Bayesian Exponential Family Harmoniums

Fan Guo **Eric P. Xing**

May 24, 2005
CMU-ML-06-103

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

A Bayesian Exponential Family Harmonium (BEFH) model is presented for topical modeling of text and multimedia data, and for "posterior" latent semantic projection of such data for subsequent data mining tasks. BEFHs are a Bayesian approach to inference and learning with the recently proposed EFH models and their variants, which enables smoothed, robust estimation of the topic-attribute coupling coefficients that are reminiscent of the smoothed topical word-probabilities in the latent Dirichlet Allocation (LDA) model. The Langevin algorithm conjoint with an MCMC scheme is applied for posterior inference with BEFH. An empirical Bayes method is also developed to estimate the hyperparameters.

Keywords: Bayesian learning, latent semantics indexing, Markov chain Monte Carlo, undirected graphical models

1 Introduction

The vast size of the text and multimedia information available from digital libraries and world-wide-web, and large amount of knowledge contained therein, creates a need to organize and summarize topical contents of these data. In recent years, there is a growing volume of research on applying probabilistic graphical models (GMs) to develop automatic information distillation systems that can explore and exploit real-world data from diverse sources, such as texts, images and biological sequences.

Probabilistic graphical models provide a compact description of complex stochastic relationships among random variables, which can correspond to both perceivable entities (e.g., words, imageries) and abstract concepts (e.g., topics, themes); and such a formalism often facilitates flexible statistical reasoning and query answering based on efficient computational algorithms. Inspired by the classical approach of *latent semantic indexing* (Deerwester et al., 1990), recently there have been important advances in developing latent semantic GMs for large text corpus and/or multimedia data, based on either a Bayesian network (BN) or a Markov random field (MRF) formalism. For instance, the *probabilistic latent semantic indexing* (pLSI) (Hofmann, 1999) method models each document as an admixture of topic-specific distributions of words. The more recent *latent Dirichlet allocation* (LDA) technique (Blei et al., 2003) employs a hierarchical Bayesian extension of pLSI, treating both the document-specific topic-mixing coefficients and the topic-specific word probabilities as random variables, under appropriate conjugate priors. LDA can be extended to multimedia collections by assuming that the unobserved "topics" are correlated with both image variables and word variables (Barnard et al., 2003, Blei and Jordan, 2003). Recently, Welling et al. (2004) proposed another class of latent semantic GMs known as the exponential family harmonium model (EFH), which can be understood as an undirected, and non-Bayesian counterpart of the LDA model. Subsequently, Xing et al. (2005) extended EFH to a *dual-wing harmonium model* (DWH) for joint modeling of text and image. Gehler et al. (2006) proposed the *rate adapting Poisson* (RAP) model which follows the general architecture of EFH model and use conditional Poisson distributions to model observed count data. And McCallum et al. (2006) proposed a training criterion called *multiple-conditional learning* (MCL) for MRFs and EFHs. Unlike the directed GMs such as pLSI and LDA, EFH does not employ auxiliary latent variables (i.e., the imaginary topic indicators for every word) to facilitate topic mixing and simulate data generation; and it allows a more flexible representation of the latent topic aspects for documents (i.e., as a point is a Euclidean space rather than in a simplex).

An important advantage of the directed latent-topic models such as LDA is that they can be straightforwardly embedded in a Bayesian framework, and can undergo Bayesian training, smoothing and inference. To date, the MRF-based models such as EFH and DWH have been largely limited to a maximum likelihood (ML) framework, which is prone to undesirable effects such as overfitting the (small) data, high variance in sampling-based inference and parameter estimation, and indifference to prior knowledge. These limitations restrict their utilities in many realistic data mining scenarios where data are sparse and spurious. The ML framework also makes it difficult to fully exploit the modeling power of MRF in latent topic distillations and to develop future extensions. The unavailability of a Bayesian version of EFH is partly due to the remarkable technical difficulties one must overcome when working under such a formalism. It is well-known that sta-

tistical learning of EFH models from data, even under an ML framework, is technically nontrivial. As discussed in Murray and Ghahramani (2004) and Qi et al. (2005), Bayesian learning for general MRF, is even more challenging, particularly in cases that involve latent variables as in EFH. In this paper, we attempt to address some of this challenges: endowing EFH with a simple Bayesian prior, and presenting a sampling-based algorithm for Bayesian inference and learning.

We present Bayesian EFH (BEFH), in which a multivariate Gaussian prior is introduced for the weight matrix that couples the latent topics with observed attributes in EFH (and also in DWH). As detailed in the sequel, it is illuminative to view the weight matrix of EFH as the matrix of word probabilities under all topics in LDA. Under this analogy, our prior corresponds to the Dirichlet priors for the word probabilities in LDA. It is well-known that methods for Bayesian inference/learning in directed GMs such as LDA does not apply to the undirected GMs concerned here, because of the intractability and non-conjugacy arising from the partition function. In this paper, we present the Langevin algorithm conjoint with a MCMC sampling scheme for posterior inference under BEFH. We also propose an empirical Bayes method based on the Langevin algorithm for unsupervised estimation of the BEFH hyperparameter given training data. Finally we show comparisons of ML and Bayesian approaches on a synthetic dataset with known parameters and a dataset provided by TRECVID 2003 (Smeaton and Over, 2003) with both text and image data.

2 From EFH to Bayesian EFH

In this section, we outline the basic structure of a Bayesian EFH in the context of a simple instantiation of EFH for latent topic modeling of text corpus.

For completeness, we begin with a brief recap of the basic EFH, as described in (Welling et al., 2004). Consider an undirected GM defined on a complete bipartite graph containing two layers of nodes (Fig 1). Let $\mathbf{H} = \{H_j\}$ denote the set of *hidden units* in such a graph, and let $\mathbf{X} = \{X_i\}$ denote the set of *input units*. An EFH defines the following Markov random field:

$$p(\mathbf{x}, \mathbf{h}) \propto \frac{1}{Z} \exp \Big\{ \sum_{ia} \theta_{ia} f_{ia}(x_i) + \sum_{jb} \lambda_{jb} g_{jb}(h_j) + \sum_{ijab} W_{ia}^{jb} f_{ia}(x_i) g_{jb}(h_j) \Big\}, \tag{1}$$

where $\{f_{ia}(\cdot) : \forall a\}$ denotes the set of potential functions (or features) defined on each of the input units (indexed by i) in the model, and likewise $\{g_{jb}(\cdot) : \forall b\}$ for the hidden units; $\Theta = \{\theta_{ia}\} \cup \{\lambda_{jb}\} \cup \{W_{ia}^{jb}\}$ denotes the "weights" of the corresponding potentials or potential pairs; and Z stands for the partition function, which is a function of Θ.

