2017

On the instability and degeneracy of deep learning models

Andee Kaplan
Iowa State University

Daniel J. Nordman
Iowa State University, dnordman@iastate.edu

Stephen B. Vardeman
Iowa State University, vardeman@iastate.edu

On the instability and degeneracy of deep learning models

Abstract
A probability model exhibits instability if small changes in a data outcome result in large, and often unanticipated, changes in probability. This instability is a property of the probability model, rather than the fitted parameter vector. For correlated data structures found in several application areas, there is increasing interest in predicting/identifying such sensitivity in model probability structure. We consider the problem of quantifying instability for general probability models defined on sequences of observations, where each sequence of length N has a finite number of possible values. A sequence of probability models results, indexed by N, that accommodates data of expanding dimension. Model instability is formally shown to occur when a certain log-probability ratio under such models grows faster than N. In this case, a one component change in the data sequence can shift probability by orders of magnitude. Also, as instability becomes more extreme, the resulting probability models are shown to tend to degeneracy, placing all their probability on potentially small portions of the sample space. These results on instability apply to large classes of models commonly used in random graphs, network analysis, and machine learning contexts.

Keywords
Degeneracy, Instability, Classification, Deep Learning, Graphical Models

Disciplines
Industrial Engineering | Industrial Technology | Statistics and Probability

Comments
This is a preprint of the article Kaplan, Andee, Daniel Nordman, and Stephen Vardeman. "A note on the instability and degeneracy of deep learning models." *arXiv preprint arXiv:1612.01159v2* (2017).

On the instability and degeneracy of deep learning models

Andee Kaplan
Department of Statistics, Iowa State University
and
Daniel Nordman
Department of Statistics, Iowa State University
and
Stephen Vardeman
Department of Statistics and Department of Industrial and Manufacturing Systems Engineering, Iowa State University

November 8, 2017

Abstract

A probability model exhibits instability if small changes in a data outcome result in large, and often unanticipated, changes in probability. This instability is a property of the probability model, rather than the fitted parameter vector. For correlated data structures found in several application areas, there is increasing interest in predicting/identifying such sensitivity in model probability structure. We consider the problem of quantifying instability for general probability models defined on sequences of observations, where each sequence of length N has a finite number of possible values. A sequence of probability models results, indexed by N, that accommodates data of expanding dimension. Model instability is formally shown to occur when a certain log-probability ratio under such models grows faster than N. In this case, a one component change in the data sequence can shift probability by orders of magnitude. Also, as instability becomes more extreme, the resulting probability models are shown to tend to degeneracy, placing all their probability on potentially small portions of the sample space. These results on instability apply to large classes of models commonly used in random graphs, network analysis, and machine learning contexts.

Keywords: Degeneracy, Instability, Classification, Deep Learning, Graphical Models

1 Introduction

We consider the behavior, and the potential impropriety, of sequences of discrete probability models built to incorporate observations of increasing sample size N. Interest is in identifying instability in such models, which is roughly characterized by probabilities with extreme sensitivity to small changes in data configuration. The concept of instability was introduced in the field of statistical physics (i.e., point processes) by Ruelle (1999) and then further extended by Schweinberger (2011) for a family of exponential models. At issue, models exhibiting instability are typically undesirable as these tend to provide poor representations of data or data-generation. As an example, such models can include near-degenerate distributions that assign essentially all probability mass to only a subset of an overall sample space. The latter issue in connection to degeneracy has been recognized as a concern in that dominant model outcomes may not resemble observed data (cf. Handcock 2003). As a compounding issue, model instability often has direct negative impacts for statistical inference and computations based on likelihood functions. Namely, volatilities in probability structure can potentially hamper the numerical evaluations required for maximum likelihood estimation as well as other model-based simulations via Markov Chain Monte Carlo (MCMC). These reasons motivate our general study of instability for a broad class of probability models, described next.

In the model framework, let $\boldsymbol{X}_N = (X_1, \ldots, X_N)$ denote a collection of discrete random variables with a finite sample space, $\mathcal{X}^N$, represented as some N-fold Cartesian product. That is, $\mathcal{X}$ with $|\mathcal{X}| < \infty$ denotes the set of potential outcomes for each single variable X_i, so that the product space $\mathcal{X}^N$ corresponds to values for the variables $\boldsymbol{X}_N = (X_1, \ldots, X_N)$. For each N, let $P_{\boldsymbol{\theta}_N}$ denote a probability model on $\mathcal{X}^N$, under which $P_{\boldsymbol{\theta}_N}(x_1, \ldots, x_N) > 0$ is the probability of the data outcome $(x_1, \ldots, x_N) \in \mathcal{X}^N$. In this, we assume that the model support of $P_{\boldsymbol{\theta}_N}$ is the sample space $\mathcal{X}^N$. This framework produces probability models $P_{\boldsymbol{\theta}_N}$, indexed by a generic sequence of parameters $\boldsymbol{\theta}_N$, to describe data $\boldsymbol{X}_N$ of any given sample size $N \geq 1$. For simplicity, we will refer to this distributional class as *Finite Outcome Everywhere Supported (FOES)* models in the following. The dimension and structure of such parameters are generic, without restriction, though natural cases will be seen to include those where $\boldsymbol{\theta}_N \in \mathbb{R}^{q(N)}$ for some arbitrary integer-valued function $q(\cdot) \geq 1$.

Section 2 provides some examples of FOES models encountered in graph/network analysis and machine learning (i.e., deep learning models). These are used as references for later illustrations. Section 3 then establishes several formal results for FOES models with regard to instability. Schweinberger (2011) originally developed instability results specific to a certain class of discrete exponential models. For similar exponential models with random networks, Handcock (2003) studied model degeneracy, where a probability model places near complete mass on modes and may thereby narrow the feasible model outcomes. As findings here and from Schweinberger

(2011) suggest, model instability and degeneracy may also be related by viewing degeneracy as an extreme, or limiting form, of instability. Our main results establish a broad characterization of model instability, appropriate across the whole FOES model class, that incorporates results of Schweinberger (2011) as a special case. We prescribe a general and simple condition for identifying instability in a FOES model sequence, which quantifies whether certain maximal probabilities in a FOES model are too extreme relative to the sample size N. When these conditions are met, the probability structure of a FOES model is shown to exhibit extreme sensitivity, with probability assignments possessing extreme peaks and troughs across nearly identical outcomes. As the measure of model instability increases, probabilities from an unstable FOES model additionally increase in volatility and provably slide into degeneracy. Section 5 then emphasizes the implications of such model instability, showing that such impropriety can be expected to numerically hinder maximum likelihood estimation and MCMC-based simulations. As one potential remedy, suggestions are given for constraining model parameterizations to avoid the most problematic regions of the parameter space. Proofs of the main results appear in Appendix A.

2 Examples

Many model families fall under the umbrella of FOES models. For illustration, this section presents three specific examples of FOES models, including models with deep architectures.

2.1 Discrete Exponential Family Models

For random variables $\boldsymbol{X} \equiv \boldsymbol{X}_N = (X_1, \ldots, X_N)$ with sample space $\mathcal{X}^N$, $|\mathcal{X}| < \infty$, consider an exponential family model for $\boldsymbol{X}$ with probability mass function given by

$$p_{N,\boldsymbol{\theta}}(\boldsymbol{x}) = \exp\left[\boldsymbol{\eta}^T(\boldsymbol{\theta})\boldsymbol{g}_N(\boldsymbol{x}) - \psi(\boldsymbol{\theta})\right], \qquad \boldsymbol{x} \in \mathcal{X}^N, \tag{1}$$

depending on parameter vector $\boldsymbol{\theta} \in \Theta_N \subset \mathbb{R}^k$ and natural parameter function $\boldsymbol{\eta} : \mathbb{R}^k \mapsto \mathbb{R}^L$ with fixed positive integers k and L denoting their dimensions. Above, $\boldsymbol{g}_N : \mathcal{X}^N \mapsto \mathbb{R}^L$ is a vector of sufficient statistics, while

$$\psi(\boldsymbol{\theta}) = \log \sum_{\boldsymbol{x} \in \mathcal{X}^N} \exp\left[\boldsymbol{\eta}^T(\boldsymbol{\theta})\boldsymbol{g}_N(\boldsymbol{x})\right], \qquad \boldsymbol{\theta} \in \Theta_N \equiv \{\boldsymbol{\theta} \in \mathbb{R}^k : \psi(\boldsymbol{\theta}) < \infty\},$$

denotes the normalizing function with parameter space Θ_N. The natural parameter function $\eta(\boldsymbol{\theta})$ has a linear form (i.e., $\eta(\boldsymbol{\theta}) = \boldsymbol{A}\boldsymbol{\theta}$ for a given $L \times k$ matrix $\boldsymbol{A}$) in many common model formulations, though may also be nonlinear (e.g., curved exponential families). In the linear case, $\eta(\boldsymbol{\theta}) = \boldsymbol{\theta}$ may be generally assumed in the exponential parameterization with a minor modification to the definition of sufficient statistics $\boldsymbol{g}_N(\boldsymbol{x})$.

Such discrete exponential family models are special cases of the FOES models, as seen by defining $P_{\boldsymbol{\theta}_N}(\boldsymbol{x}) \equiv p_{N,\boldsymbol{\theta}_N}(\boldsymbol{x}) > 0$, $\boldsymbol{x} \in \mathcal{X}^N$, based on (1) and a parameter sequence $\boldsymbol{\theta}_N \in \Theta_N \subset \mathbb{R}^k$. For example, if observations $\boldsymbol{X} = (X_1, \ldots, X_N)$ correspond to N independent and identically distributed Bernoulli random variables, each indicating a binary 0-1 outcome, the resulting probabilities have exponential form (1) given by

$$P_{\boldsymbol{\theta}_N}(\boldsymbol{x}) \propto \exp\left[\boldsymbol{\theta}_N \sum_{i=1}^{N} x_i\right], \qquad \boldsymbol{x} = (x_1, \ldots, x_N) \in \{0,1\}^N, \tag{2}$$

with sufficient statistic $\boldsymbol{g}_N(\boldsymbol{x}) \equiv \sum_{i=1}^{N} x_i$ and "log odds ratio" parameter $\boldsymbol{\theta}_N \equiv \log[P_{\boldsymbol{\theta}_N}(X_i = 1)/P_{\boldsymbol{\theta}_N}(X_i = 0)] \in \mathbb{R}$. More generally, supposing $\boldsymbol{X} = (X_1, \ldots, X_N)$ represent N independent trials, each assuming an outcome $\{1, \ldots, k\}$ among k possibilities (e.g., a die roll), a multinomial distribution is given by

$$P_{\boldsymbol{\theta}_N}(\boldsymbol{x}) \propto \exp\left[\boldsymbol{\theta}_N^T \boldsymbol{g}_N(\boldsymbol{x})\right] = \exp\left[\sum_{j=1}^{k} \theta_{j,N} \sum_{i=1}^{N} \mathbb{I}(x_i = j)\right], \qquad \boldsymbol{x} \in \{1, \ldots, k\}^N, \tag{3}$$

with sufficient statistic $\boldsymbol{g}_N(\boldsymbol{x})$ involving a count $\sum_{i=1}^{N} \mathbb{I}(x_i = j)$ for each outcome $j \in \{1, \ldots, k\}$, where $\mathbb{I}(\cdot)$ denotes the indicator function, and parameters $\boldsymbol{\theta}_N = (\theta_{1,N}, \ldots, \theta_{k,N}) \in \mathbb{R}^k$ defining log-probability ratios $\theta_{i,N} - \theta_{j,N} = \log[P_{\boldsymbol{\theta}_N}(X_1 = i)/P_{\boldsymbol{\theta}_N}(X_1 = j)]$. In addition to such standard models for discrete independent data, exponential models of FOES type commonly arise with dependent spatial data (Besag 1974) and network/relational data (Wasserman and Faust 1994; Handcock 2003). For a random graph or network with, say, n nodes, consider $N = \binom{n}{2}$ random edges where the ith edge is associated with a pair of nodes $s_i \equiv \{v_i, u_i\}$ and a binary variable $X_i \in \{0,1\}$ indicating presence/absence of an edge among the node pair s_i, $i = 1, \ldots, N$. Here the length N of the edge variable sequence $\boldsymbol{X} = (X_1, \ldots, X_N)$ increases as a function of node number n and corresponding exponential models often incorporate graph topographical features derived from $\boldsymbol{X}$. As an example, consider a graph model of exponential/FOES form prescribed by

$$P_{\boldsymbol{\theta}_N}(\boldsymbol{x}) \propto \exp\left[\sum_{j=1}^{3} \theta_{j,N} g_{j,N}(\boldsymbol{x})\right], \qquad \boldsymbol{x} = (x_1, \ldots, x_N) \in \{0,1\}^N, \tag{4}$$

$$g_{1,N}(\boldsymbol{x}) \equiv \sum_{i=1}^{N} x_i, \qquad g_{2,N}(\boldsymbol{x}) \equiv \sum_{\substack{1 \le i < j \le N, \\ s_i \cap s_j \neq \emptyset}} x_i x_j, \qquad g_{3,N}(\boldsymbol{x}) \equiv \sum_{\substack{1 \le i < j < \ell \le N, \\ s_i \cap s_j \neq \emptyset, s_i \cap s_\ell \neq \emptyset, s_j \cap s_\ell \neq \emptyset}} x_i x_j x_\ell,$$

involving the numbers of edges, 2-stars and triangles among an outcome $\boldsymbol{x}$ given by $g_{1,N}(\boldsymbol{x})$, $g_{2,N}(\boldsymbol{x})$ and $g_{3,N}(\boldsymbol{x})$, respectively, along with $k = 3$ real parameters $\boldsymbol{\theta}_N \equiv (\theta_{1,N}, \theta_{2,N}, \theta_{3,N})$. For this network model (4) in particular, as well as for more general models of form (1), Schweinberger (2011) considered instability in such exponential models with sequences of fixed parameters $\boldsymbol{\theta}_N = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k$, $N \ge 1$, of fixed dimension k.

For model sequences $P_{\boldsymbol{\theta}_N}(\boldsymbol{x}) \equiv p_{N,\boldsymbol{\theta}_N}(\boldsymbol{x})$ of the exponential type (1), such as those in (2)-(4), note that the dimension k of the parameter $\boldsymbol{\theta}_N \in \Theta \subset \mathbb{R}^k$ necessarily remains the same for all sample sizes $N \geq 1$ as the form of the natural parameter function $\eta(\cdot)$ in (1) and the number of sufficient statistics $\boldsymbol{g}_N(\boldsymbol{x})$ do not depend on N. Consequently, $\boldsymbol{\theta}_N$ lies in a parameter space of fixed Euclidean dimension k. However, this aspect need not be true for other types of FOES models considered in Sections 2.2 - 2.3, where instead the numbers of parameters and sufficient statistics commonly increase with the sample size N.

