The Annals of Statistics
2023, Vol. 51, No. 2, 519–546
https://doi.org/10.1214/23-AOS2257

LOCAL CONVEXITY OF THE TAP FREE ENERGY AND AMP CONVERGENCE FOR $\mathbb{Z}_2$-SYNCHRONIZATION

BY MICHAEL CELENTANO[1,a], ZHOU FAN[2,c] AND SONG MEI[1,b]

[1]*Department of Statistics, University of California, Berkeley,* [a]*mcelentano@berkeley.edu,* [b]*songmei@berkeley.edu*
[2]*Department of Statistics and Data Science, Yale University,* [c]*zhou.fan@yale.edu*

We study mean-field variational Bayesian inference using the TAP approach, for $\mathbb{Z}_2$-synchronization as a prototypical example of a high-dimensional Bayesian model. We show that for any signal strength $\lambda > 1$ (the weak-recovery threshold), there exists a unique local minimizer of the TAP free energy functional near the mean of the Bayes posterior law. Furthermore, the TAP free energy in a local neighborhood of this minimizer is strongly convex. Consequently, a natural-gradient/mirror-descent algorithm achieves linear convergence to this minimizer from a local initialization, which may be obtained by a constant number of iterations of Approximate Message Passing (AMP). This provides a rigorous foundation for variational inference in high dimensions via minimization of the TAP free energy.

We also analyze the finite-sample convergence of AMP, showing that AMP is asymptotically stable at the TAP minimizer for any $\lambda > 1$, and is linearly convergent to this minimizer from a spectral initialization for sufficiently large λ. Such a guarantee is stronger than results obtainable by state evolution analyses, which only describe a fixed number of AMP iterations in the infinite-sample limit.

Our proofs combine the Kac–Rice formula and Sudakov–Fernique Gaussian comparison inequality to analyze the complexity of critical points that satisfy strong convexity and stability conditions within their local neighborhoods.

1. Introduction. Variational inference is an increasingly popular method for performing approximate Bayesian inference, and is widely used in applications ranging from document classification to population genetics [25, 30, 79, 100]. For large-scale problems, variational methods provide an appealing alternative to Markov Chain Monte Carlo procedures, particularly in settings where MCMC may be computationally prohibitive to apply. We refer readers to the classical expositions [71, 121] and the recent review [24] for an introduction.

In "mean-field" models where the posterior distribution $p(\boldsymbol{x}|\boldsymbol{Y})$ of parameters $\boldsymbol{x}$ given data $\boldsymbol{Y}$ may be close to being a product measure, a common approach to variational inference is to approximate $p(\boldsymbol{x}|\boldsymbol{Y})$ by a product law. The most widely used such approximation minimizes the KL-divergence to $p(\boldsymbol{x}|\boldsymbol{Y})$ over the class $\mathcal{Q}$ of product measures,

$$\hat{q}(\boldsymbol{x}) = \arg\min_{q \in \mathcal{Q}} \mathsf{D}_{\mathrm{KL}}\big(q(\boldsymbol{x}) \,\|\, p(\boldsymbol{x}|\boldsymbol{Y})\big). \tag{1}$$

When $\boldsymbol{x} \in \mathbb{R}^n$ is high-dimensional, a problematic phenomenon may occur in which this distribution $\hat{q}(\boldsymbol{x})$ provides inconsistent approximations to the posterior marginals and posterior means, even in models where all low-dimensional marginals of $p(\boldsymbol{x}|\boldsymbol{Y})$ have approximately independent coordinates. Such a phenomenon was first investigated by Thouless, Anderson, and Palmer for the Sherrington–Kirkpatrick (SK) model of spin glasses, where a simple

Received January 2022; revised November 2022.
MSC2020 subject classifications. 62C10.
Key words and phrases. Variational inference, $\mathbb{Z}_2$ synchronization, TAP free energy, approximate message passing, natural gradient descent, nonconvex optimization, landscape analysis.

method of addressing this inaccuracy—now often called the "TAP correction"—was also proposed [118]. Manifestations of this phenomenon and analogues of the TAP free energy for several high-dimensional statistical models have been studied in [53, 59, 74, 99, 103], and we provide further discussion in Section 1.3.

The TAP approach to variational inference constructs a free energy functional $\mathcal{F}_{\mathrm{TAP}}$ by adding a correction term to the KL-divergence objective (1). This TAP correction accounts for dependences between pairs of coordinates of $\boldsymbol{x}$ in their posterior law, which are individually weak but may have a nonnegligible aggregate effect in high dimensions. Variational inference is performed by minimizing $\mathcal{F}_{\mathrm{TAP}}$, or by solving the TAP stationary equations

$$0 = \nabla \mathcal{F}_{\mathrm{TAP}}. \tag{2}$$

Since the pioneering work of [26, 47, 72], both the theory and implementation of TAP-variational inference have been closely connected to Approximate Message Passing (AMP) algorithms, which provide specific iterative procedures for solving (2). TAP-variational inference has been successfully applied via AMP to a variety of high-dimensional statistical problems. We highlight in particular the line of work [10, 45, 76, 77, 88, 91, 102] on low-rank matrix estimation, of which the $\mathbb{Z}_2$-synchronization problem is a specific example.

The goal of our current paper is to address several foundational questions regarding TAP-variational inference that, despite the above successes, remain poorly understood. First, the convergence of AMP is usually known only in a weak sense, guaranteeing $\|\sqrt{n} \cdot \nabla \mathcal{F}_{\mathrm{TAP}}\|_2^2 < \varepsilon$ in the limit $n \to \infty$ for a constant number of AMP iterations $k \equiv k(\varepsilon)$ independent of n. Such a guarantee is too weak to ensure, for example, even the high-probability existence of a critical point of $\mathcal{F}_{\mathrm{TAP}}$ to which the AMP iterates converge. It does not establish whether the minimizer of $\mathcal{F}_{\mathrm{TAP}}$ is close to the true Bayes posterior mean, and indeed, these properties remain conjectural in most models to which the AMP/TAP approach has been applied. Second, regularity properties of the landscape of $\mathcal{F}_{\mathrm{TAP}}$ are largely unknown, making it unclear whether optimization algorithms other than AMP can successfully implement the TAP-variational inference paradigm.

In this paper, we clarify these properties of $\mathcal{F}_{\mathrm{TAP}}$ and the convergence of AMP and other descent algorithms for the specific model of $\mathbb{Z}_2$-synchronization. We build upon previous results and techniques of [53], which studied this model in a regime of large signal-to-noise. Our main results will show that for any signal strength above the weak-recovery threshold, there exists a unique local minimizer $\boldsymbol{m}_\star$ of $\mathcal{F}_{\mathrm{TAP}}$ near the Bayes posterior mean, and $\mathcal{F}_{\mathrm{TAP}}$ is strongly convex in a local neighborhood (of nontrivial size) around $\boldsymbol{m}_\star$. Consequently, a generic natural-gradient-descent (NGD) algorithm exhibits linear convergence to $\boldsymbol{m}_\star$ from a local initialization, which may be obtained by a finite number of iterations of AMP. We also show that the Jacobian of the AMP map is stable at $\boldsymbol{m}_\star$, so that AMP initialized in a (potentially very) small neighborhood of $\boldsymbol{m}_\star$ will also converge for fixed n as the number of iterations $t \to \infty$. In the large signal-to-noise regime of [53], we show that both NGD and AMP exhibit linear convergence to $\boldsymbol{m}_\star$ from a spectral initialization.

Formalizing these properties of $\mathcal{F}_{\mathrm{TAP}}$ and the convergence of generic optimization algorithms has several appeals over the existing theory around AMP. First, it clarifies a concrete objective function for high-dimensional variational inference, which can serve a number of practical purposes such as assessing algorithm convergence. Second, the convergence and state evolution of AMP are tied to probabilistic aspects of the model, whereas NGD is always a strict descent algorithm (for small enough step size, even in misspecified models) and may provide a more flexible and robust approach for optimization in practice. Finally, understanding the landscape of $\mathcal{F}_{\mathrm{TAP}}$ may be useful in other contexts. For example, following the initial posting of our work, [32, 51] have used the local strong convexity of $\mathcal{F}_{\mathrm{TAP}}$ in the related

SK model to argue that its stationary point is Lipschitz in the external field. This is a central technical ingredient in these works to show the correctness of an algorithmic stochastic localization procedure for sampling from the SK measure.

We review relevant background on the $\mathbb{Z}_2$-synchronization model in Section 1.1, and we describe our results in more detail in Section 1.2.

1.1. *$\mathbb{Z}_2$-synchronization and the TAP free energy.* In $\mathbb{Z}_2$-synchronization, we wish to estimate an unknown binary vector $\boldsymbol{x} \in \{-1, +1\}^n$ having the entry-wise symmetric Bernoulli prior $x_i \overset{\text{i.i.d.}}{\sim} \text{Unif}\{-1, +1\}$. For a signal-to-noise parameter $\lambda > 0$, we observe

$$\boldsymbol{Y} = \frac{\lambda}{n}\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} + \boldsymbol{W} \quad \text{where } \boldsymbol{W} \sim \text{GOE}(n). \tag{3}$$

Thus $\boldsymbol{W}$ is symmetric Gaussian noise, having entries $(w_{ii} : i = 1, \dots, n) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 2/n)$ independent of $(w_{ij} : 1 \le i < j \le n) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/n)$. Equivalently, $\boldsymbol{W} = (\boldsymbol{Z} + \boldsymbol{Z}^{\top})/\sqrt{2n}$ where $(z_{ij} : i, j = 1, \dots, n) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$.

The parameter $\boldsymbol{x}$ is identifiable only up to ± sign, and the posterior law $p(\boldsymbol{x}|\boldsymbol{Y})$ has the corresponding sign symmetry $p(\boldsymbol{x}|\boldsymbol{Y}) = p(-\boldsymbol{x}|\boldsymbol{Y})$. Thus, we will consider estimation of the sign-invariant rank-one matrix $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^{\top}$. The Bayes posterior-mean estimate of this matrix is

$$\widehat{\boldsymbol{X}}_{\text{Bayes}} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} \mid \boldsymbol{Y}]. \tag{4}$$

The asymptotic squared-error Bayes risk of this estimator was characterized by Deshpande, Abbe, and Montanari in [44]:

$$\lim_{n\to\infty} \frac{1}{n^2}\mathbb{E}[\|\widehat{\boldsymbol{X}}_{\text{Bayes}} - \boldsymbol{x}\boldsymbol{x}^{\top}\|_{\mathsf{F}}^2] = \begin{cases} 1 - q_*(\lambda)^2 & \text{if } \lambda > 1, \\ 1 & \text{if } \lambda \le 1, \end{cases} \tag{5}$$

where $q_\star(\lambda) > 0$ is the solution to a fixed-point equation (11). Thus for $\lambda < 1$, no nontrivial estimation is possible in the large-n limit, as the optimal Bayes risk coincides with that of the trivial estimator $\widehat{\boldsymbol{X}} = \boldsymbol{0}$. In contrast, for $\lambda > 1$, the Bayes estimator achieves positive entry-wise correlation with $\boldsymbol{x}\boldsymbol{x}^{\top}$.

[44] studied also an AMP algorithm for approximately computing $\widehat{\boldsymbol{X}}_{\text{Bayes}}$. Starting from initializations $\boldsymbol{h}^0, \boldsymbol{m}^{-1} \in \mathbb{R}^n$, this algorithm takes the form

$$\begin{aligned} \boldsymbol{m}^k &= \tanh(\boldsymbol{h}^k), \\ \boldsymbol{h}^{k+1} &= \lambda\boldsymbol{Y}\boldsymbol{m}^k - \lambda^2[1 - Q(\boldsymbol{m}^k)]\boldsymbol{m}^{k-1}, \end{aligned} \tag{AMP}$$

where $Q(\boldsymbol{m}) = \|\boldsymbol{m}\|_2^2/n$. The analyses of [44] imply that for any $\lambda > 1$ and $\varepsilon > 0$, starting from an informative initialization $\boldsymbol{h}^0$, there exists an iterate $k \equiv k(\lambda, \varepsilon)$ of AMP for which $\|\boldsymbol{m}^k(\boldsymbol{m}^k)^{\top} - \widehat{\boldsymbol{X}}_{\text{Bayes}}\|_{\mathsf{F}}^2/n^2 < \varepsilon$, with high probability for all large n. More recent results of [91] imply that such a guarantee holds also for AMP with a spectral initialization.

The TAP free energy in this $\mathbb{Z}_2$-synchronization model is defined for $\boldsymbol{m} \in (-1, 1)^n$ by

$$\mathcal{F}_{\text{TAP}}(\boldsymbol{m}) = -\frac{\lambda}{2n}\langle \boldsymbol{m}, \boldsymbol{Y}\boldsymbol{m}\rangle - \frac{1}{n}\sum_{i=1}^{n} \mathsf{h}(m_i) - \frac{\lambda^2}{4}[1 - Q(\boldsymbol{m})]^2, \tag{TAP}$$

where $Q(\boldsymbol{m}) = \|\boldsymbol{m}\|_2^2/n$ as above, and $\mathsf{h}(m)$ is the binary entropy function

$$\mathsf{h}(m) = -\frac{1+m}{2}\log\frac{1+m}{2} - \frac{1-m}{2}\log\frac{1-m}{2}. \tag{6}$$

This function $\mathcal{F}_{\mathrm{TAP}}$ has the sign symmetry $\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) = \mathcal{F}_{\mathrm{TAP}}(-\boldsymbol{m})$, corresponding to the above sign symmetry of the posterior law. The first two terms of (TAP) coincide[1] with the KL-divergence $\mathsf{D}_{\mathrm{KL}}(q(\boldsymbol{x}) \parallel p(\boldsymbol{x}|\boldsymbol{Y}))$ for a product measure $q(\boldsymbol{x})$ on $\{-1,+1\}^n$, upon parameterizing q by its mean $\boldsymbol{m} = \mathbb{E}_{\boldsymbol{x}\sim q}[\boldsymbol{x}] \in (-1,1)^n$. The third term of (TAP) is the TAP correction. Applying $\mathsf{h}'(m) = -\operatorname{arctanh}(m)$, the stationary condition $0 = \nabla\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m})$ may be rearranged as the TAP mean-field equations

$$\boldsymbol{m} = \tanh(\lambda\boldsymbol{Y}\boldsymbol{m} - \lambda^2[1 - Q(\boldsymbol{m})]\boldsymbol{m}),$$

and the AMP algorithm (AMP) is an iterative scheme for computing a fixed point of these equations.

In [53], an upper bound for the expected number of critical points of $\mathcal{F}_{\mathrm{TAP}}$ in sub-regions of the domain $(-1,1)^n$ was derived for any $\lambda > 0$. Using this result, for $\lambda > \lambda_0$ a large enough absolute constant, it was shown that the global minimizer $\boldsymbol{m}_\star$ of $\mathcal{F}_{\mathrm{TAP}}$ satisfies $\mathbb{E}[\|\boldsymbol{m}_\star\boldsymbol{m}_\star^\top - \widehat{\boldsymbol{X}}_{\mathrm{Bayes}}\|_{\mathsf{F}}^2]/n^2 \to 0$, and that this holds more generally for any critical point $\boldsymbol{m}$ of $\mathcal{F}_{\mathrm{TAP}}$ in the domain

$$\mathcal{S} = \{\boldsymbol{m} \in (-1,1)^n : \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) < -\lambda^2/3\}.$$

As a consequence, it was also shown that $\mathbb{E}[\|\boldsymbol{m}_\star\boldsymbol{m}_\star^\top - \widehat{\boldsymbol{X}}_{\mathrm{Bayes}}\|_{\mathsf{F}}^2]/n^2$ must be bounded away from 0 for the minimizer $\boldsymbol{m}_\star$ of the naive mean-field objective (1) parametrized similarly by $\boldsymbol{m}$. We note that the landscape guarantees in [53] do not extend to the entire weak-recovery regime $\lambda > 1$. The analyses for large $\lambda > \lambda_0$ also fall short of showing uniqueness (up to sign) of the TAP critical point $\boldsymbol{m}_\star$ in $\mathcal{S}$, and of establishing polynomial-time convergence of AMP or other optimization algorithms for computing $\boldsymbol{m}_\star$.

1.2. *Contributions.*

Our current work establishes the following properties of $\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m})$ and of descent algorithms for minimizing this objective function.