The bipartite topology of the harmonium graph suggests that nodes within the same layer are conditionally independent given all nodes of the opposite layer. Specifically, from Eq. (1), we have the following factored form for the between-layer conditional distribution functions: $p(\mathbf{x}|\mathbf{h}) = \prod_i p(x_i|\mathbf{h})$, $p(\mathbf{h}|\mathbf{x}) = \prod_j p(h_j|\mathbf{x})$, and each of the singleton conditional has a simple exponential family form:

$$p(x_i|\mathbf{h}) = \exp\big\{ \sum_a \hat{\theta}_{ia} f_{ia}(x_i) - A_i(\{\hat{\theta}_{ia}\}) \big\}, \tag{2}$$

$$p(h_j|\mathbf{x}) = \exp\big\{ \sum_b \hat{\lambda}_{jb} g_{jb}(h_j) - B_j(\{\hat{\lambda}_{jb}\}) \big\}, \tag{3}$$

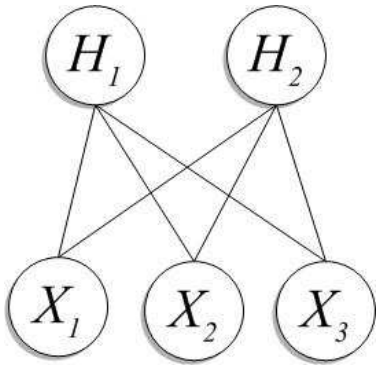


Figure 1: The graphical model representation for (a) a harmonium with 2 hidden units and 3 input units.

where $A_i(\cdot)$ and $B_j(\cdot)$ denote the respective log-partition functions; and the shifted parameters $\hat{\theta}_{ia}$ and $\hat{\lambda}_{jb}$ are defined as,

$$\hat{\theta}_{ia} = \theta_{ia} + \sum_{jb} W_{ia}^{jb} g_{jb}(h_j),$$
$$\hat{\lambda}_{jb} = \lambda_{jb} + \sum_{ia} W_{ia}^{jb} f_{ia}(x_i),$$

where the shifts are induced by the total couplings between units in the input and hidden layers. As seen from the above definition, since all the parameters in the joint distribution under EFH can be identified from the local conditional distributions, one can determine an EFH using a bottom-up strategy to by directly specifying the often easily comprehensible local conditionals. For instance, as our running example in this paper, we define the following Gaussian-Bernoulli EFH (GB-EFH) for text:

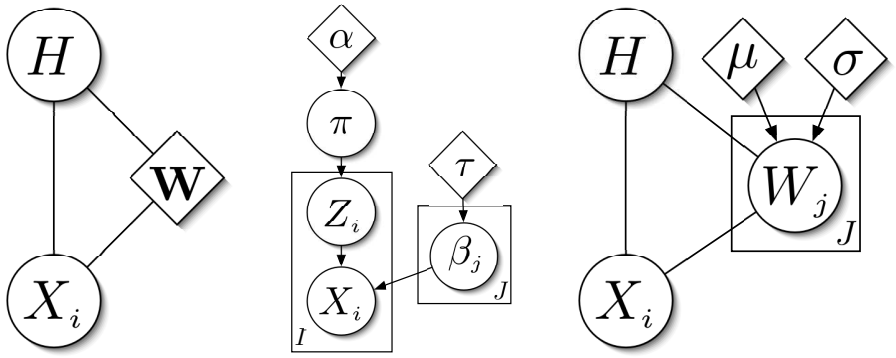
$$p(x_i|\mathbf{h}) = \text{Bernoulli}\big(x_i|\text{logit}(\theta_i + \sum_j h_j W_{ij})\big), \tag{4}$$

$$p(h_j|\mathbf{x}) = \mathcal{N}(h_j|\sum_i x_i W_{ij}, 1), \tag{5}$$

where $\text{logit}(\alpha) = (1+e^{-\alpha})^{-1}$ is the logistic function, and the shift of the logit-transformed Bernoulli rate θ_i is induced by a weighted combination of the latent units $\mathbf{h}$. It can be shown that under this construction, we obtain an EFH with the joint:

$$p(\mathbf{x}, \mathbf{h}) \propto \exp\big\{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x} - \frac{1}{2}\mathbf{h}^{\mathrm{T}}\mathbf{h} + \mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{h}\big\}. \tag{6}$$

The GB-EFH models text (represented by variables $\mathbf{x}$) as binary occurrences of words, which is suitable for sparse, short text such as video captions. When modeling long articles, one may want to directly model word counts; and in this case one can replace Eq. (4) with, e.g., a Binomial distribution.



(a) (b) (c)

Figure 2: A comparison of EFH, LDA and BEFH models over a single document. Circles represent variables, and diamond represents model parmeters. (a) EFH. For easy comparison, the hidden unit (i.e., the topic weight coefficients) $\{H_j\}$ and the input units $\{X_i\}$ are represented as vector valued variables H and X, respectively. For simplicity, only the $\mathbf{W}$ parameter of EFH is explicitly shown. (b) LDA. Note the correspondence between π in LDA and H in EFH, and the fact that β_j's are random variables rather than parameters. I denotes the length of the document. (c) BEFH. Note that $\mathbf{W} \equiv \{W_j\}$ are now lifted as random variables.

