# An Instability in Variational Inference for Topic Models

**Behrooz Ghorbani** [1] **Hamid Javadi** [2] **Andrea Montanari** [1 3]

## Abstract

Naive mean field variational methods are the state-of-the-art approach to inference in topic models. We show that these methods suffer from an instability that can produce misleading conclusions. Namely, for certain regimes of the model parameters, variational inference outputs a non-trivial decomposition into topics. However -–for the same parameter values-– the data contain no actual information about the true topic decomposition, and the output of the algorithm is uncorrelated with it. In particular, the estimated posterior mean is wrong, and estimated credible regions do not achieve the nominal coverage. We discuss how this instability is remedied by more accurate mean field approximations.

## 1. Introduction

Topic modeling (Blei, 2012) aims at extracting the latent structure from a corpus of documents (images or texts), that are represented as vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$. The key assumption is that the $n$ documents are (approximately) convex combinations of a small number $k$ of topics $\tilde{\boldsymbol{h}}_1, \dots, \tilde{\boldsymbol{h}}_k \in \mathbb{R}^d$. Conditional on the topics, documents are generated independently by letting $\boldsymbol{x}_a = (\sqrt{\beta}/d) \sum_{\ell=1}^k w_{a,\ell} \tilde{\boldsymbol{h}}_\ell + \boldsymbol{z}_a$, where the weights $\boldsymbol{w}_a = (w_{a,\ell})_{1\le \ell \le k}$ and noise vectors $\boldsymbol{z}_a$ are i.i.d. across $a \in \{1, \dots, n\}$. The coefficient $\beta \ge 0$ can be interpreted as a signal-to-noise ratio. It is also useful to introduce the matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ whose $i$-th row is $\boldsymbol{x}_i$, and therefore

$$\boldsymbol{X} = \frac{\sqrt{\beta}}{d} \boldsymbol{W} \boldsymbol{H}^{\mathsf{T}} + \boldsymbol{Z}\,, \tag{1.1}$$

where $\boldsymbol{W} \in \mathbb{R}^{n\times k}$ and $\boldsymbol{H} \in \mathbb{R}^{d\times k}$. The $a$-th row of $\boldsymbol{W}$, is the vector of weights $\boldsymbol{w}_a$, while the rows of $\boldsymbol{H}$ will be denoted by $\boldsymbol{h}_i \in \mathbb{R}^k$.

[1]Department of Electrical Engineering, Stanford University, CA [2]Digital Signal Processing Group, Rice University, TX [3]Department of Statistics, Stanford University, CA. Correspondence to: Behrooz Ghorbani <ghorbani@stanford.edu>.

*Proceedings of the $36^{th}$ International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019. 

Note that $\boldsymbol{w}_a$ belongs to the simplex $\mathsf{P}_1(k) = \{\boldsymbol{w} \in \mathbb{R}^k_{\ge 0} : \langle \boldsymbol{w}, \boldsymbol{1}_k\rangle = 1\}$. It is common to assume that its prior is Dirichlet: this class of models is known as *Latent Dirichlet Allocations*, or LDA (Blei et al., 2003). Here we will consider a symmetric Dirichlet prior, with all parameters equal to $\nu$ (which we will denote by $\mathrm{Dir}(\nu; k)$). As for the topics $\boldsymbol{H}$, their prior distribution depends on the specific application. here we will consider two simple examples: in the *Gaussian* case, we assume $(\tilde{\boldsymbol{h}}_i)_{i\le d} \sim_{iid} \mathsf{N}(0, \boldsymbol{I}_k)$; in the Dirichlet case $(\tilde{\boldsymbol{h}}_i)_{i\le d} \sim_{iid} \mathrm{Dir}(\tilde{\nu}; k)$. Most of our discussion and explicit formulas will refer for simplicity to the Gaussian case. However, we derived analogous expressions for the Dirichlet model, and will compare with numerical simulations carried out under both distributions. Our methodology is indeed general. Finally, $\boldsymbol{Z}$ will be a noise matrix with entries $(Z_{ij})_{i\in[n], j\in[d]} \sim_{iid} \mathsf{N}(0, 1/d)$.

In fully Bayesian topic models, the parameters of the Dirichlet distribution, as well as the topic distributions are themselves unknown and to be learned from data. Here we will work in an idealized setting in which they are known. We will also assume that data are in fact distributed according to the postulated generative model. Since we are studying the limitations of current approaches, our main point is only reinforced by assuming this idealized scenario.

Computing the posterior distribution of $\boldsymbol{H}$, $\boldsymbol{W}$ given the data $\boldsymbol{X}$ is computationally challenging. Since the seminal work of Blei, Ng and Jordan (Blei et al., 2003), variational inference is the method of choice for addressing this problem within topic models. The term 'variational inference' refers to a broad class of methods that aim at approximating the posterior computation by solving an optimization problem, see (Jordan et al., 1999; Wainwright et al., 2008; Blei et al., 2017) for background. A popular starting point is the Gibbs variational principle, namely the fact that the posterior solves the following convex optimization problem:

$$p_{\boldsymbol{W},\boldsymbol{H}|\boldsymbol{X}}(\,\cdot\,,\cdot\,,|\boldsymbol{X}) = \arg\min_{q\in\mathcal{P}_{n,d,k}} \mathrm{KL}(q\|p_{\boldsymbol{W},\boldsymbol{H}|\boldsymbol{X}}) \tag{1.2}$$

where $\mathrm{KL}(\,\cdot\,\|\,\cdot\,)$ denotes the Kullback-Leibler divergence. Optimization is within the space $\mathcal{P}_{n,d,k}$ of probability measures on $\boldsymbol{H}, \boldsymbol{W}$.

Even for $\boldsymbol{W}, \boldsymbol{H}$ discrete, the Gibbs principle has exponentially many decision variables. Variational methods differ in the way the problem (1.2) is approximated. The main

approach within topic modeling is *naive mean field*, which restricts the optimization problem to the space of probability measures that factorize over the rows of $\boldsymbol{W}, \boldsymbol{H}$:

$$\hat{q}(\boldsymbol{W}, \boldsymbol{H}) = \prod_{i=1}^{d} q_i(\boldsymbol{h}_i) \prod_{a=1}^{n} \tilde{q}_a(\boldsymbol{w}_a) \ . \qquad (1.3)$$

By a suitable parametrization of the marginals $q_i$, $\tilde{q}_a$, this leads to an optimization problem of dimension $O((n+d)k)$, cf. Section 3. Despite being non-convex, this problem is separately convex in the $(q_i)_{i\le d}$ and $(\tilde{q}_a)_{a\le n}$, which naturally suggests the use of an alternating minimization algorithm which has been successfully deployed in a broad range of applications ranging from computer vision to genetics (Fei-Fei & Perona, 2005; Wang & Blei, 2011; Raj et al., 2014). We will refer to this as to the *naive mean field iteration*. Following a common use in the topics models literature, we will use the terms 'variational inference' and 'naive mean field' interchangeably.

The main result of this paper is that naive mean field presents an instability for learning Latent Dirichlet Allocations. We focus on the limit $n, d \to \infty$ with $n/d = \delta$ fixed. Hence, an LDA distribution is determined by the parameters $(k, \delta, \nu, \beta)$. We will show that there are regions in this parameter space such that the following two findings hold simultaneously:

**No non-trivial estimator.** Any estimator $\widehat{\boldsymbol{H}}$, $\widehat{\boldsymbol{W}}$ of the topic or weight matrices is asymptotically uncorrelated with the real model parameters $\boldsymbol{H}, \boldsymbol{W}$. In other words, the data do not contain enough signal to perform any strong inference.

**Variational inference is randomly biased.** Given the above, one would hope the Bayesian posterior to be centered on an unbiased estimate. In particular, $p(\boldsymbol{w}_a|\boldsymbol{X})$ (the posterior distribution over weights of document $a$) should be centered around the uniform distribution $\boldsymbol{w}_a = (1/k, \dots, 1/k)$. In contrast, we will show that the posterior produced by naive mean field is centered around a random distribution that is uncorrelated with the actual weights. Similarly, the posterior over topic vectors is centered around random vectors uncorrelated with the true topics.

One key argument in support of Bayesian methods is the hope that they provide a measure of uncertainty of the estimated variables. In view of this, the failure just described is particularly dangerous because it suggests some measure of certainty, although the estimates are essentially random. While the limitation of variational methods have been pointed out in the past, ours is the first case in which such an inconsistency is estcabilished rigorously in topic models.

Is there a way to eliminate this instability by using a better mean field approximation? We show that a promising approach is provided by a classical idea in statistical physics, the Thouless-Anderson-Palmer (TAP) free energy (Thouless et al., 1977; Opper & Winther, 2001).

**Variational inference via the TAP free energy.** We show that the instability of naive mean field is remedied by using the TAP free energy instead of the naive mean field free energy. The latter can be optimized using an iterative scheme that is analogous to the naive mean field iteration and is known as approximate message passing (AMP).