1. *Existence of Bayes-optimal TAP local minimizer*. For any $\lambda > 1$, we show there exists a local minimizer $\boldsymbol{m}_\star$ of $\mathcal{F}_{\mathrm{TAP}}$ such that $\|\boldsymbol{m}_\star\boldsymbol{m}_\star^\top - \widehat{\boldsymbol{X}}_{\mathrm{Bayes}}\|_{\mathsf{F}}^2/n^2 \to 0$ in probability. This strengthens the guarantee of [53] that was shown for large $\lambda > \lambda_0$. Subject to the validity of a numerical conjecture about a deterministic low-dimensional variational problem (see Remark 4.5), our results imply that this is also the global minimizer of $\mathcal{F}_{\mathrm{TAP}}$ for any $\lambda > 1$.
2. *Local strong convexity of the TAP free energy*. For any $\lambda > 1$, we show that $\mathcal{F}_{\mathrm{TAP}}$ is strongly convex in a $\sqrt{\varepsilon n}$-neighborhood of this local minimizer $\boldsymbol{m}_\star$. Hence, this local minimizer is the unique critical point satisfying $\|\boldsymbol{m}_\star\boldsymbol{m}_\star^\top - \widehat{\boldsymbol{X}}_{\mathrm{Bayes}}\|_{\mathsf{F}}^2/n^2 < \iota(\varepsilon)$, for some constant $\iota(\varepsilon) > 0$.
3. *Local convergence of natural gradient descent*. We introduce a natural gradient descent (NGD) algorithm for minimizing $\mathcal{F}_{\mathrm{TAP}}$, which is equivalently a mirror descent procedure that adapts to the curvature of $\mathcal{F}_{\mathrm{TAP}}$ near the boundaries of $(-1,1)^n$. For any $\lambda > 1$, we prove that NGD achieves linear convergence to $\boldsymbol{m}_\star$ from an initialization within this $\sqrt{\varepsilon n}$-neighborhood. This initialization may be obtained by first performing a fixed number of iterations of AMP, thus yielding a polynomial-time algorithm for computing $\boldsymbol{m}_\star$.
4. *Stability of AMP*. For any $\lambda > 1$, we show that the AMP map is stable at $\boldsymbol{m}_\star$, in the sense of having a Jacobian with spectral radius strictly less than 1. Thus, AMP initialized in a sufficiently small neighborhood of $\boldsymbol{m}_\star$ will also linearly converge to $\boldsymbol{m}_\star$.
5. *Finite-n convergence of AMP and NGD*. Finally, for $\lambda > \lambda_0$ a large enough absolute constant, our results combine with those of [53] to show that $\boldsymbol{m}_\star$ is the global minimizer and unique critical point (up to sign) of $\mathcal{F}_{\mathrm{TAP}}$ in the domain $\{\boldsymbol{m} : \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) < -\lambda^2/3\}$. In this

[1] Up to an additive constant, and a replacement of $\mathbb{E}_{\boldsymbol{x}\sim q}[\langle\boldsymbol{x},\boldsymbol{Y}\boldsymbol{x}\rangle]$ by $\langle\boldsymbol{m},\boldsymbol{Y}\boldsymbol{m}\rangle$ which incurs negligible error

signal-to-noise regime, we prove that both AMP and NGD alone exhibit linear convergence to $\boldsymbol{m}_\star$ from a spectral initialization.

We emphasize that this convergence of AMP is established in the sense $\lim_{k\to\infty} \boldsymbol{m}^k = \boldsymbol{m}_\star$ for fixed dimension n, which is stronger than the guarantee $\limsup_{n\to\infty} \|\boldsymbol{m}^k - \boldsymbol{m}_\star\|_2^2/n < \varepsilon$ for fixed $k \equiv k(\lambda, \varepsilon)$ that is obtainable by standard analyses of the AMP state evolution.

The main challenge in understanding the landscape of $\mathcal{F}_{\text{TAP}}$ locally near $\boldsymbol{m}_\star$ is that—for any constant signal strength λ—this point $\boldsymbol{m}_\star$ does not converge to the true signal vector $\boldsymbol{x} \in \{-1, +1\}^n$ as $n \to \infty$, but rather remains random in $(-1, 1)^n$. Thus, it is not enough to study the landscape of $\mathcal{F}_{\text{TAP}}$ in a vanishing neighborhood of $\boldsymbol{x}$ using, for example, the uniform convergence arguments [82, 116]. The above results instead pertain to the geometry of $\mathcal{F}_{\text{TAP}}$ in a random region of the cube $(-1, 1)^n$.

We will prove these results using a combination of the Kac–Rice formula and Gaussian comparison inequalities. We provide a detailed overview of this proof in Section 4. The Kac–Rice formula has been successfully applied to study the complexity of critical points for various nonconvex function landscapes. However, to our knowledge, our argument for using Kac–Rice to study also the local geometry around a particular critical point is novel. We believe that this technique may be of independent interest for some recent analyses of related disordered systems [27, 46, 54], where conditioning on a sequence of AMP iterates was used as a surrogate for conditioning on an actual TAP critical point.

1.3. *Further related literature.*

1.3.1. *Variational inference.* The terminology "variational inference" encompasses a large family of methods for approximate Bayesian inference [23, 84, 95, 126], based upon approximating a variational representation to the evidence or marginal log-likelihood of the observed data. Variational inference has been incorporated into many software packages including Pyro [22], Infer.NET [83] and Edward [120].

There has been renewed interest in theoretical analyses of variational inference in recent years, focusing on a number of common desiderata: [21, 58, 62, 63, 123] study properties of consistency and asymptotic normality for estimates of low-dimensional parameters in latent variable models (i.e., of the prior "hyperparameters" in Bayesian contexts), using variational approximations for the marginal log-likelihood. In particular, [21, 58] establish such guarantees for the mean-field variational approximation in stochastic block models (SBMs), which are closely related to the $\mathbb{Z}_2$-synchronization model of our work. [92, 98, 127] study the optimization landscape and convergence properties of iterative coordinate ascent (CAVI) and block coordinate ascent (BCAVI) algorithms, with [127] showing that BCAVI achieves an optimal exponentially-vanishing rate of estimation error for the latent community membership vector in SBMs with asymptotically diverging signal strength. [1, 40, 105, 125, 128] study rates of posterior contraction for both variational Bayes and α-fractional variational Bayes methods, establishing conditions under which the variational posteriors may enjoy the same optimal rates of contraction in a frequentist Bernstein–von-Mises sense as the true Bayes posteriors. In particular, [1, 40, 125] discuss applications of these results to low-rank matrix estimation problems, including matrix completion, probabilistic PCA, and topic models.

In our work, we study the $\mathbb{Z}_2$-synchronization model with bounded signal strength, which is in a different asymptotic regime from the above posterior contraction results for SBMs and low-rank matrix estimation. Fixing the true parameter $\boldsymbol{x}$ as the all-1s vector, the Bayes estimate for $\boldsymbol{x}$ in our setting has a marginal distribution of coordinates that converges to a nondegenerate limit law, and an asymptotically nonvanishing per-coordinate Bayes risk.

Our focus on such a setting is motivated in part by our belief that in many applications, Bayesian approaches to inference may be favored because the data is in a regime of limited signal-to-noise that is far from theoretical regimes of posterior contraction. Instead, information in the hypothesized prior is important in informing inference, and the desideratum is then to obtain an accurate estimate of the posterior distribution under this prior. Our results are oriented towards this goal, showing (in a simple but illustrative model) that minimizing the TAP free energy yields a variational approximation which consistently estimates the posterior marginals, even when the posterior distribution itself does not concentrate strongly around the true parameter.

1.3.2. *TAP free energy and the naive mean-field approximation.* Thouless, Anderson and Palmer introduced in [118] the TAP equations (and the associated TAP free energy) as a system of asymptotically exact mean-field equations in the SK model. For spin glasses, the validity of the TAP equations and their relation to the Gibbs measure have been extensively studied—see, for example, [29, 31, 43, 97] in the physics literature, and [6, 18, 26, 35, 38, 39, 113, 117] for rigorous mathematical results. Direct optimization of an analogous TAP free energy (a.k.a. approximate Bethe free energy) was proposed for Bayesian linear and generalized linear models in [74, 103], which recognized that its critical points are in exact correspondence with fixed points of AMP. $\mathbb{Z}_2$-synchronization corresponds to the SK model with an added ferromagnetic bias, and the form of the TAP free energy that we study is identical to the (high-temperature) TAP free energy in the SK model with this added ferromagnetic component.

We emphasize that both the TAP approach and the "naive" mean-field approach of (1) have received significant attention in the theoretical literature. A line of work [7, 11, 36, 52, 67, 124] on the theory of nonlinear large deviations establishes that the naive mean-field approximation to the free energy (i.e., the marginal log-likelihood in Bayesian models) is asymptotically accurate to leading order, without the need for a TAP correction, under a condition that the log-density has a "low-complexity gradient." In Ising models with couplings matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times n}$ having $O(1)$ operator norm, such a condition holds when $\boldsymbol{Y}$ is nearly low-rank in the sense $\|\boldsymbol{Y}\|_F^2 = o(n)$ [11]. It does not hold for $\mathbb{Z}_2$-synchronization with any fixed signal strength λ, where [53, 59] contrasted variational inference based on the TAP and naive mean-field approximations. In particular, [59] showed that for $\lambda \in (1/2, 1)$, naive mean-field variational Bayes may yield a "falsely informative" variational posterior, and [53] showed that critical points of the naive mean-field free energy cannot correspond to consistent approximations of the posterior mean for any sufficiently large but fixed value of λ.

1.3.3. *Spiked matrix models and $\mathbb{Z}_2$-synchronization.* Spiked matrix models have been a mainstay in the statistical literature since their introduction by [70]. $\mathbb{Z}_2$-synchronization is a specific example of the spiked model with Bernoulli prior, and also of more general synchronization problems over compact groups [9, 109]. The Bayes risks in $\mathbb{Z}_2$-synchronization and other spiked matrix models were studied in [10, 44, 75, 76]. For $\mathbb{Z}_2$-synchronization, nontrivial signal estimation above the weak-recovery threshold $\lambda = 1$ can also be achieved by spectral methods [8, 96] and semidefinite programming [69, 90], although such methods do not achieve the asymptotically optimal Bayes risk (5).

$\mathbb{Z}_2$-synchronization has been studied in part as a simpler analogue of the symmetric two-component SBM that replaces the noise $\mathbf{A} - \mathbb{E}[\mathbf{A}]$ of the adjacency matrix $\mathbf{A}$ by Gaussian noise, and it is possible to make formal connections between estimation in these models via universality arguments [44, 90]. We believe that certain aspects of our analyses and results may also be extendable to the SBM via universality arguments developed for AMP in [15, 37, 50, 122] and for minimizers of optimization objective functions with random data in [64, 66, 87, 89], and this would be interesting to explore in future work.

1.3.4. *AMP algorithms.* AMP algorithms were proposed and studied in [47, 72] for Bayesian linear regression and compressed sensing. They may be derived by approximating belief propagation on dense graphical models; see, for example, [48, 86]. Various generalizations of AMP have been developed, including the Generalized AMP algorithm of [101] and the Vector AMP algorithm of [104], and we refer to [55] for a recent review. The state evolution formalism of AMP was introduced in [47] and rigorously established in [16, 26]. This has since been generalized in [20, 68, 91]. A finite-n analysis of AMP was performed in [107], which extended the validity of the state evolution to $o(\log n/\log\log n)$ iterations. Following the initial posting of our work, [78] established a different finite-n guarantee for AMP via a novel decomposition of the AMP iterates, which applies for $o(n/(\log^7 n))$ iterations in the $\mathbb{Z}_2$-synchronization problem with signal strength $\lambda \in (1, 1.2)$.

1.3.5. *Gaussian comparison inequalities.* The proofs of our main results rely heavily on Slepian's comparison inequality [110] and its later development by Sudakov–Fernique [56, 114, 115], to reduce the study of $\mathcal{F}_{\text{TAP}}$ to a simpler Gaussian process. This approach is related to a recent line of work that generalizes Gordon's inequality [60, 73] to a Convex Gaussian Minimax Theorem (CGMT) [34, 85, 94, 111, 119].

1.3.6. *Kac–Rice formula and complexity analysis.* Physics calculations of the complexity of critical points in spin glass models using the Kac–Rice formalism can be found in [28, 31, 41, 42, 57]. This method was made rigorous for spherical spin glasses in [4, 5, 112], and a more recent line of work [3, 12, 13, 19, 53, 81] has used this approach to analyze nonconvex function landscapes in other high-dimensional probabilistic and statistical models.

2. Main results.

2.1. *Local analysis of the TAP free energy.* Our first result shows the existence and uniqueness of a local minimizer of the TAP free energy $\mathcal{F}_{\text{TAP}}$ near the Bayes estimator (cf. equation (4)), for any signal strength $\lambda > 1$. We also establish strong convexity of $\mathcal{F}_{\text{TAP}}$ in a $\sqrt{\varepsilon n}$-neighborhood around this minimizer, as well as the stability of the AMP map

$$T_{\mathsf{AMP}}(\boldsymbol{m}, \boldsymbol{m}_-) = \big(\tanh(\lambda \boldsymbol{Y}\boldsymbol{m} - \lambda^2[1 - Q(\boldsymbol{m})]\boldsymbol{m}_-), \boldsymbol{m}\big) \tag{7}$$

at this local minimizer. This is the map for which the AMP iterations (AMP) may be expressed as $(\boldsymbol{m}^{k+1}, \boldsymbol{m}^k) = T_{\mathsf{AMP}}(\boldsymbol{m}^k, \boldsymbol{m}^{k-1})$.

THEOREM 2.1 (Local convexity and AMP stability). *Fix any $\lambda > 1$. There exist λ-dependent constants $\varepsilon, t > 0$ and $r \in (0, 1)$ such that for any fixed $\iota > 0$, with probability approaching 1 as $n \to \infty$, the following all occur.*

(a) (*Bayes-optimal TAP local minimizer*) *Let $\widehat{\boldsymbol{X}}_{\text{Bayes}} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top \mid \boldsymbol{Y}]$. There exists a critical point and local minimizer $\boldsymbol{m}_\star$ of $\mathcal{F}_{\text{TAP}}(\boldsymbol{m})$ such that*

$$\frac{1}{n^2}\|\boldsymbol{m}_\star \boldsymbol{m}_\star^\top - \widehat{\boldsymbol{X}}_{\text{Bayes}}\|_{\mathsf{F}}^2 < \iota. \tag{8}$$

For sufficiently small $\iota > 0$ (which is λ-dependent and n-independent), this is the unique critical point satisfying (8) up to ± sign.

(b) (*Local strong convexity of TAP free energy*) *Let $\lambda_{\min}(\cdot)$ denote the smallest eigenvalue. For this local minimizer $\boldsymbol{m}_\star$, we have*

$$\lambda_{\min}\big(n \cdot \nabla^2 \mathcal{F}_{\text{TAP}}(\boldsymbol{m})\big) > t > 0 \quad \textit{for all } \boldsymbol{m} \in (-1, 1)^n \cap \mathsf{B}_{\sqrt{\varepsilon n}}(\boldsymbol{m}_\star).$$

In particular, $\mathcal{F}_{\text{TAP}}$ is strongly convex over $(-1, 1)^n \cap \mathsf{B}_{\sqrt{\varepsilon n}}(\boldsymbol{m}_\star)$.

(c) (*Local stability of AMP*) *Let* $\mathrm{d}T_{\mathsf{AMP}} \in \mathbb{R}^{2n\times 2n}$ *be the Jacobian of the AMP map* (7), *and let* $\rho(\cdot)$ *denote the spectral radius. For this local minimizer* $\boldsymbol{m}_\star$, *we have*

$$\rho\big(\mathrm{d}T_{\mathsf{AMP}}(\boldsymbol{m}_\star, \boldsymbol{m}_\star)\big) < r < 1.$$

Combining with the global landscape analysis of [53], this implies the following immediate corollary for large enough signal strength λ.

COROLLARY 2.2 (Global landscape for large λ). *For an absolute constant* $\lambda_0 > 0$, *suppose* $\lambda > \lambda_0$. *Then with probability approaching* 1 *as* $n \to \infty$, *the local minimizers* $\pm\boldsymbol{m}_\star$ *guaranteed by Theorem* 2.1 *are the global minimizers of* $\mathcal{F}_{\mathrm{TAP}}$. *Furthermore, they are the only critical points of* $\mathcal{F}_{\mathrm{TAP}}$ *in the domain*

$$\mathcal{S} = \big\{\boldsymbol{m} \in (-1,1)^n : \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) < -\lambda^2/3\big\}.$$

A proof sketch of Theorem 2.1 can be found in Section 4, and its detailed proof can be found in Appendix B (see the Supplementary Material [33]). The proof of Corollary 2.2 can be found in Appendix C.1.

2.2. *Convergence of algorithms.* We study convergence of the AMP algorithm (AMP), with the spectral initialization

$$\text{(SI)} \qquad \boldsymbol{h}^0 = \text{principal eigenvector of } \boldsymbol{Y} \text{ with } \|\boldsymbol{h}^0\|_2 = \sqrt{n\lambda^2(\lambda^2-1)}, \qquad \boldsymbol{m}^{-1} = \lambda\boldsymbol{h}^0.$$

We choose this scaling for $\boldsymbol{h}^0$ as in [91], Section 2.4, to simplify the AMP state evolution.

We introduce also the following more "generic" first-order natural gradient descent (NGD) algorithm, with a step size parameter $\eta > 0$:

$$\text{(NGD)} \qquad \begin{aligned} \boldsymbol{m}^k &= \tanh(\boldsymbol{h}^k), \\ \boldsymbol{h}^{k+1} &= \boldsymbol{h}^k - \eta n \cdot \nabla\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}^k) \\ &= (1-\eta)\boldsymbol{h}^k + \eta\big(\lambda\boldsymbol{Y}\boldsymbol{m}^k - \lambda^2\big[1 - Q(\boldsymbol{m}^k)\big]\boldsymbol{m}^k\big). \end{aligned}$$

We call this algorithm "natural gradient descent" because we may apply $(\mathrm{d}/\mathrm{d}h)\tanh(h) = 1 - \tanh(h)^2$ to write the $\boldsymbol{m}$-gradient $\nabla\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}^k)$ equivalently as a preconditioned $\boldsymbol{h}$-gradient,

$$\nabla\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}^k) = \boldsymbol{I}(\boldsymbol{m}^k)^{-1} \cdot \nabla_{\boldsymbol{h}}\mathcal{F}_{\mathrm{TAP}}\big(\tanh(\boldsymbol{h}^k)\big), \qquad \boldsymbol{I}(\boldsymbol{m}) = \mathrm{diag}\bigg(\frac{1}{1-\boldsymbol{m}^2}\bigg),$$

where $\boldsymbol{I}(\boldsymbol{m})$ is proportional to the Fisher information matrix in a model of n independent Bernoulli $\{-1,+1\}$ variables with mean $\boldsymbol{m} \in \mathbb{R}^n$. This identifies (NGD) as a natural gradient method [2]. We note that setting the step size $\eta = 1$ yields an algorithm similar to (AMP), but with $\boldsymbol{m}^{k-1}$ replaced by $\boldsymbol{m}^k$. For simplicity, we will consider the same spectral initialization $\boldsymbol{h}^0$ for this algorithm as for AMP in (SI), although here this specific choice of initialization is less important.