It is interesting to examine side-by-side the GB-EFH and the LDA model as displayed in Fig. 2. Note that, when treating each hidden unit h_j as a representative of a latent topic aspect, Eq.(4) can be understood as a likelihood function of an observed attribute, such as a word occurrence, induced by a combination of topics. Thus, the coupling matrix $\mathbf{W} = \{W_1, \ldots, W_J\}$ in GB-EFH is reminiscent of the word probability matrix $\mathbf{B} = \{\beta_1, \ldots, \beta_J\}$ in LDA, where β_j denotes the M-dimensional vector (M denotes the size of the vocabulary) of multinomial word probabilities under topic j. In GB-EFH each M-vector W_j represents the set of "contributions" topic j has on each word in a vocabulary. Although structurally similar, it is noteworthy that the topic mixing mechanism of GB-EFH is very different from that of the LDA model. In LDA the topic mixing is achieved by marginalizing out the auxiliary topic indicator variables for each word occurrence [1]. Whereas it can be shown that in EFH the expected rates of all words are directly determined by the weighted sum of topic specific contributions $\sum_j h_j W_j \equiv \mathbf{Wh}$. In this regard EFH is closer to the classical LSI principle in which the observed rates of all words can be expressed as a weighted combinations of the eigen-topics (i.e., orthonormal topic-specific word rate vectors).

Empirically, it was noted that the performance of EFH and variants on latent semantic modeling is comparable, and sometimes superior, to LDA (Welling et al., 2004, Xing et al., 2005). But as shown in Fig. 2, structurally EFH is not yet a full undirected counterpart of LDA, which employs an elegant hierarchical strategy to incorporate priors for both the word probabilities $\mathbf{B}$ and topic mixing coefficients π. We expect that, as is the case for LDA, it is possible for EFH to also leverage on the possible extra modeling power endowed by a Bayesian formalism.

[1] As illustrated in Fig. 2b, the LDA likelihood of a word x_w, given topic mixing coefficient θ and the probabilities of this word under all J topics, $\{\beta_{w,1}, \ldots, \beta_{w,J}\}$ can be written as $p(x_w|\theta) = \sum_z p(z|\theta)p(x_w|\mathbf{B}, z) = \sum_j \theta_j \beta_{w,j}$.

Now we propose a Bayesian EFH that exploits the purported benefits. To maintain exchangability between hidden units $\{h_j\}$, we place column-iid prior on $\mathbf{W}$, that is, each column of $\mathbf{W}$ follows a multivariate Gaussian, which is a common choice for modeling continuous parameters without any additional assumption:

$$p(\mathbf{W}) = \prod_{j=1}^{J} p(W_j) = \prod_{j=1}^{J} \mathcal{N}(W_j|\mu, \Sigma). \tag{7}$$

A full covariance matrix in the above prior would have size M^2, which is prohibitively expensive for modeling large vocabulary. For simplicity, we consider a further simplification where: $\Sigma = \text{diag}(\boldsymbol{\sigma})$, i.e. $\Sigma_{ij} = \sigma_i \sigma_j \delta(i, j)$. (Modulo computational cost, upgrading to the full covariance matrix is straightforward with the same algorithm developed in the sequel.) This means that each element of $\mathbf{W}$ follows an independent normal distribution. Note that although we omit correlations between the topic-word coupling coefficients, the expressiveness of this prior is comparable to the Dirichlet prior for columns in the $\mathbf{B}$ matrix of LDA, which captures little correlation behavior of the word-probabilities sampled from a simplex.

Now we are left with two remaining sets of parameters of EFH: θ and λ. It turns out that in many practical settings (e.g., GB-EFH and DWH), λ is vacuous, i.e., $\lambda = 0$, which essentially "centers" the conditional distribution $p(\mathbf{h}|\mathbf{x})$ at the shifts induced by the input units. For θ, in EFH it lacks an intuitive semantics, such as being a prior for topic coefficients as in LDA. Therefore we choose to leave θ as fixed parameters to be estimated via an ML principle.

Now, putting things together, we arrive at a Bayesian EFH model with the following joint density function

$$p(\mathbf{x}, \mathbf{h}, \mathbf{W}|\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = p(\mathbf{W}|\boldsymbol{\mu}, \boldsymbol{\sigma}) p(\mathbf{x}, \mathbf{h}|\boldsymbol{\theta}, \mathbf{W}). \tag{8}$$

The hyperparamters in the model are μ and σ, which we treat as fixed quantities presumably known or to be estimated.

3 Posterior Inference via MCMC

Given the prior distribution on $\mathbf{W}$ with presumably known hyperparameters and a collection of N *iid*-sampled data $\mathrm{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$, also suppose that parameters θ are known or already estimated by an alternative learning method such as ML learning, we need to compute or approximate the posterior

$$p(\mathbf{W}|\mathrm{X}) \propto p(\mathrm{X}|\mathbf{W}) p(\mathbf{W}) = \frac{1}{(Z(\mathbf{W}))^N} \tilde{p}(\mathrm{X}|\mathbf{W}) p(\mathbf{W}) \tag{9}$$

and the predictive posterior density over hidden variables

$$p(\mathbf{h}|\mathbf{x}, \mathrm{X}) = \int_{\mathbf{W}} p(\mathbf{h}|\mathbf{x}, \mathbf{W}) p(\mathbf{W}|\mathrm{X}) d\mathbf{W}, \tag{10}$$

where $\tilde{p}(\cdot)$ in Eq. (9) represents the unnormalized density function corresponding to $p(\cdot)$.

We can take a Monte Carlo approach to obtain a set of m samples $\{\mathbf{W}_1, \ldots, \mathbf{W}_m\}$ by simulating an ergodic Markov chain whose stationary distribution is the posterior $p(\mathbf{W}|\mathrm{X})$. The difficulty here is due to the presence of an intractable term $(1/Z(\mathbf{W}))^N$ in the posterior distribution, which is *a function* of the target random parameters $\mathbf{W}$. Therefore, unlike simple posterior inference settings in which there is a *normalization constant* that will be canceled out by computing the ratio of two posterior densities or taking the derivative, in Bayesian inference with MRFs using MCMC we have to seek an efficient approximation of the intractable random partition function in posterior distribution.