The rest of the paper is organized as follows. Section 2 discusses a simpler example, $\mathbb{Z}_2$-synchronization, which shares important features with latent Dirichlet allocations. Since calculations are fairly straightforward, this example allows to explain the main mathematical points in a simple context. We then present our main results about instability of naive mean field in Section 3, and discuss the use of TAP free energy to overcome the instability in Section 4. As mentioned above, all formal statements will refer to the Gaussian case, although we obtained analogous results for the Dirichlet case. The corresponding theoretical predictions will be compared with numerical results in the plots. Proofs will be deferred to the Supplementary Material (SM), which also contain further numerical illustrations and technical results.

### 1.1. Related Literature

Over the last fifteen years, topic models have been generalized to cover an impressive number of applications, including mixed membership models (Erosheva et al., 2004; Airoldi et al., 2008), dynamic topic models (Blei & Lafferty, 2006b), correlated topic models (Blei & Lafferty, 2006a; Blei et al., 2007), spatial LDA (Wang & Grimson, 2008), relational topic models (Chang & Blei, 2009), Bayesian tensor models (Zhou et al., 2015). While other approaches have been used (e.g. Gibbs sampling), variational algorithms allow to leverage advances in optimization algorithms and architectures towards the goal of variational inference (Hoffman et al., 2010; Broderick et al., 2013).

Despite this broad empirical success, little is rigorously known about the accuracy of variational inference in concrete statistical problems. Wang and Titterington (Wang & Titterington, 2004; Wang et al., 2006) studied the missing data and Gaussian mixture models. In the context of Gaussian mixtures, the same authors prove that the covariance of the variational posterior is asymptotically smaller (in the positive semidefinite order) than the inverse of the Fisher information matrix (Wang & Titterington, 2005) (see also (Giordano et al., 2015)). All of these results are established in the classical large sample asymptotics $n \to \infty$ with $d$ fixed. In the present paper we focus instead on the high-dimensional limit $n = \Theta(d)$ and prove that also the mode (or mean) of the variational posterior is incorrect. The high-dimensional regime is particularly relevant for

the analysis of Bayesian methods. Indeed, in the classical low-dimensional asymptotics Bayesian approaches do not outperform maximum likelihood.

Naive mean field variational inference was used in (Celisse et al., 2012; Bickel et al., 2013) to estimate the parameters of the stochastic block model. The recent paper (Zhang & Zhou, 2017) also studies variational inference in the context of the stochastic block model. The work of (Celisse et al., 2012; Bickel et al., 2013; Zhang & Zhou, 2017) establishes positive results at large signal-to-noise ratio (albeit for a different model), while we prove inconsistency at low signal-to-noise ratio. General conditions for consistency of variational Bayes methods are proposed in (Pati et al., 2017).

Our work also builds on recent theoretical advances in high-dimensional low-rank models, that were mainly driven by techniques from mathematical statistical physics (more specifically, spin glass theory). An incomplete list of relevant references includes (Korada & Macris, 2009; Deshpande & Montanari, 2014; Deshpande et al., 2016; Krzakala et al., 2016; Barbier et al., 2016; Lelarge & Miolane, 2016; Miolane, 2017; Lesieur et al., 2017; Alaoui & Krzakala, 2018). These papers prove asymptotically exact characterizations of the Bayes optimal estimation error in low-rank models, to an increasing degree of generality, under the high-dimensional scaling $n, d \to \infty$ with $n/d \to \delta \in (0, \infty)$.

Related ideas also suggest an iterative algorithm for Bayesian estimation, namely Bayes Approximate Message Passing (Donoho et al., 2009; 2010). As mentioned above, Bayes AMP can be regarded as minimizing a different variational approximation known as the TAP free energy. An important advantage over naive mean field is that AMP can be rigorously analyzed using a method known as state evolution (Bayati & Montanari, 2011; Javanmard & Montanari, 2013; Berthier et al., 2017).

### 1.2. Notations

We denote by $\boldsymbol{I}_m$ the identity matrix, and by $\boldsymbol{J}_m$ the all-ones matrix in $m$ dimensions. We use $\mathbf{1}_k \in \mathbb{R}^k$ for the all-ones vector. We will use $\otimes$ for the tensor (outer) product. In particular, given vectors expressed in the canonical basis as $\boldsymbol{u} = \sum_{i=1}^{d_1} u_i \boldsymbol{e}_i \in \mathbb{R}^{d_1}$ and $\boldsymbol{v} = \sum_{i=j}^{d_2} v_j \boldsymbol{e}_j \in \mathbb{R}^{d_2}$, $\boldsymbol{u} \otimes \boldsymbol{v} \in \mathbb{R}^{d_1} \otimes \mathbb{R}^{d_2}$ is the tensor with coordinates $(\boldsymbol{u} \otimes \boldsymbol{v})_{ij} = u_i v_j$ in the basis $\boldsymbol{e}_i \otimes \boldsymbol{e}_j$. We will identify the space of matrices $\mathbb{R}^{d_1 \times d_2}$ with the tensor product $\mathbb{R}^{d_1} \otimes \mathbb{R}^{d_2}$. Given a symmetric matrix $\boldsymbol{M} \in \mathbb{R}^{n \times n}$, we denote by $\lambda_1(\boldsymbol{M}) \geq \lambda_2(\boldsymbol{M}) \geq \cdots \geq \lambda_n(\boldsymbol{M})$ its eigenvalues in decreasing order. For a matrix (or vector) $\boldsymbol{A} \in \mathbb{R}^{d \times n}$ we denote the orthogonal projector onto the subspace spanned by the columns of $\boldsymbol{A}$ by $\boldsymbol{P}_{\boldsymbol{A}} \in \mathbb{R}^{d \times d}$, and its orthogonal complement by $\boldsymbol{P}_{\boldsymbol{A}}^{\perp} = \boldsymbol{I}_d - \boldsymbol{P}_{\boldsymbol{A}}$. When the subscript is omitted, it is understood that $\boldsymbol{P} = \mathbf{1}_d \mathbf{1}_d / d$ and $\boldsymbol{P}_{\perp} = \boldsymbol{I}_d - \boldsymbol{P}$.

## 2. A Toy Example: $\mathbb{Z}_2$-Synchronization

Before passing to the main results, it is useful to present the main ideas on a toy example. In $\mathbb{Z}_2$ synchronization we are interested in estimating a vector $\boldsymbol{\sigma} \in \{+1, -1\}^n$ from observations $\boldsymbol{X} \in \mathbb{R}^{n \times n}$, generated according to

$$\boldsymbol{X} = \frac{\lambda}{n} \boldsymbol{\sigma} \boldsymbol{\sigma}^{\mathsf{T}} + \boldsymbol{Z} \,, \tag{2.1}$$

where $\boldsymbol{Z} = \boldsymbol{Z}^{\mathsf{T}} \in \mathbb{R}^{n \times n}$ is a noise matrix from the Gaussian Orthogonal Ensemble $\mathsf{GOE}(n)$, namely $(Z_{ij})_{i<j\leq n} \sim_{iid} \mathsf{N}(0, 1/n)$ are independent of $(Z_{ii})_{i \leq n} \sim_{iid} \mathsf{N}(0, 2/n)$. The parameter $\lambda \geq 0$ corresponds to the signal-to-noise ratio.

It is known that for $\lambda \leq 1$ no algorithm can estimate $\boldsymbol{\sigma}$ from data $\boldsymbol{X}$ with positive correlation in the limit $n \to \infty$. The following is an immediate consequence of (Korada & Macris, 2009; Deshpande et al., 2016), see Supplementary Material (SM).

**Lemma 2.1.** *Under model (2.1), for $\lambda \leq 1$ and any estimator $\hat{\boldsymbol{\sigma}} : \mathbb{R}^{n \times n} \to \mathbb{R}^n \setminus \{\mathbf{0}\}$, the following limit holds in probability:*

$$\limsup_{n \to \infty} \frac{|\langle \hat{\boldsymbol{\sigma}}(\boldsymbol{X}), \boldsymbol{\sigma} \rangle|}{\|\hat{\boldsymbol{\sigma}}(\boldsymbol{X})\|_2 \|\boldsymbol{\sigma}\|_2} = 0 \,. \tag{2.2}$$

How does variational inference perform on this problem? Any product probability distribution $\hat{q}(\boldsymbol{\sigma}) = \prod_{i=1}^n q_i(\sigma_i)$ can be parametrized by the means $m_i = \sum_{\sigma_i \in \{+1,-1\}} q_i(\sigma_i)\, \sigma_i$, and it is immediate to get $\mathrm{KL}(\hat{q} \| p_{\boldsymbol{\sigma}|\boldsymbol{X}}) = \mathcal{F}(\boldsymbol{m}) + \text{const.}$, where

$$\mathcal{F}(\boldsymbol{m}) \equiv -\frac{\lambda}{2} \langle \boldsymbol{m}, \boldsymbol{X}_0 \boldsymbol{m} \rangle - \sum_{i=1}^n \mathsf{h}(m_i) \,. \tag{2.3}$$

Here $\boldsymbol{X}_0$ is obtained from $\boldsymbol{X}$ by setting the diagonal entries to 0, and $\mathsf{h}(x) = -\frac{(1+x)}{2} \log \frac{(1+x)}{2} - \frac{(1-x)}{2} \log \frac{(1-x)}{2}$ is the binary entropy function. In view of Lemma 2.1, the correct posterior distribution should be essentially uniform, resulting in $\boldsymbol{m}$ vanishing. Indeed, $\boldsymbol{m}_* = 0$ is a stationary point of the mean field free energy $\mathcal{F}(\boldsymbol{m})$: $\nabla \mathcal{F}(\boldsymbol{m})|_{\boldsymbol{m}=\boldsymbol{m}_*} = 0$. We refer to this as the 'uninformative fixed point'.

