Proof of The TAP Free Energy for High-Dimensional Linear Regression with Spherical Priors at All Temperatures

Zhiyuan Yu
UIUC

Jingbo Liu
UIUC

Abstract

Approximate inference is central to Bayesian learning, with variational inference (VI) providing a scalable framework for posterior approximation. While mean-field VI often fails in high dimensions, the more refined Bethe approximation, equivalent to the Thouless-Anderson-Palmer (TAP) free energy in statistical physics, has long been conjectured to capture Bayes-optimal behavior. We prove that the TAP formula holds for Bayesian linear regression with a uniform spherical prior at all noise levels ($\Delta > 0$), extending the result of Qiu and Sen (2023) in the high-noise regime. Our argument constructs a ridge regression functional that dominates the TAP free energy, yielding the first rigorous analysis of the global optimizer of the non-concave TAP functional for a planted inference model at an arbitrary noise level. This verifies that TAP, rather than mean-field, is the correct variational description in this setting.

1 INTRODUCTION

Variational inference (VI) has become a central tool in statistics and machine learning for approximating intractable posteriors (Wainwright et al., 2008). While classical VI is usually formulated in terms of mean-field approximations,recent statistical results indicate that naive mean-field VI can be asymptotically consistent only under restrictive conditions (Mukherjee and Sen, 2022), and may fail in genuinely high-dimensional regimes, motivating the study of refined free energies. A more general principle is given by the Bethe free energy, originally developed in statistical physics (Bethe, 1935). The Bethe free energy is a variational functional whose stationary points coincide with the fixed points of the belief propagation algorithm (Yedidia et al., 2001, 2003). Recent work has provided rigorous analyses of the convergence of belief propagation to Bethe free energy stationary points (Koehler, 2019). In spin glass theory, the corresponding variational principle is the Thouless–Anderson–Palmer (TAP) free energy (Thouless et al., 1977). Their connection has also been established by Welling and Teh (2013), who demonstrate that in binary networks the Bethe free energy agrees with the TAP free energy up to second order in the weights. In high-dimensional inference, Krzakala et al. (2014) formulated the functional under the name Bethe free energy, while Qiu and Sen (2023) analyzed the same functional under the TAP terminology in their rigorous study of linear regression.

Recent progress has advanced the theoretical understanding of the TAP free energy. Qiu and Sen (2023) derived asymptotic formulas in certain high-temperature regimes, while Celentano et al. (2023) analyzed the local geometry of the TAP functional, establishing that the fixed points of Approximate Message Passing (AMP) correspond to approximate local maximizers. Nevertheless, these results do not provide a complete characterization of the variational landscape. This naturally leads to the central question:

> *How to characterize the global maximizers of the TAP free energy?*

Addressing this question is not only of intrinsic interest in statistical physics and Bayesian inference, but also has implications for modern machine learning. In particular, whether the global optimizer of the TAP free energy correctly characterizes posterior means is an important question, especially due to its implications for score-based generative modeling (Song et al., 2021; Montanari and Wu, 2023; Cui et al., 2024). Thus, clarifying the behavior of TAP maximizers provides both a rigorous foundation for classical inference methods and contemporary relevance for generative modeling.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

The Thouless–Anderson–Palmer (TAP) free energy provides a refined variational principle that incorporates Onsager correction terms (Thouless et al., 1977; Mézard et al., 1987). Unlike purely algorithmic approaches, TAP defines an explicit objective that can be optimized by gradient-based methods, which is often regarded as providing greater robustness compared to AMP (El Alaoui et al., 2022). From a statistical perspective, global maximizers of TAP correspond to centers of high-probability states, thereby offering a geometric characterization of posterior clustering (Subag, 2017; Mézard et al., 1987; Fan et al., 2021). From a mathematical perspective, TAP free energies have become fundamental objects of study in probability theory, leading to rigorous formulations in spin glass models (Chen et al., 2023; Subag, 2023) and underpinning major advances such as the resolution of the Ising perceptron capacity problem (Huang, 2024) and Talagrand's operator norm conjecture (Brennecke et al., 2023). These developments demonstrate the broad significance of TAP as a unifying framework in statistical physics, Bayesian inference, and modern machine learning.