2.2 Restricted Boltzmann Machines

A restricted Boltzmann machine (RBM) is an undirected graphical model specified for discrete or continuous random variables, with binary variables being most common (cf. Smolensky 1986). A RBM architecture has two layers, hidden ($\mathcal{H}$) and visible ($\mathcal{V}$), with conditional independence within each layer. Let $\boldsymbol{X} = (X_1, \ldots, X_N)$ denote the N random variables for visibles with support $\mathcal{X}^N$ and $\boldsymbol{H} = (H_1, \ldots, H_{N_\mathcal{H}})$ denote the $N_\mathcal{H}$ random variables for hiddens with support $\mathcal{X}^{N_\mathcal{H}}$ where $\mathcal{X} = \{-1, 1\}$. For parameters $\boldsymbol{\theta}_N^\mathcal{H} \in \mathbb{R}^{N_\mathcal{H}}$, $\boldsymbol{\theta}_N^\mathcal{V} \in \mathbb{R}^N$, and $\boldsymbol{\theta}_N^{\mathcal{H}\mathcal{V}}$ as a real matrix with dimension $N_\mathcal{H} \times N$, the RBM model for $\tilde{\boldsymbol{X}} = (\boldsymbol{X}, \boldsymbol{H})$ has the joint probability mass function

$$\tilde{P}_{\boldsymbol{\theta}_N}(\tilde{\boldsymbol{x}}) = \exp\left[(\boldsymbol{\theta}_N^\mathcal{H})^T \boldsymbol{h} + (\boldsymbol{\theta}_N^\mathcal{V})^T \boldsymbol{x} + \boldsymbol{h}^T \boldsymbol{\theta}_N^{\mathcal{H}\mathcal{V}} \boldsymbol{x} - \psi(\boldsymbol{\theta}_N)\right], \quad \tilde{\boldsymbol{x}} = (\boldsymbol{x}, \boldsymbol{h}) \in \{\pm 1\}^{N+N_\mathcal{H}} \tag{5}$$

with normalizing function

$$\psi(\boldsymbol{\theta}_N) = \log \sum_{\tilde{\boldsymbol{x}} \in \{\pm 1\}^{N+N_H}} \exp\left[(\boldsymbol{\theta}_N^\mathcal{H})^T \boldsymbol{h} + (\boldsymbol{\theta}_N^\mathcal{V})^T \boldsymbol{x} + \boldsymbol{h}^T \boldsymbol{\theta}_N^{\mathcal{H}\mathcal{V}} \boldsymbol{x}\right].$$

Let $\boldsymbol{\theta}_N = (\boldsymbol{\theta}_N^\mathcal{H}, \boldsymbol{\theta}_N^\mathcal{V}, \boldsymbol{\theta}_N^{\mathcal{H}\mathcal{V}}) \in \Theta_N \equiv \mathbb{R}^{q(N)}$, with $q(N) = N + N_\mathcal{H} + N * N_\mathcal{H}$, denote the parameter vector for the RBM, as indexed by the number N of visible random variables (which may differ from the actual lengths of these parameter vectors). The probability mass function for the visible variables $\boldsymbol{X} = (X_1, \ldots, X_N)$ follows from marginalizing the joint specification to yield

$$P_{\boldsymbol{\theta}_N}(\boldsymbol{x}) = \sum_{\boldsymbol{h} \in \{\pm 1\}^{N_\mathcal{H}}} \tilde{P}_{\boldsymbol{\theta}_N}(\boldsymbol{x}, \boldsymbol{h}), \qquad \boldsymbol{x} \in \{\pm 1\}^N \equiv \mathcal{X}^N. \tag{6}$$

Here the baseline model (5) for hidden/visible variables is a linear exponential one in sufficient statistics $(\tilde{\boldsymbol{X}}, \boldsymbol{X}^T \boldsymbol{H})$ using $\tilde{\boldsymbol{X}} = (\boldsymbol{X}, \boldsymbol{H})$ from (5), but the form differs from the previous exponential models in (1) in that the lengths of parameters $\boldsymbol{\theta}_N$ and statistics $(\tilde{\boldsymbol{X}}, \boldsymbol{X}^T \boldsymbol{H})$ increase to incorporate more visible variables. That is, in contrast to (1), the natural parameter function involved in the RBM model (5), as the identity mapping of the parameters $\boldsymbol{\theta}_N \in \mathbb{R}^{q(N)}$, naturally grows in dimension $q(N) \to \infty$ to accommodate visible variables $X_1, \ldots, X_N$ of increasing sample size $N \to \infty$. Additionally, one may further arbitrarily choose the number $N_\mathcal{H}$ of hidden variables $\boldsymbol{H}$ in the joint RBM model (5) to define a marginal model (6) for the N visible variables $\boldsymbol{X}$,

and the number $N_{\mathcal{H}}$ of hiddens may also potentially increase with N. Because $|\mathcal{X}| = 2$ and $P_{\boldsymbol{\theta}_N}(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathcal{X}^N$, the RBM specification (6) for visibles corresponds to a FOES model, while the joint distribution (5) for $(\boldsymbol{X}, \boldsymbol{H})$ is also a FOES model. As this example also indicates, any model formed by marginalizing a base FOES model class, such as the RBM joint specification (5), is again a FOES model.

2.3 Deep Learning

Consider two models with "deep architecture" that contain multiple hidden (or latent) layers in addition to a visible layer of data, namely a deep Boltzmann machine (Salakhutdinov and Hinton 2009) and a deep belief network Hinton, Osindero, and Teh (2006)]. Let M denote the number of hidden layers included in the model and let $N_{(H,1)}, \ldots, N_{(H,M)}$ denote the numbers of hidden variables within each hidden layer. Then the random vector $\tilde{\boldsymbol{X}} = \{H_1^{(1)}, \ldots, H_{N_{(H,1)}}^{(1)}, \ldots, H_1^{(M)}, \ldots, H_{N_{(H,M)}}^{(M)}, \boldsymbol{X}\}$ collects both the hidden variables $\{H_i^{(j)} : i = 1, \ldots, N_{(H,j)}, j = 1, \ldots, M\}$ and visible variables $\boldsymbol{X} = (X_1, \ldots, X_N)$ in a deep probabilistic model. Each variable outcome will again lie in $\mathcal{X} = \{-1, 1\}$.

Deep Boltzmann machine (DBM). The DBM class of models maintains conditional independence within all layers in the model by stacking RBM models and only allowing conditional dependence between neighboring layers. The joint probability mass function for a DBM is

$$\tilde{P}_{\boldsymbol{\theta}_N}(\tilde{\boldsymbol{x}}) = \exp\left[\sum_{i=1}^{M} \boldsymbol{\alpha}^{(i)T}\boldsymbol{h}^{(i)} + \boldsymbol{\beta}^T\boldsymbol{x} + \boldsymbol{h}^{(1)T}\Gamma^{(0)}\boldsymbol{x} + \sum_{i=1}^{M-1} \boldsymbol{h}^{(i)T}\Gamma^{(i)}\boldsymbol{h}^{(i+1)} - \psi(\boldsymbol{\theta}_N)\right],$$

for $\tilde{\boldsymbol{x}} = (\boldsymbol{h}^{(1)}, \ldots, \boldsymbol{h}^{(M)}, \boldsymbol{x}) \in \mathcal{X}^{N_{(H,1)}+\cdots+N_{(H,M)}+N}$ where

$$\psi(\boldsymbol{\theta}_N) = \log \sum_{\tilde{\boldsymbol{x}} \in \mathcal{X}^{N_{(H,1)}+\cdots+N_{(H,M)}+N}} \exp\left[\sum_{i=1}^{M} \boldsymbol{\alpha}^{(i)T}\boldsymbol{h}^{(i)} + \boldsymbol{\beta}^T\boldsymbol{x} + \boldsymbol{h}^{(1)T}\Gamma^{(0)}\boldsymbol{x} + \sum_{i=1}^{M-1} \boldsymbol{h}^{(i)T}\Gamma^{(i)}\boldsymbol{h}^{(i+1)}\right],$$

is the normalizing function for $\boldsymbol{\theta}_N = (\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(M)}, \boldsymbol{\beta}, \Gamma^{(0)}, \ldots, \Gamma^{(M-1)}) \in \Theta_N \subset \mathbb{R}^{q(N)}$, consisting of model parameters $\boldsymbol{\beta} \in \mathbb{R}^N$, $\boldsymbol{\alpha}^{(i)} \in \mathbb{R}^{N_{(H,i)}}$, $i = 1, \ldots, M$, along with a matrix $\Gamma^{(0)}$ of dimension $N_{(H,1)} \times N$, and matrices $\Gamma^{(i)}$ of dimension $N_{(H,i)} \times N_{(H,i+1)}$ for $i = 1, \ldots, M-1$. The combined parameter vector $\boldsymbol{\theta}_N$ has total length $q(N) = N_{(H,1)} + \cdots N_{(H,M)} + N + N_{(H,1)} * N + N_{H,2} * H_{(H,1)} + \cdots + N_{(H,M)} * H_{(H,M)-1}$. The probability mass function for the visible random variables $X_1, \ldots, X_N$ follows from this joint specification as

$$P_{\boldsymbol{\theta}_N}(\boldsymbol{x}) = \sum_{(\boldsymbol{h}^{(1)}, \ldots, \boldsymbol{h}^{(M)}) \in \mathcal{X}^{N_{(H,1)}+\cdots+N_{(H,M)}}} \tilde{P}_{\boldsymbol{\theta}_N}(\boldsymbol{h}^{(1)}, \ldots, \boldsymbol{h}^{(M)}, \boldsymbol{x}), \qquad \boldsymbol{x} \in \mathcal{X}^N.$$

Again like the RBM case, the DBM model specification is an example of a FOES model.

Deep belief network (DBN). A DBN resembles a DBM in that there are multiple layers of latent random variables stacked in a deep architecture with no conditional dependence between

layers. The difference between the DBM and DBN models is that all but the last stacked layer in a DBN are Bayesian networks (see Pearl 1985), rather than RBMs. A Bayesian network is a probabilistic graphical model that defines conditional dependence to be directed, rather than undirected (as with the RBM). Thus for visibles $X_1, \ldots, X_N$ with support $\mathcal{X}^N, |\ \mathcal{X}\ | < \infty$, a DBN is also a FOES model with $q(N)$ the length of parameter vector is dependent on the dimension of the visibles because $P_{\boldsymbol{\theta}_N}(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathcal{X}^N$. Commonly, as in logistic belief nets (Neal 1992), a "weight" parameter is placed on each interaction between visibles, $X_1, \ldots, X_N$, and the first layer of latent variables, $H_1^{(1)}, \ldots, H_{N_{(H,1)}}^{(1)}$, satisfying the definition of a FOES model.

3 Main Results on Model Instability

We now present a formal definition for instability of FOES models as well as a simple condition for identifying instability in a FOES model sequence.

3.1 A Criterion for Instability

To define a measure of instability in FOES models, it is useful to consider the behavior of data models P_{θ_N}, again supported on a set $\mathcal{X}^N$ of outcomes for $\boldsymbol{X} \equiv \boldsymbol{X}_N = (X_1, \ldots, X_N)$, in connection to the sample size N. A relevant quantity to this end is a log-ratio of extremal probabilities (LREP), defined as

$$\text{LREP}(\boldsymbol{\theta}_N) = \log \left[\frac{\max\limits_{\boldsymbol{x} \in \mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\boldsymbol{x})}{\min\limits_{\boldsymbol{x} \in \mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\boldsymbol{x})} \right], \tag{7}$$

based on maximum and minimal model probabilities. In what follows, the main idea is that instability, and other negative model features, can be associated with a FOES model formulation for N random variables where the LREP (7) is overly large relative to the sample size N. That is, a sequence of FOES probability models $P_{\boldsymbol{\theta}_N}$ results in specifying the distribution of observations $\boldsymbol{X} = (X_1, \ldots, X_N)$ for each sample size $N \geq 1$ and instability will generally occur among these models whenever the corresponding LREP (7) grows faster than N. This leads to the following definition.

Definition 1 (S-unstable FOES model)**.** A FOES model formulation for $\boldsymbol{X}_N = (X_1, \ldots, X_N)$ is *Schweinberger-unstable* or *S-unstable* if

$$\lim_{N \to \infty} \frac{1}{N} \text{LREP}(\boldsymbol{\theta}_N) \equiv \lim_{N \to \infty} \frac{1}{N} \log \left[\frac{\max\limits_{\boldsymbol{x} \in \mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\boldsymbol{x})}{\min\limits_{\boldsymbol{x} \in \mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\boldsymbol{x})} \right] = \infty \tag{8}$$

as the number of variables increases ($N \to \infty$).

In other words, a model is S-unstable if $\text{LREP}(\boldsymbol{\theta}_N)/N$ is an unbounded sequence of sample size N; namely, given any $C > 0$, there exists an integer $N_C > 0$ so that $\text{LREP}(\boldsymbol{\theta}_N)/N > C$ holds for all $N \geq N_C$. A FOES model formulation may be termed S-stable if it fails to be S-unstable, i.e., if $\sup_{N\geq 1} \text{LREP}(\boldsymbol{\theta}_N)/N$ is bounded.

This definition of S-unstable is a generalization or reinterpretation of "unstable" used in Schweinberger (2011) by allowing possibly non-exponential family models (e.g., RBM and DBM models in Sections 2.2-2.3 as well as a potentially increasing number $q(N)$ of parameters through the parameter sequence $\boldsymbol{\theta}_N \in \mathbb{R}^{q(N)}$. While this definition differs in form and scope from the original, it does match that in Schweinberger (2011) for the special case of exponential models (cf. Section 2.1 considered there. Section 4 provides several examples of unstable models as well as causes for model instability, where the latter may often be traced to issues in model form (i.e., data functions) and/or parameterization. We next describe several potentially undesirable features associated with S-unstable FOES models.

Remark 1. In the definition (8) of S-instability, we note that the numerical measure $\text{LREP}(\boldsymbol{\theta}_N)/N$ of model instability is invariant to *independent replications* of data. That is, let $M \geq 1$ denote a possible number of replications and consider data $\boldsymbol{Y}_{N,M} \equiv (\boldsymbol{X}_N^{(1)}, \ldots, \boldsymbol{X}_N^{(M)})$ formed by $\{\boldsymbol{X}_N^{(j)}\}_{j=1}^M$ as M iid replications of a random vector $\boldsymbol{X}_N = (X_1, \ldots, X_N)$, where the latter follows a FOES model with probabilities $P_{\boldsymbol{\theta}_N}(\boldsymbol{x}) > 0$, $\boldsymbol{x} \in \mathcal{X}^N$. This leads to a joint model, say $P_{\boldsymbol{\theta}_N}(\boldsymbol{y})$, $\boldsymbol{y} \in \mathcal{X}^{NM}$, for $\boldsymbol{Y}_{N,M}$ consisting of $N * M$ random variables in total. Then, the LREP for $\boldsymbol{Y}_{N,M}$, scaled by associated size, is given by

$$\begin{aligned}\frac{1}{NM}\text{LREP}_{\boldsymbol{Y}_{N,M}}(\boldsymbol{\theta}_N) &\equiv \frac{1}{NM}\log\left[\frac{\max_{\boldsymbol{y}\in\mathcal{X}^{NM}} P_{\boldsymbol{\theta}_N}(\boldsymbol{y})}{\min_{\boldsymbol{y}\in\mathcal{X}^{NM}} P_{\boldsymbol{\theta}_N}(\boldsymbol{y})}\right] \\ &= \frac{1}{NM}\log\left[\frac{\max_{\boldsymbol{x}\in\mathcal{X}^{N}} P_{\boldsymbol{\theta}_N}(\boldsymbol{x})}{\min_{\boldsymbol{x}\in\mathcal{X}^{N}} P_{\boldsymbol{\theta}_N}(\boldsymbol{x})}\right]^M \equiv \frac{1}{N}\text{LREP}_{\boldsymbol{X}_N}(\boldsymbol{\theta}_N),\end{aligned}$$

where $\text{LREP}_{\boldsymbol{X}_N}(\boldsymbol{\theta}_N) \equiv \text{LREP}(\boldsymbol{\theta}_N)$ denotes the log-ratio of extremal probabilities for $\boldsymbol{X}_N$ defined from (7). That is, due to iid properties, the sample-size corrected LREP for $\boldsymbol{Y}_{N,M}$ equals the analog, $\text{LREP}(\boldsymbol{\theta}_N)/N$, from the underlying common data model for $\boldsymbol{X}_N$ alone, regardless of the level $M \geq 1$ of independent replication. Consequently, the definition of an S-unstable model is unaffected by independent replication and all instability properties may be characterized by those of one observation from the common FOES model. For computational purposes, this aspect also implies that if the original data $\boldsymbol{X}_N = (X_1, \ldots, X_N)$ in a FOES model consist of N iid random variables, then the size-scaled log-ratio (7) may be calculated as

$$\frac{1}{N}\text{LREP}(\boldsymbol{\theta}_N) \equiv \frac{1}{N}\text{LREP}_{\boldsymbol{X}_N}(\boldsymbol{\theta}_N) = \log\left[\frac{\max_{x\in\mathcal{X}} P_{\boldsymbol{\theta}_N}(X_1 = x)}{\min_{x\in\mathcal{X}} P_{\boldsymbol{\theta}_N}(X_1 = x)}\right]$$

based on the extremal probabilities of just one random variable X_1.