*Is $\boldsymbol{m}_*$ a local minimum?* Computing the Hessian at the uninformative fixed point yields

$$\nabla^2 \mathcal{F}(\boldsymbol{m})\big|_{\boldsymbol{m}=\boldsymbol{m}_*} = -\lambda \boldsymbol{X}_0 + \boldsymbol{I} \,. \tag{2.4}$$

The matrix $\boldsymbol{X}_0$ is a rank-one deformation of a Wigner matrix and its spectrum is well understood (Baik et al., 2005; Féral & Péché, 2007; Benaych-Georges & Nadakuditi, 2011). For

$\lambda \le 1$, its eigenvalues are contained with high probability in the interval $[-2, 2]$, with $\lambda_{\min}(\boldsymbol{X}) \to -2$, $\lambda_{\max}(\boldsymbol{X}) \to 2$ as $n \to \infty$. For $\lambda > 1$, $\lambda_{\max}(\boldsymbol{X}) \to \lambda + \lambda^{-1}$, while the other eigenvalues are contained in $[-2, 2]$. This implies

$$\lim_{n\to\infty} \lambda_{\min}(\nabla^2 \mathcal{F}\big|_{\boldsymbol{m}_*}) = \begin{cases} 1 - 2\lambda & \text{if } \lambda \le 1, \\ -\lambda^2 & \text{if } \lambda > 1. \end{cases} \quad (2.5)$$

In other words, $\boldsymbol{m}_* = 0$ is a local minimum for $\lambda < 1/2$, but becomes a saddle point for $\lambda > 1/2$. In particular, for $\lambda \in (1/2, 1)$, variational inference will produce an estimate $\hat{\boldsymbol{m}} \neq 0$, although the posterior should be essentially uniform. In fact, it is possible to make this conclusion more quantitative. In the Supplementary Material we prove that any local minimum $\hat{\boldsymbol{m}}$ has norm $\|\hat{\boldsymbol{m}}\|_2^2 \ge c_0 n$, with high probability.

The above mathematical phenomenon implies that naive mean field leads to incorrect inferential statements for $\lambda \in (1/2, 1)$. In order to formalize this point, given any estimators $\{\hat{q}_i(\,\cdot\,)\}_{i\le n}$ of the posterior marginals, we define the per-coordinate expected coverage as

$$\mathscr{Q}(\hat{q}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\big(\sigma_i = \arg\max_{\tau_i \in \{+1,-1\}} \hat{q}_i(\tau_i)\big)\,. \quad (2.6)$$

This is the expected fraction of coordinates that are estimated correctly by choosing $\boldsymbol{\sigma}$ according to the estimated posterior. On the other hand, if the $\hat{q}_i$ were accurate, Bayesian theory would suggest claiming the coverage

$$\widehat{\mathscr{Q}}(\hat{q}) \equiv \frac{1}{n}\sum_{i\le n} \max_{\tau_i} \hat{q}_i(\tau_i)\,. \quad (2.7)$$

The following result shows that mean field overestimates the coverage achieved.

**Theorem 1.** *Let $\hat{\boldsymbol{m}} \in [-1, 1]^n$ be any local minimum of the mean field free energy $\mathcal{F}(\boldsymbol{m})$, under the $\mathbb{Z}_2$-synchronization model (2.1), and consider the corresponding posterior marginal estimates $\hat{q}_i(\sigma_i) = (1 + \hat{m}_i\sigma_i)/2$. Then, there exists a numerical constant $c_0 > 0$ such that, with high probability, for $\lambda \in (1/2, 1)$,*

$$\mathscr{Q}(\hat{q}) \le \frac{1}{2} + o_n(1)\,, \quad \widehat{\mathscr{Q}}(\hat{q}) \ge \frac{1}{2} + c_0 \min\big((2\lambda - 1), 1\big)\,.$$

## 3. Instability for Topic Models

### 3.1. Information-Theoretic Limit

As in the case of $\mathbb{Z}_2$ synchronization discussed in Section 2, we expect it to be impossible to estimate the factors $\boldsymbol{W}, \boldsymbol{H}$ with strictly positive correlation for small enough signal-to-noise ratio $\beta$ (or small enough sample size $\delta$). The exact threshold was characterized recently in (Miolane, 2017) (but see also (Deshpande & Montanari, 2014; Barbier et al., 2016; Lelarge & Miolane, 2016; Lesieur et al., 2017) for closely related results). The characterization in (Miolane, 2017) is given in terms of a variational principle over $k \times k$ matrices.

**Theorem 2** (Special case of (Miolane, 2017))**.** *Let $\mathrm{I}_n(\boldsymbol{X}; \boldsymbol{W}, \boldsymbol{H})$ denote the mutual information between the data $\boldsymbol{X}$ and the factors $\boldsymbol{H}, \boldsymbol{W}$ under the LDA model (1.1). Then, the following limit holds almost surely*

$$\lim_{n,d\to\infty} \frac{1}{d}\mathrm{I}_n(\boldsymbol{X}; \boldsymbol{W}, \boldsymbol{H}) = \inf_{\boldsymbol{M}\in\mathbb{S}_k} \mathsf{RS}(\boldsymbol{M}; k, \delta, \nu)\,, \quad (3.1)$$

*where $\mathbb{S}_k$ is the cone of $k \times k$ positive semidefinite matrices and $\mathsf{RS}(\cdots)$ is a function given explicitly in SM.*

It is also shown in SM that $\boldsymbol{M}^* = (\delta\beta/k^2)\boldsymbol{J}_k$ is a stationary point of the free energy $\mathsf{RS}(\boldsymbol{M}; k, \delta, \nu)$. We shall refer to $\boldsymbol{M}^*$ as the uninformative point. Let $\beta_{\text{Bayes}} = \beta_{\text{Bayes}}(k, \delta, \nu)$ be the supremum value of $\beta$ such that the infimum in Eq. (3.1) is uniquely achieved at $\boldsymbol{M}^*$ (namely, the supremum $\beta$ such that $\mathsf{RS}(\boldsymbol{M}; k, \delta, \nu) > \mathsf{RS}(\boldsymbol{M}_*; k, \delta, \nu)$ for all $\boldsymbol{M} \neq \boldsymbol{M}_*$.

As formalized below, for $\beta < \beta_{\text{Bayes}}$ the data $\boldsymbol{X}$ do not contain sufficient information for estimating $\boldsymbol{H}$, $\boldsymbol{W}$ in a non-trivial manner.

**Proposition 3.1.** *Let $\boldsymbol{M}_* = \delta\beta\boldsymbol{J}_k/k^2$. Then $\boldsymbol{M}^*$ is a stationary point of the function $\boldsymbol{M} \mapsto \mathsf{RS}(\boldsymbol{M}; \beta, k, \delta, \nu)$. Further, it is a local minimum provided $\beta < \beta_{\text{spect}}(k, \delta, \nu)$ where the spectral threshold is given by $\beta_{\text{spect}} \equiv k(k\nu + 1)/\sqrt{\delta}$.*

*Finally, if $\beta < \beta_{\text{Bayes}}(k, \delta, \nu)$, there is no estimator $\boldsymbol{X} \mapsto \widehat{\boldsymbol{F}}_n(\boldsymbol{X})$ whose mean square error $\mathbb{E}\{\|\boldsymbol{W}\boldsymbol{H}^{\mathsf{T}} - \widehat{\boldsymbol{F}}_n(\boldsymbol{X})\|_F^2\}$ is asymptotically smaller than the mean square error of the trivial estimator $\widehat{\boldsymbol{F}}_n(\boldsymbol{X}) = c\mathbf{1}_n(\boldsymbol{X}^{\mathsf{T}}\mathbf{1}_n)^{\mathsf{T}}$, for $c \equiv \sqrt{\beta}/(k + \beta\delta)$ a constant.*

This result compares the mean square error of an arbitrary estimator $\widehat{\boldsymbol{F}}_n$, to the mean square error of the trivial estimator that replaces each column of $\boldsymbol{X}$ by its average. Of course, $\beta_{\text{Bayes}} \le \beta_{\text{spect}}$. However, this upper bound appears to be tight for small $k$.

**Remark 3.1.** Solving numerically the $k(k+1)/2$-dimensional problem (3.1) indicates that $\beta_{\text{Bayes}}(k, \nu, \delta) = \beta_{\text{spect}}(k, \nu, \delta)$ for $k \in \{2, 3\}$ and $\nu = 1$.