Alternatively, the iterations (NGD) may be understood as a mirror-descent/Bregman-gradient method in the $\boldsymbol{m}$-parameterization [17, 93]. Recalling the binary entropy function h from (6), we define

$$\text{(9)} \qquad \begin{aligned} L &= \frac{1}{\eta}, \qquad H(\boldsymbol{m}) = \frac{1}{n}\sum_{i=1}^n \mathsf{h}(m_i), \\ D_{-H}(\boldsymbol{m}, \boldsymbol{m}') &= -H(\boldsymbol{m}) + H(\boldsymbol{m}') + \big\langle\nabla H(\boldsymbol{m}'), \boldsymbol{m} - \boldsymbol{m}'\big\rangle, \end{aligned}$$

where L is the inverse step size, $-H(\boldsymbol{m})$ is a separable convex prox function, and $D_{-H}(\boldsymbol{m}, \boldsymbol{m}')$ is its associated Bregman divergence. Then it may be checked that (NGD) takes the equivalent mirror-descent form

$$\boldsymbol{m}^{k+1} = \arg \min_{\boldsymbol{m} \in (-1,1)^n} \mathcal{F}_{\text{TAP}}(\boldsymbol{m}^k) + \langle \nabla \mathcal{F}_{\text{TAP}}(\boldsymbol{m}^k), \boldsymbol{m} - \boldsymbol{m}^k \rangle + L \cdot D_{-H}(\boldsymbol{m}, \boldsymbol{m}^k). \tag{10}$$

One motivation for studying this algorithm, rather than ordinary gradient descent in the $\boldsymbol{m}$-parameterization, is that the Hessian $\nabla^2 \mathcal{F}_{\text{TAP}}(\boldsymbol{m})$ is not uniformly bounded over $(-1, 1)^n$, and instead diverges as $\boldsymbol{m}$ approaches the boundaries of the cube. The form (10) naturally adapts to this nonuniform curvature of $\mathcal{F}_{\text{TAP}}$, allowing for a convergence analysis using techniques of [14, 80] for minimizing functions that are not strongly smooth in the Euclidean metric.

Combining the local strong convexity of Theorem 2.1, the state evolution of spectrally-initialized AMP, and this type of convergence analysis for NGD, we deduce the following result, whose proof can be found in Section 5.1 and Appendix C.

THEOREM 2.3 (Computation of Bayes-optimal TAP minimizer). *Fix any $\lambda > 1$. There exist λ-dependent constants $C, \mu, \eta_0 > 0$ and $T \geq 1$ such that with probability approaching 1 as $n \to \infty$, the following occurs.*

Fix any step size $\eta \in (0, \eta_0)$, let $\boldsymbol{m}^T \in (-1, 1)^n$ be the Tth iteration of (AMP) from the spectral initialization (SI), and let $\boldsymbol{m}^{T+k} \in (-1, 1)^n$ be obtained by k iterations of (NGD) with step size η from the initialization $\boldsymbol{m}^T$. Let $\boldsymbol{m}_\star$ be the Bayes-optimal local minimizer of $\mathcal{F}_{\text{TAP}}$ in Theorem 2.1. Then for some choice of sign ± and every $k \geq 1$,

$$\mathcal{F}_{\text{TAP}}(\boldsymbol{m}^{T+k}) - \mathcal{F}_{\text{TAP}}(\pm \boldsymbol{m}_\star) < C(1 - \mu\eta)^k,$$
$$\|\boldsymbol{m}^{T+k} - (\pm \boldsymbol{m}_\star)\|_2 < C(1 - \mu\eta)^k \sqrt{n}.$$

In particular, $\lim_{k \to \infty} \boldsymbol{m}^{T+k} \in \{+\boldsymbol{m}_\star, -\boldsymbol{m}_\star\}$.

This theorem implies that for any fixed value of $\lambda > 1$, the Bayes-optimal local minimizer $\boldsymbol{m}_\star$ of $\mathcal{F}_{\text{TAP}}$ guaranteed by Theorem 2.1 may be computed in time that is polynomial in the problem size n (in the usual sense of linear convergence). Let us remark that the convergence analysis of NGD in this result is purely geometric, relying only on the smoothness and local convexity properties of $\mathcal{F}_{\text{TAP}}$. We hence expect that a similar convergence analysis may be performed for momentum-accelerated or stochastic variants of NGD, such as those developed recently in [49, 61, 65].

For sufficiently large signal strength λ, where the more global landscape of $\mathcal{F}_{\text{TAP}}$ is clarified by Corollary 2.2, our next result Theorem 2.4 verifies that the hybrid AMP/NGD approach in Theorem 2.3 is not needed, and that either algorithm alone can achieve linear convergence to the global TAP minimizer $\boldsymbol{m}_\star$ from a spectral initialization. The proof of Theorem 2.4 can be found in Sections 5.2 and 5.3, and Appendix C.

THEOREM 2.4 (Convergence of AMP and NGD for large λ). *For an absolute constant $\lambda_0 > 0$, suppose $\lambda > \lambda_0$ and let $\boldsymbol{m}_\star$ be the global minimizer of $\mathcal{F}_{\text{TAP}}$ in Corollary 2.2. Then there exist λ-dependent constants $C, \mu, \eta_0 > 0$ and $\alpha \in (0, 1)$ such that with probability approaching 1 as $n \to \infty$, the following all occur.*

(a) (*Convergence of AMP*) *Let $\boldsymbol{m}^k$ be the kth iterate of* (AMP) *from the spectral initialization* (SI). *For some choice of sign ± and every $k \geq 1$,*

$$\mathcal{F}_{\text{TAP}}(\boldsymbol{m}^k) - \mathcal{F}_{\text{TAP}}(\pm \boldsymbol{m}_\star) < C\alpha^k, \qquad \|\boldsymbol{m}^k - (\pm \boldsymbol{m}_\star)\|_2 < C\alpha^k \sqrt{n}.$$

(b) (*Convergence of NGD*) *Fix any step size* $\eta \in (0, \eta_0)$, *and let* $\boldsymbol{m}^k$ *be the kth iterate of* (NGD) *from the spectral initialization* (SI) *with step size* η. *For some choice of sign* $\pm$ *and every* $k \geq 1$,

$$\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}^k) - \mathcal{F}_{\mathrm{TAP}}(\pm\boldsymbol{m}_\star) < C(1-\mu\eta)^k, \qquad \|\boldsymbol{m}^k - (\pm\boldsymbol{m}_\star)\|_2 < C(1-\mu\eta)^k\sqrt{n}.$$

In particular, for both algorithms, $\lim_{k\to\infty} \boldsymbol{m}^k \in \{+\boldsymbol{m}_\star, -\boldsymbol{m}_\star\}$.

REMARK 2.5. We believe that the requirement $\lambda > \lambda_0$ sufficiently large in Theorem 2.4 is artificial, and that this result also holds for all $\lambda > 1$. This is supported by numerical simulations in Section 3 below. Let us clarify that such a guarantee for AMP does not follow from its state evolution combined with its local stability shown in Theorem 2.1(c): The state evolution ensures convergence to a $\sqrt{\varepsilon n}$-neighborhood of $\boldsymbol{m}_\star$, for any fixed $\varepsilon > 0$, in a finite number of AMP iterations. However, the local stability in Theorem 2.1(c) does not quantify the size of the neighborhood of $\boldsymbol{m}_\star$ in which AMP is then guaranteed to converge to $\boldsymbol{m}_\star$.

REMARK 2.6. Part of our analysis of Theorem 2.4(a) still uses the state evolution for AMP with spectral initialization developed in [91]. This result would hold equally if AMP is initialized with a vector $\boldsymbol{m}_1$ that is independent of the noise matrix $\boldsymbol{W}$ and has nonvanishing correlation with $\boldsymbol{m}_\star$, by the validity of the AMP state evolution also in this setting. For a random initialization that is uncorrelated with $\boldsymbol{m}_\star$, we note that an analysis of AMP seems challenging even in this setting of large but fixed $\lambda > \lambda_0$, as the algorithm would still require $O(\log(n))$ iterations to achieve a nonnegligible correlation with $\boldsymbol{m}_\star$, and existing finite-n analyses of AMP [78, 106] do not seem to immediately apply to describe this early phase of optimization. In Theorem 2.4(b), the spectral initialization is used to ensure that NGD is initialized in a basin of attraction of $\boldsymbol{m}_\star$, and analyses of the global landscape of $\mathcal{F}_{\mathrm{TAP}}$ in [53] are also insufficient to show that this basin of attraction includes random initializations.

3. Numerical simulations.

3.1. *Convergence of algorithms.* We perform numerical simulations to confirm the global convergence of AMP and NGD for all $\lambda > 1$, and to compare their convergence rates. We initialize both AMP and NGD using the spectral initialization (SI).

In Figure 1(a), we plot the residual squared error $\min\{\|\boldsymbol{m}^k - \boldsymbol{m}_\star\|_2^2/n, \|\boldsymbol{m}^k + \boldsymbol{m}_\star\|_2^2/n\}$, where $\boldsymbol{m}^k$ is the kth iterate of AMP or NGD with different step sizes, and $\boldsymbol{m}_\star = \arg\min_{\boldsymbol{m}} \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m})$. (We first compute $\boldsymbol{m}_\star$ up to high numerical accuracy using AMP.) For each algorithm, we simulated 10 random instances of $\boldsymbol{Y} \in \mathbb{R}^{n\times n}$ according to the $\mathbb{Z}_2$-synchronization model (3), with $n = 500$ and $\lambda = 1.5$. Figure 1(a) shows that AMP and NGD with step sizes 0.1 and 0.5 all consistently achieve convergence to $\boldsymbol{m}_\star$, where AMP has the fastest rate of convergence.

In Figure 1(b), we report the success probability of NGD for achieving convergence to $\boldsymbol{m}_\star$, for various step sizes η (horizontal axis) and signal-to-noise ratios $\lambda > 1$ (vertical axis). The success probability is defined as the fraction of the 10 random instances of $\boldsymbol{Y}$ for which NGD achieved residual squared error 10^{-4} within $k = 12{,}000$ iterations. Figure 1(b) suggests that NGD with step size $\eta < 0.4$ converges for any $\lambda > 1$, and illustrates that as λ increases, NGD allows for a larger step size in achieving this convergence.

3.2. *Universality with respect to the noise distribution.* Although we analyze AMP and NGD for Gaussian noise, we expect the properties of these estimators and of the TAP free energy landscape to be robust under sufficiently light-tailed distributions of noise entries. Here, we verify this numerically for three examples of symmetric non-Gaussian noise matrices $\boldsymbol{W}$:

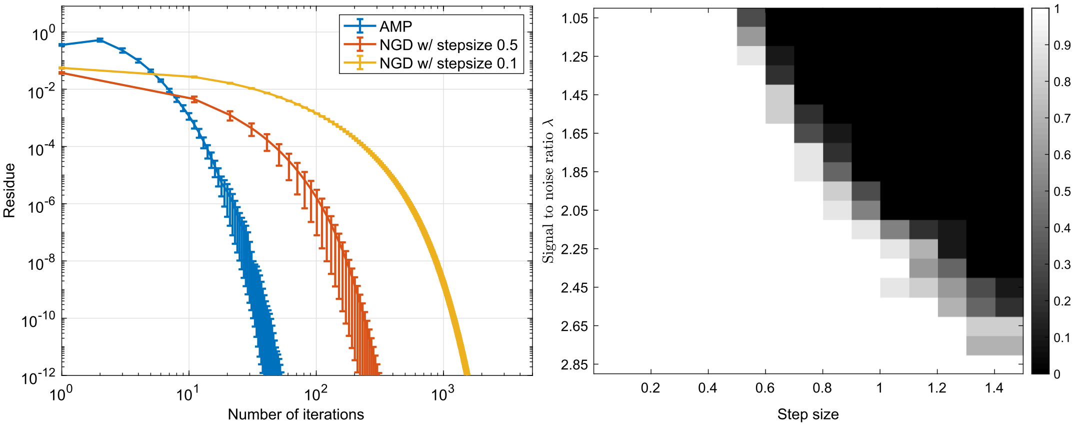

FIG. 1. *Convergence of AMP and NGD from a spectral initialization. Left*: *Residual squared error* $\min\{\|\boldsymbol{m}^k - \boldsymbol{m}_\star\|_2^2/n, \|\boldsymbol{m}^k + \boldsymbol{m}_\star\|_2^2/n\}$ *versus number of iterations* k (*both on a log-scale*), *for signal-to-noise ratio* $\lambda = 1.5$. *The mean curve is averaged over* 10 *independent instances, and the error bars report* $1/\sqrt{10}$ *times the standard deviation across instances. Right*: *Success probability of NGD for convergence to* $\boldsymbol{m}_\star$, *for varying signal-to-noise ratios* λ *and step sizes* η. *In both panels*, $n = 500$.

- Rademacher: $(w_{ij} : 1 \le i \le j \le n) \overset{\text{i.i.d.}}{\sim} \text{Unif}\{-1/\sqrt{n}, 1/\sqrt{n}\}$.
- Double-exponential (Laplace): $\boldsymbol{W} = (\boldsymbol{G} + \boldsymbol{G}^\top)/\sqrt{2n}$, where $(G_{ij} : 1 \le i, j \le n) \overset{\text{i.i.d.}}{\sim} (1/\sqrt{2}) \exp\{-\sqrt{2} \cdot |x|\}$.
- Student's t: $\boldsymbol{W} = (\boldsymbol{G} + \boldsymbol{G}^\top)/\sqrt{2n}$, where $(G_{ij} : 1 \le i, j \le n) \overset{\text{i.i.d.}}{\sim} t(\nu)/\sqrt{\nu/(\nu-2)}$ and the degrees-of-freedom is $\nu = 4$.

In all three examples, all entries w_{ij} have mean 0, and all off-diagonal entries w_{ij} have variance $1/n$.

In Figure 2(a), we report the estimation mean squared error (MSE) $\min\{\|\boldsymbol{m}_\star - \boldsymbol{x}\|_2^2/n, \|\boldsymbol{m}_\star + \boldsymbol{x}\|_2^2/n\}$ versus λ, where $\boldsymbol{m}_\star = \arg\min_{\boldsymbol{m}} \mathcal{F}_{\text{TAP}}(\boldsymbol{m})$ is computed from AMP up to high accuracy as before, and the noise matrix $\boldsymbol{W}$ is generated from either the assumed Gaussian (GOE) model or from the above three non-Gaussian ensembles. In Figure 2(b), we report the

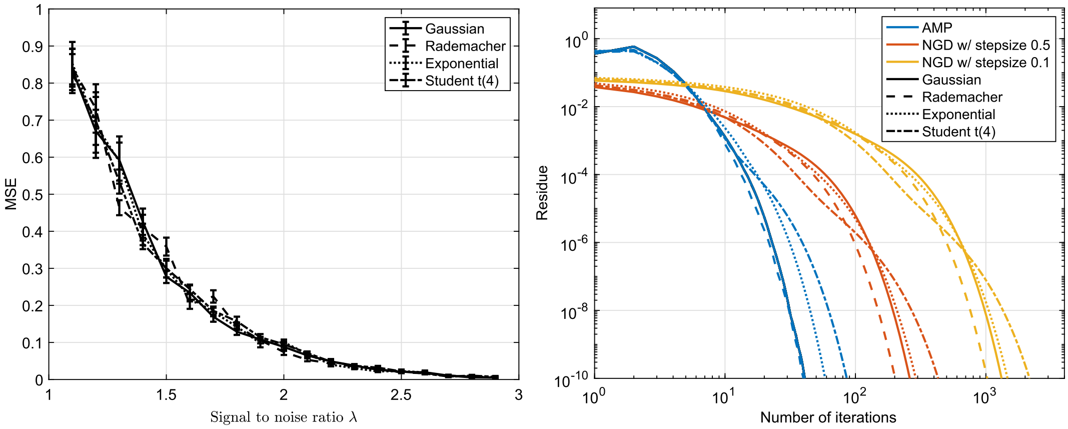

FIG. 2. *Universality with respect to the noise distribution. Left*: *Estimation mean squared error* $\min\{\|\boldsymbol{m}_\star - \boldsymbol{x}\|_2^2/n, \|\boldsymbol{m}_\star - \boldsymbol{x}\|_2^2/n\}$ *versus the signal-to-noise ratio* λ, *for different noise ensembles. The mean curve is averaged over* 10 *independent instances, and the error bars report* $1/\sqrt{10}$ *times the standard deviation across instances. Right*: *Residual squared error* $\min\{\|\boldsymbol{m}^k - \boldsymbol{m}_\star\|_2^2/n, \|\boldsymbol{m}^k + \boldsymbol{m}_\star\|_2^2/n\}$ *versus the number of iterations* k, *for different noise ensembles and signal-to-noise ratio* $\lambda = 1.5$. *In both panels*, $n = 500$.

residual squared error $\min\{\|\boldsymbol{m}^k - \boldsymbol{m}_\star\|_2^2/n, \|\boldsymbol{m}^k + \boldsymbol{m}_\star\|_2^2/n\}$ versus the number of algorithm iterations k, for the same four noise ensembles. These figures show that properties of the TAP minimizers and of the AMP and NGD iterates are indeed robust to these distributions of the noise entries, even for some heavy-tailed distributions.