3.2 Characterizations and Consequences of Instability

As a basic characteristic, S-unstable FOES model sequences have extremely sensitive probability structures. One aspect is that small changes in data configuration can lead to very large changes in probability. Consider, for example, the quantity given by

$$\Delta_N(\boldsymbol{\theta}_N) \equiv \max\left\{\log\frac{P_{\boldsymbol{\theta}_N}(\boldsymbol{x})}{P_{\boldsymbol{\theta}_N}(\boldsymbol{x}^*)} : \boldsymbol{x} \;\&\; \boldsymbol{x}^* \in \mathcal{X}^N \text{ differ in exactly one component}\right\},$$

which represents the biggest log-probability ratio for a one-component change in data outcomes in a FOES model with parameter $\boldsymbol{\theta}_N$. We then have the following result prescribing the behavior of $\Delta_N(\boldsymbol{\theta}_N)$ for S-unstable FOES models.

Theorem 1. *Let $P_{\boldsymbol{\theta}_N}$, with support $\mathcal{X}^N$, $N \geq 1$, be a sequence of FOES models.*

(i) For any integer $N \geq 1$ and any given $C > 0$, if $\mathrm{LREP}(\boldsymbol{\theta}_N)/N > C$ in (7), then

$$\Delta_N(\boldsymbol{\theta}_N) > C,$$

or probabilities from a one-component change in some outcome have log-ratio exceeding C.

(ii) Suppose the FOES model sequence is S-unstable. Then, for all large N and given any arbitrary $C > 0$, there exist outcomes $\boldsymbol{x}, \boldsymbol{x}^* \in \mathcal{X}^N$, differing by one component, such that

$$\frac{P_{\boldsymbol{\theta}_N}(\boldsymbol{x})}{P_{\boldsymbol{\theta}_N}(\boldsymbol{x}^*)} > \exp[NC].$$

Theorem 1(i) is a non-asymptotic result, which connects to the definition of instability in a FOES model through a log-ratio of extreme model probabilities (7) being too large relative to the associated sample size N. If so, Theorem 1(i) guarantees the FOES model must also exhibit correspondingly large changes in probability for very small differences among some data configurations, a property that intuitively captures a notion of instability. Furthermore, and perhaps more seriously under Theorem 1(ii), S-unstable models can never have universally bounded changes in probability among single component variations in data configurations. While not all one-component changes in data may produce massive changes in probability, unstable models must have some such data outcomes with this property. As a consequence, unstable probability structures may exhibit extreme sensitivity through large peaks and troughs over the sample space.

Additionally, S-unstable FOES model sequences are also connected to degenerate models, where *degeneracy* involves assigning essentially all probability to modes within the sample space, which could potentially represent a small subset among the totality of outcomes. For perspective, note that differing sizes of the scaled log-ratio $\mathrm{LREP}(\boldsymbol{\theta}_N)/N$ from (7) induce a spectrum of levels of instability/stability and Theorem 1 indicates increasing sensitivity of model probabilities as

(7) increases. Furthermore, as the instability measure grows and the log-ratio $\text{LREP}(\boldsymbol{\theta}_N)/N$ diverges, as in the definition (8) of S-unstable models, then a FOES model sequence will become degenerate. Theorem 2 provides a formal statement of such degeneracy due to S-instability. For a given $0 < \epsilon < 1$, define a ϵ-modal set of outcomes as

$$\mathcal{M}_{\epsilon,\boldsymbol{\theta}_N} \equiv \left\{ \boldsymbol{x} \in \mathcal{X}^N : \log P_{\boldsymbol{\theta}_N}(\boldsymbol{x}) > (1-\epsilon) \max_{\boldsymbol{y}\in\mathcal{X}^N} \log P_{\boldsymbol{\theta}_N}(\boldsymbol{y}) + \epsilon \min_{\boldsymbol{y}\in\mathcal{X}^N} \log P_{\boldsymbol{\theta}_N}(\boldsymbol{y}) \right\}. \tag{9}$$

Theorem 2. *For any arbitrarily small $0 < \epsilon < 1$, an S-unstable FOES model sequence $P_{\boldsymbol{\theta}_N}$, $N \geq 1$, for $\boldsymbol{X}_N = (X_1, \ldots, X_N)$ satisfies*

$$P_{\boldsymbol{\theta}_N}\left(\boldsymbol{X}_N \in \mathcal{M}_{\epsilon,\boldsymbol{\theta}_N}\right) \to 1 \text{ as } N \to \infty.$$

In other words, as the sample size grows in S-unstable FOES models, all probability tends to concentrate mass on an ϵ-modal set, where ϵ can be made arbitrarily small. Intuitively, the occurrence of such degeneracy can be explained by a type of "reverse" pigeonhole principle for unstable FOES models: if all outcomes should receive positive probability but the maximal probability far exceeds the minimal one in the model, then little probability remains for distribution among remaining model outcomes (i.e., if nearly all available pigeons are stuffed into one hole, the remaining pigeonholes must have few occupants). Degeneracy in unstable models can pose dangers in data modeling as well, particularly when a mode set represents a narrow collection of outcomes among those realistically possible for adequately describing data. In which case, model outcomes may fail to look like data of interest.

Connected to degeneracy, S-unstable FOES models may also exhibit additional kinds of extreme and undesirable sensitivity in probabilities if model parameters $\boldsymbol{\theta}_N$ can further be "dialed" between positive and negative values. That is, some FOES models naturally involve parameter spaces covering a positive-negative spectrum of parameter possibilities, where the signs of parameters provide a standard device for increasing or decreasing probabilities of outcomes in the model formulation. In fact, for many models, the switch of a parameter sign serves to produce reciprocal probabilities, as outlined in the following model assumption about parameter sign reversal (PSR).

Model Condition PSR *(Reciprocal Probabilities from Parameter Sign Reversal)*: Let $P_{\boldsymbol{\theta}_N}$, with support $\mathcal{X}^N$, $N \geq 1$, represent a sequence of FOES models. For each $N \geq 1$ and any outcome $\boldsymbol{x} \in \mathcal{X}^N$, suppose it holds that

$$P_{\boldsymbol{\theta}_N}(\boldsymbol{x}) \cdot P_{-\boldsymbol{\theta}_N}(\boldsymbol{x}) \;=\; \max_{\boldsymbol{y}\in\mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\boldsymbol{y}) \cdot \min_{\boldsymbol{y}\in\mathcal{X}^N} P_{-\boldsymbol{\theta}_N}(\boldsymbol{y}),$$

where $\max_{\boldsymbol{y}\in\mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\boldsymbol{y})$ and $\min_{\boldsymbol{y}\in\mathcal{X}^N} P_{-\boldsymbol{\theta}_N}(\boldsymbol{y})$ denote the maximum and minimum probabilities under parameters $\boldsymbol{\theta}_N$ and $-\boldsymbol{\theta}_N$, respectively.

The above model condition incorporates many standard parameterizations and follows, for instance, whenever $P_{\boldsymbol{\theta}_N}(\boldsymbol{x})/P_{\boldsymbol{\theta}_N}(\boldsymbol{y}) = [P_{-\boldsymbol{\theta}_N}(\boldsymbol{x})/P_{-\boldsymbol{\theta}_N}(\boldsymbol{y})]^{-1}$ holds for outcomes $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}^N$ in a

FOES model. For instance, this latter condition is fulfilled for all linear exponential families from Section 2.1 (e.g., (2)-(4)) as well as all network models from Sections 2.2-2.3 (e.g., (5)-(6)). When parameters can be tuned in sign with effects prescribed in the model condition PSR, unstable FOES models will exhibit further probability sensitivities, as outlined in the following extension of Theorem 2.

Corollary 1. *Let $P_{\boldsymbol{\theta}_N}$, with support $\mathcal{X}^N$, $N \geq 1$, be a sequence of FOES models satisfying model condition PSR. If the models $P_{\boldsymbol{\theta}_N}$ are additionally S-unstable, then*

(i) the models $P_{-\boldsymbol{\theta}_N}$ defined by $-\boldsymbol{\theta}_N$ are also S-unstable;

(ii) and for the complement $\mathcal{M}^c_{\epsilon,\boldsymbol{\theta}_N} \equiv \mathcal{X}^N \setminus \mathcal{M}_{\epsilon,\boldsymbol{\theta}_N}$ of any mode-set $\mathcal{M}_{\epsilon,\boldsymbol{\theta}_N}$ under $\boldsymbol{\theta}_N$ from (9), with $0 < \epsilon < 1$, it holds under $-\boldsymbol{\theta}_N$ that

$$P_{-\boldsymbol{\theta}_N}(\boldsymbol{X}_N \in \mathcal{M}^c_{\epsilon,\boldsymbol{\theta}_N}) \to 1 \qquad \text{as } N \to \infty,$$

while, by Theorem 2, $P_{\boldsymbol{\theta}_N}(\boldsymbol{X}_N \in \mathcal{M}_{\epsilon,\boldsymbol{\theta}_N}) \to 1$ holds for $\boldsymbol{X}_N = (X_1, \ldots, X_N)$ under $\boldsymbol{\theta}_N$.

For unstable models, Corollary 1 shows that shifts in parameters around zero (i.e., from $\boldsymbol{\theta}_N$ to $-\boldsymbol{\theta}_N$) can induce extreme changes in probability among subsets of the sample space, as another manifestation of instability and hyper-sensitivity in probability structure. For one-parameter exponential families, involving a fixed real-valued linear parameter $\boldsymbol{\theta}_N = \theta \in \mathbb{R}$ and sufficient statistic $\boldsymbol{g}_N(\boldsymbol{x}) \in \mathbb{R}$ in (1), Schweinberger (2011 Theorem 3) proved a result similar in spirit, though based on a characterization there in terms of maximum $U_N \equiv \max_{\boldsymbol{x}\in\mathcal{X}^N} g_N(\boldsymbol{x})$ and minimal $L_N \equiv \min_{\boldsymbol{x}\in\mathcal{X}^N} g_N(\boldsymbol{x})$ values of the sufficient statistic. For this case in particular, mode sets have specific, and essentially complementary, forms over positive and negative parameters, namely, $\mathcal{M}_{\epsilon,\boldsymbol{\theta}_N} = \{\boldsymbol{x} \in \mathcal{X}^N : g_N(\boldsymbol{x}) > (1-\epsilon)U_N + \epsilon L_N\}$ and $\mathcal{M}_{\epsilon,-\boldsymbol{\theta}_N} = \{\boldsymbol{x} \in \mathcal{X}^N : g_N(\boldsymbol{x}) < \epsilon U_N + (1-\epsilon)L_N\}$ for any $\boldsymbol{\theta}_N > 0$, and Schweinberger (2011 Theorem 3) showed each mode set collects all mass, under positive and negative parameters, respectively, with unstable models of this exponential type. However, for all unstable FOES models, Corollary 1 generalizes the same principle that unstable models can push all probability to different, and in fact disjoint, parts of the sample space, depending on how parameters fall with respect to zero. This feature can numerically complicate likelihood manipulations, such as maximization or MCMC-based Bayes posterior sampling, as further discussed in Section 5.

Remark 2. Under the model condition PSR, Corollary 1 can also be extended to cases where parameter components $\boldsymbol{\theta}_N = (\boldsymbol{\theta}_{1,N}, \boldsymbol{\theta}_{N,2})$ (say) are not all changed in sign (e.g., $-\boldsymbol{\theta}_N$) but, more generally, are instead altered to another parameter configuration $\boldsymbol{\theta}^A_N = (\boldsymbol{\theta}^A_{1,N}, \boldsymbol{\theta}^A_{2,N})$ involving a switch in sign only among some dominating model parameters $\boldsymbol{\theta}^A_{N,2} = -\boldsymbol{\theta}_{N,2}$ with remaining parameters $\boldsymbol{\theta}^A_{1,N}$ being arbitrarily chosen. If a change sign occurs among parameters $(\pm\boldsymbol{\theta}_{N,2})$ which

dominate the probability structure of the model, then the results of Corollary 1 can still hold with $\boldsymbol{\theta}_N^A$ replacing $-\boldsymbol{\theta}_N$; as an example of one sufficient condition, if $\lim_{N\to\infty} \max_{\boldsymbol{x}\in\mathcal{X}^N} |G_N(\boldsymbol{x}, -\boldsymbol{\theta}_N) - G_N(\boldsymbol{x}, \boldsymbol{\theta}_N^A)| = 0$ holds in addition to Corollary 1 assumptions, where

$$G_N(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\log P_{\boldsymbol{\theta}}(\boldsymbol{x}) - \min_{\boldsymbol{y}\in\mathcal{X}^N} \log P_{\boldsymbol{\theta}}(\boldsymbol{y})}{\max_{\boldsymbol{y}\in\mathcal{X}^N} \log P_{\boldsymbol{\theta}}(\boldsymbol{y}) - \min_{\boldsymbol{y}\in\mathcal{X}^N} \log P_{\boldsymbol{\theta}}(\boldsymbol{y})}, \quad \boldsymbol{x} \in \mathcal{X}^N,$$

represents a standardized form of $\boldsymbol{\theta}$-model probabilities, then the results of Corollary 1 apply to $\boldsymbol{\theta}_N^A$ in addition to $-\boldsymbol{\theta}_N$. As a consequence, an unstable model under $\boldsymbol{\theta}_N$ can then imply that many more unstable models exist over a broader spectrum of possibilities for variations $\boldsymbol{\theta}_N^A$ of $\boldsymbol{\theta}_N$, which involves some amount of sign change among components of $\boldsymbol{\theta}_N$.

4 Illustrations

Model instability can depend intricately on how functions of parameters and data $\boldsymbol{X}_N = (X_1, \ldots, X_n)$ are combined in the formulation of the model probabilities, though some general causes may be identified. As one issue, a broad parameter space (or wide interpretation of this space) may admit some parameters as technically valid that have an undue and often undesirable impact on the model structure for a prescribed data size N. In this case, both the size and dimension of model parameters can be problematic and induce instability. In combination to this last point, further causes of instability may also be traced to the magnitude of statistics in the model. Potentially massive, and thereby unstable, statistics were the primary focus of instability studies of Schweinberger (2011) for certain discrete exponential models having parameters/statistics of fixed dimension. However, as shown in the following, bounded statistics may still lead to instability if the parameter dimension is high. We next provides some examples to illustrate S-instability in FOES models, which also suggest some potential strategies for preventing unstable models.

4.1 Equi-probability Models

As a baseline for comparisons, consider a simplistic model for $\boldsymbol{X}_N = (X_1, \ldots, X_N)$ with uniform probabilities over the sample space, say $P_{\boldsymbol{\theta}_N}(\boldsymbol{x}) = |\mathcal{X}|^{-N}$, $\boldsymbol{x} \in \mathcal{X}^N$, where each random variable has $|\mathcal{X}| \geq 1$ outcomes. In contrast to instability, model probabilities here are completely insensitive to changes in data outcomes across the sample space, and the associated log-ratio of extreme probabilities (7) is

$$\frac{1}{N}\text{LREP}(\boldsymbol{\theta}_N) = 0 \quad \text{(uniform probability model)},$$

which is as small as possible. In fact, a LREP value of zero can only occur for a FOES model having uniform probabilities, and such equi-probability models are always S-stable.

4.2 One-parameter Exponential Models

A fundamental model considered in the instability work of Schweinberger (2011) involves a one-parameter exponential model corresponding to (1) with a real-valued parameter, say $\boldsymbol{\theta}_N = \eta(\boldsymbol{\theta}_N) \in \mathbb{R}$, and sufficient statistic $\boldsymbol{g}_N(\boldsymbol{x}) \in \mathbb{R}$. For such models, upon scaling by sample size N, the log-ratio of extreme probabilities in (7) for assessing instability becomes

$$\frac{1}{N}\text{LREP}(\boldsymbol{\theta}_N) \equiv |\boldsymbol{\theta}_N|\frac{(U_N - L_N)}{N} \quad \text{(one-parameter exponential model)}, \tag{10}$$

where $U_N \equiv \max_{\boldsymbol{x}\in\mathcal{X}^N} g_N(\boldsymbol{x})$ and $L_N \equiv \min_{\boldsymbol{x}\in\mathcal{X}^N} g_N(\boldsymbol{x})$ denote the maximal and minimal values of the single sufficient statistic. In this case, an S-unstable model results, by definition (8), whenever $\lim_{N\to\infty} |\boldsymbol{\theta}_N|(U_N - L_N)/N = \infty$ holds or, in other words, if the combined magnitudes of parameter $|\boldsymbol{\theta}_N|$ and maximal difference $U_N - L_N$ in statistic values are overwhelmingly large relative to the sample size N. If we further assume that $\boldsymbol{\theta}_N = \theta \in \mathbb{R} \setminus \{0\}$ is a fixed (non-zero) parameter for all $N \geq 1$, as considered in Schweinberger (2011), then an S-unstable model results solely if the sufficient statistic admits a value $U_N - L_N$ too large relative to number N of observations, i.e., if $(U_N - L_N)/N \to \infty$ as $N \to \infty$. The latter aspect reflects the definition of Schweinberger (2011), for this setting, that a real-valued *statistic* $g_N(\boldsymbol{x})$ may be classified as *unstable* when $\lim_{N\to\infty} |(U_N - L_N)/N = \infty$ holds and as *stable* otherwise (e.g., if $\sup_{N\geq 1}(U_N - L_N)/N < \infty$).