### 3.2. Naive Mean Field Free Energy

We consider a trial joint distribution that factorizes according to rows of $\boldsymbol{W}$ and $\boldsymbol{H}$ according to Eq. (1.3). It turns out (see SM) that, for any stationary point of $\mathrm{KL}(\hat{q}\|p_{\boldsymbol{H},\boldsymbol{W}|\boldsymbol{X}})$ over such product distributions, the marginals take the form

$$q_i(\boldsymbol{h}) = e^{\langle \boldsymbol{m}_i, \boldsymbol{h}\rangle - \frac{1}{2}\langle \boldsymbol{h}, \boldsymbol{Q}_i\boldsymbol{h}\rangle - \phi(\boldsymbol{m}_i, \boldsymbol{Q}_i)} q_0(\boldsymbol{h})\,, \quad (3.2)$$

where $q_0(\,\cdot\,)$ is the prior distribution of $\boldsymbol{h}_i$ (the $i$-th row of $\boldsymbol{H}$), and $\phi : \mathbb{R}^k \times \mathbb{R}^{k\times k} \to \mathbb{R}$ is defined implicitly by the normalization condition $\int q_i(\mathrm{d}\boldsymbol{h}_i) = 1$. A similar form holds for $\tilde{q}_a(\boldsymbol{w})$, with parameters $\tilde{\boldsymbol{m}}_a$, $\widetilde{\boldsymbol{Q}}_a$, and normalization factor $\tilde{\phi}(\tilde{\boldsymbol{m}}_a, \widetilde{\boldsymbol{Q}}_a)$. In the following we let $\boldsymbol{m} = (\boldsymbol{m}_i)_{i\le d}$, $\tilde{\boldsymbol{m}} = (\tilde{\boldsymbol{m}}_a)_{a\le n}$ denote the set of parameters in these distributions; these can also be viewed as matrices $\boldsymbol{m} \in \mathbb{R}^{d\times k}$ and $\tilde{\boldsymbol{m}} \in \mathbb{R}^{d\times k}$ whose $i$-th row is $\boldsymbol{m}_i$ (in the former case) or $\tilde{\boldsymbol{m}}_i$ (in the latter).

It is useful to define the functions $\mathsf{F}, \widetilde{\mathsf{F}} : \mathbb{R}^k \times \mathbb{R}^{k\times k} \to \mathbb{R}^k$ and $\mathsf{G}, \widetilde{\mathsf{G}} : \mathbb{R}^k \times \mathbb{R}^{k\times k} \to \mathbb{R}^{k\times k}$ as (proportional to) expectations with respect to the approximate posteriors (3.2)

$$\mathsf{F}(\boldsymbol{m}_i; \boldsymbol{Q}) \equiv \sqrt{\beta} \int \boldsymbol{h}\; q_i(\mathrm{d}\boldsymbol{h})\,, \tag{3.3}$$

$$\mathsf{G}(\boldsymbol{m}_i; \boldsymbol{Q}) \equiv \beta \int \boldsymbol{h}^{\otimes 2}\; q_i(\mathrm{d}\boldsymbol{h})\;. \tag{3.4}$$

Similarly $\widetilde{\mathsf{F}}(\tilde{\boldsymbol{m}}_a; \widetilde{\boldsymbol{Q}})$, $\widetilde{\mathsf{G}}(\tilde{\boldsymbol{m}}_a; \widetilde{\boldsymbol{Q}})$ will denote the first and second moments of $\tilde{q}_a(\boldsymbol{w})$. For $\boldsymbol{m} \in \mathbb{R}^{d\times k}$, we overload the notation and denote by $\mathsf{F}(\boldsymbol{m}; \boldsymbol{Q}) \in \mathbb{R}^{d\times k}$ the matrix whose $i$-th row is $\mathsf{F}(\boldsymbol{m}_i; \boldsymbol{Q})$ (and similarly for $\widetilde{\mathsf{F}}(\tilde{\boldsymbol{m}}; \widetilde{\boldsymbol{Q}})$).

When restricted to a product-form ansatz with parametrization (3.2), the mean field free energy takes the form (see SM) $\mathrm{KL}(\hat{q}\|p_{\boldsymbol{W},\boldsymbol{H}|\boldsymbol{X}}) = \mathcal{F}(\boldsymbol{r}, \tilde{\boldsymbol{r}}, \boldsymbol{\Omega}, \tilde{\boldsymbol{\Omega}}) + d\|\boldsymbol{X}\|_F^2/2 + \log p_{\boldsymbol{X}}(\boldsymbol{X})$, where

$$\mathcal{F}(\boldsymbol{r}, \tilde{\boldsymbol{r}}, \boldsymbol{\Omega}, \tilde{\boldsymbol{\Omega}}) = \sum_{i=1}^{d} \psi_*(\boldsymbol{r}_i, \boldsymbol{\Omega}_i) + \sum_{a=1}^{n} \tilde{\psi}_*(\tilde{\boldsymbol{r}}_a, \tilde{\boldsymbol{\Omega}}) - \sqrt{\beta}\mathsf{Tr}\left(\boldsymbol{X}\boldsymbol{r}\tilde{\boldsymbol{r}}^{\mathsf{T}}\right) + \frac{\beta}{2d}\sum_{i=1}^{d}\sum_{a=1}^{n}\langle \boldsymbol{\Omega}_i, \tilde{\boldsymbol{\Omega}}_a\rangle\,, \tag{3.5}$$

and $\psi_*$, $\tilde{\psi}_*$ are the Legendre duals of $\phi$, $\tilde{\phi}$, e.g.

$$\psi_*(\boldsymbol{r}, \boldsymbol{\Omega}) \equiv \sup_{\boldsymbol{m},\boldsymbol{Q}} \left\{\langle \boldsymbol{r}, \boldsymbol{m}\rangle - \frac{1}{2}\langle \boldsymbol{\Omega}, \boldsymbol{Q}\rangle - \phi(\boldsymbol{m}, \boldsymbol{Q})\right\}.$$

This equation implies a convex duality relation between $(\boldsymbol{r}, \tilde{\boldsymbol{r}}, \boldsymbol{\Omega}, \tilde{\boldsymbol{\Omega}})$ and $(\boldsymbol{m}, \tilde{\boldsymbol{m}}, \boldsymbol{Q}, \widetilde{\boldsymbol{Q}})$. Namely

$$\boldsymbol{r}_i \equiv \frac{1}{\sqrt{\beta}}\mathsf{F}(\boldsymbol{m}_i; \boldsymbol{Q})\,, \quad \boldsymbol{\Omega}_i \equiv \frac{1}{\beta}\mathsf{G}(\boldsymbol{m}_i; \boldsymbol{Q})\,, \tag{3.6}$$

and similarly for $\tilde{\boldsymbol{r}}_a, \tilde{\boldsymbol{\Omega}}_a$ and $\tilde{\boldsymbol{m}}_a, \widetilde{\boldsymbol{Q}}_a$. By strict convexity of $\phi(\boldsymbol{m}, \boldsymbol{Q})$, $\tilde{\phi}(\tilde{\boldsymbol{m}}, \widetilde{\boldsymbol{Q}})$ (the latter is strongly convex on the hyperplane $\langle \boldsymbol{1}, \tilde{\boldsymbol{m}}\rangle = 0$, $\langle \boldsymbol{1}, \widetilde{\boldsymbol{Q}}\boldsymbol{1}\rangle = 0$) we can view $\mathcal{F}(\cdots)$ as a function of $(\boldsymbol{r}, \tilde{\boldsymbol{r}}, \boldsymbol{\Omega}, \tilde{\boldsymbol{\Omega}})$ or $(\boldsymbol{m}, \tilde{\boldsymbol{m}}, \boldsymbol{Q}, \widetilde{\boldsymbol{Q}})$. With an abuse of notation, we will write $\mathcal{F}(\boldsymbol{r}, \tilde{\boldsymbol{r}}, \boldsymbol{\Omega}, \tilde{\boldsymbol{\Omega}})$ or $\mathcal{F}(\boldsymbol{m}, \tilde{\boldsymbol{m}}, \boldsymbol{Q}, \widetilde{\boldsymbol{Q}})$ interchangeably.

A critical (stationary) point of the free energy (3.5) is a point at which $\nabla\mathcal{F}(\boldsymbol{m}, \tilde{\boldsymbol{m}}, \boldsymbol{Q}, \widetilde{\boldsymbol{Q}}) = \boldsymbol{0}$. It turns out that the mean field free energy always admits a point that does not distinguish between the $k$ latent factors, and in particular $\boldsymbol{m} = \boldsymbol{v}\boldsymbol{1}_k^{\mathsf{T}}$, $\tilde{\boldsymbol{m}} = \tilde{\boldsymbol{v}}\boldsymbol{1}_k^{\mathsf{T}}$, as stated in detail below. We will refer to this as the *uninformative critical point* (or *uninformative fixed point*).