Despite these advances, in Bayesian linear regression an essential step remains to be established: whether the global maximum of the TAP functional converges to the true free energy. Constructing this equality would be an essential step toward showing that the global maximizer coincides with the posterior mean. Fixed point calculations of (Krzakala et al., 2014) suggest a positive answer (see Conjecture 1 below), but a rigorous proof is difficult because the TAP functional is non-concave and may admit spurious stationary points.

A first step towards rigorously addressing this open problem was taken by Qiu and Sen (2023), who focused on the high-temperature regime $\Delta > \Delta_0$ (for a certain threshold Δ_0) under uniform spherical priors. Consider a Gaussian random linear estimation framework, with data points $(y_i, x_i) : 1 \le i \le n$, where $\mathbf{y} = (y_i)_{i=1}^N \in \mathbb{R}^N$ are responses and $\mathbf{x} = (x_{ij})_{i=1,j=1}^{N,p} \in \mathbb{R}^{N\times p}$ are feature vectors. The model assumes:

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \sqrt{\Delta}\mathbf{z}, \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the coefficient vector and $\mathbf{z} = (z_i)_{i=1}^N \sim \mathcal{N}(0, I_N)$ are independent and identically distributed errors. Using a prior distribution π for $\boldsymbol{\beta}$, the posterior distribution is constructed as follows:

$$\frac{\mathrm{d}P_{\boldsymbol{\beta}|\mathbf{y},\mathbf{x}}}{\mathrm{d}\pi}(\boldsymbol{\beta}|\mathbf{y},\mathbf{x}) = \frac{1}{\mathcal{Z}_p} e^{-\frac{1}{2\Delta}\|\mathbf{y}-\mathbf{x}\boldsymbol{\beta}\|^2}. \tag{2}$$

The posterior normalization $\mathcal{Z}_p$ is also called the partition function in statistical physics. We also refer to

$$F_p = \frac{1}{p}\ln(\mathcal{Z}_p) \tag{3}$$

as the free energy. While Qiu and Sen (2023) showed that the TAP approximation correctly characterizes F_p when π is a uniform spherical distribution and $\Delta > \Delta_0$, extending its validity to the low-temperature regime has remained an open question:

Conjecture 1. (Krzakala et al., 2014; Qiu and Sen, 2023) *For Bayesian linear regression with a uniform prior on the sphere $\mathbb{S}^{p-1}(\sqrt{p})$, the TAP free energy provides an asymptotically exact characterization of the true free energy in the high-dimensional limit, for all noise levels $\Delta > 0$. That is,*

$$\lim_{p\to\infty}\frac{1}{p}\log\mathcal{Z}_p = \lim_{p\to\infty}\max_{\|a\|\le\sqrt{p}}\{-\frac{1}{2\Delta p}\|\mathbf{y}-\mathbf{x}a\|^2 - \frac{\alpha}{2}\ln(1+\frac{1-\|a\|^2/p}{\Delta\alpha}) + \frac{1}{2}\ln(1-\frac{\|a\|^2}{p})\}. \tag{4}$$

Throughout this paper, we refer to the quantity in (4) as the TAP free energy, which corresponds to the log-partition function or the free entropy in the statistical physics literature, up to a sign.