For illustration, consider the iid Bernoulli model (2) for $\boldsymbol{X}_N = (X_1, \ldots, X_N)$ with log-odds ratio parameter $\boldsymbol{\theta}_N = \log[P_{\boldsymbol{\theta}_N}(X_1 = 1)/P_{\boldsymbol{\theta}_N}(X_1 = 0)] \in \mathbb{R}$. Remark 1 (Section 3.1) then gives the model instability measure (8) directly as

$$\frac{1}{N}\text{LREP}(\boldsymbol{\theta}_N) = |\boldsymbol{\theta}_N| \quad \text{(iid Bernoulli model)},$$

so that an unstable (or stable) model results for a divergent (or bounded) parameter sequence $|\boldsymbol{\theta}_N|$. The above instability expression for the Bernoulli model follows as well from the N-scaled LREP value (10) for a one-parameter exponential distribution, using that the sufficient statistic involved $g_N(\boldsymbol{x}) = \sum_{i=1}^N x_i$, $\boldsymbol{x} = (x_1, \ldots, x_n) \in \{0, 1\}^N$, has maximum and minimum values $U_N = N$ and $L_N = 0$. In this case, Schweinberger (2011) has noted that the statistic is stable (i.e., bounded $(U_N - L_N)/N = 1$) and the Bernoulli model is as well when, in particular, $\boldsymbol{\theta}_N = \theta \in \mathbb{R}$ is fixed for $N \geq 1$.

Alternatively, considering a random graph with $N = \binom{n}{2}$ edges among n nodes, the exponential graph model from (4), when based purely on the number of $g_{2,N}(\boldsymbol{x})$ of 2-stars or solely the number $g_{3,N}(\boldsymbol{x})$ of triangles, $\boldsymbol{x} \in \{0, 1\}^N$, has an measure of instability from (8) as

$$\frac{1}{N}\text{LREP}(\boldsymbol{\theta}_N) = \begin{cases} |\boldsymbol{\theta}_N|(n-2) & \text{(2-star graph model)} \\ |\boldsymbol{\theta}_N|(n-2)/3 & \text{(triangle graph model)}, \end{cases}$$

by using the (one-parameter exponential) LREP formula (10) with statistic maximums $U_N = N(n-2)$ for 2-stars or $U_N = N(n-2)/3$ for triangles and with minimums $L_N = 0$ in both cases. Because the variable number $N \to \infty$ as the node number $n \to \infty$, both counts of 2-stars and triangles are unstable statistics in the sense of Schweinberger (2011) (i.e., $\lim_{N\to\infty}(U_N - L_N)/N = \infty$). Furthermore, both types of graph models are always S-unstable for all possible of parameter sequences $\boldsymbol{\theta}_N \in \mathbb{R}$ that are bounded away from zero (i.e., $\lim_{N\to\infty} \mathrm{LREP}(\boldsymbol{\theta}_N)/N = \infty$ then holds, including the fixed parameter case $\boldsymbol{\theta}_N = \theta \in \mathbb{R} \setminus \{0\}$ from Schweinberger (2011)).

4.3 Fixed-dimensional Linear Exponential Models

As a generalization of the one-parameter exponential case, we next consider linear exponential families (1) with k parameters $\boldsymbol{\theta}_N = (\theta_{1,N}, \ldots, \theta_{k,N})'$ and k sufficient statistics $\boldsymbol{g}_N(\boldsymbol{x}) = (g_{1,N}(\boldsymbol{x}), \ldots, g_{k,N}(\boldsymbol{x}))'$. Here the dimension k of model parameters/statistics is fixed, and we next prescribe a condition helpful to avoiding instability in such models. For this, define $U_{i,N} = \max_{\boldsymbol{x}\in\mathcal{X}^N} g_{i,N}(\boldsymbol{x})$ and $L_{i,N} = \min_{\boldsymbol{x}\in\mathcal{X}^N} g_{i,N}(\boldsymbol{x})$ as the maximal and minimal values of the ith statistic, $i = 1, \ldots, k$, based on observations $\boldsymbol{X}_N = (X_1, \ldots, X_N)$.

Proposition 1. *Let $P_{\boldsymbol{\theta}_N}$, $N \geq 1$, denote linear exponential models (1) with parameters $\boldsymbol{\theta}_N = (\theta_{1,N}, \ldots, \theta_{k,N})' \in \mathbb{R}^k$ and statistics $\boldsymbol{g}_N(\boldsymbol{x}) = (g_{1,N}(\boldsymbol{x}), \ldots, g_{k,N}(\boldsymbol{x}))' \in \mathbb{R}^k$, for fixed $k \geq 1$. Then, the models $P_{\boldsymbol{\theta}_N}$ are S-stable if*

$$\sup_{N\geq 1} \frac{1}{N} \max_{1\leq i\leq k} |\theta_{i,N}|(U_{i,N} - L_{i,N}) < \infty \tag{11}$$

holds, i.e., if $\max_{1\leq i\leq k} |\theta_{i,N}|(U_{i,N} - L_{i,N})/N$ is bounded sequence of sample size N.

Remark 3. In the one-parameter exponential case $k = 1$, recall the exponential model is stable/unstable depending on whether $\mathrm{LREP}(\boldsymbol{\theta}_N)/N = |\theta_{1,N}|(U_{1,N} - L_{1,N})/N \equiv |\boldsymbol{\theta}_N|(U_N - L_N)/N$ in (10) is convergent/divergent. Hence, for $k = 1$, the condition (11) of Proposition 1 captures the same notion of S-stability based on (10).

Proposition 1 provides a sufficient condition for the stability of linear exponential models with fixed parameter dimension $k \geq 1$, whereby an S-stable model is guaranteed if the compounded magnitude of each combination of parameter $\theta_{i,N}$ and sufficient statistic value $(U_{i,N} - L_{i,N})$ is bounded by the sample size N, $i = 1, \ldots, k$. This supports the findings of Schweinberger (2011), who showed degeneracy follows in such models under one type of violation of the condition (11) in Proposition 1 (namely, involving $k > 1$ non-zero parameters with $k - 1$ statistics being $O(N)$ bounded while one statistic diverges in maximal size faster than the number N of observations). To further illustrate the result in Proposition 1, consider the multinomial distribution (3) for $\boldsymbol{X}_N = (X_1, \ldots, X_N)$ having $k \geq 2$ categories $\{1, \ldots, k\}$ and k parameters $\boldsymbol{\theta}_N = (\theta_{1,N}, \ldots, \theta_{k,N})'$.

The variables are iid under this model so that Remark 1 (Section 3.1 yields the corresponding N-scaled log-ratio of extreme probabilities (7) as

$$\begin{aligned}\frac{1}{N}\mathrm{LREP}(\boldsymbol{\theta}_N) &= \frac{\max_{1\le i\le k} P_{\boldsymbol{\theta}_N}(X_1 = i)}{\min_{1\le i\le k} P_{\boldsymbol{\theta}_N}(X_1 = i)} \\ &= \max_{1\le i\le k} \theta_{i,N} - \min_{1\le i\le k} \theta_{i,N} \qquad \text{(iid multinomial model)}.\end{aligned}$$

Hence, a multinomial model sequence is unstable (or stable) depending on whether (or not) the maximal parameter difference $\max_{1\le i\le k}\theta_{i,N} - \min_{1\le i\le k}\theta_{i,N}$ diverges. Furthermore, using that each of the k sufficient (count) statistics from the multinomial model (3) satisfies $(U_{i,N} - L_{i,N})/N = 1$, we see that (11) of Proposition 1 becomes purely a parameter condition, $\sup_{N\ge 1}\max_{1\le i\le k}|\theta_{i,N}| < \infty$, for ensuring that $\mathrm{LREP}(\boldsymbol{\theta}_N)/N = \max_{1\le i\le k}\theta_{i,N} - \min_{1\le i\le k}\theta_{i,N}$ is bounded and stability follows for the multinomial distribution. Additionally, a stable multinomial sequence (i.e., bounded $\mathrm{LREP}(\boldsymbol{\theta}_N)/N$) turns out to be nearly equivalent to (11) (e.g., these are the same if the smallest parameter $\min_{1\le i\le k}|\theta_{i,N}|$ remains bounded).

When the condition (11) of Proposition 1 is violated, this aspect suggests a potentially unstable model that may be investigated more closely. For example, consider the exponential graph model from (4) involving counts of edges, 2-stars and triangles with fixed parameters $\boldsymbol{\theta}_N = (\theta_1, \theta_2, \theta_3)' \in \mathbb{R}^3$ for $N \ge 1$. If either the 2-star parameter $\theta_2 \ne 0$ or triangle parameter $\theta_3 \ne 0$ is non-zero, then $\max_{1\le i\le 3}|\theta_i|(U_{i,N} - L_{i,N})/N \propto (n-2) \to \infty$ holds in (11) by $(U_{2,N} - L_{2,N})/N = 3(U_{3,N} - L_{3,N})/N = (n-2)$ for 2-star and triangle statistics ($i = 2, 3$), so that Proposition 1 hints that an unstable model may result when $|\theta_2| + |\theta_3| \ne 0$. Relatedly, a result from Schweinberger (2011 Result 3) states that this model is unstable for all fixed parameters excluding cases $\theta_2 = \theta_3 = 0$ or $\theta_2 = -\theta_3/3$. However, more is true in line with the instability suggested by Proposition 1 whenever $|\theta_2| + |\theta_3| \ne 0$ (i.e., excluding $\theta_2 = \theta_3 = 0$).

To see this, consider an even number $n > 2$ of nodes and let $\boldsymbol{x}_0$ denote the data outcome in $\mathcal{X}^N \equiv \{0,1\}^N$ with all $N = \binom{n}{2}$ edges being zero, let $\boldsymbol{x}_1$ denote the outcome with all edges being 1, and let $\boldsymbol{x}_2$ denote the edge configuration from dividing the nodes into two equal groups, with no edges within a group and all edges between the groups (so that no triangles exist in $\boldsymbol{x}_2$). Then, the N-scaled log-ratio (7) for the exponential graph model (4) can, by definition, be bounded below by

$$\begin{aligned}\frac{1}{N}\mathrm{LREP}_N(\boldsymbol{\theta}_N) &\ge \max_{i=1,2}\frac{1}{N}\left|\log\left[\frac{P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_i)}{P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_0)}\right]\right| \\ &= (n-2)\max\left\{\left|\theta_2 + \frac{\theta_3}{3} + \frac{\theta_1}{n-2}\right|, \frac{n}{4(n-1)}\left|\theta_2 + \frac{8\theta_1}{n-2}\right|\right\};\end{aligned}$$

a similar expression also holds for an odd node number $n > 2$. Consequently, for all fixed parameters excluding $\theta_2 = \theta_3 = 0$, $\lim_{N\to\infty}\mathrm{LREP}_N(\boldsymbol{\theta}_N)/N = \infty$ then follows and the graph model with 2-stars and triangles is S-unstable, as suggested by the breach of Proposition 1 for

this model when $|\theta_2| + |\theta_3| \neq 0$. That is, instability holds even under $\theta_2 = -\theta_3/3$ case potentially allowed by Schweinberger's (2011) results.

4.4 Latent Variable Models of Increasing Parameter Dimension

We next consider instability of discrete data models based on exponential formulations involving hidden, or latent, variables, such as those probabilistic graphical models described in Sections 2.2-2.3. We will focus on restricted Boltzmann machine (RBM) models (Section 2.2, having one layer of latent variables for simplicity, though the same instability concepts may be extended to other deep learning models (Section 2.3. For N visible variables $\boldsymbol{X} \equiv \boldsymbol{X}_N = (X_1, \ldots, X_n)$ as data, each observation $X_i \in \{\pm 1\}$ being binary, the RBM-based model (6) for $\boldsymbol{X}$ is again of FOES-type, though not an exponential model. However, the distribution of visible variables is induced by an underlying joint exponential model (5) for both visible and latent variables $(\boldsymbol{X}, \boldsymbol{H})$, where $\boldsymbol{H} = (H_1, \ldots, H_{N_\mathcal{H}})$ denotes a vector of $N_\mathcal{H}$ hidden variables (similarly binary). The joint model is of linear exponential form involving $q(N) \equiv N + N_\mathcal{H} + N * N_\mathcal{H}$ sufficient statistics given by $(\boldsymbol{X}, \boldsymbol{H}, \boldsymbol{X}^T\boldsymbol{H})$ and parameters $\boldsymbol{\theta}_N = (\boldsymbol{\theta}_N^\mathcal{V}, \boldsymbol{\theta}_N^\mathcal{H}, \boldsymbol{\theta}_N^{\mathcal{V}\mathcal{H}}) \in \mathbb{R}^{q(N)}$ corresponding to the N visible variables $\boldsymbol{X}$ (i.e., $\boldsymbol{\theta}_N^\mathcal{V} \in \mathbb{R}^N$), the $N_\mathcal{H}$ hidden variables $\boldsymbol{H}$ (i.e., $\boldsymbol{\theta}_N^\mathcal{H} \in \mathbb{R}^{N_\mathcal{H}}$), and the $N * N_\mathcal{H}$ cross-product variables $\boldsymbol{X}^T\boldsymbol{H}$ (i.e., $\boldsymbol{\theta}_N^{\mathcal{V}\mathcal{H}} \in \mathbb{R}^{N * N_\mathcal{H}}$). However, unlike some previous exponential models considered in Sections 4.2-4.3 (cf. Proposition 1, note that the RBM formulation always associates parameters with *bounded* statistics (i.e., the components of $(\boldsymbol{X}, \boldsymbol{H}, \boldsymbol{X}^T\boldsymbol{H})$) so that model instability cannot arise here due to the magnitude of sufficient statistics exceeding the sample size N. Instead, RBM instability may be linked solely to parameter configuration and the fact that the number $q(N) \geq N$ of parameters necessarily increases with the number N of observations $\boldsymbol{X}$, in contrast to previous exponential cases of fixed parameter dimension.

To highlight the instability issues for the RBM model, consider a simple model for N visibles $\boldsymbol{X}$ with no hidden variables ($N_\mathcal{H} = 0$), for which model statements (5)-(6) coincide. An independence model then results for variables in $\boldsymbol{X}$, which has $q(N) = N$ parameters $\boldsymbol{\theta}_N^\mathcal{V} = (\theta_{1,N}^\mathcal{V}, \ldots, \theta_{N,N}^\mathcal{V}) \in \mathbb{R}^N$, and the measure of model instability becomes

$$\frac{1}{N}\text{LREP}(\boldsymbol{\theta}_N) = \frac{2}{N}\sum_{i=1}^{N} |\theta_{i,N}^\mathcal{V}| \quad \text{(RBM model, no hiddens)}.$$

Hence, this model sequence for $\boldsymbol{X}$ will be S-unstable model if the aggregation of absolute parameters grows faster than the number N of parameters/visible variables. Consequently, even for a simplest RBM model involving independence, preventing instability requires careful choice of parameters, particularly with regard to how a parameter configuration differs from zero. For more general RBM models, the number $N_\mathcal{H}$ of hidden variables $\boldsymbol{H}$ can also be chosen arbitrarily (i.e., as some function $N_\mathcal{H} \equiv N_{N,\mathcal{H}}$ of N), which can substantially inflate the number $q(N)$ of

model parameters and further impact model instability through accumulated parameters. To better understand the effects of instability in the RBM structure, Proposition 2 next frames the general behavior of extreme probabilities in the joint RBM model (5) for $(\boldsymbol{X}, \boldsymbol{H})$ and the implied RBM data model (6) for $\boldsymbol{X}$ alone. Specifically, critical measures of instability may be closely connected in both models through tight bounds on their respective LREP values (7). As a result, Proposition 2 shows how an unstable distribution for observations $\boldsymbol{X}$ may be traced to sources of instability in the original joint distribution for $(\boldsymbol{X}, \boldsymbol{H})$. This also suggests a device for avoiding instability, as provided next.