We also tested Student's t-distribution with degrees-of-freedom $\nu = 3$, and observed that when $\lambda \in (1, 2)$ and $n = 500$, AMP oscillates between two points rather than converging to a fixed point. Instead, the NGD algorithm with a sufficiently small step size continues to converge to the global minimizer.

3.3. *Comparing TAP and mean-field variational Bayes.* We compare the TAP approach to naive mean-field variational Bayes (mean-field VB), under both a correctly specified noise model and a misspecified model that lies outside of the preceding universality class.

For $\mathbb{Z}_2$-synchronization, parametrizing (1) by the mean vector $\boldsymbol{m} = \mathbb{E}_{\boldsymbol{x}\sim q}[\boldsymbol{x}]$ gives the mean-field VB free energy

$$\mathcal{F}_{\mathrm{VB}}(\boldsymbol{m}) = -\frac{1}{n}\sum_{i=1}^{n} \mathsf{h}(m_i) - \frac{\lambda}{2n}\langle \boldsymbol{m}, \boldsymbol{Y}\boldsymbol{m}\rangle.$$

This coincides with (TAP) upon removing the TAP correction term.

In Figure 3, we compare the mean squared errors $\min\{\|\boldsymbol{m}_\star - \boldsymbol{x}\|_2^2/n, \|\boldsymbol{m}_\star + \boldsymbol{x}\|_2^2/n\}$ for the minimizers $\boldsymbol{m}_\star$ of $\mathcal{F}_{\mathrm{TAP}}$ and of $\mathcal{F}_{\mathrm{VB}}$, when $\boldsymbol{Y}$ is generated according to the following two models:

- The correctly specified $\mathbb{Z}_2$-synchronization model (3).
- A misspecified model $\boldsymbol{Y} = (\lambda/n)\boldsymbol{x}\boldsymbol{x}^\mathsf{T} + \boldsymbol{W}$, where $\boldsymbol{x} \sim \mathrm{Unif}(\{-1, +1\}^n)$ has the assumed discrete uniform prior, but $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\mathsf{T}$ does not have independent entries. We choose $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ as a uniformly sampled orthogonal matrix, and $\boldsymbol{D} = \mathrm{diag}(d_1, \dots, d_n)$ where $(d_i : 1 \le i \le n) \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([-\sqrt{3}, \sqrt{3}])$. By this scaling of $\boldsymbol{W}$, we have $\|\boldsymbol{W}\|_\mathsf{F}^2/n \approx 1$ which matches the scaling of $\boldsymbol{W} \sim \mathrm{GOE}(n)$.

For both free energies, we compute their (possibly local) minimizers using the NGD iterations $\boldsymbol{h}^{k+1} = \boldsymbol{h}^k - \eta n \nabla_{\boldsymbol{m}} \mathcal{F}(\boldsymbol{m}^k)$, with step size $\eta = 0.1$ and a spectral initialization. We observe that NGD typically converged within $k = 8000$ iterations (in the sense of achieving a small gradient), despite the lack of a theoretical convergence guarantee in certain settings. Under this model misspecification, the minimizer $\boldsymbol{m}_\star$ of $\mathcal{F}_{\mathrm{TAP}}$ defined according to (TAP) is

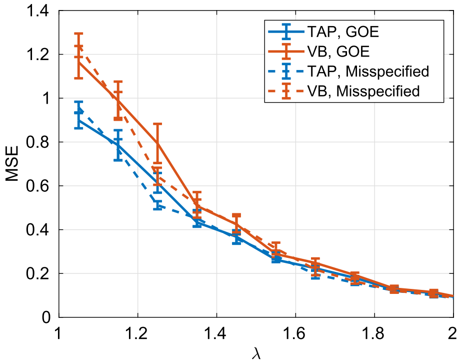

FIG. 3. *Comparison of TAP with mean-field VB. The plot shows mean squared errors of the TAP and VB minimizers in both a correctly specified and a misspecified model, for signal-to-noise ratio* $\lambda \in [1, 2]$ *and* $n = 500$. *The mean curve is averaged over* 10 *independent instances, and the error bars report* $1/\sqrt{10}$ *times the standard deviation across instances.*

no longer expected to be asymptotically exact for the Bayes posterior mean in the true generating model. Nonetheless, we observe that $\boldsymbol{m}_\star$ which minimizes $\mathcal{F}_{\mathrm{TAP}}$ achieves lower mean squared error than that which minimizes $\mathcal{F}_{\mathrm{VB}}$, in both the well-specified and misspecified examples. For larger values of λ, the difference in mean squared error between these approaches becomes harder to discern, although the theory implies (in the well-specified setting) that this difference is asymptotically nonvanishing for any fixed $\lambda > 1$.

4. Local analysis of the TAP free energy.

In this section, we describe the main ideas and steps in the proof of Theorem 2.1.

We will prove that each statement of the theorem holds with probability approaching 1 conditional on the signal vector $\boldsymbol{x} \in \{-1, +1\}^n$. By symmetry, this conditional probability is the same for any given vector $\boldsymbol{x} \in \{-1, +1\}^n$, so we may assume without loss of generality

$$\boldsymbol{x} = \mathbf{1} = (1, 1, \dots, 1).$$

Conditional on $\boldsymbol{x}$, the only remaining randomness is in the noise matrix $\boldsymbol{W} \sim \mathrm{GOE}(n)$, and $\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m})$ is a Gaussian process indexed by $\boldsymbol{m} \in (-1, 1)^n$.

The proof combines information derived from the Kac–Rice formula for the expected number of critical points of Gaussian processes, the Sudakov–Fernique Gaussian comparison inequality, and the AMP state evolution. It is helpful to summarize the type of information each of these tools will provide:

Kac–Rice formula. We use the Kac–Rice formula to upper bound the expected number of critical points of the TAP free energy in certain regions of the domain $(-1, 1)^n$, or for which the TAP Hessian or AMP Jacobian violate the stated properties of Theorem 2.1. In particular, by establishing upper bounds that are vanishing as $n \to \infty$, we prove the nonexistence of such critical points with high probability.

Sudakov–Fernique inequality. We use the Sudakov–Fernique inequality to lower bound the infima of Gaussian processes defined by $\boldsymbol{W} \sim \mathrm{GOE}(n)$ with the infima of Gaussian processes defined by a standard Gaussian vector $\boldsymbol{g} \in \mathbb{R}^n$. We then analyze the latter to obtain variational lower bounds for large n. There are three Gaussian processes to which we apply this technique:

- The TAP free energy itself, to obtain lower bounds on its minimum value over regions of $(-1, 1)^n$.
- A Gaussian process whose infimum gives the minimum eigenvalue of the TAP Hessian over subsets of $(-1, 1)^n$, to show local strong convexity of the TAP free energy.
- A Gaussian process whose infimum is related to the spectral radius of the AMP Jacobian, to show local stability of the AMP map.

AMP state evolution. We use the AMP state evolution to evaluate the TAP free energy at the iterates of AMP, giving upper bounds for the TAP free energy value near the Bayes estimator.

The information provided by each of these three tools is distinct, and the proof of Theorem 2.1 combines the information we can extract from each.

We outline the four main steps of the proof in Section 4.1. These steps are discussed in Sections 4.2 through 4.5, and the technical arguments that execute each step are deferred to Appendix B.

4.1. *Proof outline.* For small parameters $\delta, \eta > 0$, we define two deterministic subsets $\mathcal{B}_\delta, \mathcal{D}_\eta \subset (-1, 1)^n$ based on the empirical distribution of coordinates of $\boldsymbol{m} \in (-1, 1)^n$. These subsets will contain the desired TAP local minimizer $\boldsymbol{m}_\star$ with high probability (conditional on $\boldsymbol{x} = \mathbf{1}$).

For $\lambda > 1$, let $q_\star = q_\star(\lambda)$ be the unique solution in $(0, 1)$ (cf. Proposition A.2) to the fixed-point equation

$$q_\star = \mathbb{E}_{G\sim\mathcal{N}(0,1)}\big[\tanh\big(\lambda^2 q_\star + \lambda\sqrt{q_\star}G\big)^2\big]. \tag{11}$$

Define

$$h_\star = \mathbb{E}_{G\sim\mathcal{N}(0,1)}\big[\log 2\cosh\big(\lambda^2 q_\star + \lambda\sqrt{q_\star}G\big)\big] - \lambda^2 q_\star, \tag{12}$$

$$e_\star = -\frac{\lambda^2}{4}\big(1 - 2q_\star - q_\star^2\big) - \mathbb{E}_{G\sim\mathcal{N}(0,1)}\big[\log 2\cosh\big(\lambda^2 q_\star + \lambda\sqrt{q_\star}G\big)\big]. \tag{13}$$

For any point $\boldsymbol{m} \in (-1, 1)^n$, denote

$$Q(\boldsymbol{m}) = \frac{1}{n}\|\boldsymbol{m}\|_2^2, \qquad M(\boldsymbol{m}) = \frac{1}{n}\boldsymbol{m}^\mathsf{T}\mathbf{1}, \qquad H(\boldsymbol{m}) = \frac{1}{n}\sum_{i=1}^n \mathsf{h}(m_i),$$

where $\mathsf{h}(\cdot)$ is the binary entropy function from (6). We define the first subset $\mathcal{B}_\delta$ as

$$\mathcal{B}_\delta = \big\{\boldsymbol{m} \in (-1, 1)^n : \big|Q(\boldsymbol{m}) - q_\star\big|, \big|M(\boldsymbol{m}) - q_\star\big|, \big|H(\boldsymbol{m}) - h_\star\big| < \delta\big\}. \tag{14}$$

Let

$$\mu_\star = \text{distribution of } \tanh\big(\lambda^2 q_\star + \lambda\sqrt{q_\star}G\big) \text{ when } G \sim \mathcal{N}(0, 1), \tag{15}$$

which will be the limiting empirical distribution of coordinates of $\boldsymbol{m}_\star$. For $\boldsymbol{m} \in (-1, 1)^n$, let $\hat{\mu}_{\boldsymbol{m}}$ be the empirical distribution of coordinates of $\boldsymbol{m}$, that is,

$$\hat{\mu}_{\boldsymbol{m}} = \frac{1}{n}\sum_{i=1}^n \delta_{m_i}. \tag{16}$$

Denote by $W(\mu, \mu')$ the Wasserstein-2 distance between $\operatorname{arctanh}\mu$ and $\operatorname{arctanh}\mu'$, where $\operatorname{arctanh}\mu$ is shorthand for the law of $\operatorname{arctanh} m$ when $m \sim \mu$. That is, we have

$$\begin{aligned} W(\mu, \mu') &\equiv W_2\big(\operatorname{arctanh}\mu, \operatorname{arctanh}\mu'\big) \\ &= \bigg(\inf_{\text{couplings } \nu \text{ of } (\mu,\mu')} \int \big(\operatorname{arctanh} m - \operatorname{arctanh} m'\big)^2 \, d\nu(m, m')\bigg)^{1/2}. \end{aligned} \tag{17}$$

We review properties of this distance in Appendix A.3. We define the second subset $\mathcal{D}_\eta$ as

$$\mathcal{D}_\eta = \big\{\boldsymbol{m} \in (-1, 1)^n : W(\hat{\mu}_{\boldsymbol{m}}, \mu_\star) < \eta\big\}. \tag{18}$$

The proof of Theorem 2.1 then consists of four steps (all conditional on $\boldsymbol{x} = \mathbf{1}$):

1. For sufficiently small $\delta > 0$, we use the Sudakov–Fernique inequality to lower bound the value of $\mathcal{F}_{\text{TAP}}$ on $\mathcal{B}_\delta \setminus \mathcal{B}_{\delta/2}$. Comparing with the value of $\mathcal{F}_{\text{TAP}}$ achieved by an iterate $\boldsymbol{m}^k \in \mathcal{B}_{\delta/2}$ of AMP, we show that $\mathcal{F}_{\text{TAP}}$ must have a local minimizer $\boldsymbol{m}_\star$ in $\mathcal{B}_\delta$, and $\mathcal{F}_{\text{TAP}}(\boldsymbol{m}_\star) \approx e_\star$.
2. For any fixed $\eta > 0$, we use a Kac–Rice upper bound to show that with high probability, any such local minimizer $\boldsymbol{m}_\star$ cannot belong to $\mathcal{B}_\delta \setminus \mathcal{D}_\eta$. Thus, it must belong to $\mathcal{B}_\delta \cap \mathcal{D}_\eta$.
3. For $\varepsilon, t > 0$ sufficiently small, we apply a second Kac–Rice upper bound to show that for all critical points $\boldsymbol{m}_\star \in \mathcal{D}_\eta$, $\lambda_{\min}(n \cdot \nabla^2\mathcal{F}_{\text{TAP}}) \geq t$ everywhere in a $\sqrt{\varepsilon n}$-ball around $\boldsymbol{m}_\star$. We analyze the Kac–Rice bound by representing $\lambda_{\min}(n \cdot \nabla^2\mathcal{F}_{\text{TAP}})$ over this ball as the infimum of a Gaussian process, and lower bounding its value by a second application of the Sudakov–Fernique inequality.

This implies that $\mathcal{F}_{\text{TAP}}$ is strongly convex near any local minimizer $\boldsymbol{m}_\star \in \mathcal{B}_\delta \cap \mathcal{D}_\eta$ of Steps 1 and 2. This convexity then ensures that there exists a unique such local minimizer satisfying (8), establishing Theorem 2.1(a–b).

4. To show Theorem 2.1(c), we relate each (possibly complex) eigenvalue μ of $\mathrm{d}T_{\mathsf{AMP}}(\boldsymbol{m}_\star, \boldsymbol{m}_\star)$ to a zero eigenvalue of a corresponding "Bethe Hessian" of $\mathcal{F}_{\mathrm{TAP}}$ [108]. We extend the Kac–Rice/Sudakov–Fernique argument of Step 3 from $\nabla^2 \mathcal{F}_{\mathrm{TAP}}$ to this Bethe Hessian, and show that it is positive definite whenever $|\mu|$ exceeds some constant $r(\lambda) \in (0,1)$. Thus all eigenvalues of $\mathrm{d}T_{\mathsf{AMP}}$ satisfy $|\mu| \le r(\lambda)$.

The next four sections describe these steps in greater detail.

4.2. *Sudakov–Fernique lower bound for the TAP free energy.* We record here the following application of the Slepian/Sudakov–Fernique comparison inequality for Gaussian processes.

LEMMA 4.1. *Let $\mathcal{X}$ be a separable metric space, and let $f : \mathcal{X} \to \mathbb{R}$ and $\boldsymbol{v} : \mathcal{X} \to \mathbb{R}^n$ be bounded measurable functions on $\mathcal{X}$. Let $\boldsymbol{W} \sim \mathrm{GOE}(n)$ and $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I}_n)$. Then*

$$\mathbb{E}\Big[\sup_{x \in \mathcal{X}} \boldsymbol{v}(x)^\top \boldsymbol{W} \boldsymbol{v}(x) + f(x)\Big] \le \mathbb{E}\Big[\sup_{x \in \mathcal{X}} \frac{2}{\sqrt{n}} \|\boldsymbol{v}(x)\|_2 \langle \boldsymbol{g}, \boldsymbol{v}(x)\rangle + f(x)\Big].$$

Note that (conditional on $\boldsymbol{x} = \mathbf{1}$) $-\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m})$ is a Gaussian process of this form, where $\mathcal{X} = (-1,1)^n$ and $\boldsymbol{v}(\boldsymbol{m}) = \sqrt{\lambda/2n} \cdot \boldsymbol{m}$. Then applying this comparison lemma and an analysis of the comparison process, we obtain the following lower bound for $\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m})$ in terms of a low-dimensional, deterministic variational formula.

LEMMA 4.2. *Fix any $\lambda > 1$, and suppose $\boldsymbol{x} = \mathbf{1}$. Fix any $\varepsilon > 0$ and two compact sets $K \subseteq [0,1]^2 \times [0, \log 2]$ and $K' \subset \mathbb{R}^3$. Then for some $(\lambda, K', \varepsilon)$-dependent constant $c > 0$ and all large n, with probability at least $1 - e^{-cn}$,*

$$\inf_{\boldsymbol{m} \in (-1,1)^n : (Q(\boldsymbol{m}), M(\boldsymbol{m}), H(\boldsymbol{m})) \in K} \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) > \inf_{(q,\varphi,h) \in K} \sup_{(\gamma,\tau,\nu) \in K'} E_\lambda(q, \varphi, h; \gamma, \tau, \nu) - \varepsilon, \tag{19}$$

where

$$\begin{aligned} E_\lambda(q, \varphi, h; \gamma, \tau, \nu) &= -\frac{\lambda^2}{2}\varphi^2 - \frac{\lambda^2}{4}(1-q)^2 - h + \frac{q\gamma}{2} + \varphi\tau + \nu h \\ &\quad - \mathbb{E}_{G \sim \mathcal{N}(0,1)}\Big\{\sup_{m \in (-1,1)} \Big[\lambda\sqrt{q} \cdot Gm + \frac{\gamma m^2}{2} + \tau m + \nu \mathsf{h}(m)\Big]\Big\}. \end{aligned} \tag{20}$$

Lemma 4.2 makes precise the statement that

$$\bar{E}_\lambda(q, \varphi, h) = \sup_{(\gamma,\tau,\nu) \in K'} E_\lambda(q, \varphi, h; \gamma, \tau, \nu) \tag{21}$$

is a lower bound for $\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m})$ when $Q(\boldsymbol{m}) \approx q$, $M(\boldsymbol{m}) \approx \varphi$, and $H(\boldsymbol{m}) \approx h$. We may show that $\bar{E}_\lambda(q, \varphi, h)$ has a local minimizer at $(q, \varphi, h) = (q_\star, q_\star, h_\star)$ and is strongly convex around this minimizer, and hence give a more explicit lower bound for $\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m})$ when $\boldsymbol{m} \in \mathcal{B}_\delta$ for sufficiently small $\delta > 0$.