The spherical prior assumption enables the geometric arguments of Qiu and Sen (2023), while a similar analysis for spherical spin models was previously provided by Belius (2022). In addition, variants of spherical or normalized-signal models arise in applied contexts such as signal processing and compressed sensing on spherical domains (Cornelius et al., 2016) underscoring the relevance of spherical assumptions beyond purely theoretical considerations. In the spherical case, the posterior distribution is concentrated in a sphere cap, and the high-temperature assumption is necessary for the arguments of Qiu and Sen (2023) in justifying that the free energy in *any* sphere cap is that of a pure state. It is worth noting that for spherical Sherrington–Kirkpatrick(SK) models, the TAP expression was established for any temperature, if the TAP optimization is constrained to satisfy the so-called Plefka condition to ensure high-temperature phase of the sphere cap (Belius, 2022), but it is not obvious whether a similar condition is needed for the Bayes regression problem since $\mathbf{y}$ is a "planted" signal. If $\Delta < \Delta_0$, then some large sphere caps are not in a pure state, but these large spheres may not be relevant for the maximizers of the TAP free energy, hence the validity of the TAP free energy representation is elusive. A more recent work Celentano et al. (2023) established a version of Conjecture 1 for i.i.d. priors, by showing that there exists a local maximum value of the TAP free energy that converges to the free energy. This does

not establish Conjecture 1, because there may exist local maximizers of the TAP free energy that are not global. Indeed, we provide an example in Appendix A where the TAP function is non-concave, so that spurious local maximizers is possible.

In this work, we prove that the TAP approximation is valid for all $\Delta > 0$ under uniform spherical priors, establishing Conjecture 1. As noted before, the key challenge in establishing Conjecture 1 is to exclude the possibility of a larger global maximum value. While the TAP free energy functional may not be globally concave, we choose a surrogate functional based on ridge regression that can be shown to globally dominate it. The surrogate functional is globally concave, so that its global maximum can be easily controlled. Another key ingredient of our proof is the replica symmetric (RS) free energy expression established by Reeves and Pfister (2019) and Barbier et al. (2020). Our result shows that, unlike in Belius (2022), there is no need to impose the Plefka condition that lower bounds $\|a\|$, for TAP optimization in the Bayes-optimal setting ($\mathbf{y}$ is generated by signal from a matching prior), which aligns with the current practice. Our result also implies that any near global maximizer of the TAP free energy must be close to the solution of a certain ridge regression.

The remainder of this paper is structured as follows: Section 2 introduces the model setup, notations, and defines the TAP free energy in the context of spherical priors. Section 3 presents our main result, which establishes the asymptotic equivalence between the TAP free energy and the true free energy. Section 4 provides a detailed proof, which proceeds by establishing a Gaussian-spherical equivalence, analyzing the ridge regression surrogate, and connecting it to the RS potential. Section 5, presents numerical simulations that empirically investigate the geometry of the TAP free energy landscape. Section 6 concludes with a discussion of implications and directions for future work.

2 PRELIMINARIES

We assume that π is a uniform prior on $S_{p-1}(\sqrt{p}) := \left\{\beta \in \mathbb{R}^p : \|\beta\| = \sqrt{p}\right\}$. The matrix $\mathbf{x}$ has i.i.d. Gaussian entries $x_{i,j} \sim \mathcal{N}(0, \frac{1}{N})$, so that its operator norm is $O(1)$. The measurement rate N/p is denoted by $\alpha \in (0, \infty)$, which is assumed to be a constant. Therefore, the model is characterized by two parameters, Δ and α. A random variable converging to zero in probability is denoted by $o_P(1)$.

2.1 TAP Free Energy

Under the aforementioned settings, let a be a vector in S_{p-1} with the norm $\|a\|^2 = pq$. We can then define the following TAP function for a, which was derived by Qiu and Sen (2023):

$$f_{\mathrm{TAP}}(a) = -\frac{1}{2\Delta p}\|\mathbf{y} - \mathbf{x}a\|^2 - \frac{\alpha}{2}\ln(1 + \frac{1-q}{\Delta\alpha}) + \frac{1}{2}\ln(1-q). \tag{5}$$

The TAP free energy is subsequently obtained by evaluating the supremum of f_{TAP}, or equivalently

$$F_{TAP} = \sup_{\|a\|^2 \le p} f_{\mathrm{TAP}}(a). \tag{6}$$

In fact, Krzakala et al. (2014) derived a general version of (5) for i.i.d. priors (as we will see, spherical prior is asymptotically equivalent to i.i.d. Gaussian prior). They showed that the stationary points of f_{TAP} coincides with the fixed points of the AMP algorithm, suggesting the conjecture that F_{TAP} is asymptotically equivalent to the true free energy. Qiu and Sen (2023) established that $F_{TAP} - F_p$ converges in probability to 0 as $p \to \infty$, provided that Δ is sufficiently large. And in this work, we will extend their result by proving that this convergence holds for any $\Delta > 0$, thereby removing the restriction on Δ.