To state the result, let $\mathrm{LREP}_{\boldsymbol{X}}(\boldsymbol{\theta}_N) \equiv \mathrm{LREP}(\boldsymbol{\theta}_N)$ denote the LREP value (7) from the marginal distribution $P_{\boldsymbol{\theta}_N}$ of visibles $\boldsymbol{X}$ in (6) and write the LREP for the joint distribution $\tilde{P}_{\boldsymbol{\theta}_N}$ of $(\boldsymbol{X}, \boldsymbol{H})$ from (5) as

$$\begin{aligned}\mathrm{LREP}_{(\boldsymbol{X},\boldsymbol{H})}(\boldsymbol{\theta}_N) &= \log\left[\frac{\max_{(\boldsymbol{x},\boldsymbol{h})\in\{\pm 1\}^{N+N_{\mathcal{H}}}} \tilde{P}_{\boldsymbol{\theta}_N}(\boldsymbol{x},\boldsymbol{h})}{\min_{(\boldsymbol{x},\boldsymbol{h})\in\{\pm 1\}^{N+N_{\mathcal{H}}}} \tilde{P}_{\boldsymbol{\theta}_N}(\boldsymbol{x},\boldsymbol{h})}\right] \qquad \text{(joint RBM model)} \\ &= \left(\max_{(\boldsymbol{x},\boldsymbol{h})\in\{\pm 1\}^{N+N_{\mathcal{H}}}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x},\boldsymbol{h}) - \min_{(\boldsymbol{x},\boldsymbol{h})\in\{\pm 1\}^{N+N_{\mathcal{H}}}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x},\boldsymbol{h})\right),\end{aligned}$$

written as a function

$$f_{\boldsymbol{\theta}_N}(\boldsymbol{x},\boldsymbol{h}) \equiv \sum_{i=1}^{N} x_i \theta^{\mathcal{V}}_{i,N} + \sum_{j=1}^{N_{\mathcal{H}}} h_j \theta^{\mathcal{H}}_{j,N} + \sum_{i=1}^{N}\sum_{j=1}^{N_{\mathcal{H}}} x_i h_j \theta^{\mathcal{VH}}_{ij,N} \tag{12}$$

of outcomes $\boldsymbol{x} = (x_1, \ldots, x_N) \in \{\pm 1\}^N$, $\boldsymbol{h} = (h_1, \ldots, h_{N_{\mathcal{H}}}) \in \{\pm 1\}^{N_{\mathcal{H}}}$ and parameters $\boldsymbol{\theta}_N \equiv (\boldsymbol{\theta}^{\mathcal{V}}_N, \boldsymbol{\theta}^{\mathcal{H}}_N, \boldsymbol{\theta}^{\mathcal{VH}}_N)$, with $\theta^{\mathcal{V}}_{i,N}$, $\theta^{\mathcal{H}}_{j,N}$ and $\theta^{\mathcal{VH}}_{ij,N}$ denoting respective parameter components, $1 \le i \le N$, $1 \le j \le N_{\mathcal{H}}$. Due to the marginalization steps in defining the distribution (6) of $\boldsymbol{X}$, note that $\mathrm{LREP}_{\boldsymbol{X}}(\boldsymbol{\theta}_N)$ has no immediate analytical expression similar to that of $\mathrm{LREP}_{(\boldsymbol{X},\boldsymbol{H})}(\boldsymbol{\theta}_N)$. For clarity, recall also that S-instability (8) in each model type refers to a respective divergence (i.e., $\lim_{N\to\infty} \mathrm{LREP}_{(\boldsymbol{X},\boldsymbol{H})}(\boldsymbol{\theta}_N)/(N + N_{\mathcal{H}}) = \infty$, $\lim_{N\to\infty} \mathrm{LREP}(\boldsymbol{\theta}_N)/N = \infty$) upon scaling by the corresponding number of variables in a distribution. In the following, let $|\boldsymbol{y}|_1 = \sum_{i=1}^{d} |y_i|$ denote the L1 norm of a generic vector $\boldsymbol{y} = (y_1, \ldots, y_d)$, $d \ge 1$.

Proposition 2. *Let $P_{\boldsymbol{\theta}_N}$ denote a RBM-based data model (6) for $N \ge 1$ visible variables $\boldsymbol{X} \equiv \boldsymbol{X}_N$ derived from $\tilde{P}_{\boldsymbol{\theta}_N}$ as the joint RBM distribution (5) of $(\boldsymbol{X}, \boldsymbol{H})$ involving some number $N_{\mathcal{H}} \equiv N_{N,\mathcal{H}} \ge 0$ of hidden variables $\boldsymbol{H} \equiv \boldsymbol{H}_N$ and parameters $\boldsymbol{\theta}_N \equiv (\boldsymbol{\theta}^{\mathcal{V}}_N, \boldsymbol{\theta}^{\mathcal{H}}_N, \boldsymbol{\theta}^{\mathcal{VH}}_N) \in \mathbb{R}^N \times \mathbb{R}^{N_{\mathcal{H}}} \times \mathbb{R}^{N * N_{\mathcal{H}}}$. Then,*

(i) the instability measure $\mathrm{LREP}(\boldsymbol{\theta}_N)$ for the marginal model $P_{\boldsymbol{\theta}_N}$ of $\boldsymbol{X}$ satisfies

$$|\mathrm{LREP}(\boldsymbol{\theta}_N) - A_N(\boldsymbol{\theta}_N)| \le N_{\mathcal{H}} \log 2$$

for

$$A_N(\boldsymbol{\theta}_N) \equiv \max_{\boldsymbol{x}} \max_{\boldsymbol{h}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x},\boldsymbol{h}) - \min_{\boldsymbol{x}} \max_{\boldsymbol{h}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x},\boldsymbol{h})$$

based on $f_{\boldsymbol{\theta}_N}$ from (12) with components $\boldsymbol{x} \in \{\pm 1\}^N, \boldsymbol{h} \in \{\pm 1\}^{N_{\mathcal{H}}}$.

(ii) The instability measure $\mathrm{LREP}_{(\boldsymbol{X},\boldsymbol{H})}(\boldsymbol{\theta}_N) \equiv (\max_{\boldsymbol{x}} \max_{\boldsymbol{h}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x},\boldsymbol{h}) - \min_{\boldsymbol{x}} \min_{\boldsymbol{h}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x},\boldsymbol{h}))$ *for the joint model* $\tilde{P}_{\boldsymbol{\theta}_N}$ *of* $(\boldsymbol{X},\boldsymbol{H})$ *satisfies*

$$\begin{aligned} 2B_N(\boldsymbol{\theta}_N) + 2|\boldsymbol{\theta}_N^{\mathcal{H}}|_1 \;\geq\; \mathrm{LREP}_{(\boldsymbol{X},\boldsymbol{H})}(\boldsymbol{\theta}_N) &\geq 2\max\left\{B_N(\boldsymbol{\theta}_N), |\boldsymbol{\theta}_N^{\mathcal{H}}|_1\right\} \\ &\geq 2B_N(\boldsymbol{\theta}_N) \\ &\geq A_N(\boldsymbol{\theta}_N) \\ &\geq \max\left\{C_N(\boldsymbol{\theta}_N),\, B_N(\boldsymbol{\theta}_N) - 2|\boldsymbol{\theta}_N^{\mathcal{H}}|_1\right\} \end{aligned}$$

for

$$B_N(\boldsymbol{\theta}_N) \equiv \max_{\boldsymbol{h}} k_{\boldsymbol{\theta}_N}(\boldsymbol{h}) \geq |\boldsymbol{\theta}_N^{\mathcal{V}}|_1, \qquad k_{\boldsymbol{\theta}_N}(\boldsymbol{h}) \equiv \sum_{i=1}^{N} \left| \theta_{i,N}^{\mathcal{V}} + \sum_{j=1}^{N_{\mathcal{H}}} h_j \theta_{ij,N}^{\mathcal{VH}} \right|,$$

and $C_N(\boldsymbol{\theta}_N) \equiv \min_{\boldsymbol{h}} k_{\boldsymbol{\theta}_N}(\boldsymbol{h})$ *based on a function* $k_{\boldsymbol{\theta}_N}(\boldsymbol{h})$ *of hidden variable outcomes* $\boldsymbol{h} = (h_1, \ldots, h_{N_{\mathcal{H}}})$ *and visible-related parameters* $\boldsymbol{\theta}_N^{\mathcal{V}}$ *and* $\boldsymbol{\theta}_N^{\mathcal{VH}}$.

(iii) Assuming $\sup_{N\geq 1} N_{\mathcal{H}}/N < \infty$ *additionally, then the following properties 1.-7. hold:*

1. *an S-unstable visible model* $P_{\boldsymbol{\theta}_N}$ *is equivalent to the condition* $\lim_{N\to\infty} A_N(\boldsymbol{\theta}_N)/N = \infty$*; further,* $P_{\boldsymbol{\theta}_N}$ *is stable when* $A_N(\boldsymbol{\theta}_N)/N$, $N \geq 1$, *is bounded.*
2. *an S-unstable joint model* $P_{\boldsymbol{\theta}_N}$ *is equivalent to the condition* $\lim_{N\to\infty} \max\{|\boldsymbol{\theta}_N^{\mathcal{H}}|_1, B_N(\boldsymbol{\theta}_N)\}/N = \infty$*; further,* $\tilde{P}_{\boldsymbol{\theta}_N}$ *is stable when* $[|\boldsymbol{\theta}_N^{\mathcal{H}}|_1 + B_N(\boldsymbol{\theta}_N)]/N$, $N \geq 1$, *is bounded.*
3. *if the visible model* $P_{\boldsymbol{\theta}_N}$ *is S-unstable, then the joint model* $\tilde{P}_{\boldsymbol{\theta}_N}$ *is also S-unstable.*
4. *when* $\lim_{N\to\infty}(|\boldsymbol{\theta}_N^{\mathcal{V}}|_1 - 2|\boldsymbol{\theta}_N^{\mathcal{H}}|_1)/N = \infty$, *both* $P_{\boldsymbol{\theta}_N}$ *and* $\tilde{P}_{\boldsymbol{\theta}_N}$ *are necessarily S-unstable.*
5. *when* $\lim_{N\to\infty} |\boldsymbol{\theta}_N^{\mathcal{H}}|_1/N = \infty$, *the joint model* $\tilde{P}_{\boldsymbol{\theta}_N}$ *is necessarily S-unstable.*
6. *when* $\sup_{N\geq 1} |\boldsymbol{\theta}_N^{\mathcal{H}}|_1/N < \infty$, *the visible model* $P_{\boldsymbol{\theta}_N}$ *being S-stable or S-unstable is equivalent to the joint model* $\tilde{P}_{\boldsymbol{\theta}_N}$ *being stable or unstable.*
7. *an S-stable visible model* $P_{\boldsymbol{\theta}_N}$ *results if*

$$|\boldsymbol{\theta}_N^{\mathcal{V}}|_1 + |\boldsymbol{\theta}_N^{\mathcal{VH}}|_1 \leq CN, \quad N \geq 1,$$

for some $C > 0$, *while an S-stable joint model* $\tilde{P}_{\boldsymbol{\theta}_N}$ *results if*

$$|\boldsymbol{\theta}_N|_1 \equiv |\boldsymbol{\theta}_N^{\mathcal{V}}|_1 + |\boldsymbol{\theta}_N^{\mathcal{H}}|_1 + |\boldsymbol{\theta}_N^{\mathcal{VH}}|_1 \leq CN, \quad N \geq 1.$$

Remark 4. The condition $\sup_{N\geq 1} N_{\mathcal{H}}/N < \infty$ in Proposition 2(iii) is often mild in practice (i.e., the number $N_{\mathcal{H}}$ of hidden variables is typically not excessively larger than the number N of visible observations). This allows instability results for both marginal and joint RBM models to be more readily stated together, as the numbers N and $N + N_{\mathcal{H}}$ of variables in these models become asymptotically equivalent.

In Proposition 2(iii), the relationships between RBM models with regard to instability, and the effects of different parameter types, follow from the bounds on model instability measures in

Proposition 2(i)-(ii). Generally speaking, all instability in the marginal RBM model for the data $\boldsymbol{X}$ can be attributed to an excessively large model quantity $A_N(\boldsymbol{\theta}_N)$, which predominantly follows when main $\boldsymbol{\theta}_N^{\mathcal{V}}$ and interaction $\boldsymbol{\theta}_N^{\mathcal{VH}}$ parameters related to visible variables are too large in magnitude (e.g., upon accumulation in terms such as $|\boldsymbol{\theta}_N^{\mathcal{V}}|_1$, $B_N(\boldsymbol{\theta}_N)$ or $C_N(\boldsymbol{\theta}_N)$). For example, for any bounded sequence $|\boldsymbol{\theta}_N^{\mathcal{H}}|/N$ of hidden parameters, if main visible parameters $\boldsymbol{\theta}_N^{\mathcal{V}}$ are too extreme ($|\boldsymbol{\theta}_N^{\mathcal{V}}|_1/N \to \infty$), this aspect will guarantee instability in the visible model ($A_N(\boldsymbol{\theta}_N)/N \to \infty$). In fact, the instability measure $A_N(\boldsymbol{\theta}_N) \equiv \max_{\boldsymbol{x}} \max_{\boldsymbol{h}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x}, \boldsymbol{h}) - \min_{\boldsymbol{x}} \max_{\boldsymbol{h}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x}, \boldsymbol{h})$ for marginal/visible model represents a clearly smaller portion of the instability measure $\text{LREP}_{(\boldsymbol{X},\boldsymbol{H})}(\boldsymbol{\theta}_N) \equiv \max_{\boldsymbol{x}} \max_{\boldsymbol{h}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x}, \boldsymbol{h}) - \min_{\boldsymbol{x}} \min_{\boldsymbol{h}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x}, \boldsymbol{h})$ in the joint RBM model, implying that an unstable marginal model (i.e., due to $\boldsymbol{\theta}_N^{\mathcal{V}}$, $\boldsymbol{\theta}_N^{\mathcal{VH}}$) must always translate to an unstable joint model and that further potential causes of instability exist for the joint model, often due to the size $|\boldsymbol{\theta}_N^{\mathcal{H}}|_1$. For example, while the joint RBM model for $(\boldsymbol{X}, \boldsymbol{H})$ must always be unstable due to a diverging combination of visible and/or interaction parameters ($|\boldsymbol{\theta}_N^{\mathcal{V}}|_1/N \to \infty$ or $B_N(\boldsymbol{\theta}_N)/N \to \infty$) (Proposition 2(iii.2)), instability for the joint model can also result when the main hidden parameters $\boldsymbol{\theta}_N^{\mathcal{H}}$ become too large relative to sample size ($|\boldsymbol{\theta}_N^{\mathcal{H}}|_1/N \to \infty$ in Proposition 2(iii.5)). However, under Proposition 2, the main hidden parameters $\boldsymbol{\theta}_N^{\mathcal{H}}$ do not necessarily entail a source of instability for the marginal visible model. To explain this distinction, consider a joint model where all parameters related to visibles are zero, $\boldsymbol{\theta}_N^{\mathcal{V}} = \boldsymbol{\theta}_N^{\mathcal{VH}} = \boldsymbol{0}$, but the hidden-related parameters diverge in sum $|\boldsymbol{\theta}_N^{\mathcal{H}}|_1/N \to \infty$. Here, the explosive behavior among parameters $\boldsymbol{\theta}_N^{\mathcal{H}}$ induces instability in the joint model for $(\boldsymbol{X}, \boldsymbol{H})$ but the marginal model for $\boldsymbol{X}$, however, has a perfectly stable (and in fact uniform) distribution in this case. When the hidden parameters are bounded relative to the sample size ($\sup_{N \geq 1} |\boldsymbol{\theta}_N^{\mathcal{H}}|_1/N < \infty$), then all instability in both the joint and marginal RBM models can be directly linked to excessively large visible $\boldsymbol{\theta}_N^{\mathcal{V}}$ and/or interaction parameters $\boldsymbol{\theta}_N^{\mathcal{VH}}$ so that features of stability/instability must be the same across both models (Proposition 2(iii.6)). Hence, to prevent instability in the joint model, the combined magnitudes of all parameters θ_N must be controlled (cf. Proposition 2(iii.7)), while a stable visible data model technically results in constraining only the sizes of visible-related parameters $\boldsymbol{\theta}_N^{\mathcal{V}}$, $\boldsymbol{\theta}_N^{\mathcal{VH}}$. Nevertheless, because the joint model is often employed in practice for purposes of simulation and simulation-based inference, it is still reasonable to consider parameter choices for ensuring a stable joint model (and, consequently, a stable visible model as well). Further evidence of this is seen in the following numerical example.