**Lemma 3.2.** *The naive mean field free energy of Eq. (3.5) admits a stationary point whereby, for all $i \in [d]$, $a \in [n]$,*

$$\boldsymbol{m}_i^* = \frac{\sqrt{\beta}}{k}\,(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{1}_n)_i\,\boldsymbol{1}_k\,, \tag{3.7}$$

$$\tilde{\boldsymbol{m}}_a^* = \frac{\beta}{k(1+q_1^*+kq_2^*)}\,(\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{1}_n)_a\,\boldsymbol{1}_k\,, \tag{3.8}$$

*and further $\boldsymbol{Q}_i^* = q_1^*\boldsymbol{I}_k + q_2^*\boldsymbol{J}_k$, $\widetilde{\boldsymbol{Q}}_a^* = \tilde{q}_1^*\boldsymbol{I}_k + \tilde{q}_2^*\boldsymbol{J}_k$. The parameters $q_i^*, \tilde{q}_i^*$ are explicitly given in the Supplementary Material.*

We note that there appear always to be a unique stationary point of the form given by this lemma. Although we do not have a proof of uniqueness, in the SM we prove that the solution is unique conditional on a certain inequality that can be easily checked numerically.

### 3.3. Naive Mean Field Iteration

As mentioned in the introduction, the variational approximation of the free energy is often minimized by alternating minimization over the marginals $(q_i)_{i\le d}$, $(\tilde{q}_a)_{a\le n}$ of Eq. (1.3). Using the parametrization (3.2), we obtain the following naive mean field iteration for $\boldsymbol{m}^t, \tilde{\boldsymbol{m}}^t, \boldsymbol{Q}^t, \widetilde{\boldsymbol{Q}}^t$ (see SM):

$$\boldsymbol{m}^{t+1} = \boldsymbol{X}^{\mathsf{T}}\,\widetilde{\mathsf{F}}(\tilde{\boldsymbol{m}}^t; \widetilde{\boldsymbol{Q}}^t)\,, \quad \boldsymbol{Q}^{t+1} = \frac{1}{d}\sum_{a=1}^{n}\widetilde{\mathsf{G}}(\tilde{\boldsymbol{m}}_a^t; \widetilde{\boldsymbol{Q}}^t)\,,$$

$$\tilde{\boldsymbol{m}}^t = \boldsymbol{X}\,\mathsf{F}(\boldsymbol{m}^t; \boldsymbol{Q}^t)\,, \quad \widetilde{\boldsymbol{Q}}^t = \frac{1}{d}\sum_{i=1}^{d}\mathsf{G}(\boldsymbol{m}_i^t; \boldsymbol{Q}^t)\,.$$

Note that, while the free energy naturally depends on the $(\boldsymbol{Q}_i)_{i\le d}$, $(\widetilde{\boldsymbol{Q}}_a)_{a\le n}$, the iteration sets $\boldsymbol{Q}_i^t = \boldsymbol{Q}^t$, $\widetilde{\boldsymbol{Q}}_a^t = \widetilde{\boldsymbol{Q}}^t$, independent of the indices $i, a$. In fact, any stationary point of $\mathcal{F}(\boldsymbol{m}, \tilde{\boldsymbol{m}}, \boldsymbol{Q}, \widetilde{\boldsymbol{Q}})$ can be shown to be of this form.

The state of the naive mean field iteration is given by the pair $(\boldsymbol{m}^t, \boldsymbol{Q}^t) \in \mathbb{R}^{d\times k} \times \mathbb{R}^{k\times k}$, and $(\tilde{\boldsymbol{m}}^t, \widetilde{\boldsymbol{Q}}^t)$ can be viewed as derived variables. The iteration hence defines a mapping $\mathcal{M}_{\boldsymbol{X}} : \mathbb{R}^{d\times k} \times \mathbb{R}^{k\times k} \to \mathbb{R}^{d\times k} \times \mathbb{R}^{k\times k}$, and we can write it in the form $(\boldsymbol{m}^{t+1}, \boldsymbol{Q}^{t+1}) = \mathcal{M}_{\boldsymbol{X}}(\boldsymbol{m}^t, \boldsymbol{Q}^t)$. Any critical point of the free energy (3.5) is a fixed point of the naive mean field iteration and vice-versa, as shown in the SM. In particular, the uninformative critical point $(\boldsymbol{m}^*, \tilde{\boldsymbol{m}}^*, \boldsymbol{Q}^*, \widetilde{\boldsymbol{Q}}^*)$ is a fixed point of the naive mean field iteration.

### 3.4. Instability

In view of Section 3.1, for $\beta < \beta_{\text{Bayes}}(k, \delta, \nu)$, the real posterior should be centered around a point symmetric under

permutations of the topics. In particular, the posterior $\tilde{q}(\boldsymbol{w}_a)$ over the weights of document $a$ should be centered around the symmetric distribution $\boldsymbol{w}_a = (1/k, \dots, 1/k)$.

A minimum consistency condition for variational inference is that the uninformative stationary point is a local minimum of the posterior for $\beta < \beta_{\text{Bayes}}$. The next theorem provides a necessary condition for stability of the uninformative point, which we expect to be tight. As discussed below, it implies that this point is a saddle in an interval of $\beta$ below $\beta_{\text{Bayes}}$. We recall that the index of a smooth function $f$ at stationary point $\boldsymbol{x}_*$ is the number of the negative eigenvalues of the Hessian $\nabla^2 f(\boldsymbol{x}_*)$.

**Theorem 3.** *Define $q_1^*$, $q_2^*$ as in Lemma 3.2, and let*

$$L(\beta, k, \delta, \nu) \equiv \frac{\beta(1+\sqrt{\delta})^2}{1+q_1^*}\left(\frac{q_1^*}{\delta\beta} + k\left[\frac{q_2^*}{1+q_1^*+kq_2^*}\left(\frac{1}{\delta\beta}+\frac{1}{k}\right) - \frac{1}{k^2}\right]_+\right).$$

*If $L(\beta, k, \delta, \nu) > 1$, then there exists $\varepsilon_1, \varepsilon_2 > 0$ such that the uninformative critical point of Lemma 3.2, $(\boldsymbol{m}^*, \tilde{\boldsymbol{m}}^*, \boldsymbol{Q}^*, \widetilde{\boldsymbol{Q}}^*)$ is, with high probability, a saddle point, with index at least $n\varepsilon_1$ and $\lambda_{\min}(\mathcal{F}|_{\boldsymbol{m}^*, \tilde{\boldsymbol{m}}^*, \boldsymbol{Q}^*, \widetilde{\boldsymbol{Q}}^*}) \le -\varepsilon_2$.*

*Correspondingly $(\boldsymbol{m}^*, \boldsymbol{Q}^*)$ is an unstable critical point of the mapping $\mathcal{M}_{\boldsymbol{X}}$ in the sense that the Jacobian $\boldsymbol{D}\mathcal{M}_{\boldsymbol{X}}$ has spectral radius larger than one at $(\boldsymbol{m}^*, \boldsymbol{Q}^*)$.*

**Remark 3.2.** We established an analogous instability phase transition for the Dirichlet case. The corresponding prediction is reported in Figure 3. Explicit formulas are reported in the SM.

In the following, we will say that a fixed point $(\boldsymbol{m}^*, \boldsymbol{Q}^*)$ is stable if the linearization of $\mathcal{M}_{\boldsymbol{X}}(\,\cdot\,)$ at $(\boldsymbol{m}^*, \boldsymbol{Q}^*)$ (i.e. the Jacobian matrix $\boldsymbol{D}\mathcal{M}_{\boldsymbol{X}}(\boldsymbol{m}^*, \boldsymbol{Q}^*)$) has spectral radius smaller than one. By the Hartman-Grobman linearization theorem (Perko, 2013), this implies that $(\boldsymbol{m}^*, \boldsymbol{Q}^*)$ is an attractive fixed point. Vice-versa, we say that $(\boldsymbol{m}^*, \boldsymbol{Q}^*)$ is unstable if the Jacobian $\boldsymbol{D}\mathcal{M}_{\boldsymbol{X}}(\boldsymbol{m}^*, \boldsymbol{Q}^*)$ has spectral radius larger than one. In this case, for any neighborhood of $(\boldsymbol{m}^*, \boldsymbol{Q}^*)$, and a generic initialization in that neighborhood, $(\boldsymbol{m}^t, \boldsymbol{Q}^t)$ does not converge to the fixed point.

Motivated by Theorem 3, we define the instability threshold $\beta_{\text{inst}} = \beta_{\text{inst}}(k, \delta, \nu)$ as the infimum $\beta \ge 0$ such that $L(\beta, k, \delta, \nu) > 1$. Let us emphasize that, while we discuss the consequences of the instability at $\beta_{\text{inst}}$ on the naive mean field iteration, this is a problem of the variational free energy (3.5) and not of the specific optimization algorithm.