In the following, we investigate two MCMC approximation schemes and show that in both cases the intractable term can be written as expectations under the data distribution $p(\mathbf{x}|\mathbf{W})$. Then we show that these terms can be approximated efficiently by minimizing the contrastive divergence (CD) (Hinton, 2002), or equivalently, by performing Gibbs sampling for only very few steps starting from data (the empirical distribution). The derivation is in parallel with that in Murray and Ghahramani (2004); here we provide a more detailed discussion on the comparison of the two schemes

3.1 Approximation schemes

3.1.1 Metropolis-Hasting algorithms

Consider simulating a Markov chain using a Metropolis-Hasting algorithm with the proposal distribution $q(\mathbf{W}'|\mathbf{W})$. The acceptance probability of the transition $\mathbf{W} \to \mathbf{W}'$ is

$$\rho(\mathbf{W}, \mathbf{W}') = \min\left(\frac{p(\mathbf{W}'|\mathrm{X})}{p(\mathbf{W}|\mathrm{X})}\,\frac{q(\mathbf{W}|\mathbf{W}')}{q(\mathbf{W}'|\mathbf{W})}, 1\right) \tag{11}$$

Suppose the proposal distribution is easy to draw sample from and is tractable, then the only difficulty in implementing Metropolis-Hasting algorithms is to approximate the intractable term $\left(\frac{Z(\mathbf{W})}{Z(\mathbf{W}')}\right)^N$, where N is the size of the dataset. The ratio of two partition functions can be written as an expectation over the data distribution $p(\mathbf{x}|\mathbf{W}')$.

$$\begin{aligned}\frac{Z(\mathbf{W})}{Z(\mathbf{W}')} &= \sum_{\mathbf{x}} e^{\frac{1}{2}\mathbf{x}^{\mathrm{T}}(\mathbf{W}\mathbf{W}^{\mathrm{T}} - \mathbf{W}'\mathbf{W}'^{\mathrm{T}})\mathbf{x}}\,\frac{e^{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x} + \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{W}'\mathbf{W}'^{\mathrm{T}}\mathbf{x}}}{Z(\mathbf{W}')} \\ &= \left\langle \exp\left\{\frac{1}{2}\mathbf{x}^{\mathrm{T}}\left(\mathbf{W}\mathbf{W}^{\mathrm{T}} - \mathbf{W}'\mathbf{W}'^{\mathrm{T}}\right)\mathbf{x}\right\}\right\rangle_{p(\mathbf{x}|\mathbf{W}')}\end{aligned} \tag{12}$$

3.1.2 The Langevin algorithm

We also investigate the *Langevin algorithm* as an alternative approximate MCMC scheme. The Markov chain simulated by the Langevin algorithm is characterized by the following stochastic transition equation

$$\mathbf{W}' = \mathbf{W} + \frac{\epsilon^2}{2}\nabla \log p(\mathbf{W}|\mathrm{X}) + \epsilon N_{\mathbf{W}} \tag{13}$$

where $N_{\mathbf{W}}$ are randomly generated from $\mathcal{N}(\mathbf{0}, I_{|\mathbf{W}|})$. This is a discrete version of the Langevin diffusion and ϵ^2 corresponds to the discretization size [2]. The Markov chain converges when ϵ is reasonbly small and has the desired density $p(\mathbf{W}|\mathrm{X})$ as $\epsilon^2 \to 0$. The gradient of the posterior is the sum of three terms

$$\nabla \log p(\mathbf{W}|\mathrm{X}) = \nabla \log p(\mathrm{X}, \mathbf{W}) = \nabla \log p(\mathbf{W}) + \nabla \log \tilde{p}(\mathrm{X}|\mathbf{W}) + (-N \nabla \log Z(\mathbf{W})) \quad (14)$$

where in the GB-EFH model

$$\{\nabla \log p(\mathbf{W})\}_{ij} \triangleq \frac{\partial \log p(\mathbf{W})}{\partial W_{ij}} = \frac{\partial \log p_{ij}(\mathbf{W})}{\partial W_{ij}} = -\frac{1}{\sigma_i^2}(W_{ij} - \mu_i) \quad (15)$$

and $\nabla \log \tilde{p}(\mathrm{X}|\mathbf{W})$ is also tractable

$$\nabla \log \tilde{p}(\mathrm{X}|\mathbf{W}) = \sum_i \nabla \log \tilde{p}(\mathbf{x}_i|\mathbf{W}) = \sum_i \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \mathbf{W} = X X^{\mathrm{T}} \mathbf{W}. \quad (16)$$

Hence, the only intractable term involved in the Langevin algorithm is $N \nabla \log Z(\mathbf{W})$, in which $\nabla \log Z(\mathbf{W})$ can be written as an expectation over the data distribution $p(\mathbf{x}|\mathbf{W})$

$$\begin{aligned} \nabla \log Z(\mathbf{W}) &= \frac{1}{Z(\mathbf{W})} \sum_{\mathbf{x}} \nabla \tilde{p}(\mathbf{x}|\mathbf{W}) = \frac{\tilde{p}(\mathbf{x}|\mathbf{W})}{Z(\mathbf{W})} \sum_{\mathbf{x}} \nabla \log \tilde{p}(\mathbf{x}|\mathbf{W}) \\ &= \sum_{\mathbf{x}} p(\mathbf{x}|\mathbf{W}) \nabla \log \tilde{p}(\mathbf{x}|\mathbf{W}) = \left\langle \mathbf{x}\mathbf{x}^{\mathrm{T}} \mathbf{W} \right\rangle_{p(\mathbf{x}|\mathbf{W})} \end{aligned} \quad (17)$$

3.1.3 Discussion on the two schemes

The straightforward approach of estimating $Z(\mathbf{W})$ itself often fails to provide reliable estimates. To provide some intuition of the nature of this difficulty, we give a brief illustration as follows: with some mathematical manipulation which is included in Appendix A, the partition function in the BG-EFH model equals the expectation of the following random variable

$$Z(\mathbf{t}) = \prod_i (1 + t_i) \quad (18)$$

under the multivariate lognormal distribution of $\mathbf{t}$

$$\mathbf{t} \sim \text{LogNormal}(\boldsymbol{\theta}, \mathbf{W}\mathbf{W}^T)$$

Thus under the Bayesian framework in which $\mathbf{W}$ is considered a random matrix, we should expect $Z(\mathbf{W})$ to have *exponential* mean and variance.

[2] A diffusion is a continuous time process which can be defined by a stochastic differential equation. The Langevin diffusion is characterized by

$$d\mathbf{W}(t) = \frac{1}{2} \nabla \log p(\mathbf{W}(t)|\mathrm{X}) dt + dB(t),$$

where $B(t)$ is a $|\mathbf{W}|$-dimensional Brownian motion.