LEMMA 4.3. *Fix any $\lambda > 1$, and let $E_\lambda(q, \varphi, h; \gamma, \tau, \nu)$ be as defined in Lemma 4.2. Then*

$$\sup_{(\gamma,\tau,\nu) \in \mathbb{R}^3} E_\lambda(q_\star, q_\star, h_\star; \gamma, \tau, \nu) = E_\lambda(q_\star, q_\star, h_\star; 0, \lambda^2 q_\star, 1) = e_\star. \tag{22}$$

Fix any subset $K' \subseteq \mathbb{R}^3$ containing $(0, \lambda^2 q_\star, 1)$ in its interior, and define $\bar{E}_\lambda$ by (21). Then for some λ, K'-dependent constants $\delta, c > 0$ and all (q, φ, h) satisfying $|q - q_\star|, |\varphi - q_\star|, |h - h_\star| \le \delta$,

$$\bar{E}_\lambda(q, \varphi, h) \ge e_\star + c(q - q_\star)^2 + c(\varphi - q_\star)^2 + c(h - h_\star)^2. \tag{23}$$

Lemmas 4.2 and 4.3 together imply that the energy value $\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m})$ is bounded away from $e_\star$ on the domain $\boldsymbol{m} \in \mathcal{B}_\delta \setminus \mathcal{B}_{\delta/2}$. The AMP state evolution may be applied to show that AMP iterates eventually enter $\mathcal{B}_{\delta/2}$, and achieve a TAP free energy value arbitrarily close to $e_\star$ (cf. Lemma A.7). Combined, these yield the following corollary.

COROLLARY 4.4. *Fix any $\lambda > 1$ and $\delta > 0$, and suppose $\boldsymbol{x} = \mathbf{1}$. Then with probability approaching 1 as $n \to \infty$, there exists a critical point and local minimizer $\boldsymbol{m}_\star$ of $\mathcal{F}_{\mathrm{TAP}}$ belonging to $\mathcal{B}_\delta$ and satisfying $|\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}_\star) - e_\star| < \delta$.*

The detailed proofs of this section are contained in Appendix B.1.

REMARK 4.5. We conjecture, based on numerical evidence, that $(q_\star, q_\star, h_\star)$ is in fact the global minimizer of $\bar{E}_\lambda(q, \varphi, h)$ for all $\lambda > 1$: We may first restrict E_λ to $\nu = 1$ and $\tau = \lambda^2\varphi$, to obtain the further lower bound

$$\bar{E}_\lambda(q, \varphi, h) \geq \bar{E}_\lambda(q, \varphi) = \sup_\gamma E_\lambda(q, \varphi; \gamma), \tag{24}$$

where

$$\begin{aligned} E_\lambda(q, \varphi; \gamma) = {} & \frac{\lambda^2}{2}\varphi^2 - \frac{\lambda^2}{4}(1-q)^2 + \frac{q\gamma}{2} \\ & - \mathbb{E}_{G\sim\mathcal{N}(0,1)}\Big[\sup_{m\in(-1,1)} \lambda\sqrt{q}\cdot Gm + \frac{\gamma m^2}{2} + \lambda^2\varphi m + \mathsf{h}(m)\Big]. \end{aligned}$$

Numerical evaluations of this function $\bar{E}_\lambda(q, \varphi)$ over the relevant domain $q \in (0, 1)$ and $|\varphi| < \sqrt{q}$ are presented in Figure 4. For all tested values of $\lambda > 1$, these evaluations support the claim that $\bar{E}_\lambda(q, \varphi)$ has the unique *global* minimizer $(q, \varphi) = (q_\star, q_\star)$. This claim then implies that $(q_\star, q_\star, h_\star)$ is also the unique global minimizer of $\bar{E}_\lambda(q, \varphi, h)$, by the global convexity of $h \mapsto \bar{E}_\lambda(q_\star, q_\star, h)$ and its strong convexity near its minimizer $h_\star$.

Subject to the validity of this numerical conjecture, Lemma 4.2 may be used to show that $\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m})$ is also bounded away from $e_\star$ for all $\boldsymbol{m} \in (-1, 1)^n \setminus \mathcal{B}_\delta$. Our subsequent arguments will then imply that for any $\lambda > 1$, with probability approaching 1, the (unique) local minimizer $\boldsymbol{m}_\star$ described by Corollary 4.4 and Theorem 2.1 is in fact the global minimizer of $\mathcal{F}_{\mathrm{TAP}}$. (All theoretical results stated in this work will be established using only that $\boldsymbol{m}_\star$ is a local minimizer of $\mathcal{F}_{\mathrm{TAP}}$, and they will not require the validity of this conjecture.)

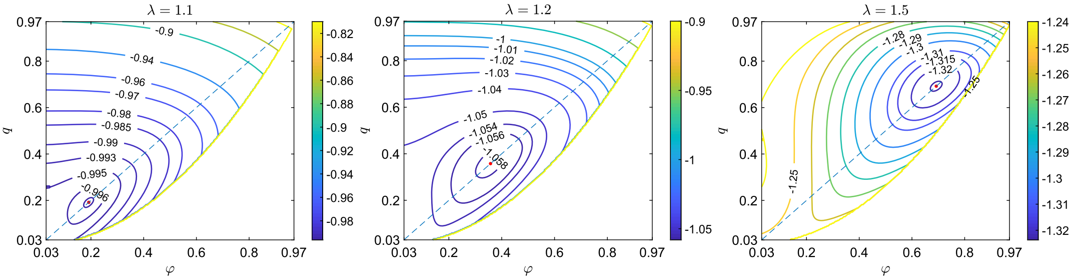

FIG. 4. *The contour plot of the function $\bar{E}_\lambda(q, \varphi)$ as defined in equation (24). Here we take $\lambda = 1.1, 1.2, 1.5$. The global minimum is at $(q, \varphi) = (q_\star(\lambda), q_\star(\lambda))$ where $q_\star(1.1) \approx 0.1917$, $q_\star(1.2) \approx 0.3577$, $q_\star(1.5) \approx 0.6923$. The dashed line is $q = \varphi$.*

4.3. *Kac–Rice localization of critical points.* We now use a Kac–Rice upper bound to show that the critical point(s) $\boldsymbol{m}_\star$ described by Corollary 4.4 must belong to the more restrictive set $\mathcal{B}_\delta \cap \mathcal{D}_\eta$ (cf. equation (14) and (18)).

Define functions $\boldsymbol{g}$ and $\boldsymbol{H}$, which are the gradient and Hessian of the renormalized TAP free energy

$$\boldsymbol{g}(\boldsymbol{m}) = n \cdot \nabla \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) = -\lambda \boldsymbol{Y}\boldsymbol{m} + \operatorname{arctanh}(\boldsymbol{m}) + \lambda^2[1 - Q(\boldsymbol{m})]\boldsymbol{m}, \tag{25}$$

$$\begin{aligned} \boldsymbol{H}(\boldsymbol{m}) &= n \cdot \nabla^2 \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) \\ &= -\lambda \boldsymbol{Y} + \operatorname{diag}\left(\frac{1}{1 - \boldsymbol{m}^2}\right) + \lambda^2[1 - Q(\boldsymbol{m})]\mathbf{I} - \frac{2\lambda^2}{n}\boldsymbol{m}\boldsymbol{m}^\mathsf{T}. \end{aligned} \tag{26}$$

We apply the following Kac–Rice upper bound from [53].

LEMMA 4.6. *Fix any $\lambda > 0$, suppose $\boldsymbol{x} = \mathbf{1}$, and let $T \subseteq (-1, 1)^n \setminus \{\mathbf{0}\}$ be any (deterministic) Borel-measurable set. Then*

$$\mathbb{E}\big[\big|\{\boldsymbol{m} \in T : \boldsymbol{g}(\boldsymbol{m}) = \mathbf{0}\}\big|\big] \le \int_T \mathbb{E}\big[\big|\det \boldsymbol{H}(\boldsymbol{m})\big| \mid \boldsymbol{g}(\boldsymbol{m}) = \mathbf{0}\big] p_{\boldsymbol{g}(\boldsymbol{m})}(\mathbf{0})\, \mathrm{d}\boldsymbol{m},$$

where $p_{\boldsymbol{g}(\boldsymbol{m})}(\mathbf{0})$ is the Lebesgue-density of the distribution of $\boldsymbol{g}(\boldsymbol{m})$ at $\boldsymbol{g}(\boldsymbol{m}) = \mathbf{0}$.

Applying this bound, we eliminate the possibility that the critical point(s) described by Corollary 4.4 belong to $\mathcal{B}_\delta \setminus \mathcal{D}_\eta$, as stated in the following lemma. Thus they belong to $\mathcal{B}_\delta \cap \mathcal{D}_\eta$ as desired.

LEMMA 4.7. *Fix any $\lambda > 1$ and $\eta > 0$, and suppose $\boldsymbol{x} = \mathbf{1}$. Then for some (λ, η)-dependent constants $c, \delta > 0$ and all large n,*

$$\mathbb{P}\big[\textit{there exists } \boldsymbol{m} \in \mathcal{B}_\delta : \boldsymbol{g}(\boldsymbol{m}) = \mathbf{0}, \big|\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) - e_\star\big| < \delta, \boldsymbol{m} \notin \mathcal{D}_\eta\big] < e^{-cn}.$$

Let us make two high-level clarifications regarding the proof: First, to show Lemma 4.7, we wish to apply Lemma 4.6 with T being the set

$$\big\{\boldsymbol{m} \in \mathcal{B}_\delta \setminus \mathcal{D}_\eta : \big|\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) - e_\star\big| < \delta\big\}.$$

We cannot do so directly, because $\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m})$ is random, and hence this is not a deterministic subset of $(-1, 1)^n$. However, restricted to points $\boldsymbol{m}$ where $\boldsymbol{g}(\boldsymbol{m}) = \mathbf{0}$, the identity $0 = \boldsymbol{m}^\top \boldsymbol{g}(\boldsymbol{m})$ allows us to re-express $\boldsymbol{m}^\top \boldsymbol{Y}\boldsymbol{m}$ and $\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m})$ as deterministic functions of $\boldsymbol{m}$. Lemma 4.7 is then obtained by replacing $|\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) - e_\star| < \delta$ with an equivalent deterministic condition to define T.

Second, we remark that the Sudakov–Fernique argument of the preceding section cannot be used here to similarly localize $\boldsymbol{m}_\star$ to $\mathcal{D}_\eta$, by bounding the TAP free energy value outside $\mathcal{B}_\delta \setminus \mathcal{D}_\eta$. This is because there exists $\boldsymbol{m} \in (-1, 1)^n$ with one coordinate very close to ±1, so that $W(\hat{\mu}_{\boldsymbol{m}}, \mu_\star)$ is arbitrarily large (cf. equation (17)) and $\boldsymbol{m} \notin \mathcal{D}_\eta$, but $\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m})$ is arbitrarily close to $e_\star$ in value. Thus, we use this separate Kac–Rice argument, and the condition $\boldsymbol{m} \in \mathcal{B}_\delta$ as an input to the Kac–Rice analysis, to establish the localization to $\mathcal{D}_\eta$ in Lemma 4.7.

The detailed proofs of this section are contained in Appendix B.2.

4.4. *Sudakov–Fernique lower bound for local strong convexity.* We now show that the TAP free energy is strongly convex in a local neighborhood of any critical point $\boldsymbol{m}_\star \in \mathcal{D}_\eta$. For a parameter $\varepsilon > 0$, define

$$\ell_\varepsilon^+(\boldsymbol{m}, \boldsymbol{W}) = \inf\big\{\lambda_{\min}\big(\boldsymbol{H}(\boldsymbol{u})\big) : \boldsymbol{u} \in (-1, 1)^n \cap \mathsf{B}_{\sqrt{\varepsilon n}}(\boldsymbol{m})\big\}. \tag{27}$$

The dependence of ℓ_ε^+ on $\boldsymbol{W}$ is via $\boldsymbol{H}(\boldsymbol{u})$. We will make this dependence implicit in what follows, and write simply $\ell_\varepsilon^+(\boldsymbol{m}) = \ell_\varepsilon^+(\boldsymbol{m}, \boldsymbol{W})$. If $\ell_\varepsilon^+(\boldsymbol{m}_\star) \geq t > 0$, then the TAP free energy is strongly convex on a $\sqrt{\varepsilon n}$-ball around $\boldsymbol{m}_\star$, as desired. We use a Kac–Rice upper bound to show, with high probability, no critical points of $\mathcal{F}_{\text{TAP}}$ belong to the set

$$\{\boldsymbol{m} \in \mathcal{D}_\eta : \ell_\varepsilon^+(\boldsymbol{m}) < t\}$$

for some sufficiently small constant $t > 0$.

The condition $\ell_\varepsilon^+(\boldsymbol{m}) < t$ is again random, so this is not a deterministic subset of $(-1, 1)^n$. We address this using the following extension of the Kac–Rice upper bound in Lemma 4.6.

LEMMA 4.8. *Fix any $\lambda > 0$, and suppose $\boldsymbol{x} = \boldsymbol{1}$. Let Sym_n be the space of real symmetric $n \times n$ matrices, $T \subseteq (-1, 1)^n \setminus \{\boldsymbol{0}\}$ any (deterministic) Borel-measurable set, and $\ell : T \times \text{Sym}_n \to \mathbb{R}$ any Borel-measurable function. Let $c > 0$ and $t \in \mathbb{R}$ be any (possibly n-dependent) values, and let $U \sim \text{Unif}([-c, c])$ be a uniform random variable independent of $\boldsymbol{W}$. Define*

$$\mathcal{C} = \{\boldsymbol{m} \in T : \boldsymbol{g}(\boldsymbol{m}) = \boldsymbol{0} \text{ and } \ell(\boldsymbol{m}, \boldsymbol{W}) + U < t\}.$$

Then

$$\mathbb{E}[|\mathcal{C}|] \leq \int_T \mathbb{E}\big[|\det \boldsymbol{H}(\boldsymbol{m})| \cdot \boldsymbol{1}\{\ell(\boldsymbol{m}, \boldsymbol{W}) + U < t\} \mid \boldsymbol{g}(\boldsymbol{m}) = \boldsymbol{0}\big] p_{\boldsymbol{g}(\boldsymbol{m})}(\boldsymbol{0}) \, \mathrm{d}\boldsymbol{m}, \tag{28}$$

where $p_{\boldsymbol{g}(\boldsymbol{m})}(\boldsymbol{0})$ is the Lebesgue-density of the distribution of $\boldsymbol{g}(\boldsymbol{m})$ at $\boldsymbol{g}(\boldsymbol{m}) = \boldsymbol{0}$, and the expectations are over both U and $\boldsymbol{W}$.

(Introducing this auxiliary variable U alleviates the need to check a technical condition that $\ell(\boldsymbol{m}, \boldsymbol{W}) = t$ and $\boldsymbol{g}(\boldsymbol{m}) = \boldsymbol{0}$ do not simultaneously occur at any $\boldsymbol{m} \in T$, when applying the Kac–Rice lemma.)

In [53], an upper bound on the determinant $|\det \boldsymbol{H}(\boldsymbol{m})|$ was established via a spectral analysis of $\boldsymbol{H}(\boldsymbol{m})$, which shows $\mathbb{E}[|\det \boldsymbol{H}(\boldsymbol{m})|^2 \mid \boldsymbol{g}(\boldsymbol{m}) = \boldsymbol{0}] \leq e^{c(\eta)n}$ for $\boldsymbol{m} \in \mathcal{D}_\eta$ and a constant $c(\eta) \to 0$ as $\eta \to 0$. Thus, to show that (28) is vanishing, we complement this by showing an exponentially small upper bound for the probability $\mathbb{P}[\ell_\varepsilon^+(\boldsymbol{m}) + U < t \mid \boldsymbol{g}(\boldsymbol{m}) = \boldsymbol{0}]$. We do this again using the Sudakov–Fernique inequality of Lemma 4.1, to obtain the variational lower bound on the conditional mean $\mathbb{E}[\ell_\varepsilon^+(\boldsymbol{m}) \mid \boldsymbol{g}(\boldsymbol{m}) = \boldsymbol{0}]$ stated in part (a) of the following lemma. This bound is shown to be positive in part (b).