### 3.5. Numerical Results for Naive Mean Field

In order to investigate the impact of the instability described above, we carried out extensive numerical simulations with the naive mean field iteration. After any number of iterations $t$, estimates of the factors $\boldsymbol{H}, \boldsymbol{W}$ are obtained by computing expectations with respect to the marginals (3.2). This results in

$$\widehat{\boldsymbol{H}}^t = \boldsymbol{r}^t = \frac{1}{\sqrt{\beta}}\mathsf{F}(\boldsymbol{m}^t; \boldsymbol{Q}_t)\,, \quad \widehat{\boldsymbol{W}}^t = \tilde{\boldsymbol{r}}^t = \frac{1}{\sqrt{\beta}}\widetilde{\mathsf{F}}(\tilde{\boldsymbol{m}}^t; \widetilde{\boldsymbol{Q}}_t)\,.$$

We select a two-dimensional grid of $(\delta, \beta)$'s and generate 400 different instances according to the LDA model for each grid point. We report various statistics of the estimates aggregated over the 400 instances. We have performed the simulations for $\nu = \tilde{\nu} = 1$ and $k \in \{2, 3\}$, both for the Gaussian and the Dirichlet models. For space considerations, we focus here on the case $\nu = 1$, $k = 2$, and discuss other results in the SM. (Simulations for other values of $\nu$ also yield similar results.)

We initialize the naive mean field iteration near the uninformative fixed-point and iterate until a convergence criterion or the maximum number of 300 iterations is reached.

Recall the definition $\boldsymbol{P}_\perp = \boldsymbol{I}_k - \boldsymbol{1}_k\boldsymbol{1}_k^{\mathsf{T}}/k$. In order to investigate the instability of Theorem 3, we define the quantities

$$\mathsf{V}(\widehat{\boldsymbol{W}}) \equiv \frac{1}{\sqrt{n}}\|\widehat{\boldsymbol{W}}\boldsymbol{P}_\perp\|_F\,, \qquad \mathsf{V}(\widehat{\boldsymbol{H}}) \equiv \frac{1}{\sqrt{d}}\|\widehat{\boldsymbol{H}}\boldsymbol{P}_\perp\|_F$$

In Figure 1 we plot empirical results for the average $\mathsf{V}(\widehat{\boldsymbol{W}})$, $\mathsf{V}(\widehat{\boldsymbol{H}})$ for $k = 2$, $\nu = 1$ and four values of $\delta$, within the Gaussian model. In Figure 2 (left frame), we plot the empirical probability that variational inference does not converge to the uninformative fixed point or, more precisely,

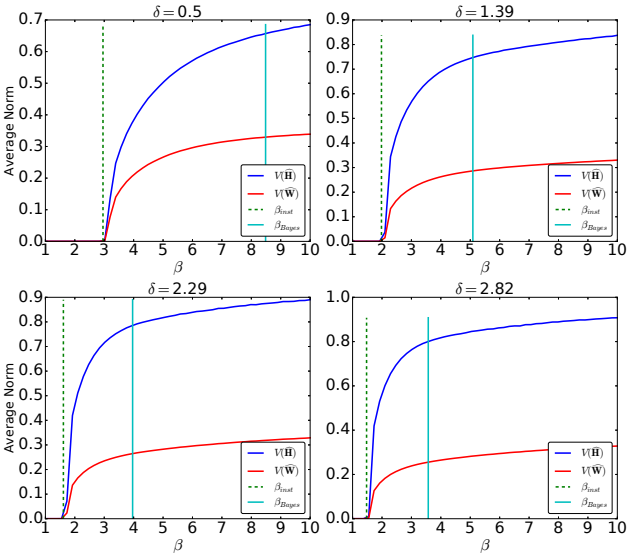


*Figure 1.* Normalized distances $\mathsf{V}(\widehat{\boldsymbol{H}})$, $\mathsf{V}(\widehat{\boldsymbol{W}})$ of the naive mean field estimates from the uninformative fixed point. Here $k = 2$, $d = 1000$ and $n = d\delta$: each data point corresponds to an average over 400 random realizations.

$\widehat{\mathbb{P}}(\mathsf{V}(\widehat{\boldsymbol{W}}) \geq \varepsilon_0)$ with $\varepsilon_0 = 10^{-4}$, evaluated on a grid of $(\beta, \delta)$ values, for the same model. We also plot the Bayes threshold $\beta_{\text{Bayes}}$ (which numerically coincides with the spectral threshold $\beta_{\text{spect}}$) and the instability $\beta_{\text{inst}}$.

It is clear from Figures 1, 2 (left frame), that variational inference stops converging to the uninformative fixed point (although we initialize close to it) when $\beta$ is still significantly smaller than the Bayes threshold $\beta_{\text{Bayes}}$ (i.e. in a regime in which the uninformative fixed point would a reasonable output). The data are consistent with the hypothesis that variational inference becomes unstable at $\beta_{\text{inst}}$, as predicted by Theorem 3.

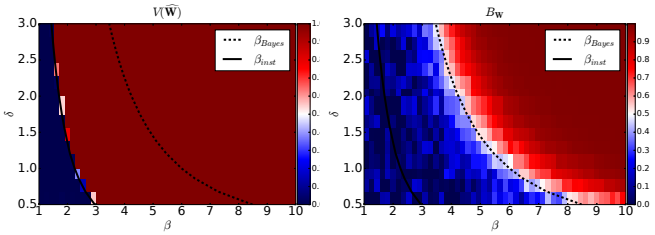


*Figure 2.* Gaussian model. Left frame: Empirical fraction of instances such that $\mathsf{V}(\widehat{\boldsymbol{W}}) \geq \varepsilon_0 = 10^{-4}$, where $\widehat{\boldsymbol{W}}$ is the naive mean field estimate. Here $k = 2$, $d = 1000$ and, for each $(\delta, \beta)$ point on a grid, we used 400 random realizations to estimate the probability of $\mathsf{V}(\widehat{\boldsymbol{W}}) \geq \varepsilon_0$. Right frame: Binder cumulant for the correlation between the naive mean field estimate $\widehat{\boldsymbol{W}}$, and the true weights $\boldsymbol{W}$, $\boldsymbol{H}$. Here $k = 2$, $d = 1000$ and $n = d\delta$, and we averaged over 400 realizations. Solid lines: The instability curve $\beta_{\text{inst}}(\delta)$. Dashed lines: The Bayes phase transition $\beta_{\text{Bayes}}$.

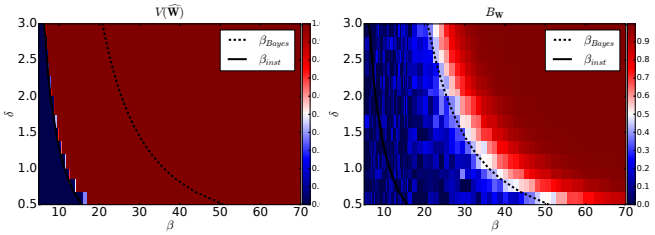


*Figure 3.* Same as for Figure 2, for the Dirichlet model.

Because of Proposition 3.1, we expect the estimates $\widehat{\boldsymbol{H}}, \widehat{\boldsymbol{W}}$ produced by variational inference to be asymptotically uncorrelated with the true factors for $\beta_{\text{inst}} < \beta < \beta_{\text{Bayes}}$. In order to test this hypothesis, we compute suitable correlation ratios between $\widehat{\boldsymbol{H}}, \widehat{\boldsymbol{W}}$ and the true parameters (known as 'Binder cumulants' $\mathsf{B}_{\boldsymbol{H}}$ and $\mathsf{B}_{\boldsymbol{W}}$). These quantities grow from 0 to 1 as $\beta$ grows, and the transition is centered around $\beta_{\text{Bayes}}$. Figure 2 (right frame) reports the results $\mathsf{B}_{\boldsymbol{W}}$ on a grid of $(\beta, \delta)$ values. Again, the transition is well predicted by the analytical curve $\beta_{\text{Bayes}}$. These data support our claim that, for $\beta_{\text{inst}} < \beta < \beta_{\text{Bayes}}$, the output of variational inference is non-uniform but uncorrelated with the true signal.

In Figure 3, we repeat the same experiment carried out in Figure 2, but for the Dirichlet model, with $\tilde{\nu} = \nu = 1$.

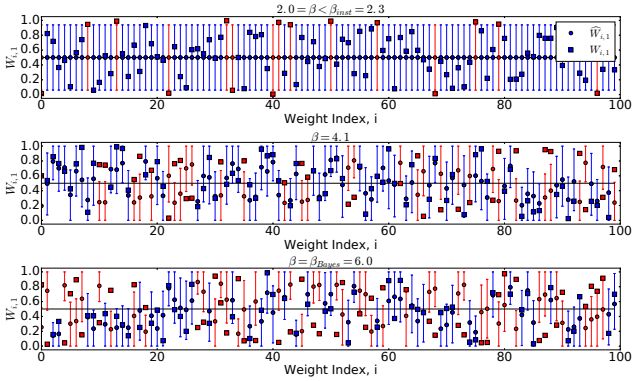


*Figure 4.* Gaussian model. Bayesian credible intervals as computed by variational inference at nominal coverage level $1 - \alpha = 0.9$. Here $k = 2$, $n = d = 5000$, $\beta \in \{2, 4.1, 6\}$ (for reference $\beta_{\text{inst}} \approx 2.3$, $\beta_{\text{Bayes}} = 6$). Circles: posterior mean. Squares: true weights. Red: coordinates on which the credible interval does not cover the true value of $w_{i,1}$.