Thus, we put more emphasis on variance in the bias-variance tradeoff of estimators in approximate Bayesian learning. Compare the approximations in the Langevin algorithm to update $\mathbf{W}$ as in Eq. 13 and in Metropolis-Hasting algorithms to compute the acceptance probability as in Eq. 11

$$\epsilon^2 \nabla \log p(\mathbf{W}|\mathrm{X}) = -\epsilon^2 N \nabla \log Z(\mathbf{W}) + C \tag{19}$$

$$\left(\frac{Z(\mathbf{W})}{Z(\mathbf{W}')}\right)^N \approx e^N e^{(\mathbf{W}-\mathbf{W}')N\nabla \log Z(\mathbf{W}')} \tag{20}$$

where $C = \epsilon^2(\nabla \log p(\mathbf{W}) + \nabla \log \tilde{p}(\mathrm{X}|\mathbf{W}))$ can be computed exactly, and Eq. 20 is obtained by first-order Taylor expansion. We expect the latter approximation has *exponential* variance compared to the former one. Therefore, we choose the Langevin algorithm conjoint with the MCMC scheme for posterior inference on BEFH model.

3.2 Approximating the expectations with brief sampling

$\nabla \log Z(\mathbf{W})$ in Eq. 17 can be estimated using a "sampling very few steps from the data" technique. It is first proposed by Hinton (2002) under the name of minimizing constrastive divergence (CD) and suggested by Murray and Ghahramani (2004) for approximate Bayesian inference in MRF in which it is named *brief sampling*. Brief sampling in GB-EFH runs multiple chains starting from the data X. Each chain performs l full step of Gibbs sampling. A total of N samples are obtained, denoted by $X_l = (\mathbf{x}_1^{(l)}, \ldots, \mathbf{x}_N^{(l)})$. Then $\nabla \log Z(\mathbf{W})$ is approximated as an expectation over the empirical distribution of X_l. This whole procedure of brief sampling is illustrated as follows where we set $l = 1$:

- Draw $\mathbf{h}_k^{(1)} \sim \mathcal{N}(\mathbf{W}^{\mathrm{T}}\mathbf{x}_k, I_J)$ for $k = 1, \ldots, N$;
- Draw $\mathbf{x}_k^{(1)} \sim \mathrm{Bernoulli}\left(\mathrm{logit}((\boldsymbol{\theta} + \mathbf{W}\mathbf{h}_k^{(1)}))\right)$ for $k = 1, \ldots, N$;
- $\nabla \log Z(\mathbf{W}) \approx \frac{1}{N}\sum_k \mathbf{x}_k^{(1)}(\mathbf{x}_k^{(1)})^{\mathrm{T}}\mathbf{W} = \frac{1}{N}\mathrm{X}_1\mathrm{X}_1^T\mathbf{W}$

Brief sampling has been previously shown to provide low variance estimation with a small bias in ML learning (Carreira-Perpinan and Hinton, 2005). The intractable term in ML learning of MRF is just the same term $\nabla \log Z(\mathbf{W})$, therefore we expect similar low-variance behavior of brief sampling estimation in the Langevin algorithm. Fig. 4 in the experiment section provides an empirical demonstration.

3.3 Computing the predictive posterior density

Given m samples $\{\mathbf{W}_1, \ldots, \mathbf{W}_m\}$ obtained by the Langevin algorithm with brief sampling described above, the predictive conditional distribution is approximated by

$$p(\mathbf{h}|\mathbf{x}, \mathrm{X}) = \frac{1}{m}\sum_{k=1}^{m} p(\mathbf{h}|\mathbf{x}, \mathbf{W}_k). \tag{21}$$

More specifically, in GB-EFH we are interested in the conditional expectation of $\mathbf{h}$ given $\mathbf{x}$, this is computed as

$$E\left(\mathbf{h}|\mathbf{x},\mathrm{X}\right)=\frac{1}{m}\sum_{k=1}^{m}E\left(\mathbf{h}|\mathbf{x},\mathbf{W}_k\right)=\frac{1}{m}\sum_{k=1}^{m}\mathbf{W}_k^{\mathrm{T}}\mathbf{x}. \tag{22}$$

4 Hyperparameter Estimation

Now we briefly outline how to compute the maximum likelihood estimates of the hyperparameters μ and σ of BEFH from training data, based on an empirical Bayes principle. (ML estimation for other model parameters such as θ and λ roughly follows the same scheme and hence omitted for simplicity.) We employ a Monte Carlo EM scheme. In the "E"-step, we impute the hidden variables in BEFH, specifically, $\mathbf{W}$, from its posterior distribution; and in Sec. 3 we have developed the Langevin algorithm for this step. Given a set of K imputed $\mathbf{W}$ from iteration t, we proceed to the "M"-step, in which now we are essentially back to the standard ML learning scenario for fully observed MRF, and compute an estimate of the hyperparameters as follow: we seek:

$$\begin{aligned}(\mu^{(t+1)},\sigma^{(t+1)}) &= \arg\max_{\mu,\sigma}\sum_{k=1}^{K}\log p(X,\mathbf{W}_k^{(t)}|\mu,\sigma)\\ &= \arg\max_{\mu,\sigma}\sum_{k=1}^{K}\left(\log p(\mathbf{W}_k^{(t)}|\mu,\sigma)+\log p(X|\mathbf{W}_k^{(t)})\right)\\ &= \arg\max_{\mu,\sigma}\sum_{k=1}^{K}\log p(\mathbf{W}_k^{(t)}|\mu,\sigma),\end{aligned} \tag{23}$$

where $\mathbf{W}_k^{(t)}$ denotes the k-th imputed sample at iteration t.

It can be shown that, the ML estimate of each element of μ and σ is:

$$\mu_i^{(t+1)}=\frac{1}{JK}\sum_j\sum_k W_{ij,k}^{(t)} \tag{24}$$

$$\sigma_i{}^{(t+1)}=\sqrt{\frac{1}{JK-1}\sum_j\sum_k(W_{ij,k}^{(t)}-\mu_i^{(t+1)})^2} \tag{25}$$

where $W_{ij,k}^{(t)}$ denotes the ij-th element of $\mathbf{W}_k^{(t)}$.