2.2 RS Formula

We mostly follow the presentation of Barbier et al. (2020), with some adaptations of their notations to be consistent with (1). Consider the model in equation (1), where the prior distribution of $\boldsymbol{\beta}$ is given by $P_0^{\otimes p}$. We define

$$\Sigma(E;\Delta)^{-2} := \frac{\alpha}{\alpha\Delta + E}. \tag{7}$$

Let $i(\tilde{S};\tilde{Y})$ denote the mutual information (MI) for the denoising model $\tilde{y} = \tilde{s} + \tilde{z}\Sigma$, where $\tilde{s} \sim P_0$ and $\tilde{z} \sim \mathcal{N}(0,1)$. The corresponding expression for $i(\tilde{S};\tilde{Y})$ is given by:

$$i(\tilde{S};\tilde{Y}) = -\mathbb{E}_{\tilde{S},\tilde{Z}}\left[\ln \mathbb{E}_{\tilde{\beta}}\left[\exp\left(-\frac{(\tilde{\beta} - (\tilde{S} + \tilde{Z}\Sigma))^2}{2\Sigma^2}\right)\right]\right] - \frac{1}{2}. \tag{8}$$

The replica method then yields the RS potential for model (1) with prior $P_0^{\otimes p}$:

$$i^{\mathrm{RS}}(E;\Delta) := i(\tilde{S};\tilde{Y}) + \frac{1}{2}(\alpha\ln(1 + E/(\alpha\Delta)) - \frac{E}{\Sigma^2}). \tag{9}$$

This formula was first derived by Tanaka (2002) and later rigorously proved by Reeves and Pfister (2019). Barbier et al. (2020) subsequently provided a different proof, assuming that the prior is discrete and i^{RS} has at most three stationary points as a function of E. Reeves and Pfister also explored this relationship under the assumption that the prior has finite fourth moment and a "Single-Crossing Property". Additionally, as noted by Barbier et al. (2020, Remark 3.6), the discrete prior can be extended to more generalized distributions, where the constraint is relaxed as discussed in the work of Barbier et al. (2019). Specifically, under the assumptions, the following holds:

Lemma 1. (Reeves and Pfister, 2019) *Let $i^{\mathrm{RS}}(E;\Delta)$ be the RS potential defined in* (9)*, and let $\mathcal{Z}_p$ be the partition function defined in* (2)*. Under the assumption that P_0 has finite fourth moment and satisfies the single-crossing property, we have:*

$$-\frac{\alpha}{2}-\min_{E\geq 0} i^{\mathrm{RS}}(E;\Delta)=\lim_{p\to\infty}\{\frac{1}{p}\mathbb{E}\ln \mathcal{Z}_p\} \tag{10}$$

Thus, if we can establish that our prior satisfies the RS formula and $\min_{E\geq 0} i^{\mathrm{RS}}(E;\Delta) = -\max_{\|a\|^2\leq p} f_{\mathrm{TAP}}(a)-\frac{\alpha}{2}$ holds, it would confirm the correctness of the TAP free energy in the uniform spherical setting for all $\Delta > 0$. And we defer the verification of the RS assumptions to Appendix B.

3 MAIN RESULT

In this section, we formally state our main theoretical result, which establishes the asymptotic correctness of the TAP free energy approximation for high-dimensional Bayesian linear regression under a uniform spherical prior.

Theorem 2. *Fix $\alpha > 0$ and $\Delta > 0$. Consider the Bayesian linear regression model*

$$y=\mathbf{x}\beta+\sqrt{\Delta}\,\epsilon, \tag{11}$$

where $\mathbf{x}\in\mathbb{R}^{N\times p}$ has i.i.d. $\mathcal{N}(0,\frac{1}{N})$ entries, $\epsilon\sim\mathcal{N}(0,I_N)$, and the prior on $\beta\in\mathbb{R}^p$ is uniform on the sphere $\mathbb{S}^{p-1}(\sqrt{p})$. Assume that $N/p\to\alpha$ as $p\to\infty$.