Once more we observe a large region of model parameters for which the variational posterior is not centered on the uninformative point, but is uncorrelated with the ground truth. The instability phase transition is well captured by our theoretical prediction $\beta_{\text{inst}}$ also in the Dirichlet case. Finally, in Figure 4 we plot the estimates obtained for 100 entries of the weights vector $w_{i,1}$ for three instances with $n = d = 5000$ and $\beta = 2 < \beta_{\text{inst}}$, $\beta = 4.1 \in (\beta_{\text{inst}}, \beta_{\text{Bayes}})$ and $\beta = 6 = \beta_{\text{Bayes}}$. The interval for $w_{a,1}$ is the form $\{w_{a,1} \in [0,1] : \tilde{q}_a(w_{a,1}) \geq t_a(\alpha)\}$ and are constructed to achieve nominal coverage level $1 - \alpha = 0.9$. It is visually clear that the claimed coverage level is not verified in these simulations for $\beta > \beta_{\text{inst}}$, confirming our analytical results. Indeed, for the three simulations in Figure 4 we achieve coverage 0.87 (for $\beta = 2 < \beta_{\text{inst}}$), 0.65 (for $\beta = 4.1 \in (\beta_{\text{inst}}, \beta_{\text{Bayes}})$), and 0.51 (for $\beta = 6 = \beta_{\text{Bayes}}$). Further results of this type are reported in the SM.

## 4. Fixing the Instability

The fact that naive mean field is not accurate for certain classes of random high-dimensional probability distributions is well understood within statistical physics. In particular, in the context of mean field spin glasses (Mezard et al., 1988), naive mean field is known to lead to an asymptotically incorrect expression for the free energy. We expect the same mechanism to be relevant for topic models.

Namely, the product-form expression (1.3) only holds asymptotically in the sense of finite-dimensional marginals. However, when computing the term $\mathbb{E}_{\hat{q}} \log p_{\boldsymbol{X}|\boldsymbol{W},\boldsymbol{H}}(\boldsymbol{X}|\boldsymbol{H}, \boldsymbol{W})$ in the KL divergence (1.2), the error due to the product form approximation is non-negligible. Keeping track of this error leads to the so-called TAP free energy.

The TAP approach replaces the free energy (3.5) with

$\mathcal{F}_{\text{TAP}} = \mathcal{F}_{\text{TAP}}(\boldsymbol{r}, \tilde{\boldsymbol{r}})$ defined as follows (see SM)

$$\mathcal{F}_{\text{TAP}}(\boldsymbol{r}, \tilde{\boldsymbol{r}}) = -\sqrt{\beta}\mathsf{Tr}\left(\boldsymbol{X}\boldsymbol{r}\tilde{\boldsymbol{r}}^{\mathsf{T}}\right) - \frac{\beta}{2d}\sum_{i=1}^{d}\sum_{a=1}^{n}\langle \boldsymbol{r}_i, \tilde{\boldsymbol{r}}_a\rangle^2 + \sum_{i=1}^{d}\psi\left(\boldsymbol{r}_i, \frac{\beta}{d}\sum_{a=1}^{n}\tilde{\boldsymbol{r}}_a^{\otimes 2}\right) + \sum_{a=1}^{n}\tilde{\psi}\left(\tilde{\boldsymbol{r}}_a, \frac{\beta}{d}\sum_{i=1}^{d}\boldsymbol{r}_i^{\otimes 2}\right),$$

where $\tilde{\boldsymbol{r}}\mathbf{1}_k = \mathbf{1}_n$, and we defined the partial Legendre transforms $\psi(\boldsymbol{r}, \boldsymbol{Q}) \equiv \sup_{\boldsymbol{m}}\{\langle \boldsymbol{r}, \boldsymbol{m}\rangle - \phi(\boldsymbol{m}, \boldsymbol{Q})\}$ and similarly for $\tilde{\psi}(\tilde{\boldsymbol{r}}, \widetilde{\boldsymbol{Q}})$.

Calculus shows that stationary points of this free energy are in one-to-one correspondence with the fixed points of the following iteration:

$$\begin{aligned}
\boldsymbol{m}^{t+1} &= \boldsymbol{X}^{\mathsf{T}}\widetilde{\mathsf{F}}(\tilde{\boldsymbol{m}}^t; \widetilde{\boldsymbol{Q}}^t) - \mathsf{F}(\boldsymbol{m}^t; \boldsymbol{Q}^t)\tilde{\boldsymbol{\Omega}}_t\,,\\
\tilde{\boldsymbol{m}}^t &= \boldsymbol{X}\,\mathsf{F}(\boldsymbol{m}^t; \boldsymbol{Q}^t) - \widetilde{\mathsf{F}}(\tilde{\boldsymbol{m}}^{t-1}; \widetilde{\boldsymbol{Q}}^{t-1})\boldsymbol{\Omega}_t\,,\\
\boldsymbol{Q}^{t+1} &= \frac{1}{d}\sum_{a=1}^{n}\widetilde{\mathsf{F}}(\tilde{\boldsymbol{m}}_a^t; \widetilde{\boldsymbol{Q}}^t)^{\otimes 2}\,,\quad \widetilde{\boldsymbol{Q}}^t = \frac{1}{d}\sum_{i=1}^{d}\mathsf{F}(\boldsymbol{m}_i^t; \boldsymbol{Q}^t)^{\otimes 2}\,.
\end{aligned}$$

where $\boldsymbol{\Omega}_t, \tilde{\boldsymbol{\Omega}}_t \in \mathbb{R}^{k\times k}$ are matrices defined in the SM. The stationarity conditions for the TAP free energy $\mathcal{F}_{\text{TAP}}$ are known as TAP equations, and the above iterative algorithm is a special case of approximate message passing (AMP), with Bayesian updates.

Estimates of the factors $\boldsymbol{W}$, $\boldsymbol{H}$ are computed following the same recipe as for naive mean field, cf. Eq. (**??**), namely $\widehat{\boldsymbol{H}}^t = \boldsymbol{r}^t = \mathsf{F}(\boldsymbol{m}^t; \boldsymbol{Q}_t)/\sqrt{\beta}$, $\widehat{\boldsymbol{W}}^t = \tilde{\boldsymbol{r}}^t = \widetilde{\mathsf{F}}(\tilde{\boldsymbol{m}}^t; \widetilde{\boldsymbol{Q}}_t)/\sqrt{\beta}$.

It is not hard to see that the AMP iteration admits an uninformative fixed point, which is a stationary point of the TAP free energy. This construction is analogous to the one for naive mean field, and we omit it here.

The next theorem establishes that the uninformative fixed point of the TAP free energy is a local minimum for all $\beta$ below the spectral threshold $\beta_{\text{spect}}(k, \nu, \delta)$. Since $\beta_{\text{Bayes}}(k, \nu, \delta) \le \beta_{\text{spect}}(k, \nu, \delta)$, this shows that the instability we discovered in the case of naive mean field is corrected by the TAP free energy.

**Theorem 4.** *Let $(\boldsymbol{r}_*, \tilde{\boldsymbol{r}}_*)$ be the uninformative stationary point of the TAP free energy. If $\beta < \beta_{\text{spect}}(k, \nu, \delta)$, then there exists $\varepsilon > 0$ such that, with high probability $\lambda_{\min}\left(\nabla^2\mathcal{F}_{\text{TAP}}|_{(\boldsymbol{r}_*, \tilde{\boldsymbol{r}}_*)}\right) \ge \varepsilon$.*

In order to confirm the stability analysis at the previous section, we carried out numerical simulations analogous to the ones of Section 3.5. We initialize the iteration as for naive mean field, and monitor the same quantities, as in Section 3.5. In Figure 5 we report results on the distance from the uninformative subspace$\mathsf{V}(\widehat{\boldsymbol{W}})$, (left frame), and the Binder cumulant $\mathsf{B}_{\boldsymbol{W}}$, measuring the correlation between AMP estimates and the true factors $\boldsymbol{W}, \boldsymbol{H}$ (right frame). We repeat the same experiment in Figure 6 for the Dirichlet model.

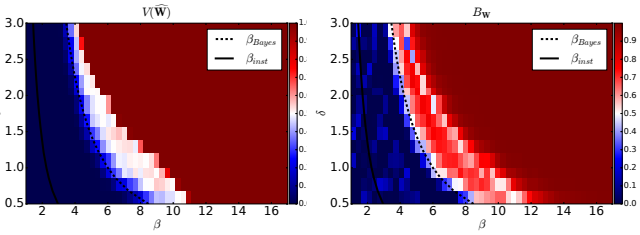


*Figure 5.* Left frame: Empirical fraction of instances such that $\mathsf{V}(\widehat{\boldsymbol{W}}) \ge \varepsilon_0 = 5\cdot 10^{-3}$, where $\widehat{\boldsymbol{W}}$ is the AMP estimate. Here $k = 2$, $d = 1000$, and for each $(\delta, \beta)$ point on the grid we ran AMP on 400 random realizations. Right frame: Binder cumulant for the correlation between AMP estimates $\widehat{\boldsymbol{W}}$, $\widehat{\boldsymbol{H}}$ and the true weights and topics $\boldsymbol{W}$, $\boldsymbol{H}$. Here $k = 2$,$d = 1000$ and estimates are obtained by averaging over 400 realizations.