To initialize the EM procedure, we can make use of the ML estimate of $\mathbf{W}$, denoted by $\mathbf{W}^{MLE}$, and let

$$\mu_i^{(0)}=\frac{1}{J}\sum_j W_{ij}^{MLE} \tag{26}$$

$$\sigma_i{}^{(0)}=\sqrt{\frac{1}{J-1}\sum_j(W_{ij}^{MLE}-\mu_i^{(0)})^2} \tag{27}$$

5 Experiments

5.1 Synthetic EFH parameter estimation

The dataset is generated for a GB-EFH model with $\theta = 0$. The model contains $M = 100$ observed variables and $J = 10$ hidden variables, so the number of parameters in W is $M \times J = 1000$. We vary the size of the training dataset from 25 to 200 and compare the performance of ML estimation via gradient ascent and the Langevin algorithm proposed in Sec. 3.

Generating *iid* samples from a general MRF is known to be nontrivial. However, for a GB-EFH model exact samples can be genereated fairly efficiently by employing the perfect sampling technique (Childs et al., 2001) when all the elements of the matrix $V = WW^{\mathrm{T}}$ are non-negative. To ensure this property, we first generate an $M \times M$ matrix whose elements are uniformly distributed in the $[0, 0.1]$ interval. Then W is determined by performing an SVD on this matrix so that V is the best rank-J approximation.

There is an inidentifiability issue here because the data distribution

$$p(\mathbf{x}|\mathbf{W}) = \frac{1}{Z} \exp\left(\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{x}\right) \tag{28}$$

is a function of V and is invariant if W is right-multiplied by an orthogonal matrix Q because $(WQ)(WQ)^{\mathrm{T}} = WW^{\mathrm{T}}$. Also it can be shown the prior of W defined in Eq. 7 is also invariant under this transformation. Therefore our evaluation criteria are based on the matrix V instead of W. We define two error measures: *mean averaged error* (*mae*) and *mean relative error* (*mre*) to evaluate an estimate $\hat{V}$

$$mae = \frac{1}{M^2} \sum_i \sum_j |V_{ij} - \hat{V}_{ij}| \tag{29}$$

$$mre = \frac{1}{M^2} \sum_i \sum_j \frac{|V_{ij} - \hat{V}_{ij}|}{\max\{|V_{ij}|, |\hat{V}_{ij}|\}} \tag{30}$$

Two tunable parameters in the Langevin algorithm are yet to be determined: the step size ϵ in Eq. 13 and the number of steps l to sample from data in brief sampling. We choose an appropriate ϵ by investigating the evolution of a number of elements of W during the simulation of the Markov chain. Under a too large step size the chain goes to infinity in a few steps, and under a too small one the burn-in time is undesirably long. Fig 3 shows a simulation of the Langevin algorithm using the step size we choose.

Fig. 4 shows the estimate of the gradient using brief sampling versus the number of sampling steps l. We also generate the same number of samples using the perfect sampling technique to provide an approximately correct version for comparison. Brief sampling provides biased estimation compared to the exact sampling approach, but the bias is relatively small considering the difficulty of dealing with intratable partition function. Note that the bias is not decreased by increasing l. The variance of the estimation, on the other hand, is minimized when $l = 1$. Therefore, we let $l = 1$ in the subsequent experiments.

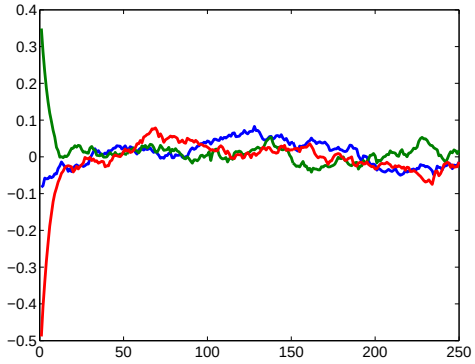


Figure 3: Details of Monte Carlo simulations of the Langevin algorithm, with y-axis corresponds to the value of W_{11}. Three chains of different starting points are shown. The burn-in time to reach convergence is approximately 50 transition.

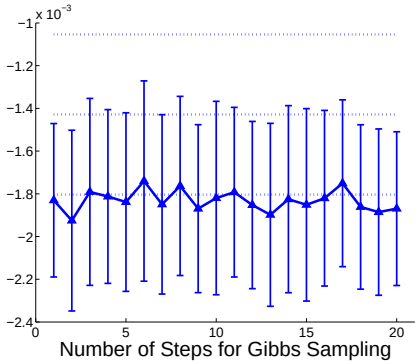
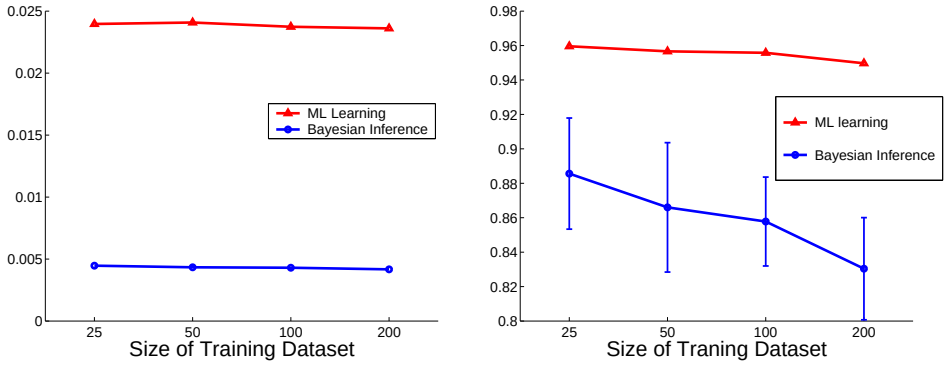


Figure 4: The estimation versus the number of sampling steps in brief sampling (solid line) compared with the estimation perfect sampling (dash line), with y-axis corresponds to an estimated derivative of log-partition function $\partial \log Z(\mathrm{W})/\partial \mathrm{W}_{11}$ averaged over 50 runs. Both sampling schemes generate 100 samples in each run. The standard error bars are scaled by 1.64, indicating 95% significance of the difference in estimation.

