,1)}} \approx -(\boldsymbol{h}^{(1)} - \langle\boldsymbol{h}^{(1)}\rangle)\boldsymbol{x}^{\mathrm{T}}$ the rest is the same as in Bernoulli-Bernoulli DBN