In our numerical experiment, we allow the two types of terms (main effects terms corresponding to visible and hidden parameters $\boldsymbol{\theta}_{main} = (\boldsymbol{\theta}_N^{\mathcal{V}}, \boldsymbol{\theta}_N^{\mathcal{H}})$ and interaction parameters $\boldsymbol{\theta}_N^{\mathcal{VH}}$) to have varying average magnitudes, $||\boldsymbol{\theta}_{main}||/(N_{\mathcal{H}} + N_{\mathcal{V}})$ and $||\boldsymbol{\theta}_{interaction}||/(N_{\mathcal{H}} * N_{\mathcal{V}})$ for a RBM with $N_{\mathcal{V}} = 9$ visibles and $N_{\mathcal{H}} = 5$ hiddens. These average magnitudes vary on a grid between 0.001 and 3 with 20 breaks, yielding 400 grid points. At each point in the grid, 100 vectors ($\boldsymbol{\theta}_{main}$) are sampled uniformly on a sphere with radius corresponding to the first coordinate in the grid and

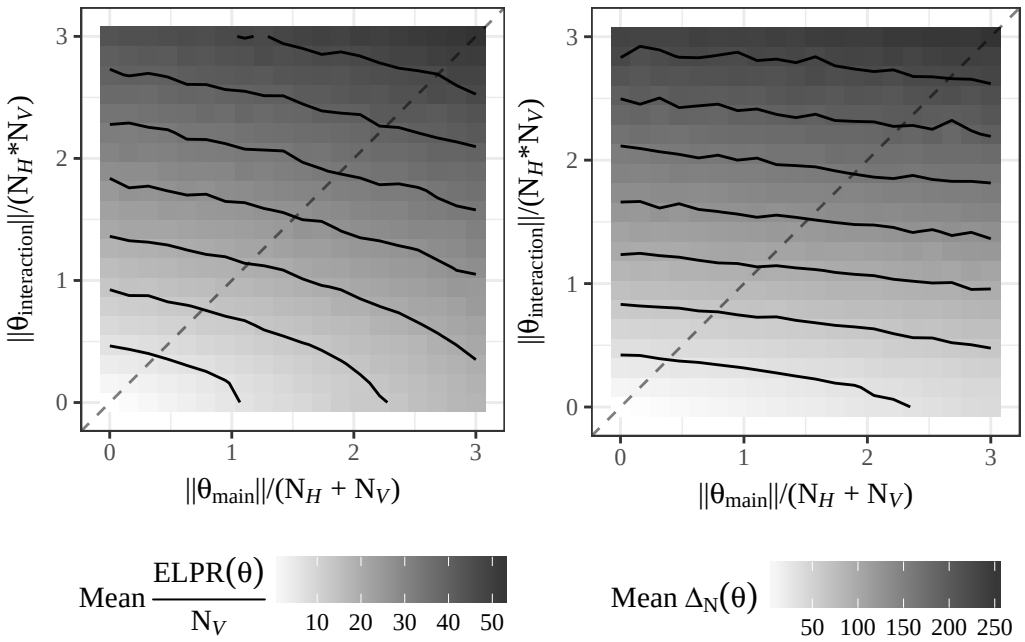

Figure 1: The sample mean value of $\text{ELPR}(\boldsymbol{\theta})/N_{\mathcal{V}}$ (left) and $\Delta_N(\boldsymbol{\theta})$ at each grid point for each combination of magnitude of $\boldsymbol{\theta}$. As the magnitude of $\boldsymbol{\theta}$ grows, so does the value of these metrics, indicating typical instability in the model.

100 vectors ($\boldsymbol{\theta}_{interaction}$) are sampled uniformly on a sphere with radius corresponding to the second coordinate in the grid via sums of squared and scaled iid Normal$(0, 1)$ variables. These vectors are then paired to create 100 values of $\boldsymbol{\theta}_N$ with magnitudes at each point in the grid. The values $\text{LREP}(\boldsymbol{\theta}_N)/N_{\mathcal{V}}$ and $\Delta_N(\boldsymbol{\theta}_N)$ are then calculated for each $\boldsymbol{\theta}_N$ and then summarized for each point in the grid using the sample mean. The results of this numerical study are shown in Figure 1. From these two plots, it is clear that for larger magnitudes of the parameter vectors, there is evidence of S-instability in that the log-ratio of extremal probabilities scaled by $N_{\mathcal{V}}$ and the the biggest log-probability ratio for a one-component change in data outcomes are both increasing away from $\boldsymbol{\theta}_N = \mathbf{0}$, further supporting 2(iii.2 and iii.5).

In more complicated graphical models involving further or deeper hidden layers, the same issues and causes of instability similarly exist, but are compounded by a greater number of model parameters. S-unstable joint models will similarly follow if the combined sizes of all parameters are too great relative to the total number of variables, while instability in the data model for visible variables will depend only on the main or interaction parameters directly related to visibles and how their accumulated magnitude compares to the observation sample size N.

5 Statistical Consequences of Instability

Due to their induced sensitivity in probability structure, S-instability in FOES models may often translate to numerical complications, and in fact obstructions, in both simulation and statistical inference based on likelihoods. We describe these aspects in Sections 5.1-5.3 with regard to data simulation, maximum likelihood estimation and Bayes inference, respectively.

5.1 Implications for Simulation

Suppose one aims to apply MCMC to simulate data $\boldsymbol{X} = (X_1, \ldots, X_n)$ from a FOES model $P_{\boldsymbol{\theta}_N}(\boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{X}^N$ whereby a chain is constructed with $P_{\boldsymbol{\theta}_N}$ as the stationary distribution. For an unstable FOES model, one-component changes in outcomes may produce radically different probabilities, which may then impose numerical barriers to MCMC. For example, consider implementation of a Gibbs sampler from the full conditional distributions $P_{\boldsymbol{\theta}_N}(X_i = x|\boldsymbol{x}_{-i})$, $x \in \mathcal{X}$, of each variable X_i based on values $\boldsymbol{x}_{-i} \in \mathcal{X}^{N-1}$ for the remaining variables, say $\boldsymbol{X}_{-i}$, in $\boldsymbol{X}$. If a single change in X_i from one value x_1 to another x_2 may produce two outcomes $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$ for $\boldsymbol{X}$ with vastly different probabilities under the joint distribution $P_{\boldsymbol{\theta}_N}$, then the Gibbs sampler can have extreme log-ratios in its transition probabilities,

$$\log \left| \frac{P_{\boldsymbol{\theta}_N}(X_i = x_1|\boldsymbol{x}^{(1)}_{-i})}{P_{\boldsymbol{\theta}_N}(X_i = x_2|\boldsymbol{x}^{(2)}_{-i})} \right| = \log \left| \frac{P_{\boldsymbol{\theta}_N}(\boldsymbol{x}^{(1)})}{P_{\boldsymbol{\theta}_N}(\boldsymbol{x}^{(2)})} \right|,$$

as conditional probabilities are proportional to joint probabilities that, with unstable models, can have unbounded log-probability ratios in one-component changes (Theorem 1). This can hinder the ability of a chain to effectively explore the sample space of the observations $\boldsymbol{X}$, as the chain may mix poorly by moving rapidly to, and slowly away from, sections of the sample space. In this case, for example, the Markov chain may become entrapped within a mode of the probability function, with rare chance of escaping to adequately mimic the occupation frequencies in the overall sample space. If modes of the unstable model are not unique, then important outcomes may be missed without multiple chains or impractically enormous numbers of MCMC samples. This mixing problem is due to the unstable stationary distribution (unbounded ratios of probabilities under the joint model), rather than in any particulars of the MCMC algorithm, and similar complications can also arise for Metropolis-Hastings algorithms for MCMC. Hence, while an unstable FOES model $P_{\boldsymbol{\theta}_N}$ may be valid and technically open to simulation by MCMC, the aspect of instability can render applications of MCMC as numerically infeasible for simulation purposes. This result is in line with conclusions of Handcock (2003) and Schweinberger (2011) for other exponential models.

5.2 Implications for Maximum Likelihood Inference

Volatility in the probability structure of an unstable model can also hamper efforts to maximize likelihood functions in statistical inference. When a FOES model is unstable along a parameter sequence $\boldsymbol{\theta}_N$, the same model can further be unstable along parameters $-\boldsymbol{\theta}_N$ in an opposite direction from the origin (model condition PSR and Corollary 1). This can translate into potential sensitivity of the likelihood function around zero, and lead to numerical complications in maximizing the objective function. We next provide a discussion of this issue in a way that builds upon and extends related findings by Schweinberger (2011), who largely focused on the case of one-parameter exponential models.

With many probability models, the modes and anti-modes in the probability structure under one parameter $\boldsymbol{\theta}_N$ are reversed in role when the parameter sign changes $-\boldsymbol{\theta}_N$. Because unstable models tend to degeneracy, the opposite signed parameters further push unstable models to assign nearly all probability to extremely opposite data configurations, given by modes/anti-modes. This is made concrete in Theorem 3, relating the degeneracy from unstable models to the expected behavior of log-likelihood functions.

Theorem 3. *Let $P_{\boldsymbol{\theta}_N}$, $N \geq 1$, denote an S-unstable FOES model sequence, which additionally satisfies model condition PSR. Let $\boldsymbol{x}_{\max,\boldsymbol{\theta}_N}, \boldsymbol{x}_{\min,\boldsymbol{\theta}_N} \in \mathcal{X}^N$ denote, respectively, a mode and anti-mode of the model $P_{\boldsymbol{\theta}_N}(\boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{X}^N$, for N observations $\boldsymbol{X} = (X_1, \ldots, X_N)$, whereby $P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_{\max,\boldsymbol{\theta}_N}) = \max_{\boldsymbol{y}\in\mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\boldsymbol{y})$ and $P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_{\min,\boldsymbol{\theta}_N}) = \min_{\boldsymbol{y}\in\mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\boldsymbol{y})$.*

Then, letting $\xrightarrow{p,\mathrm{E}}$ denote convergence in probability and expectation, as $N \to \infty$,

$$\frac{1}{\mathrm{LREP}(\boldsymbol{\theta}_N)} \log\left[\frac{P_{\boldsymbol{\theta}_N}(\boldsymbol{X})}{\min\limits_{\boldsymbol{y}\in\mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\boldsymbol{y})}\right] = \frac{\log P_{\boldsymbol{\theta}_N}(\boldsymbol{X}) - \log P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_{\min,\boldsymbol{\theta}_N})}{\log P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_{\max,\boldsymbol{\theta}_N}) - \log P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_{\min,\boldsymbol{\theta}_N})} \xrightarrow{p,\mathrm{E}} 1$$

under $\boldsymbol{\theta}_N$ while

$$\frac{1}{\mathrm{LREP}(-\boldsymbol{\theta}_N)} \log\left[\frac{P_{-\boldsymbol{\theta}_N}(\boldsymbol{X})}{\min\limits_{\boldsymbol{y}\in\mathcal{X}^N} P_{-\boldsymbol{\theta}_N}(\boldsymbol{y})}\right] = \frac{\log P_{-\boldsymbol{\theta}_N}(\boldsymbol{X}) - \log P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_{\max,\boldsymbol{\theta}_N})}{\log P_{-\boldsymbol{\theta}_N}(\boldsymbol{x}_{\min,\boldsymbol{\theta}_N}) - \log P_{-\boldsymbol{\theta}_N}(\boldsymbol{x}_{\max,\boldsymbol{\theta}_N})} \xrightarrow{p,\mathrm{E}} 1,$$

under $-\boldsymbol{\theta}_N$, where

$$\mathrm{LREP}(\boldsymbol{\theta}_N) \equiv \log \frac{P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_{\max,\boldsymbol{\theta}_N})}{P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_{\min,\boldsymbol{\theta}_N})} = \log \frac{P_{-\boldsymbol{\theta}_N}(\boldsymbol{x}_{\min,\boldsymbol{\theta}_N})}{P_{-\boldsymbol{\theta}_N}(\boldsymbol{x}_{\max,\boldsymbol{\theta}_N})} = \mathrm{LREP}(-\boldsymbol{\theta}_N), \quad N \geq 1.$$

Theorem 3 entails log-likelihood functions based on unstable models are both inversely related and degenerate at opposited signed parameters $\boldsymbol{\theta}_N$ or $-\boldsymbol{\theta}_N$, so that likelihoods are highest at different extremes in data configuration (e.g., $\boldsymbol{x}_{\max,\boldsymbol{\theta}_N}$ under $\boldsymbol{\theta}$-probabilities or $\boldsymbol{x}_{\min,\boldsymbol{\theta}_N}$ under $-\boldsymbol{\theta}$-probabilities). If the observed outcome $\boldsymbol{x}$ for data $\boldsymbol{X}$ is not a mode/anti-mode, then probabilities for the outcome may be small under both parameters $\boldsymbol{\theta}_N$ and $-\boldsymbol{\theta}_N$, in which case associated optimization steps may then shift around zero and struggle to converge.

In many model formulations, the zero parameter $\boldsymbol{\theta}_N = \mathbf{0}$ is a "safe" position among parameters, representing a guaranteed stable model (having uniform probabilities among outcomes), which can also tether a broad parameter search attempted among unstable models. Handcock (2003) describes similar results for degenerate exponential models, and Theorem 3 also supports an important finding of Schweinberger (2011 Corollary 1) for one-parameter linear exponential models (1). In the latter case, the likelihood score function at $\boldsymbol{\theta}_N$ is the expected value $\mu(\boldsymbol{\theta}_N) \equiv \mathrm{E}_{\boldsymbol{\theta}_N} g(\boldsymbol{X})$ of the sufficient statistic $g(\cdot)$, and optimization involves solving $\mu(\cdot) = g(\boldsymbol{x})$ for an observed outcome $\boldsymbol{x}$. For unstable models in this exponential class, Schweinberger (2011 Corollary 1) shows that

$$\lim_{N\to\infty} \frac{\mu(\boldsymbol{\theta}_N) - L_n}{U_n - L_N} = \begin{cases} 1 & \text{for } \boldsymbol{\theta}_N > 0, \\ 0 & \text{for } \boldsymbol{\theta}_N < 0, \end{cases}$$

where again U_N and L_N denote the maximum and minimum values of the statistic $g(\boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{X}^N$. As described by Schweinberger (2011), the implication for maximum likelihood estimation is that, unless an observed outcome $\boldsymbol{x}$ falls at an extreme U_N, L_N (i.e., modes/anti-modes), optimization steps in the parameter space can iterate in relatively small increments around zero and fail to converge. For unstable one-parameter exponential models, the maximum likelihood results of Schweinberger (2011) turn out to be a special case of Theorem 3 and the LREP expansion (10) in this setting; namely, for an unstable model with $\boldsymbol{\theta}_N > 0$,

$$\frac{1}{\mathrm{LREP}(\boldsymbol{\theta}_N)} \log \left[\frac{P_{\boldsymbol{\theta}_N}(\boldsymbol{X})}{\min\limits_{\boldsymbol{y}\in\mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\boldsymbol{y})} \right] = \frac{\boldsymbol{g}(\boldsymbol{X}) - L_N}{U_N - L_N} \xrightarrow{p,\mathrm{E}} 1$$

holds as $N \to \infty$ by Theorem 3, while under $-\boldsymbol{\theta}_N < 0$

$$\frac{1}{\mathrm{LREP}(-\boldsymbol{\theta}_N)} \log \left[\frac{P_{-\boldsymbol{\theta}_N}(\boldsymbol{X})}{\min\limits_{\boldsymbol{y}\in\mathcal{X}^N} P_{-\boldsymbol{\theta}_N}(\boldsymbol{y})} \right] = \frac{U_N - \boldsymbol{g}(\boldsymbol{X})}{U_N - L_N} = 1 - \frac{\boldsymbol{g}(\boldsymbol{X}) - L_N}{U_N - L_N} \xrightarrow{p,\mathrm{E}} 1.$$

Again, when all probability in unstable models may be pushed to opposite extremes in the sample space, due to a combination of degeneracy and parameter sign, numerical complications in likelihood maximization may occur.