In Fig. 5 we compare the performance of ML estimation via gradient ascent and the Bayesian approach using the Langevin algorithm. The Langevin algorithm consistently achieves lower errors under both measures and with different sizes of the training set. As more data are available, the performance of ML estimation improves little; it appears that the gradient ascent procedure gets stuck into a local minimum. On the other hand, the Langevin algorithm does benifit from more data, which is possibly the consequence of the uninformative prior we placed for this problem by setting $\mu_i = 0, \sigma_i = d = 0.1$ for $i = 1, \ldots, M$. The estimation by both methods has a non-neglible bias from the true value, and we conjecture that it is due to the sparsity of the data. We also observe that the performance difference of ML estimation and the Langevin algorithm is much larger as measured by *mean absolute error* than *mean relative error*, which suggests that the latter algorithm provides better estimates for parameters with larger values.



(a) (b)

Figure 5: The Performance of ML learning and Bayesian inference using the brief Langevin algorithm under two different error measures (a) mean absolute error; (b) mean relative error. The results are averged over 10 runs. The error bar is shown only for Bayesian inference in (a), in other cases the standard error are too small to be distinguishable from the figure.

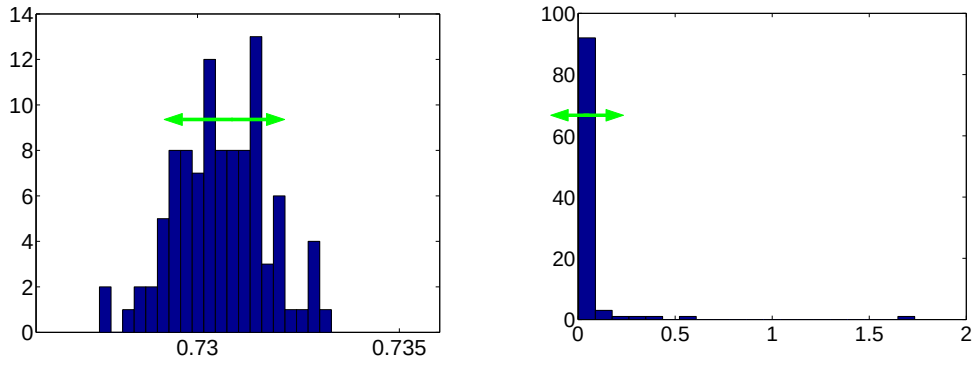
5.2 Classification of Text and Image Data

The dataset is from the complied TRECVID'03 news video collection in (Xing et al., 2005). It contains 1078 video shots with captions; each one can be treated as a document and belongs to one of five pre-defined categories. 1894 binary word occurence features and 166 continuous features for key images are extracted from each document. We extend the dual-wing harmonium (DWH) developed in (Xing et al., 2005), which was previously trained by ML estimation, to Bayesian DWH (BDWH) in which column-*iid* multivariate normal priors are placed on the coupling matrices for word and image features respectively. The hyperparameters in the priors are estimated using the empirical Bayes method developed in Sec. 4.

To give a hint on the difficulty of performing Bayesian learning in a real dataset discussed in Sec. 3, we implement the naïve Monte Carlo estimation of the partition function in Eq. 18 for both GB-EFH with synthetic dataset and DWH with real world dataset. The histograms of the estimated $\mathcal{Z}$ over 100 runs are shown in Fig. 6. In the synthetic dataset the estimated values approximately fit to a normal distribution. However, in the real dataset, there are a few spurious outliers, which shift the mean estimated values over all the runs significantly, leading to generally biased, high variance estimates. In Fig. 6(b) the variance of the estimation is three times as large as the estimated mean.

We evaluate the performance of four different models LSI (Deerwester et al., 1990), GM-LDA, DWH and BDWH for classification task on the news video collection. For each algorithm, the parameters are estimated using all data, without reference to their labels. Once the model are learned, every document in the data are projected into the lower-dimensional latent semantic space. The data are then randomly splited to a training set and a testing set with the same size. We show the result of using one nearest neighbor (1-NN) classifier to predict the category of each test data given the training data.

Fig. 7 compares the performance obtained at different dimensions of latent semantic space, or equivalently different numbers of latent topics ranging from 4 to 32. BDWH and DWH achieve comparable classification accuracy consistently, and outperform LSI and GM-LDA with a good



(a) (b)

Figure 6: Histogram of 100 estimations of partition function using a naïve Monte Carlo approximation on (a) synthetic dataset; (b) real dataset. Arrows are centered at the mean and indicate an interval of length of 2 times the standard deviation. Each estimation computes the expectation using 1000 samples. The displayed values in (b) are scaled by a factor of 2×10^{-4}.

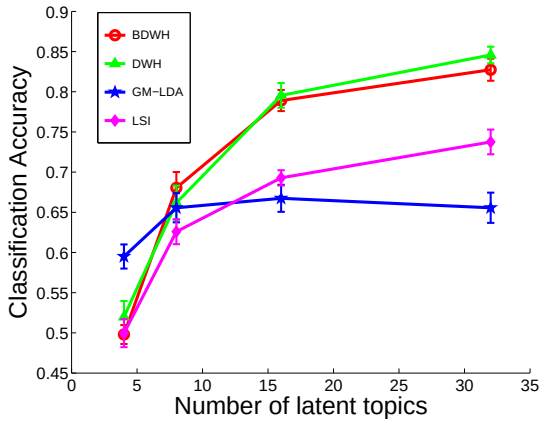


Figure 7: Classification accuracy versus number of latent topics.

margin when the number of latent topics are 16 and 32. LSI, DWH and BDWH all get better performances in higher dimensional semantic space with less dimensionality-reduction. In the constrast, GM-LDA outperforms other methods when the number of latent topics are 4 but the performance curve goes down when the number of latent topics increases from 16 to 32, which may reflect a low-dimensionality bias from the modeling.

6 Conclusion

We have proposed a new Bayesian formalism of EFH model and variants for latent semantic modeling of text and multimedia data. The Langevin algorithm conjoint with an MCMC scheme was applied to carry out approximate posterior inference, and an empirical Bayes method is also developed for esimating the parameters. The Bayesian approach achieves superior performance of parameter estimation on a synthetic data set and comparable classfication accuracy on a real dataset

of both text and image data.