LEMMA 4.9. *Suppose $\lambda > 1$ and $\boldsymbol{x} = \boldsymbol{1}$. Define*

$$\begin{aligned} &H_\lambda^+(p, u; \alpha, \kappa, \gamma) \\ &\quad = -\big[2\lambda^2 p^2 + \lambda^2 u^2 - 2\lambda^2(1 - q_\star)p^2/q_\star - \alpha u - \kappa p\big] + \lambda^2(1 - q_\star) + \gamma \\ &\qquad - \mathbb{E}_{m \sim \mu_\star}\bigg[\big(4\lambda^2(1 - p^2/q_\star) + (2z(m)p/q_\star + \alpha + \kappa m)^2\big)\Big/\Big(\frac{4}{1 - m^2} - 4\gamma\Big)\bigg], \end{aligned} \tag{29}$$

where $z(m) = \operatorname{arctanh} m - \lambda^2 q_\star + \lambda^2(1 - q_\star)m$.

(a) *Fix any $t > 0$ and compact domain $K' \subset \mathbb{R}^2 \times (-\infty, 1)$. For some (λ, K', t)-dependent constants $\varepsilon, \eta > 0$, and all large n,*

$$\inf_{\boldsymbol{m} \in \mathcal{D}_\eta} \mathbb{E}\big[\ell_\varepsilon^+(\boldsymbol{m}) \mid \boldsymbol{g}(\boldsymbol{m}) = \boldsymbol{0}\big] \geq \inf_{\substack{u \in [-1,1] \\ p \in [-\sqrt{q_\star}, \sqrt{q_\star}]}} \sup_{(\alpha, \kappa, \gamma) \in K'} H_\lambda^+(p, u; \alpha, \kappa, \gamma) - t.$$

(b) *Suppose K' contains* $(0,0,0)$ *in its interior. Then there is a* (λ, K')*-dependent constant* $t_0 > 0$ *for which*

$$\inf_{\substack{u\in[-1,1]\\ p\in[-\sqrt{q_\star},\sqrt{q_\star}]}} \sup_{(\alpha,\kappa,\gamma)\in K'} H_\lambda^+(p,u;\alpha,\kappa,\gamma) > t_0 > 0.$$

The desired upper bound for $\mathbb{P}[\ell_\varepsilon^+(\boldsymbol{m}) + U < t \mid \boldsymbol{g}(\boldsymbol{m}) = \boldsymbol{0}]$ then follows by concentration of $\ell_\varepsilon^+(\boldsymbol{m})$ around its mean. Applying this to (28) yields the following corollary on local strong convexity.

COROLLARY 4.10. *Fix any* $\lambda > 1$, *and suppose* $\boldsymbol{x} = \boldsymbol{1}$. *Then there exist* λ*-dependent constants* $\varepsilon, \eta, t, c > 0$ *such that, for all large* n,

$$\begin{aligned}\mathbb{P}\big[&\textit{there exist } \boldsymbol{m} \in \mathcal{D}_\eta \textit{ and } \boldsymbol{u} \in (-1,1)^n \cap \mathsf{B}_{\sqrt{\varepsilon n}}(\boldsymbol{m}) : \\ &\boldsymbol{g}(\boldsymbol{m}) = \boldsymbol{0} \textit{ and } \lambda_{\min}\big(\boldsymbol{H}(\boldsymbol{u})\big) < t\big] < e^{-cn}.\end{aligned} \tag{30}$$

Finally, this convexity implies Theorem 2.1(a–b) by the following argument: Letting $\boldsymbol{m}^k$ be a sufficiently large iterate of AMP, we may pick a local minimizer $\boldsymbol{m}_\star$ in Corollary 4.4 such that there is a strict descent path from $\boldsymbol{m}^k$ to $\boldsymbol{m}_\star$. Strong convexity of $\mathcal{F}_{\mathrm{TAP}}$ around $\boldsymbol{m}_\star$ and an upper bound on $\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}^k) - \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}_\star)$ then imply an upper bound on the Euclidean distance $\|\boldsymbol{m}_\star - \boldsymbol{m}^k\|_2$. Then this point $\boldsymbol{m}_\star$ must satisfy (8) by the Bayes-optimality of the AMP iterate $\boldsymbol{m}^k$. Furthermore, the local convexity of $\mathcal{F}_{\mathrm{TAP}}$ implies that such a point $\boldsymbol{m}_\star$ is unique. We provide the details of this argument in Appendix B.3.

4.5. *Local stability of AMP.* We now describe the proof of Theorem 2.1(c). Let us write the input and output of T_{AMP} in (7) as

$$(\boldsymbol{m}_+, \boldsymbol{m}) = T_{\mathsf{AMP}}(\boldsymbol{m}, \boldsymbol{m}_-).$$

Differentiating by the chain rule, the Jacobian of T_{AMP} may be expressed as

$$\begin{aligned}&\mathrm{d}T_{\mathsf{AMP}}(\boldsymbol{m}, \boldsymbol{m}_-)\\ &\quad = \begin{pmatrix} \mathrm{diag}(1-\boldsymbol{m}_+^2)\cdot[\lambda\boldsymbol{Y} + 2\lambda^2\boldsymbol{m}_-\boldsymbol{m}^\top/n] & -\mathrm{diag}(1-\boldsymbol{m}_+^2)\cdot\lambda^2[1-Q(\boldsymbol{m})] \\ \mathbf{I} & \mathbf{0}\end{pmatrix}.\end{aligned} \tag{31}$$

At any point $\boldsymbol{m}_\star \in (-1,1)^n$ where $\boldsymbol{g}(\boldsymbol{m}_\star) = \boldsymbol{0}$, we have $T_{\mathsf{AMP}}(\boldsymbol{m}_\star, \boldsymbol{m}_\star) = (\boldsymbol{m}_\star, \boldsymbol{m}_\star)$. Thus $\mathrm{d}T_{\mathsf{AMP}}(\boldsymbol{m}_\star, \boldsymbol{m}_\star) = \boldsymbol{B}(\boldsymbol{m}_\star)$ for the matrix

$$\boldsymbol{B}(\boldsymbol{m}) = \begin{pmatrix} \mathrm{diag}(1-\boldsymbol{m}^2)\cdot[\lambda\boldsymbol{Y} + 2\lambda^2\boldsymbol{m}\boldsymbol{m}^\top/n] & -\mathrm{diag}(1-\boldsymbol{m}^2)\cdot\lambda^2[1-Q(\boldsymbol{m})] \\ \mathbf{I} & \mathbf{0}\end{pmatrix}.$$

In Appendix B.4, we first verify the simple algebraic identity that the eigenvalues $\mu \in \mathbb{C}$ of this matrix $\boldsymbol{B}(\boldsymbol{m})$, for any $\boldsymbol{m} \in (-1,1)^n$, are exactly those values $\mu \in \mathbb{C}$ for which the "Bethe Hessian" matrix

$$\mu\Big(-\lambda\boldsymbol{Y} - \frac{2\lambda^2}{n}\boldsymbol{m}\boldsymbol{m}^\top\Big) + \lambda^2[1-Q(\boldsymbol{m})]\boldsymbol{I} + \mu^2\,\mathrm{diag}\Big(\frac{1}{1-\boldsymbol{m}^2}\Big) \tag{32}$$

is singular. Applying this relation, we then show the following deterministic lemma relating the spectral radius of $\boldsymbol{B}(\boldsymbol{m})$ to the smallest eigenvalue of the above matrix for real arguments $\mu = \pm r$.

LEMMA 4.11. *Fix any $\lambda > 1$. There exist λ-dependent constants $\delta > 0$ and $r_0 \in (0, 1)$ such that for any $r \in (r_0, 1)$ and $\boldsymbol{m} \in (-1, 1)^n$ with $|Q(\boldsymbol{m}) - q_\star| < \delta$, if we have*

$$\lambda_{\min}\left[\pm r\left(-\lambda \boldsymbol{Y} - \frac{2\lambda^2}{n}\boldsymbol{m}\boldsymbol{m}^\mathsf{T}\right) + \lambda^2[1 - Q(\boldsymbol{m})]\mathbf{I} + r^2 \operatorname{diag}\left(\frac{1}{1 - \boldsymbol{m}^2}\right)\right] > 0 \tag{33}$$

for both choices of sign $\pm$, then $\rho(\boldsymbol{B}(\boldsymbol{m})) < r < 1$.

To prove Theorem 2.1(c), by a simple continuity argument, it will suffice to consider exactly $r = 1$ in (33) and to show that (33) holds with high probability at $\boldsymbol{m} = \boldsymbol{m}_\star$ for both choices of sign $\pm$. For $r = 1$ and sign $+$, the matrix in (33) is precisely the Hessian $\boldsymbol{H}(\boldsymbol{m})$, whose smallest eigenvalue at $\boldsymbol{m} = \boldsymbol{m}_\star$ was bounded in the preceding section. The case of sign $-$ is a minor extension of these arguments: Define

$$\boldsymbol{H}^-(\boldsymbol{m}) = \left(\lambda \boldsymbol{Y} + \frac{2\lambda^2}{n}\boldsymbol{m}\boldsymbol{m}^\mathsf{T}\right) + \operatorname{diag}\left(\frac{1}{1 - \boldsymbol{m}^2}\right) + \lambda^2[1 - Q(\boldsymbol{m})]\mathbf{I},$$

$$\ell_\varepsilon^-(\boldsymbol{m}) = \inf\{\lambda_{\min}(\boldsymbol{H}^-(\boldsymbol{u})) : \boldsymbol{u} \in (-1, 1)^n \cap \mathsf{B}_{\sqrt{\varepsilon n}}(\boldsymbol{m})\}.$$

We show the following lemma using the Sudakov–Fernique inequality, analogously to Lemma 4.9.

LEMMA 4.12. *Suppose $\lambda > 1$ and $\boldsymbol{x} = \mathbf{1}$. Define*

$$\begin{aligned} &H_\lambda^-(p, u; \alpha, \kappa, \gamma) \\ &\quad = [2\lambda^2 p^2 + \lambda^2 u^2 - 2\lambda^2(1 - q_\star)p^2/q_\star - \alpha u - \kappa p] + \lambda^2(1 - q_\star) + \gamma \\ &\qquad - \mathbb{E}_{m \sim \mu_\star}\left[\left(4\lambda^2(1 - p^2/q_\star) + (2z(m)p/q_\star + \alpha + \kappa m)^2\right)\Big/\left(\frac{4}{1 - m^2} - 4\gamma\right)\right], \end{aligned}$$

where $z(m) = \operatorname{arctanh} m - \lambda^2 q_\star + \lambda^2(1 - q_\star)m$. Then the statements of Lemma 4.9 hold also with $\ell_\varepsilon^+(\boldsymbol{m})$ and H_λ^+ replaced by $\ell_\varepsilon^-(\boldsymbol{m})$ and H_λ^-.

Now applying this result in the Kac–Rice upper bound of Lemma 4.8 for $\ell(\boldsymbol{m}, \boldsymbol{W}) = \ell_\varepsilon^-(\boldsymbol{m})$, we obtain that (33) also holds with high probability for $r = 1$ and sign $-$, implying Theorem 2.1(c).

The detailed proofs of this section are contained in Appendix B.4.

5. Convergence of optimization algorithms.

In this section, we describe the main ideas in the proofs of Theorems 2.3 and 2.4. It again suffices to show that the results hold with high probability conditional on $\boldsymbol{x} = \mathbf{1}$. The detailed proofs of this section are contained in Appendix C.

5.1. *Convergence of NGD with local initialization.*

Theorem 2.3 is a consequence of the local strong convexity of $\mathcal{F}_{\text{TAP}}$ established in Theorem 2.1(b) and the following local convergence result for the natural gradient algorithm (NGD).

LEMMA 5.1. *Fix any $\lambda > 1$, $t > 0$, and $\varepsilon \in (0, 1)$. Consider the event where $\boldsymbol{m}_\star$ in Theorem 2.1(a) exists and is unique up to sign, and $\|\boldsymbol{W}\|_{\text{op}} < 3$ and $\lambda_{\min}(n \cdot \nabla^2 \mathcal{F}_{\text{TAP}}(\boldsymbol{m})) > t$ for every $\boldsymbol{m} \in (-1, 1)^n \cap \mathsf{B}_{\sqrt{\varepsilon n}}(\boldsymbol{m}_\star)$. Consider any initialization $\boldsymbol{m}^0 = \tanh(\boldsymbol{h}^0)$ such that*

$$\mathcal{F}_{\text{TAP}}(\boldsymbol{m}^0) < \mathcal{F}_{\text{TAP}}(\boldsymbol{m}_\star) + t\varepsilon/8, \qquad \|\boldsymbol{m}^0 - \boldsymbol{m}_\star\|_2 < \sqrt{\varepsilon n}. \tag{34}$$

There exist $(\lambda, t, \varepsilon)$*-dependent constants* $C, \mu, \eta_0 > 0$ *such that if* (NGD) *with any step size* $\eta \in (0, \eta_0)$ *is initialized at* $\boldsymbol{m}^0$*, then on this event, for every* $k \geq 1$ *we have*

$$\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}^k) < \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}_\star) + C\left(1 + \frac{\|\operatorname{arctanh}(\boldsymbol{m}^0)\|_2}{\sqrt{n}}\right)(1 - \mu\eta)^k, \tag{35}$$

$$\|\boldsymbol{m}^k - \boldsymbol{m}_\star\|_2 < C\sqrt{n}\left(1 + \frac{\|\operatorname{arctanh}(\boldsymbol{m}^0)\|_2}{\sqrt{n}}\right)(1 - \mu\eta)^k \tag{36}$$

The proof of this lemma applies the mirror-descent form of NGD given in (10), together with an observation that on the above event, $\mathcal{F}_{\mathrm{TAP}}$ is strongly smooth and strongly convex over $(-1, 1)^n \cap \mathsf{B}_{\sqrt{\varepsilon n}}(\boldsymbol{m}_\star)$ relative to the prox function $-H(\boldsymbol{m})$, in the sense of [14, 80]

$$\mu \cdot \nabla^2\big(-H(\boldsymbol{m})\big) \preceq \nabla^2 \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) \preceq L \cdot \nabla^2\big(-H(\boldsymbol{m})\big)$$

for some constants $L, \mu > 0$. We may then adapt a convergence analysis of [80] to show that, for the above initialization, NGD with sufficiently small step size $\eta > 0$ must remain in this strongly convex neighborhood and exhibit the above linear convergence to $\boldsymbol{m}_\star$. For any $\lambda > 1$, the event in Lemma 5.1 holds with high probability by Theorem 2.1. The required initial condition (34) is also with high probability achieved by a sufficiently large iteration of AMP, as may be deduced from the AMP state evolution. Combined, this yields Theorem 2.3. The detailed proofs of Lemma 5.1 and Theorem 2.3 are contained in Appendix C.2.

5.2. *Convergence of NGD from spectral initialization.* For large λ, to show the result of Theorem 2.4(b) that NGD alone converges to $\pm\boldsymbol{m}_\star$ from a spectral initialization, recall the domain

$$\mathcal{S} = \{\boldsymbol{m} \in (-1, 1)^n : \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) < -\lambda^2/3\}$$

as defined in Corollary 2.2. For a parameter $q \in (0, 1)$, define the deterministic subset

$$\mathcal{M}_q = \{\boldsymbol{m} \in (-1, 1)^n : M(\boldsymbol{m}) > q\},$$

where recall $M(\boldsymbol{m}) = \boldsymbol{m}^\top \mathbf{1}/n$. We first establish the following more quantitative characterization of the landscape of $\mathcal{F}_{\mathrm{TAP}}$.

LEMMA 5.2. *Fix any integer* $a \geq 5$, *and set* $q = 1 - \lambda^{-a}$. *Suppose* $\boldsymbol{x} = \mathbf{1}$. *For a constant* $\lambda_0(a) > 0$, *if* $\lambda > \lambda_0(a)$, *then there are* (a, λ)*-dependent constants* $C, c, t > 0$ *such that with probability at least* $1 - Ce^{-cn}$:

(a) *Every point* $\boldsymbol{m} \in \mathcal{S} \setminus \mathcal{M}_q$ *satisfies*

$$\|\sqrt{n} \cdot \nabla \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m})\|_2^2 > t.$$

(b) *Every point* $\boldsymbol{m} \in \mathcal{S} \cap \mathcal{M}_q$ *satisfies*

$$n \cdot \nabla^2 \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) \succ \frac{1}{2}\operatorname{diag}\left(\frac{1}{1 - \boldsymbol{m}^2}\right) \succeq \frac{1}{2}\mathbf{I}.$$

Part (b) of this lemma is sufficient to imply Theorem 2.4(b) on the convergence of NGD: The initialization $\boldsymbol{m}^0 = \tanh(\boldsymbol{h}^0)$ defined by (SI) will belong to the region $\mathcal{S} \cap \mathcal{M}_q$ with high probability, so Lemma 5.1 may again be used to show linear convergence to $\pm\boldsymbol{m}_\star$. The detailed proof of Lemma 5.2 is contained in Appendix C.1 and the detailed proof of Theorem 2.4(b) is contained in Appendix C.2.