5.3 Implications for Bayes Inference

The potential numerical difficulties with maximum likelihood with unstable models, as described in the previous section, can naturally carry over to Bayes inference. Considering that the degeneracy issues related to unstable models can cause likelihoods can be flat (e.g., near zero) for many parameters under a given data outcome and that sign changes in parameters can shift tremendous probability to extreme and opposite outcomes in the sample space (e.g., Corollary 1, Theorem 3), then numerical complications may arise with Bayes inference in sampling a posterior parameter space based on MCMC. The potential challenges in chain mixing are similar to those

presented in Section 5.1, though in chain movements through the parameter space as opposed to the sample space in data generation. That is, in the Bayes setting for sampling a posterior distribution for $\boldsymbol{\theta}_N$, a chain may unstable to effectively explore the parameter space due partly to extreme and potentially unbounded probability ratios from parameter sign changes, which represents a parameter space analog to how one-component changes in the sample space may impact data simulation with unstable models. For example, if $\pi(\cdot)$ denotes a prior density for $\boldsymbol{\theta}_N$ and $q(\cdot|\cdot)$ denotes a proposal distribution for use in a Metropolis-Hastings (MH) sampler, then MH acceptance probability becomes

$$\alpha\left(\boldsymbol{\theta}^{(1)} \mid \boldsymbol{\theta}^{(2)}\right) = \min\left\{1, \frac{q(\boldsymbol{\theta}_N^{(2)} \mid \boldsymbol{\theta}_N^{(1)})}{q(\boldsymbol{\theta}_N^{(1)} \mid \boldsymbol{\theta}_N^{(2)})} \frac{P_{\boldsymbol{\theta}_N^{(1)}}(\boldsymbol{x})\pi(\boldsymbol{\theta}_N^{(1)})}{P_{\boldsymbol{\theta}_N^{(2)}}(\boldsymbol{x})\pi(\boldsymbol{\theta}_N^{(2)})}\right\},$$

which indicates how parameter sensitivity in the likelihood $P_{\boldsymbol{\theta}_N}(\boldsymbol{x})$ may complicate sampling of the posterior $P_{\boldsymbol{\theta}_N}(\boldsymbol{x})\pi(\boldsymbol{\theta}_N)$ (i.e., moving from $\boldsymbol{\theta}^{(1)}$ to $\boldsymbol{\theta}^{(2)}$ in the parameter space). Furthermore, the potential for model instability and the size of the parameter space can also become greater with the introduction of latent variables to existing data variables, as involved in some model formulations described in Sections 2.2-2.3. As latent variables are often sampled with parameters in a Bayes MCMC approach, this aspect may further compound numerical problems in chain mixing.

6 Concluding Remarks

For a large class of models that covers a broad range of applications (including "deep learning"), we have developed a formal definition of instability in model probability structure and elucidated multiple consequences of instability. We have shown for FOES models that instability manifests through small changes in data leading to potentially large changes in probability as well as the potential to place all probability on certain modal subsections of the sample space, which potentially could be small. Such instability is often due to a complex interaction between the model statistics used (i.e., how numerous and large these may become) and the number and magnitudes of parameters in the model formulation. For many FOES models, the possibility exists, at least in principle, to constraint parameters in a way balances their potential contributions against those of model statistics in order to prevent probability instabilities. The FOES model class is quite broad and, in developing such models for large data sets, some caution should be used in parameter specification to control effects of model instability.

A Proofs of instability results

Proof of Proposition 1. For part (i), we prove the contrapositive, supposing that $\Delta_N(\boldsymbol{\theta}_N) \leq C$ holds for some $C > 0$ and show $\text{LREP}(\boldsymbol{\theta}_N) \leq NC$. Let $\boldsymbol{x}_{min} \equiv \arg\min_{\boldsymbol{x}\in\mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\boldsymbol{x})$ and $\boldsymbol{x}_{max} \equiv \arg\max_{\boldsymbol{x}\in\mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\boldsymbol{x})$. Note there exists a sequence $\boldsymbol{x}_{min} \equiv \boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \equiv \boldsymbol{x}_{max}$ in $\mathcal{X}^N$ of component-wise switches to move from $\boldsymbol{x}_{min}$ to $\boldsymbol{x}_{max}$ in the sample space (i.e. $\boldsymbol{x}_i, \boldsymbol{x}_{i+1} \in \mathcal{X}^N$ differ in exactly 1 component, $i = 0, \ldots, k$) for some integer $k \in \{0, 1, \ldots, N\}$. Under the FOES model, recall $P_{\boldsymbol{\theta}_N}(\boldsymbol{x}) > 0$ holds so that $\log P_{\boldsymbol{\theta}_N}(\boldsymbol{x})$ is well-defined for each outcome $\boldsymbol{x} \in \mathcal{X}^N$. Then, if $k > 0$, it follows that

$$\begin{aligned}\text{LREP}(\boldsymbol{\theta}_N) = \log\left[\frac{P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_{max})}{P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_{min})}\right] &= \left|\sum_{i=1}^{k} \log\left(\frac{P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_i)}{P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_{i-1})}\right)\right| \\ &\leq \sum_{i=1}^{k}\left|\log\left(\frac{P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_i)}{P_{\boldsymbol{\theta}}(\boldsymbol{x}_{i-1})}\right)\right| \leq k\Delta_N(\boldsymbol{\theta}_N) \leq NC,\end{aligned}$$

using $k \leq N$ and $\Delta(\boldsymbol{\theta}_N) \leq C$. If $k = 0$, then $\boldsymbol{x}_{max} = \boldsymbol{x}_{min}$ and the same bound above holds. This establishes part (i). To show part (ii), note the definition of S-instability (i.e., $\lim_{N\to\infty} \text{LREP}(\boldsymbol{\theta}_N)/N = \infty$) combined with part (i) implies that $\lim_{N\to\infty} \Delta_N(\boldsymbol{\theta}_N) = \infty$. □

Proof of Proposition 2. As $|\mathcal{X}| < \infty$ holds in the FOES model, we may suppose $|\mathcal{X}| > 1$; otherwise, $\mathcal{X}^N$ has one outcome and the model is trivially degenerate for all $N \geq 1$. Fix $0 < \epsilon < 1$ and write $\boldsymbol{x}_{min} \equiv \arg\min_{\boldsymbol{x}\in\mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\boldsymbol{x})$ and $\boldsymbol{x}_{max} \equiv \arg\max_{\boldsymbol{x}\in\mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\boldsymbol{x})$. Then, $\boldsymbol{x}_{max} \in M_{\epsilon,\boldsymbol{\theta}_N}$, so $P_{\boldsymbol{\theta}_N}(M_{\epsilon,\boldsymbol{\theta}_N}) \geq P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_{max}) > 0$. If $\boldsymbol{x} \in \mathcal{X}^N \setminus M_{\epsilon,\boldsymbol{\theta}_N}$, then by definition $P_{\boldsymbol{\theta}_N}(\boldsymbol{x}) \leq [P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_{max})]^{1-\epsilon}[P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_{min})]^{\epsilon}$ holds so that

$$1 - P_{\boldsymbol{\theta}_N}(M_{\epsilon,\boldsymbol{\theta}_N}) = \sum_{\boldsymbol{x}\in\mathcal{X}^N\setminus M_{\epsilon,\boldsymbol{\theta}_N}} P_{\boldsymbol{\theta}_N}(\boldsymbol{x}) \leq (|\mathcal{X}|^N)[P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_{max})]^{1-\epsilon}[P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_{min})]^{\epsilon}.$$

From the lower bound on $P_{\boldsymbol{\theta}_N}(M_{\epsilon,\boldsymbol{\theta}_N})$ and the upper bound on $1 - P_{\boldsymbol{\theta}_N}(M_{\epsilon,\boldsymbol{\theta}_N})$, it follows that

$$\begin{aligned}\frac{1}{N}\log\left[\frac{P_{\boldsymbol{\theta}_N}(M_{\epsilon,\boldsymbol{\theta}_N})}{1 - P_{\boldsymbol{\theta}_N}(M_{\epsilon,\boldsymbol{\theta}_N})}\right] &\geq \frac{1}{N}\log\left[\frac{P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_{max})}{(|\mathcal{X}|^N)[P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_{max})]^{1-\epsilon}[P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_{min})]^{\epsilon}}\right] \\ &= \frac{\epsilon}{N}\log\left[\frac{P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_{max})}{P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_{min})}\right] - \log|\mathcal{X}| \to \infty\end{aligned}$$

as $N \to \infty$ by the definition of an S-unstable FOES model (8). Consequently, $P_{\boldsymbol{\theta}_N}(M_{\epsilon,\boldsymbol{\theta}_N}) \to 1$ as $N \to \infty$ as claimed. □

Proof of Corollary 1. The model condition PSR implies that

$$\frac{\max_{\boldsymbol{y}\in\mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\boldsymbol{y})}{\min_{\boldsymbol{y}\in\mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\boldsymbol{y})} = \frac{\max_{\boldsymbol{y}\in\mathcal{X}^N} P_{-\boldsymbol{\theta}_N}(\boldsymbol{y})}{\min_{\boldsymbol{y}\in\mathcal{X}^N} P_{-\boldsymbol{\theta}_N}(\boldsymbol{y})} \tag{13}$$

so that the log-ratio $\text{LREP}(\boldsymbol{\theta}_N) = \text{LREP}(-\boldsymbol{\theta}_N)$ is the same for both $\boldsymbol{\theta}_N$ and $-\boldsymbol{\theta}_N$ in (7). Now part (i) of Corollary 1 follows from $\text{LREP}(\boldsymbol{\theta}_N)/N = \text{LREP}(-\boldsymbol{\theta}_N)/N \to \infty$ as $N \to \infty$ in (8).

To show part (ii), fix $0 < \epsilon < 1$ and consider a ϵ-mode set $\mathcal{M}_{\epsilon,\boldsymbol{\theta}_N}$ under $\boldsymbol{\theta}_N$ from (9). If $\boldsymbol{x} \in \mathcal{M}^c_{\epsilon,\boldsymbol{\theta}_N} \equiv \mathcal{X}^N \setminus \mathcal{M}_{\epsilon,\boldsymbol{\theta}_N}$, then, by definition,

$$\frac{P_{\boldsymbol{\theta}_N}(\boldsymbol{x})}{\min_{\boldsymbol{y}\in\mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\boldsymbol{y})} \leq \left[\frac{\max_{\boldsymbol{y}\in\mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\boldsymbol{y})}{\min_{\boldsymbol{y}\in\mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\boldsymbol{y})}\right]^{1-\epsilon}$$

holds, which is equivalent to

$$\frac{\max_{\boldsymbol{y}\in\mathcal{X}^N} P_{-\boldsymbol{\theta}_N}(\boldsymbol{y})}{P_{-\boldsymbol{\theta}_N}(\boldsymbol{x})} \leq \left[\frac{\max_{\boldsymbol{y}\in\mathcal{X}^N} P_{-\boldsymbol{\theta}_N}(\boldsymbol{y})}{\min_{\boldsymbol{y}\in\mathcal{X}^N} P_{-\boldsymbol{\theta}_N}(\boldsymbol{y})}\right]^{1-\epsilon}$$

by model condition PSR and (13). The latter is in turn equivalent to

$$\log P_{-\boldsymbol{\theta}_N}(\boldsymbol{x}) \geq \epsilon \max_{\boldsymbol{y}\in\mathcal{X}^N} \log P_{-\boldsymbol{\theta}_N}(\boldsymbol{y}) + (1-\epsilon) \min_{\boldsymbol{y}\in\mathcal{X}^N} \log P_{-\boldsymbol{\theta}_N}(\boldsymbol{y}), \tag{14}$$

so that $\boldsymbol{x} \in \mathcal{M}^c_{\epsilon,\boldsymbol{\theta}_N}$ if and only if (14) holds. Next consider the $(1-\epsilon)$-mode set $\mathcal{M}_{1-\epsilon,-\boldsymbol{\theta}_N}$ under $-\boldsymbol{\theta}_N$ from (9). If $\boldsymbol{x} \in \mathcal{M}_{1-\epsilon,-\boldsymbol{\theta}_N}$, then by definition $\boldsymbol{x}$ fulfills (14) and so $\boldsymbol{x} \in \mathcal{M}^c_{\epsilon,\boldsymbol{\theta}_N}$, showing that $\mathcal{M}_{1-\epsilon,-\boldsymbol{\theta}_N} \subset \mathcal{M}^c_{\epsilon,\boldsymbol{\theta}_N}$. By this and the fact that that Theorem 2 and Corollary 1(i) entail that $P_{-\boldsymbol{\theta}_N}(\boldsymbol{X}_N \in \mathcal{M}_{1-\epsilon,-\boldsymbol{\theta}_N}) \to 1$ as $N \to \infty$ (i.e., $P_{-\boldsymbol{\theta}_N}$ is S-unstable), we have

$$1 = \lim_{N\to\infty} P_{-\boldsymbol{\theta}_N}(\boldsymbol{X}_N \in \mathcal{M}_{1-\epsilon,-\boldsymbol{\theta}_N}) \leq \lim_{N\to\infty} P_{-\boldsymbol{\theta}_N}(\boldsymbol{X}_N \in \mathcal{M}^c_{\epsilon,\boldsymbol{\theta}_N}) \leq 1,$$

proving Corollary 1(ii) □

Proof of Proposition 1. For any two outcomes $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}^N$, the log-ratio of probabilities from the linear exponential model (1) with k parameters/statistics satisfies

$$\left|\log\left[\frac{P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_1)}{P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_2)}\right]\right| = \left|\sum_{i=1}^k \theta_{i,N}[g_{i,k}(\boldsymbol{x}_1) - g_{i,k}(\boldsymbol{x}_2)]\right| \leq \sum_{i=1}^k |\theta_{i,N}|(U_{i,N} - L_{i,N});$$

consequently, $\mathrm{LREP}(\boldsymbol{\theta}_N) \leq \sum_{i=1}^k |\theta_{i,N}|(U_{i,N} - L_{i,N})$ holds in (7) and model stability in Proposition 1 follows from (8). □