Our experiments presented in this paper focus on binary occurences of words which is suitable for short texts. In ongoing work, we are building an BEFH to directly model word counts. Also, the independent Gaussian prior we used can be replaced by an more informative one, while the inference and learning algorithm can straightforwardly apply to the new formalism. Finally, the discretization scheme in the Langevin algorithm can be more elaborate, such as incorporating the idea suggested in (Sexton and Weingarten, 1992).

References

K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

D. Blei and M. Jordan. Modeling annotated data. In *Proc. of the 26th Intl. ACM SIGIR Conference*, 2003.

D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 2003.

M. A. Carreira-Perpinan and G. E. Hinton. On contrastive divergence learning. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, 2005.

A. M. Childs, R. B. Patterson, and D. J. C. MacKay. Exact sampling from non-attractive distributions using summary states. *Physical Review E*, 63:036113, 2001.

S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6): 391–407, 1990.

P. V. Gehler, A. D. Holub, and M. Welling. The rate adapting poisson model for information retrieval and object recognition. In *Proceedings of 23rd International Conference on Machine Learning (ICML'06)*, 2006.

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

T. Hofmann. Probabilistic latent semantic indexing. In *Proc. of the 22nd Intl. ACM SIGIR conference*, pages 50–57, 1999.

A. McCallum, C. Pal, G. Druck, and X. Wang. Multi-conditional learning: Generative/discriminative training for clustering and classification. In *Proceedings of 21st National Conference on Artificial Intelligence (AAAI'06)*, 2006.

I. Murray and Z. Ghahramani. Bayesian learning in undirected graphical models: Approximate mcmc algorithms. In *Proceedings of the 20th Annual Conference on Uncertainty in AI*, 2004.

Y. Qi, M. Szummer, and T. Minka. Bayesian conditional random fields. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, 2005.

J. C. Sexton and D. H. Weingarten. Hamiltonian evolution for the hybrid monte carlo algorithm. *Nuclear Physics B*, 380:665–677, 1992.

A. F. Smeaton and P. Over. TRECVID: Benchmarking the effectiveness of information retrieval tasks on digital video. In *Proc. of the Intl. Conf. on Image and Video Retrieval*, 2003.

M. Welling, M. Rosen-Zvi, and G. Hinton. Exponenetial family harmoniums with an application to information retrieval. In *Advance Neural Information Processing Systems*, volume 17, 2004.

E. P. Xing, R. Yan, and A. G. Hauptmann. Mining associated text and images with dual-wing harmoniums. In *Proceedings of the 21st Annual Conference on Uncertainty in AI*, 2005.

A Partition Function of the GB-EFH Model

The joint probability distribution of observed variables $\mathbf{x}$ and hidden variables $\mathbf{h}$ in the GB-EFH model is

$$p(\mathbf{x}, \mathbf{h}) = \frac{1}{Z(\boldsymbol{\theta}, \mathbf{W})} \exp\left(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x} - \frac{1}{2}\mathbf{h}^{\mathrm{T}}\mathbf{h} + \mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{h}\right) \tag{31}$$

where the corresponding partition function is

$$Z(\boldsymbol{\theta}, \mathbf{W}) = \sum_{\mathbf{x},\mathbf{h}} \exp\left(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x} - \frac{1}{2}\mathbf{h}^{\mathrm{T}}\mathbf{h} + \mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{h}\right) \tag{32}$$

And the marginal pdfs are

$$p(\mathbf{x}) = \frac{1}{Z_{\mathbf{x}}(\boldsymbol{\theta}, \mathbf{W})} \exp\left(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x} + \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{x}\right) \tag{33}$$

$$p(\mathbf{h}) = \frac{1}{Z_{\mathbf{h}}(\boldsymbol{\theta}, \mathbf{W})} \exp\left(-\frac{1}{2}\mathbf{h}^{\mathrm{T}}\mathbf{h} + \mathbf{1}^{\mathrm{T}}\log(\mathbf{1} + \exp(\boldsymbol{\theta} + \mathbf{W}\mathbf{h}))\right) \tag{34}$$

where

$$Z_{\mathbf{x}}(\boldsymbol{\theta}, \mathbf{W}) = (2\pi)^{-\frac{J}{2}} Z(\boldsymbol{\theta}, \mathbf{W})$$
$$Z_{\mathbf{h}}(\boldsymbol{\theta}, \mathbf{W}) = Z(\boldsymbol{\theta}, \mathbf{W})$$

here $J = |\mathbf{h}|$ is the number of latent topics (hidden variables).

Therefore,

$$\begin{aligned} Z_{\mathbf{x}}(\boldsymbol{\theta}, \mathbf{W}) &= (2\pi)^{-\frac{J}{2}} Z_{\mathbf{h}}(\boldsymbol{\theta}, \mathbf{W}) \\ &= \sum_{\mathbf{h}} \exp\left(\left(-\frac{J}{2}\log 2\pi - \frac{1}{2}\mathbf{h}^{\mathrm{T}}\mathbf{h}\right) + \mathbf{1}^{\mathrm{T}}\log(\mathbf{1} + \exp(\boldsymbol{\theta} + \mathbf{W}\mathbf{h}))\right) \\ &= \sum_{\mathbf{h}} q(\mathbf{h}) \exp\left(\mathbf{1}^{\mathrm{T}}\log(\mathbf{1} + \exp(\boldsymbol{\theta} + \mathbf{W}\mathbf{h}))\right) \\ &\equiv \left\langle \exp\left(\mathbf{1}^{\mathrm{T}}\log(\mathbf{1} + \exp(\boldsymbol{\theta} + \mathbf{W}\mathbf{h}))\right) \right\rangle_{q(\mathbf{h})} \end{aligned} \tag{35}$$

where

$$q(\mathbf{h}) = (2\pi)^{-\frac{J}{2}} \exp\left(\sum_{j=1}^{J} -\frac{1}{2}h_j^2\right) \sim \mathcal{N}(\mathbf{0}, I_J)$$

Thus by introducing a vector of random variable $\mathbf{t} = \boldsymbol{\theta} + \mathbf{W}\mathbf{h}$, the partition function of the GB-EFH model equals the expectation of the following random variable

$$Z(\mathbf{t}) = \prod_i (1 + t_i) \tag{36}$$

under the multivariate lognormal distribution $\mathbf{t} \sim \text{LogNormal}(\boldsymbol{\theta}, \mathbf{W}\mathbf{W}^T)$.