5.3. *Convergence of AMP from spectral initialization.* For Theorem 2.4(a) on the convergence of AMP, we directly prove contractivity of the map T_{AMP} defined in (7) locally near $\boldsymbol{m}_\star$, in a parameterization by coordinates $\boldsymbol{p}$ that lie "between" $\boldsymbol{h}$ and $\boldsymbol{m} = \tanh(\boldsymbol{h})$: Define two strictly increasing functions $\Gamma, \Lambda : \mathbb{R} \to \mathbb{R}$ as

$$\Gamma(h) = \int_0^h \sqrt{1 - \tanh(s)^2}\, \mathrm{d}s, \qquad \Lambda(p) = \tanh\bigl(\Gamma^{-1}(p)\bigr),$$

and consider $\boldsymbol{p} = \Gamma(\boldsymbol{h})$. Then $\boldsymbol{m} = \tanh(\boldsymbol{h}) = \Lambda(\boldsymbol{p})$. We write as shorthand

$$\frac{\mathrm{d}\boldsymbol{m}}{\mathrm{d}\boldsymbol{h}} = \mathrm{d}_h \tanh(\boldsymbol{h}), \qquad \frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}\boldsymbol{h}} = \mathrm{d}_h \Gamma(\boldsymbol{h}), \qquad \frac{\mathrm{d}\boldsymbol{m}}{\mathrm{d}\boldsymbol{p}} = \mathrm{d}_p \Lambda(\boldsymbol{p}),$$

where these are vectors in $\mathbb{R}^n$, and the derivatives are applied entry-wise. These definitions of Γ and Λ are designed so as to factor the identity $1 - \boldsymbol{m}^2 = \mathrm{d}\boldsymbol{m}/\mathrm{d}\boldsymbol{h}$ into the pair of identities

$$\sqrt{1 - \boldsymbol{m}^2} = \frac{\mathrm{d}\boldsymbol{m}}{\mathrm{d}\boldsymbol{p}} = \frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}\boldsymbol{h}}.$$

(This reparameterization by $\boldsymbol{p}$ may seem mysterious, and is carefully chosen to precondition the Jacobian of the AMP map and enable an operator norm bound for this Jacobian. We provide a heuristic motivation for this reparametrization in Remark C.3 in Appendix C.3.)

The range of $\boldsymbol{p} = \Gamma(\boldsymbol{h})$ is the cube $\Omega^{(p)} = (-\pi/2, \pi/2)^n$. We denote the AMP map (7) in the $\boldsymbol{p}$-parameterization as $T_{\mathsf{AMP}}^{(p)} : \Omega^{(p)} \times \Omega^{(p)} \to \Omega^{(p)} \times \Omega^{(p)}$, defined by

$$T_{\mathsf{AMP}}^{(p)}(\boldsymbol{p}, \boldsymbol{p}_-) = (\Lambda \otimes \Lambda)^{-1} \circ T_{\mathsf{AMP}}\bigl((\Lambda \otimes \Lambda)(\boldsymbol{p}, \boldsymbol{p}_-)\bigr).$$

Thus, reparameterizing by $\boldsymbol{p}^k = \Gamma(\boldsymbol{h}^k)$, the AMP iterations (AMP) take the form $(\boldsymbol{p}^{k+1}, \boldsymbol{p}^k) = T_{\mathsf{AMP}}^{(p)}(\boldsymbol{p}^k, \boldsymbol{p}^{k-1})$.

LEMMA 5.3. *Consider the metric* $\|(\boldsymbol{p}, \boldsymbol{p}')\|_\lambda = \|\boldsymbol{p}\|_2 + \lambda^{-1/5}\|\boldsymbol{p}'\|_2$. *Fix* $q = 1 - \lambda^{-5}$ *and* $\boldsymbol{x} = \mathbf{1}$. *For an absolute constant* $\lambda_0 > 0$, *suppose* $\lambda > \lambda_0$. *Then with probability at least* $1 - Ce^{-cn}$ *for* λ*-dependent constants* $C, c > 0$, *the following holds*: *If there exists a critical point* $\boldsymbol{m}_\star \in \mathcal{M}_q$ *of* $\mathcal{F}_{\mathrm{TAP}}$, *then for* $\boldsymbol{p}_\star = \Lambda^{-1}(\boldsymbol{m}_\star)$, *any* $\boldsymbol{p}, \boldsymbol{p}_- \in \mathsf{B}_{\lambda^{-7}\sqrt{n}}(\boldsymbol{p}_\star) \cap \Omega^{(p)}$, *and* $(\boldsymbol{p}_+, \boldsymbol{p}) = T_{\mathsf{AMP}}^{(p)}(\boldsymbol{p}, \boldsymbol{p}_-)$, *we have* $\boldsymbol{p}_+ \in \mathsf{B}_{\lambda^{-7}\sqrt{n}}(\boldsymbol{p}_\star) \cap \Omega^{(p)}$ *and*

$$\bigl\|(\boldsymbol{p}_+, \boldsymbol{p}) - (\boldsymbol{p}_\star, \boldsymbol{p}_\star)\bigr\|_\lambda \le 2\lambda^{-1/5}\bigl\|(\boldsymbol{p}, \boldsymbol{p}_-) - (\boldsymbol{p}_\star, \boldsymbol{p}_\star)\bigr\|_\lambda. \tag{37}$$

The AMP state evolution guarantees that with probability approaching 1 as $n \to \infty$, $\boldsymbol{p}^{k-1}, \boldsymbol{p}^k \in \mathsf{B}_{\lambda^{-7}\sqrt{n}}(\boldsymbol{p}_\star)$ for a sufficiently large iteration k. Then the contractivity guaranteed in Lemma 5.3 implies Theorem 2.4(a). The detailed proofs of Lemma 5.3 and Theorem 2.4(a) are contained in Appendix C.3.

6. Discussion.

In this paper, we showed the local strong convexity of the TAP free energy for $\mathbb{Z}_2$-synchronization around its Bayes-optimal local minimizer, and studied the finite-n convergence of optimization algorithms for computing this minimizer. Numerical simulations confirm that the TAP free energy can be efficiently optimized, and that properties of its minimizer are robust to model misspecification. Our results provide theoretical justification for using the TAP free energy to perform variational inference in this model.

In terms of proof techniques, our work introduced a method of using the Kac–Rice formula to study the local geometry of a nonconvex function around its critical points. Some intermediate results in the proof, for example the convergence of the empirical distribution of coordinates of the TAP minimizer, are of independent interest. We note that an analogous

TAP free energy function may be defined in broader contexts, such as for spiked matrix models with more general priors or for linear and generalized linear models, and some of our techniques may be useful also for analyzing the local geometries of these TAP free energy functions around their informative fixed points. However, the Rademacher $\{+1, -1\}$ prior in $\mathbb{Z}_2$-synchronization does have several conveniences, including a fixed second moment, an explicit form for both its entropy and its posterior mean function, and a unique fixed-point for the equation (11) that defines $q_\star$. Analyses of models having priors that lack these properties would have additional technical hurdles, and we leave the exploration of such extensions to future work.

Finally, we proved the finite-n convergence of a well-studied AMP algorithm for this problem, which is not implied by analysis of the AMP state evolution alone. Our proof of this result required sufficiently large λ, but we conjecture that the result holds for any $\lambda > 1$. This conjecture is supported by our numerical simulations and also by the stability of the AMP map around its fixed point, which indeed holds for any $\lambda > 1$. We leave this conjecture as an open question, and hope that the techniques developed in this paper can perhaps inspire a proof.

Funding. M. Celentano is supported by the Miller Institute for Basic Research in Science, University of California Berkeley. Z. Fan is supported in part by NSF Grants DMS-1916198 and DMS-2142476. S. Mei is supported in part by NSF Grant DMS-2210827.

SUPPLEMENTARY MATERIAL

Supplementary appendices (DOI: 10.1214/23-AOS2257SUPP; .pdf). Proofs of the main results are contained in the supplementary appendices.

REFERENCES

[1] ALQUIER, P. and RIDGWAY, J. (2020). Concentration of tempered posteriors and of their variational approximations. *Ann. Statist.* **48** 1475–1497. MR4124331 https://doi.org/10.1214/19-AOS1855

[2] AMARI, S.-I. (1998). Natural gradient works efficiently in learning. *Neural Comput.* **10** 251–276.

[3] AROUS, G. B., BOURGADE, P. and MCKENNA, B. (2021). Landscape complexity beyond invariance and the elastic manifold. Preprint. Available at arXiv:2105.05051.

[4] AUFFINGER, A. and BEN AROUS, G. (2013). Complexity of random smooth functions on the high-dimensional sphere. *Ann. Probab.* **41** 4214–4247. MR3161473 https://doi.org/10.1214/13-AOP862

[5] AUFFINGER, A., BEN AROUS, G. and ČERNÝ, J. (2013). Random matrices and complexity of spin glasses. *Comm. Pure Appl. Math.* **66** 165–201. MR2999295 https://doi.org/10.1002/cpa.21422

[6] AUFFINGER, A. and JAGANNATH, A. (2019). Thouless–Anderson–Palmer equations for generic p-spin glasses. *Ann. Probab.* **47** 2230–2256. MR3980920 https://doi.org/10.1214/18-AOP1307

[7] AUGERI, F. (2020). Nonlinear large deviation bounds with applications to Wigner matrices and sparse Erdős–Rényi graphs. *Ann. Probab.* **48** 2404–2448. MR4152647 https://doi.org/10.1214/20-AOP1427

[8] BAIK, J., BEN AROUS, G. and PÉCHÉ, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* **33** 1643–1697. MR2165575 https://doi.org/10.1214/009117905000000233

[9] BANDEIRA, A. S., CHEN, Y., LEDERMAN, R. R. and SINGER, A. (2020). Non-unique games over compact groups and orientation estimation in cryo-EM. *Inverse Probl.* **36** 064002, 39 pp. MR4105332 https://doi.org/10.1088/1361-6420/ab7d2c

[10] BARBIER, J., DIA, M., MACRIS, N., KRZAKALA, F., LESIEUR, T. and ZDEBOROVÁ, L. (2016). Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. In *Neural Information Processing Systems*.

[11] BASAK, A. and MUKHERJEE, S. (2017). Universality of the mean-field for the Potts model. *Probab. Theory Related Fields* **168** 557–600. MR3663625 https://doi.org/10.1007/s00440-016-0718-0

[12] BASKERVILLE, N. P., KEATING, J. P., MEZZADRI, F. and NAJNUDEL, J. (2021). The loss surfaces of neural networks with general activation functions. *J. Stat. Mech. Theory Exp.* **2021** Paper No. 064001, 71 pp. MR4309626 https://doi.org/10.1088/1742-5468/abfa1e

[13] BASKERVILLE, N. P., KEATING, J. P., MEZZADRI, F. and NAJNUDEL, J. (2022). A spin glass model for the loss surfaces of generative adversarial networks. *J. Stat. Phys.* **186** Paper No. 29, 45 pp. MR4368914 https://doi.org/10.1007/s10955-022-02875-w

[14] BAUSCHKE, H. H., BOLTE, J. and TEBOULLE, M. (2017). A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. *Math. Oper. Res.* **42** 330–348. MR3651994 https://doi.org/10.1287/moor.2016.0817

[15] BAYATI, M., LELARGE, M. and MONTANARI, A. (2015). Universality in polytope phase transitions and message passing algorithms. *Ann. Appl. Probab.* **25** 753–822. MR3313755 https://doi.org/10.1214/14-AAP1010

[16] BAYATI, M. and MONTANARI, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Inf. Theory* **57** 764–785. MR2810285 https://doi.org/10.1109/TIT.2010.2094817

[17] BECK, A. and TEBOULLE, M. (2003). Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.* **31** 167–175. MR1967286 https://doi.org/10.1016/S0167-6377(02)00231-6

[18] BELIUS, D. and KISTLER, N. (2019). The TAP-Plefka variational principle for the spherical SK model. *Comm. Math. Phys.* **367** 991–1017. MR3943486 https://doi.org/10.1007/s00220-019-03304-y

[19] BEN AROUS, G., MEI, S., MONTANARI, A. and NICA, M. (2019). The landscape of the spiked tensor model. *Comm. Pure Appl. Math.* **72** 2282–2330. MR4011861 https://doi.org/10.1002/cpa.21861

[20] BERTHIER, R., MONTANARI, A. and NGUYEN, P.-M. (2020). State evolution for approximate message passing with non-separable functions. *Inf. Inference* **9** 33–79. MR4079177 https://doi.org/10.1093/imaiai/iay021

[21] BICKEL, P., CHOI, D., CHANG, X. and ZHANG, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Statist.* **41** 1922–1943. MR3127853 https://doi.org/10.1214/13-AOS1124

[22] BINGHAM, E., CHEN, J. P., JANKOWIAK, M., OBERMEYER, F., PRADHAN, N., KARALETSOS, T., SINGH, R., SZERLIP, P., HORSFALL, P. et al. (2019). Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.* **20** 973–978.

[23] BLEI, D. M. (2012). Probabilistic topic models. *Commun. ACM* **55** 77–84.

[24] BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* **112** 859–877. MR3671776 https://doi.org/10.1080/01621459.2017.1285773

[25] BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.

[26] BOLTHAUSEN, E. (2014). An iterative construction of solutions of the TAP equations for the Sherrington–Kirkpatrick model. *Comm. Math. Phys.* **325** 333–366. MR3147441 https://doi.org/10.1007/s00220-013-1862-3

[27] BOLTHAUSEN, E. (2019). A Morita type proof of the replica-symmetric formula for SK. In *Statistical Mechanics of Classical and Disordered Systems*. *Springer Proc. Math. Stat.* **293** 63–93. Springer, Cham. MR4015008 https://doi.org/10.1007/978-3-030-29077-1_4

[28] BRAY, A. and MOORE, M. A. (1980). Metastable states in spin glasses. *J. Phys. C, Solid State Phys.* **13** L469.

[29] BRAY, A., MOORE, M. A. and YOUNG, A. P. (1984). Weighted averages of TAP solutions and Parisi's $q(x)$. *J. Phys. C, Solid State Phys.* **17** L155.

[30] CARBONETTO, P. and STEPHENS, M. (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Anal.* **7** 73–107. MR2896713 https://doi.org/10.1214/12-BA703

[31] CAVAGNA, A., GIARDINA, I., PARISI, G. and MÉZARD, M. (2003). On the formal equivalence of the TAP and thermodynamic methods in the SK model. *J. Phys. A* **36** 1175–1194. MR1960081 https://doi.org/10.1088/0305-4470/36/5/301

[32] CELENTANO, M. (2022). Sudakov–Fernique post-AMP, and a new proof of the local convexity of the TAP free energy. Preprint. Available at arXiv:2208.09550.

[33] CELENTANO, M., FAN, Z. and MEI, S. (2023). Supplement to "Local convexity of the TAP free energy and AMP convergence for $\mathbb{Z}_2$-synchronization." https://doi.org/10.1214/23-AOS2257SUPP

[34] CELENTANO, M., MONTANARI, A. and WEI, Y. (2020). The Lasso with general Gaussian designs with applications to hypothesis testing. Preprint. Available at arXiv:2007.13716.

[35] CHATTERJEE, S. (2010). Spin glasses and Stein's method. *Probab. Theory Related Fields* **148** 567–600. MR2678899 https://doi.org/10.1007/s00440-009-0240-8

[36] CHATTERJEE, S. and DEMBO, A. (2016). Nonlinear large deviations. *Adv. Math.* **299** 396–450. MR3519474 https://doi.org/10.1016/j.aim.2016.05.017

[37] CHEN, W.-K. and LAM, W.-K. (2021). Universality of approximate message passing algorithms. *Electron. J. Probab.* **26** Paper No. 36, 44 pp. MR4235487 https://doi.org/10.1214/21-EJP604

[38] CHEN, W.-K. and PANCHENKO, D. (2018). On the TAP free energy in the mixed p-spin models. *Comm. Math. Phys.* **362** 219–252. MR3833609 https://doi.org/10.1007/s00220-018-3143-7

[39] CHEN, W.-K., PANCHENKO, D. and SUBAG, E. (2018). Generalized TAP free energy. *Comm. Pure Appl. Math.*

[40] CHÉRIEF-ABDELLATIF, B.-E. (2019). Consistency of ELBO maximization for model selection. In *Symposium on Advances in Approximate Bayesian Inference*. *Proc. Mach. Learn. Res.* (*PMLR*) **96** 11–31. PMLR. MR3980812

[41] CRISANTI, A., LEUZZI, L., PARISI, G. and RIZZO, T. (2003). Complexity in the Sherrington–Kirkpatrick model in the annealed approximation. *Phys. Rev. B* **68** 174401.

[42] CRISANTI, A., LEUZZI, L. and RIZZO, T. (2005). Complexity in mean-field spin-glass models: Ising p-spin. *Phys. Rev. B* **71** 094202.

[43] DE DOMINICIS, C. and YOUNG, A. P. (1983). Weighted averages and order parameters for the infinite range Ising spin glass. *J. Phys. A* **16** 2063–2075. MR0712998

[44] DESHPANDE, Y., ABBE, E. and MONTANARI, A. (2017). Asymptotic mutual information for the balanced binary stochastic block model. *Inf. Inference* **6** 125–170. MR3671474 https://doi.org/10.1093/imaiai/iaw017

[45] DESHPANDE, Y. and MONTANARI, A. (2014). Information-theoretically optimal sparse PCA. In 2014 *IEEE International Symposium on Information Theory* 2197–2201. IEEE, New York.

[46] DING, J. and SUN, N. (2019). Capacity lower bound for the Ising perceptron. In *STOC'*19—*Proceedings of the* 51*st Annual ACM SIGACT Symposium on Theory of Computing* 816–827. ACM, New York. MR4003386 https://doi.org/10.1145/3313276.3316383

[47] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2009). Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci.* **106** 18914–18919.

[48] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2010). Message passing algorithms for compressed sensing: I. motivation and construction. In 2010 *IEEE Information Theory Workshop on Information Theory* (*ITW* 2010, *Cairo*) 1–5. IEEE, New York.