Proof of Proposition 2. Writing $\boldsymbol{x} = (x_1, \ldots, x_N)$ and $\boldsymbol{h} = (h_1, \ldots, h_{N_\mathcal{H}})$ with all components $x_i, h_j \in \{\pm 1\}$, probabilities in the joint RBM model (5) can be written as $\tilde{P}_{\boldsymbol{\theta}_N}(\boldsymbol{x}, \boldsymbol{h}) = c(\boldsymbol{\theta}_N)\exp[f_{\boldsymbol{\theta}_N}(\boldsymbol{x}, \boldsymbol{h})]$ in terms of the function $f_{\boldsymbol{\theta}_N}(\boldsymbol{x}, \boldsymbol{h})$ from (12) and the normalizing constant $c(\boldsymbol{\theta}_N) = \exp[-\psi(\boldsymbol{\theta}_N)]$ from (5). Let $\boldsymbol{x}_M, \boldsymbol{x}_m \in \{\pm 1\}^N$ be such that $P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_M) = \max_{\boldsymbol{x}} P_{\boldsymbol{\theta}_N}(\boldsymbol{x})$ and $P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_m) = \min_{\boldsymbol{x}} P_{\boldsymbol{\theta}_N}(\boldsymbol{x})$ under the marginal RBM model $P_{\boldsymbol{\theta}_N}(\boldsymbol{x}) = c(\boldsymbol{\theta}_N)\sum_{\boldsymbol{h}\in\{\pm 1\}^{\mathcal{N}_H}} \tilde{P}_{\boldsymbol{\theta}_N}(\boldsymbol{x}, \boldsymbol{h}) = c(\boldsymbol{\theta}_N)\sum_{\boldsymbol{h}\in\{\pm 1\}^{\mathcal{N}_H}} \exp[f_{\boldsymbol{\theta}_N}(\boldsymbol{x}, \boldsymbol{h})]$ from (6). Also, $\boldsymbol{x}_0, x_1 \in \{\pm 1\}^N$ be such that $\max_{\boldsymbol{h}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x}_0, \boldsymbol{h}) = \max_{\boldsymbol{x}} \max_{\boldsymbol{h}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x}, \boldsymbol{h})$ and $\max_{\boldsymbol{h}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x}_1, \boldsymbol{h}) = \min_{\boldsymbol{x}} \max_{\boldsymbol{h}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x}, \boldsymbol{h})$. Then, Proposition 2(i) follows from $\mathrm{LREP}_{(\boldsymbol{X})}(\boldsymbol{\theta}_N) = \log[P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_M)/P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_m)]$ and the lower/upper bounds on $P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_M)$ and $P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_m)$ as

$$c(\boldsymbol{\theta}_N)\exp[\max_{\boldsymbol{h}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x}_0, \boldsymbol{h})] \leq P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_0) \leq P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_M) \leq 2^{N_\mathcal{H}} c(\boldsymbol{\theta}_N)\exp[\max_{\boldsymbol{x}} \max_{\boldsymbol{h}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x}, \boldsymbol{h})]$$

and

$$
\begin{aligned}
c(\boldsymbol{\theta}_N)\exp[\min_{\boldsymbol{x}}\max_{\boldsymbol{h}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x},\boldsymbol{h})] &\leq c(\boldsymbol{\theta}_N)\exp[\max_{\boldsymbol{h}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x}_m,\boldsymbol{h})] \\
&\leq P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_m) \\
&\leq P_{\boldsymbol{\theta}_N}(\boldsymbol{x}_1) \\
&\leq 2^{N_{\mathcal{H}}} c(\boldsymbol{\theta}_N)\exp[\max_{\boldsymbol{h}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x}_1,\boldsymbol{h})] \\
&= 2^{N_{\mathcal{H}}} c(\boldsymbol{\theta}_N)\exp[\min_{\boldsymbol{x}}\max_{\boldsymbol{h}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x},\boldsymbol{h})]
\end{aligned}
$$

To prove Proposition 2, we next expand the function $f_{\boldsymbol{\theta}_N}(\boldsymbol{x},\boldsymbol{h})$ from (12) as

$$
f_{\boldsymbol{\theta}_N}(\boldsymbol{x},\boldsymbol{h}) = \sum_{j=1}^{N_{\mathcal{H}}} h_j\theta^{\mathcal{H}}_{j,N} + \sum_{i=1}^{N}\left(\theta^{\mathcal{V}}_{i,N} + \sum_{j=1}^{N_{\mathcal{H}}} h_j\theta^{\mathcal{VH}}_{ij,N}\right)x_i = \sum_{i=1}^{N} x_i\theta^{\mathcal{V}}_{i,N} + \sum_{j=1}^{N_{\mathcal{H}}}\left(\theta^{\mathcal{H}}_{j,N} + \sum_{i=1}^{N} x_i\theta^{\mathcal{VH}}_{ij,N}\right)h_j.
$$

By this and the fact that $x_i, h_j \in \{\pm 1\}$, we then have

$$
\begin{aligned}
\max_{\boldsymbol{x}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x},\boldsymbol{h}) &= \sum_{j=1}^{N_{\mathcal{H}}} h_j\theta^{\mathcal{H}}_{j,N} + a_{\boldsymbol{\theta}_N,\mathcal{H}}(\boldsymbol{h}), \min_{\boldsymbol{x}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x},\boldsymbol{h}) = \sum_{j=1}^{N_{\mathcal{H}}} h_j\theta^{\mathcal{H}}_{j,N} - a_{\boldsymbol{\theta}_N,\mathcal{H}}(\boldsymbol{h}), \\
\max_{\boldsymbol{h}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x},\boldsymbol{h}) &= \sum_{i=1}^{N} x_i\theta^{\mathcal{V}}_{i,N} + b_{\boldsymbol{\theta}_N,\mathcal{V}}(\boldsymbol{x}), \min_{\boldsymbol{h}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x},\boldsymbol{h}) = \sum_{i=1}^{N} x_i\theta^{\mathcal{V}}_{i,N} - b_{\boldsymbol{\theta}_N,\mathcal{V}}(\boldsymbol{x}), \\
a_{\boldsymbol{\theta}_N,\mathcal{H}}(\boldsymbol{h}) &\equiv \sum_{i=1}^{N}\left|\theta^{\mathcal{V}}_{i,N} + \sum_{j=1}^{N_{\mathcal{H}}} h_j\theta^{\mathcal{VH}}_{ij,N}\right|, b_{\boldsymbol{\theta}_N,\mathcal{V}}(\boldsymbol{x}) \equiv \sum_{j=1}^{N_{\mathcal{H}}}\left|\theta^{\mathcal{H}}_{j,N} + \sum_{i=1}^{N} x_i\theta^{\mathcal{VH}}_{ij,N}\right|,
\end{aligned} \tag{15}
$$

where $\boldsymbol{h}^T\boldsymbol{\theta}^{\mathcal{H}}_N = \sum_{j=1}^{N_{\mathcal{H}}} h_j\theta^{\mathcal{H}}_{j,N}$, $\boldsymbol{x}^T\boldsymbol{\theta}^{\mathcal{V}}_N = \sum_{i=1}^{N} x_i\theta^{\mathcal{V}}_{i,N}$ and $B_N(\boldsymbol{\theta}_N) \equiv \max_{\boldsymbol{h}} a_{\boldsymbol{\theta}_N,\mathcal{H}}(\boldsymbol{h})$. From this, it follows that

$$
\begin{aligned}
\mathrm{LREP}_{(\boldsymbol{X},\boldsymbol{H})}(\boldsymbol{\theta}_N) &= \max_{\boldsymbol{h}}\max_{\boldsymbol{x}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x},\boldsymbol{h}) - \min_{\boldsymbol{h}}\min_{\boldsymbol{x}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x},\boldsymbol{h}) \\
&= \max_{\boldsymbol{h}_1}\max_{\boldsymbol{h}_2}\left[(\boldsymbol{h}_1-\boldsymbol{h}_2)^T\boldsymbol{\theta}^{\mathcal{H}}_N + a_{\boldsymbol{\theta}_N,\mathcal{H}}(\boldsymbol{h}_1) + a_{\boldsymbol{\theta}_N,\mathcal{H}}(\boldsymbol{h}_2)\right],
\end{aligned}
$$

which leads to the upper bound $\mathrm{LREP}_{(\boldsymbol{X},\boldsymbol{H})}(\boldsymbol{\theta}_N) \leq 2B_N(\boldsymbol{\theta}_N) + 2|\boldsymbol{\theta}^{\mathcal{H}}_N|_1$. Then, taking $\boldsymbol{h}_1 = \boldsymbol{h}_2$ (i.e., before maximization) gives $\mathrm{LREP}_{(\boldsymbol{X},\boldsymbol{H})}(\boldsymbol{\theta}_N) \geq 2B_N(\boldsymbol{\theta}_N)$ and taking $\boldsymbol{h}_1 = -\boldsymbol{h}_2$, such that $\boldsymbol{h}_1^T\boldsymbol{\theta}^{\mathcal{H}}_N = |\theta^{\mathcal{H}}_N|_1$, gives $\mathrm{LREP}_{(\boldsymbol{X},\boldsymbol{H})}(\boldsymbol{\theta}_N) \geq 2|\theta^{\mathcal{H}}_N|_1$; this yields the lower bound $\mathrm{LREP}_{(\boldsymbol{X},\boldsymbol{H})}(\boldsymbol{\theta}_N) \geq 2\max\{B_N(\boldsymbol{\theta}_N), |\boldsymbol{\theta}^{\mathcal{H}}_N|_1\}$.

We next consider $A_N(\boldsymbol{\theta}_N)$ and, by (15), write

$$
\begin{aligned}
A_N(\boldsymbol{\theta}_N) &= \max_{\boldsymbol{h}}\max_{\boldsymbol{x}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x},\boldsymbol{h}) - \max_{\boldsymbol{h}}\min_{\boldsymbol{x}} f_{\boldsymbol{\theta}_N}(\boldsymbol{x},\boldsymbol{h}) \\
&= \max_{\boldsymbol{h}_1}\min_{\boldsymbol{h}_2}\left[(\boldsymbol{h}_1-\boldsymbol{h}_2)^T\boldsymbol{\theta}^{\mathcal{H}}_N + a_{\boldsymbol{\theta}_N,\mathcal{H}}(\boldsymbol{h}_1) + a_{\boldsymbol{\theta}_N,\mathcal{H}}(\boldsymbol{h}_2)\right].
\end{aligned}
$$

Taking $\boldsymbol{h}_1 = \boldsymbol{h}_2$ and maximizing over both $\boldsymbol{h}_1, \boldsymbol{h}_2$ produces the upper bound $A_N(\boldsymbol{\theta}_N) \leq 2B_N(\boldsymbol{\theta}_N)$. Then, using $(\boldsymbol{h}_1-\boldsymbol{h}_2)^T\boldsymbol{\theta}^{\mathcal{H}}_N + a_{\boldsymbol{\theta}_N,\mathcal{H}}(\boldsymbol{h}_2) \geq -2|\boldsymbol{\theta}^{\mathcal{H}}_N|_1$ and maximizing over $\boldsymbol{h}_1$ gives $A_N(\boldsymbol{\theta}_N) \geq B_N(\boldsymbol{\theta}_N) - 2|\boldsymbol{\theta}^{\mathcal{H}}_N|_1$, while setting $\boldsymbol{h}_1 = \boldsymbol{h}^*_2$ for $\boldsymbol{h}^*_2$ such that $-(\boldsymbol{h}^*_2)^T\boldsymbol{\theta}^{\mathcal{H}}_N + a_{\boldsymbol{\theta}_N,\mathcal{H}}(\boldsymbol{h}^*_2) =$

$\min_{\boldsymbol{h}_2}[-\boldsymbol{h}_2^T\boldsymbol{\theta}_N^{\mathcal{H}} + a_{\boldsymbol{\theta}_N,\mathcal{H}}(\boldsymbol{h}_2)]$ gives $A_N(\boldsymbol{\theta}_N) \geq 2a_{\boldsymbol{\theta}_N,\mathcal{H}}(\boldsymbol{h}_2^*) \geq C_N(\boldsymbol{\theta}_N) \equiv \min_{\boldsymbol{h}} a_{\boldsymbol{\theta}_N,\mathcal{H}}(\boldsymbol{h})$. Finally, note that for any $\boldsymbol{h}$, the triangle inequality gives

$$\begin{aligned} B_N(\boldsymbol{\theta}_N) \equiv \max_{\boldsymbol{h}_1} a_{\boldsymbol{\theta}_N,\mathcal{H}}(\boldsymbol{h}_1) &\geq [a_{\boldsymbol{\theta}_N,\mathcal{H}}(\boldsymbol{h}) + a_{\boldsymbol{\theta}_N,\mathcal{H}}(-\boldsymbol{h})]/2 \\ &= 2^{-1}\sum_{i=1}^{N}\left(\left|\theta_{i,N}^{\mathcal{V}} + \sum_{j=1}^{N_{\mathcal{H}}} h_j\theta_{ij,N}^{\mathcal{V}\mathcal{H}}\right| + \left|\theta_{i,N}^{\mathcal{V}} - \sum_{j=1}^{N_{\mathcal{H}}} h_j\theta_{ij,N}^{\mathcal{V}\mathcal{H}}\right|\right) \\ &\geq \sum_{i=1}^{N}|\theta_{i,N}^{\mathcal{V}}| \equiv |\boldsymbol{\theta}_N^{\mathcal{V}}|_1. \end{aligned}$$

□

Proof of Theorem 3. Let $L_{\boldsymbol{\theta}_N}(\boldsymbol{X}) = \log[P_{\boldsymbol{\theta}_N}(\boldsymbol{X})/\min_{\boldsymbol{y}\in\mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\boldsymbol{y})]/\mathrm{LREP}(\boldsymbol{\theta}_N)$, where again $\boldsymbol{X} = (X_1,\ldots,X_N)$ and $\mathrm{LREP}(\boldsymbol{\theta}_N) = \log[\max_{\boldsymbol{y}\in\mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\boldsymbol{y})/\min_{\boldsymbol{y}\in\mathcal{X}^N} P_{\boldsymbol{\theta}_N}(\boldsymbol{y})]$. As $L_{\boldsymbol{\theta}_N}(\boldsymbol{X}) \in [0,1]$, convergence of $L_{\boldsymbol{\theta}_N}(\boldsymbol{X})$ to 1 in probability under $P_{\boldsymbol{\theta}_N}$ is equivalent to convergence to 1 in expectation under $P_{\boldsymbol{\theta}_N}$ (i.e., convergence in expectation implies probabilistic convergence by Markov's inequality while probabilistic convergence implies convergence in expectation by uniform integrability/boundedness).

For $\epsilon \in (0,1)$, let $\mathcal{M}_{\epsilon,\boldsymbol{\theta}_N}$ denote a modal set as in (9). By Theorem 2, $P_{\boldsymbol{\theta}_N}(\boldsymbol{X} \in \mathcal{M}_{\epsilon,\boldsymbol{\theta}_N}) \to 1$ holds as $N \to \infty$ and, by definition of (9), $\boldsymbol{X} \in \mathcal{M}_{\epsilon,\boldsymbol{\theta}_N}$ follows if and only if $1 - L_{\boldsymbol{\theta}_N}(\boldsymbol{X}) < \epsilon$. Hence, $L_{\boldsymbol{\theta}_N}(\boldsymbol{X}) \xrightarrow{p,\mathrm{E}} 1$ holds under $\boldsymbol{\theta}_N$ in Theorem 3. The convergence $L_{-\boldsymbol{\theta}_N}(\boldsymbol{X}) \xrightarrow{p,\mathrm{E}} 1$ under $-\boldsymbol{\theta}_N$ likewise follows from Corollary 1. □

References

Besag, Julian. 1974. "Spatial Interaction and the Statistical Analysis of Lattice Systems." *Journal of the Royal Statistical Society. Series B (Methodological)*. JSTOR, 192–236.

Handcock, Mark S. 2003. "Assessing Degeneracy in Statistical Models of Social Networks." Center for Statistics; the Social Sciences, University of Washington. http://www.csss.washington.edu/.

Hinton, Geoffrey E, Simon Osindero, and Yee-Whye Teh. 2006. "A Fast Learning Algorithm for Deep Belief Nets." *Neural Computation* 18 (7). MIT Press: 1527–54.

Neal, Radford M. 1992. "Connectionist Learning of Belief Networks." *Artificial Intelligence* 56 (1). Elsevier: 71–113.

Pearl, Judea. 1985. "Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning." UCLA Computer Science Department.

Ruelle, D. 1999. *Statistical Mechanics: Rigorous Results*. London: Imperial College Press.

Salakhutdinov, Ruslan, and Geoffrey E Hinton. 2009. "Deep Boltzmann Machines." In *International Conference on Artificial Intelligence and Statistics*, 448–55. AI & Statistics.

Schweinberger, Michael. 2011. "Instability, Sensitivity, and Degeneracy of Discrete Exponential Families." *Journal of the American Statistical Association* 106 (496). Taylor & Francis: 1361–70.

Smolensky, Paul. 1986. "Information Processing in Dynamical Systems: Foundations of Harmony Theory." DTIC Document.

Wasserman, Stanley, and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Vol. 8. Cambridge: Cambridge University Press.