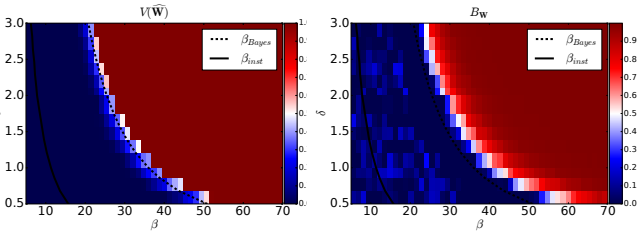


*Figure 6.* Same as in Figure 5, except for the Dirichlet model.

In the intermediate regime $\beta \in (\beta_{\text{inst}}, \beta_{\text{spect}})$, the behavior of AMP is strikingly different from the one of naive mean field. AMP remains close to the uninformative fixed point, confirming that this is a local minimum of the TAP free energy. The distance from the uninformative subspace starts growing only at the spectral threshold $\beta_{\text{spect}}$ (which coincides, in the present cases, with the Bayes threshold $\beta_{\text{Bayes}}$). At the same point, the correlation with the true factors $\boldsymbol{W}$, $\boldsymbol{H}$ also becomes strictly positive.

## 5. Discussion

Bayesian methods are particularly attractive in unsupervised learning problems such as topic modeling. Even after a low-rank factorization $\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H}^{\mathsf{T}}$ is computed, it is still unclear how to evaluate it, or to which extent it should be trusted. Bayesian approaches provide estimates of the factors $\boldsymbol{W}$, $\boldsymbol{H}$, but also a probabilistic measure of how much these estimates should be trusted. To the extent that the posterior concentrates around its mean, this can be considered as a good estimate of a true underlying signal.

It is well understood that Bayesian estimates can be unreliable if the prior is not chosen carefully. Our work points at a second reason for caution. When variational inference is used for approximating the posterior, the result can be substantially incorrect.

## Acknowledgements

H.J. and A.M. were partially supported by grants NSF CCF-1714305 and NSF IIS-1741162. B.G. was supported by Stanford's Caroline and Fabian Pease Graduate Fellowship and grants NSF-DMS 1418362 and NSF-DMS 1407813.

## References

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(Sep):1981–2014, 2008.

Alaoui, A. E. and Krzakala, F. Estimation in the spiked wigner model: A short proof of the replica formula. *arXiv preprint arXiv:1801.01593*, 2018.

Arora, S., Ge, R., Kannan, R., and Moitra, A. Computing a nonnegative matrix factorization–provably. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pp. 145–162. ACM, 2012a.

Arora, S., Ge, R., and Moitra, A. Learning topic models–going beyond svd. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pp. 1–10. IEEE, 2012b.

Baik, J., Arous, G. B., Péché, S., et al. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5): 1643–1697, 2005.

Barbier, J., Dia, M., Macris, N., Krzakala, F., Lesieur, T., and Zdeborová, L. Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 424–432. Curran Associates Inc., 2016.

Bayati, M. and Montanari, A. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57 (2):764–785, 2011.

Benaych-Georges, F. and Nadakuditi, R. R. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

Berthier, R., Montanari, A., and Nguyen, P.-M. State evolution for approximate message passing with non-separable functions. *arXiv preprint arXiv:1708.03950*, 2017.

Bickel, P., Choi, D., Chang, X., Zhang, H., et al. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, 41(4):1922–1943, 2013.

Binder, K. Finite size scaling analysis of ising model block distribution functions. *Zeitschrift für Physik B Condensed Matter*, 43(2):119–140, 1981.

Blei, D. and Lafferty, J. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006a.

Blei, D. M. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

Blei, D. M. and Lafferty, J. D. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120. ACM, 2006b.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan): 993–1022, 2003.

Blei, D. M., Lafferty, J. D., et al. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35, 2007.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Broderick, T., Boyd, N., Wibisono, A., Wilson, A. C., and Jordan, M. I. Streaming variational bayes. In *Advances in Neural Information Processing Systems*, pp. 1727–1735, 2013.

Celisse, A., Daudin, J.-J., Pierre, L., et al. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6:1847–1899, 2012.

Chang, J. and Blei, D. Relational topic models for document networks. In *Artificial Intelligence and Statistics*, pp. 81–88, 2009.

Deshpande, Y. and Montanari, A. Information-theoretically optimal sparse pca. *arXiv preprint arXiv:1402.2238*, 2014.

Deshpande, Y., Abbe, E., and Abbe, E. Asymptotic mutual information for the balanced binary stochastic block model. *Information and Inference: A Journal of the IMA*, 6(2):125–170, 2016.

Donoho, D. L., Maleki, A., and Montanari, A. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

Donoho, D. L., Maleki, A., and Montanari, A. Message passing algorithms for compressed sensing: I. motivation and construction. In *Information Theory (ITW 2010, Cairo), 2010 IEEE Information Theory Workshop on*, pp. 1–5. IEEE, 2010.

Erosheva, E., Fienberg, S., and Lafferty, J. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 101(suppl 1): 5220–5227, 2004.

Fei-Fei, L. and Perona, P. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pp. 524–531. IEEE, 2005.

Féral, D. and Péché, S. The largest eigenvalue of rank one deformation of large wigner matrices. *Communications in mathematical physics*, 272(1):185–228, 2007.

Giordano, R. J., Broderick, T., and Jordan, M. I. Linear response methods for accurate covariance estimates from mean field variational bayes. In *Advances in Neural Information Processing Systems*, pp. 1441–1449, 2015.

Hoffman, M., Bach, F. R., and Blei, D. M. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pp. 856–864, 2010.

Javanmard, A. and Montanari, A. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference: A Journal of the IMA*, 2(2):115–144, 2013.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Korada, S. B. and Macris, N. Exact solution of the gauge symmetric p-spin glass model on a complete graph. *Journal of Statistical Physics*, 136(2):205–230, 2009.

Krzakala, F., Xu, J., and Zdeborová, L. Mutual information in rank-one matrix estimation. *arXiv preprint arXiv:1603.08447*, 2016.

Lelarge, M. and Miolane, L. Fundamental limits of symmetric low-rank matrix estimation. *Probability Theory and Related Fields*, pp. 1–71, 2016.

Lesieur, T., Krzakala, F., and Zdeborová, L. Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(7): 073403, 2017.

Mezard, M., Parisi, G., Virasoro, M. A., and Thouless, D. J. Spin glass theory and beyond. *Physics Today*, 41:109, 1988.

Miolane, L. Fundamental limits of low-rank matrix estimation: the non-symmetric case. *arXiv preprint arXiv:1702.00473*, 2017.

Montanari, A. and Venkataramanan, R. Estimation of low-rank matrices via approximate message passing. *arXiv preprint arXiv:1711.01682*, 2017.

Opper, M. and Winther, O. Adaptive and self-averaging thouless-anderson-palmer mean-field theory for probabilistic modeling. *Physical Review E*, 64(5):056131, 2001.

Pati, D., Bhattacharya, A., and Yang, Y. On statistical optimality of variational bayes. *arXiv preprint arXiv:1712.08983*, 2017.

Perko, L. *Differential equations and dynamical systems*, volume 7. Springer Science & Business Media, 2013.

Raj, A., Stephens, M., and Pritchard, J. K. Variational inference of population structure in large snp datasets. *Genetics*, pp. genetics–114, 2014.

Recht, B., Re, C., Tropp, J., and Bittorf, V. Factoring nonnegative matrices with linear programs. In *Advances in Neural Information Processing Systems*, pp. 1214–1222, 2012.

Thouless, D. J., Anderson, P. W., and Palmer, R. G. Solution of'solvable model of a spin glass'. *Philosophical Magazine*, 35(3):593–601, 1977.

Wainwright, M. J., Jordan, M. I., et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Wang, B. and Titterington, D. Convergence and asymptotic normality of variational bayesian approximations for exponential family models with missing values. In *ARTIFICIAL INTELLIGENCE*. Citeseer, 2004.

Wang, B. and Titterington, D. Inadequacy of interval estimates corresponding to variational bayesian approximations. In *AISTATS*. Barbados, 2005.

Wang, B., Titterington, D., et al. Convergence properties of a general algorithm for calculating variational bayesian estimates for a normal mixture model. *Bayesian Analysis*, 1(3):625–650, 2006.

Wang, C. and Blei, D. M. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 448–456. ACM, 2011.

Wang, X. and Grimson, E. Spatial latent dirichlet allocation. In *Advances in neural information processing systems*, pp. 1577–1584, 2008.

Zhang, A. Y. and Zhou, H. H. Theoretical and computational guarantees of mean field variational inference for community detection. *arXiv preprint arXiv:1710.11268*, 2017.

Zhou, J., Bhattacharya, A., Herring, A. H., and Dunson, D. B. Bayesian factorizations of big sparse tensors. *Journal of the American Statistical Association*, 110(512): 1562–1576, 2015.