[49] DRAGOMIR, R. A., EVEN, M. and HENDRIKX, H. (2021). Fast stochastic Bregman gradient methods: Sharp analysis and variance reduction. In *International Conference on Machine Learning* 2815–2825. PMLR.

[50] DUDEJA, R., SEN, S. and LU, Y. M. (2022). Spectral universality of regularized linear regression with nearly deterministic sensing matrices. Preprint. Available at arXiv:2208.02753.

[51] EL ALAOUI, A., MONTANARI, A. and SELLKE, M. (2022). Sampling from the Sherrington–Kirkpatrick Gibbs measure via algorithmic stochastic localization. In 2022 *IEEE* 63*rd Annual Symposium on Foundations of Computer Science—FOCS* 2022 323–334. IEEE Computer Soc., Los Alamitos, CA. MR4537214

[52] ELDAN, R. (2018). Gaussian-width gradient complexity, reverse log-Sobolev inequalities and nonlinear large deviations. *Geom. Funct. Anal.* **28** 1548–1596. MR3881829 https://doi.org/10.1007/s00039-018-0461-z

[53] FAN, Z., MEI, S. and MONTANARI, A. (2021). TAP free energy, spin glasses and variational inference. *Ann. Probab.* **49** 1–45. MR4203332 https://doi.org/10.1214/20-AOP1443

[54] FAN, Z. and WU, Y. (2021). The replica-symmetric free energy for Ising spin glasses with orthogonally invariant couplings. Preprint. Available at arXiv:2105.02797.

[55] FENG, O. Y., VENKATARAMANAN, R., RUSH, C., SAMWORTH, R. J. et al. (2022). A unifying tutorial on approximate message passing. *Found. Trends Mach. Learn.* **15** 335–536.

[56] FERNIQUE, X. (1975). Regularité des trajectoires des fonctions aléatoires gaussiennes. In *École D'Été de Probabilités de Saint-Flour, IV*-1974. *Lecture Notes in Math.* **480** 1–96. Springer, Berlin. MR0413238

[57] FYODOROV, Y. V. (2004). Complexity of random energy landscapes, glass transition, and absolute value of the spectral determinant of random matrices. *Phys. Rev. Lett.* **92** 240601, 4 pp. MR2115095 https://doi.org/10.1103/PhysRevLett.92.240601

[58] GAUCHER, S. and KLOPP, O. (2021). Optimality of variational inference for stochasticblock model with missing links. *Adv. Neural Inf. Process. Syst.* **34** 19947–19959.

[59] GHORBANI, B., JAVADI, H. and MONTANARI, A. (2019). An instability in variational inference for topic models. In *International Conference on Machine Learning* 2221–2231. PMLR.

[60] GORDON, Y. (1985). Some inequalities for Gaussian processes and applications. *Israel J. Math.* **50** 265–289. MR0800188 https://doi.org/10.1007/BF02759761

[61] GUTMAN, D. H. and PEÑA, J. F. (2023). Perturbed Fenchel duality and first-order methods. *Math. Program.* **198** 443–469. MR4550955 https://doi.org/10.1007/s10107-022-01779-7

[62] HALL, P., ORMEROD, J. T. and WAND, M. P. (2011). Theory of Gaussian variational approximation for a Poisson mixed model. *Statist*. *Sinica* **21** 369–389. MR2796867

[63] HALL, P., PHAM, T., WAND, M. P. and WANG, S. S. J. (2011). Asymptotic normality and valid inference for Gaussian variational approximation. *Ann*. *Statist*. **39** 2502–2532. MR2906876 https://doi.org/10.1214/11-AOS908

[64] HAN, Q. and SHEN, Y. (2022). Universality of regularized regression estimators in high dimensions. Preprint. Available at arXiv:2206.07936.

[65] HANZELY, F., RICHTÁRIK, P. and XIAO, L. (2021). Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. *Comput*. *Optim*. *Appl*. **79** 405–440. MR4254648 https://doi.org/10.1007/s10589-021-00273-8

[66] HU, H. and LU, Y. M. (2020). Universality laws for high-dimensional learning with random features. Preprint. Available at arXiv:2009.07669.

[67] JAIN, V., KOEHLER, F. and MOSSEL, E. (2018). The mean-field approximation: Information inequalities, algorithms, and complexity. In *Conference on Learning Theory* 1326–1347. PMLR.

[68] JAVANMARD, A. and MONTANARI, A. (2013). State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Inf*. *Inference* **2** 115–144. MR3311445 https://doi.org/10.1093/imaiai/iat004

[69] JAVANMARD, A., MONTANARI, A. and RICCI-TERSENGHI, F. (2016). Phase transitions in semidefinite relaxations. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **113** E2218–E2223. MR3494080 https://doi.org/10.1073/pnas.1523097113

[70] JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann*. *Statist*. **29** 295–327. MR1863961 https://doi.org/10.1214/aos/1009210544

[71] JORDAN, M. I., GHAHRAMANI, Z., JAAKKOLA, T. S. and SAUL, L. K. (1999). An introduction to variational methods for graphical models. *Mach*. *Learn*. **37** 183–233.

[72] KABASHIMA, Y. (2003). A CDMA multiuser detection algorithm on the basis of belief propagation. *J*. *Phys*. *A* **36** 11111–11121. MR2025247 https://doi.org/10.1088/0305-4470/36/43/030

[73] KAHANE, J.-P. (1986). Une inégalité du type de Slepian et Gordon sur les processus gaussiens. *Israel J*. *Math*. **55** 109–110. MR0858463 https://doi.org/10.1007/BF02772698

[74] KRZAKALA, F., MANOEL, A., TRAMEL, E. W. and ZDEBOROVÁ, L. (2014). Variational free energies for compressed sensing. In 2014 *IEEE International Symposium on Information Theory* 1499–1503. IEEE, New York.

[75] KRZAKALA, F., XU, J. and ZDEBOROVÁ, L. (2016). Mutual information in rank-one matrix estimation. In 2016 *IEEE Information Theory Workshop* (*ITW*) 71–75. IEEE, New York.

[76] LELARGE, M. and MIOLANE, L. (2019). Fundamental limits of symmetric low-rank matrix estimation. *Probab*. *Theory Related Fields* **173** 859–929. MR3936148 https://doi.org/10.1007/s00440-018-0845-x

[77] LESIEUR, T., KRZAKALA, F. and ZDEBOROVÁ, L. (2015). Phase transitions in sparse PCA. In 2015 *IEEE International Symposium on Information Theory* (*ISIT*) 1635–1639. IEEE, New York.

[78] LI, G. and WEI, Y. (2022). A non-asymptotic framework for approximate message passing in spiked models. Preprint. Available at arXiv:2208.03313.

[79] LIANG, P., PETROV, S., JORDAN, M. I. and KLEIN, D. (2007). The infinite PCFG using hierarchical Dirichlet processes. In *Proceedings of the* 2007 *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (*EMNLP-CoNLL*) 688–697.

[80] LU, H., FREUND, R. M. and NESTEROV, Y. (2018). Relatively smooth convex optimization by first-order methods, and applications. *SIAM J*. *Optim*. **28** 333–354. MR3759881 https://doi.org/10.1137/16M1099546

[81] MAILLARD, A., AROUS, G. B. and BIROLI, G. (2020). Landscape complexity for the empirical risk of generalized linear models. In *Mathematical and Scientific Machine Learning* 287–327. PMLR.

[82] MEI, S., BAI, Y. and MONTANARI, A. (2018). The landscape of empirical risk for nonconvex losses. *Ann*. *Statist*. **46** 2747–2774. MR3851754 https://doi.org/10.1214/17-AOS1637

[83] MINKA, T., WINN, J., GUIVER, J., WEBSTER, S., ZAYKOV, Y., YANGEL, B., SPENGLER, A. and BRONSKILL, J. (2014). Infer NET 2.6. Microsoft Research Cambridge. Available at http://research.microsoft.com/infernet.

[84] MINKA, T. P. (2001). A family of algorithms for approximate Bayesian inference, PhD thesis, Massachusetts Institute of Technology. MR2717007

[85] MIOLANE, L. and MONTANARI, A. (2021). The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning. *Ann*. *Statist*. **49** 2313–2335. MR4319252 https://doi.org/10.1214/20-aos2038

[86] MONTANARI, A. (2012). Graphical models concepts in compressed sensing. In *Compressed Sensing* 394–438. Cambridge Univ. Press, Cambridge. MR2963574 https://doi.org/10.1017/CBO9780511794308.010

[87] MONTANARI, A. and NGUYEN, P.-M. (2017). Universality of the elastic net error. In 2017 *IEEE International Symposium on Information Theory* (*ISIT*) 2338–2342. IEEE, New York.

[88] MONTANARI, A. and RICHARD, E. (2016). Non-negative principal component analysis: Message passing algorithms and sharp asymptotics. *IEEE Trans*. *Inf*. *Theory* **62** 1458–1484. MR3472260 https://doi.org/10.1109/TIT.2015.2457942

[89] MONTANARI, A. and SAEED, B. (2022). Universality of empirical risk minimization. Preprint. Available at arXiv:2202.08832.

[90] MONTANARI, A. and SEN, S. (2016). Semidefinite programs on sparse random graphs and their application to community detection. In *STOC'*16*—Proceedings of the* 48*th Annual ACM SIGACT Symposium on Theory of Computing* 814–827. ACM, New York. MR3536616 https://doi.org/10.1145/2897518.2897548

[91] MONTANARI, A. and VENKATARAMANAN, R. (2021). Estimation of low-rank matrices via approximate message passing. *Ann*. *Statist*. **49** 321–345. MR4206680 https://doi.org/10.1214/20-AOS1958

[92] MUKHERJEE, S. S., SARKAR, P., WANG, Y. and YAN, B. (2018). Mean field for the stochastic blockmodel: Optimization landscape and convergence issues. *Adv*. *Neural Inf*. *Process*. *Syst*. **31**.

[93] NEMIROVSKY, A. S. and YUDIN, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. *Wiley-Interscience Series in Discrete Mathematics*. Wiley, New York. MR0702836

[94] OYMAK, S., THRAMPOULIDIS, C. and HASSIBI, B. (2013). The squared-error of generalized lasso: A precise analysis. In 2013 51*st Annual Allerton Conference on Communication*, *Control*, *and Computing* (*Allerton*) 1002–1009. IEEE, New York.

[95] PEARL, J. (1982). Reverend Bayes on inference engines: A distributed hierarchical approach. In *Proceedings of the Second AAAI Conference on Artificial Intelligence* 133–136.

[96] PÉCHÉ, S. (2006). The largest eigenvalue of small rank perturbations of Hermitian random matrices. *Probab*. *Theory Related Fields* **134** 127–173. MR2221787 https://doi.org/10.1007/s00440-005-0466-z

[97] PLEFKA, T. (1982). Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model. *J*. *Phys*. *A* **15** 1971–1978. MR0663708

[98] PLUMMER, S., PATI, D. and BHATTACHARYA, A. (2020). Dynamics of coordinate ascent variational inference: A case study in 2D Ising models. *Entropy* **22** Paper No. 1263, 33 pp. MR4222066 https://doi.org/10.3390/e22111263

[99] QIU, J. and SEN, S. (2022). The TAP free energy for high-dimensional linear regression. Preprint. Available at arXiv:2203.07539.

[100] RAJ, A., STEPHENS, M. and PRITCHARD, J. K. (2014). fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* **197** 573–589. https://doi.org/10.1534/genetics.114.164350

[101] RANGAN, S. (2011). Generalized approximate message passing for estimation with random linear mixing. In 2011 *IEEE International Symposium on Information Theory Proceedings* 2168–2172. IEEE, New York.

[102] RANGAN, S. and FLETCHER, A. K. (2012). Iterative estimation of constrained rank-one matrices in noise. In 2012 *IEEE International Symposium on Information Theory Proceedings* 1246–1250. IEEE, New York.

[103] RANGAN, S., FLETCHER, A. K., SCHNITER, P. and KAMILOV, U. S. (2017). Inference for generalized linear models via alternating directions and Bethe free energy minimization. *IEEE Trans*. *Inf*. *Theory* **63** 676–697. MR3599966 https://doi.org/10.1109/TIT.2016.2619373

[104] RANGAN, S., SCHNITER, P. and FLETCHER, A. K. (2019). Vector approximate message passing. *IEEE Trans*. *Inf*. *Theory* **65** 6664–6684. MR4009222 https://doi.org/10.1109/TIT.2019.2916359

[105] RAY, K. and SZABÓ, B. (2022). Variational Bayes for high-dimensional linear regression with sparse priors. *J*. *Amer*. *Statist*. *Assoc*. **117** 1270–1281. MR4480711 https://doi.org/10.1080/01621459.2020.1847121

[106] RUSH, C. and VENKATARAMANAN, R. (2016). Finite-sample analysis of approximate message passing. In 2016 *IEEE International Symposium on Information Theory* (*ISIT*) 755–759. https://doi.org/10.1109/ISIT.2016.7541400

[107] RUSH, C. and VENKATARAMANAN, R. (2018). Finite sample analysis of approximate message passing algorithms. *IEEE Trans*. *Inf*. *Theory* **64** 7264–7286. MR3876443 https://doi.org/10.1109/TIT.2018.2816681

[108] SAADE, A., KRZAKALA, F. and ZDEBOROVÁ, L. (2014). Spectral clustering of graphs with the Bethe Hessian. In *Neural Information Processing Systems*.

[109] SINGER, A. (2011). Angular synchronization by eigenvectors and semidefinite programming. *Appl*. *Comput*. *Harmon*. *Anal*. **30** 20–36. MR2737931 https://doi.org/10.1016/j.acha.2010.02.001

[110] SLEPIAN, D. (1962). The one-sided barrier problem for Gaussian noise. *Bell Syst*. *Tech*. *J*. **41** 463–501. MR0133183 https://doi.org/10.1002/j.1538-7305.1962.tb02419.x

[111] STOJNIC, M. (2013). A framework to characterize performance of lasso algorithms. Preprint. Available at arXiv:1303.7291.

[112] SUBAG, E. (2017). The complexity of spherical p-spin models—A second moment approach. *Ann. Probab.* **45** 3385–3450. MR3706746 https://doi.org/10.1214/16-AOP1139

[113] SUBAG, E. (2021). The free energy of spherical pure p-spin models—Computation from the TAP approach. Preprint. Available at arXiv:2101.04352.

[114] SUDAKOV, V. N. (1971). Gaussian random processes, and measures of solid angles in Hilbert space. *Dokl. Akad. Nauk SSSR* **197** 43–45. MR0288832

[115] SUDAKOV, V. N. (1979). Geometric problems in the theory of infinite-dimensional probability distributions. *Proc. Steklov Inst. Math.* **2** i–v, 1–178. Cover to cover translation of Trudy Mat. Inst. Steklov **141** (1976). MR0530375

[116] SUN, J., QU, Q. and WRIGHT, J. (2018). A geometric analysis of phase retrieval. *Found. Comput. Math.* **18** 1131–1198. MR3857907 https://doi.org/10.1007/s10208-017-9365-9

[117] TALAGRAND, M. (2011). *Mean Field Models for Spin Glasses. Volume I: Basic Examples. Ergebnisse der Mathematik und Ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics* [*Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics*] **54**. Springer, Berlin. MR2731561 https://doi.org/10.1007/978-3-642-15202-3

[118] THOULESS, D. J., ANDERSON, P. W. and PALMER, R. G. (1977). Solution of 'solvable model of a spin glass'. *Philos. Mag.* **35** 593–601.

[119] THRAMPOULIDIS, C., OYMAK, S. and HASSIBI, B. (2015). Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory* 1683–1709. PMLR.

[120] TRAN, D., KUCUKELBIR, A., DIENG, A. B., RUDOLPH, M., LIANG, D. and BLEI, D. M. (2016). Edward: A library for probabilistic modeling, inference, and criticism. Preprint. Available at arXiv:1610.09787.

[121] WAINWRIGHT, M. J. and JORDAN, M. I. (2008). *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers, Hanover.

[122] WANG, T., ZHONG, X. and FAN, Z. (2022). Universality of approximate message passing algorithms and tensor networks. Preprint. Available at arXiv:2206.13037.

[123] WANG, Y. and BLEI, D. M. (2019). Frequentist consistency of variational Bayes. *J. Amer. Statist. Assoc.* **114** 1147–1161. MR4011769 https://doi.org/10.1080/01621459.2018.1473776

[124] YAN, J. (2020). Nonlinear large deviations: Beyond the hypercube. *Ann. Appl. Probab.* **30** 812–846. MR4108123 https://doi.org/10.1214/19-AAP1516

[125] YANG, Y., PATI, D. and BHATTACHARYA, A. (2020). α-variational inference with statistical guarantees. *Ann. Statist.* **48** 886–905. MR4102680 https://doi.org/10.1214/19-AOS1827

[126] YEDIDIA, J. S., FREEMAN, W. T. and WEISS, Y. (2003). Understanding belief propagation and its generalizations. In *Exploring Artificial Intelligence in the New Millennium* **8** 236–239.

[127] ZHANG, A. Y. and ZHOU, H. H. (2020). Theoretical and computational guarantees of mean field variational inference for community detection. *Ann. Statist.* **48** 2575–2598. MR4152113 https://doi.org/10.1214/19-AOS1898

[128] ZHANG, F. and GAO, C. (2020). Convergence rates of variational posterior distributions. *Ann. Statist.* **48** 2180–2207. MR4134791 https://doi.org/10.1214/19-AOS1883