Then, as $p\to\infty$, we have convergence in probability (with respect to the randomness in $\mathbf{x}$, β, and ϵ):

$$\lim_{p\to\infty}\max_{\|a\|\leq\sqrt{p}} f_{\mathrm{TAP}}(a)=\lim_{p\to\infty}F_p, \tag{12}$$

where $f_{\mathrm{TAP}}(a)$ denotes the TAP free energy functional defined in (5) *and $F_p=\frac{1}{p}\ln\mathcal{Z}_p$ is the free energy shown in* (3)*.*

Proof. The result follows from Lemma 6 and Lemma 8 ahead. □

As a corollary, any near maximizer of the TAP functional must be close to a ridge regressor:

Corollary 1. *In the setting of Theorem 2, define $R_{p,r} := \max\{\frac{1}{p}\|a-\hat{a}_{\mathrm{ridge}}\|_2^2\colon f_{\mathrm{TAP}}(a)\geq \sup_{\|a\|\leq\sqrt{p}} f_{\mathrm{TAP}}(a)-r\}$, for any $r\geq 0$, where $\hat{a}_{\mathrm{ridge}} := \operatorname{argmin}_{a\in\mathbb{R}^p}\{\frac{1}{\Delta}\|\mathbf{y}-\mathbf{x}a\|^2+\|a\|^2\}$. Then we have the convergence in probability $\lim_{p\to\infty}R_{p,r}\leq 2r$.*

Proof. Suppose that a satisfies $f_{\mathrm{TAP}}(a)\geq \sup_{\|a\|\leq\sqrt{p}} f_{\mathrm{TAP}}(a)-r$. Then for C defined in (51), we have $-\frac{1}{2\Delta p}\|\mathbf{y}-\mathbf{x}a\|^2-\frac{1}{2p}\|a\|^2+C\geq f_{\mathrm{TAP}}(a)\geq \sup_{\|a\|\leq\sqrt{p}} f_{\mathrm{TAP}}(a)-r\geq \lim_{p\to\infty}F_p-r+o_p(1)$. On the other hand, in the proof of Lemma 8 we obtained $\lim_{p\to\infty}\sup_{a\in\mathbb{R}^p}\{-\frac{1}{2\Delta p}\|\mathbf{y}-\mathbf{x}a\|^2-\frac{1}{2p}\|a\|^2\}+C\leq \lim_{p\to\infty}F_p$. Since $\nabla_a^2\{-\frac{1}{2\Delta p}\|\mathbf{y}-\mathbf{x}a\|^2-\frac{1}{2p}\|a\|^2\}\preceq \nabla_a^2\{-\frac{1}{2p}\|a\|^2\} = -\frac{1}{p}I_p$ (in terms of the ordering of positive semidefinite matrices), we see that $\frac{1}{2p}\|a-\hat{a}_{\mathrm{ridge}}\|^2\leq r+o_p(1)$, and the result follows. □

4 PROOF OF THE MAIN RESULT

4.1 Equivalence of i.i.d. Normal Prior and the Spherical Prior

In this section, we show that the asymptotic free energy under the spherical prior can be calculated using the result of Reeves and Pfister (2019) with i.i.d. Gaussian prior. Intuitively this is simply because a high-dimensional gaussian vector is concentrated near a sphere, but the actual proof also relies strongly on the Bayes-optimal assumption; we expect that the result will no longer be true in a mismatched setting (misspecified model) where the true prior differs from the prior used in defining the TAP expression.

Let $\tilde{\pi} := \mathcal{N}(0,I_p)$, and

$$\mathbf{y}:=\mathbf{x}\boldsymbol{\beta}_0+\sqrt{\Delta}\mathbf{z} \tag{13}$$

where $\boldsymbol{\beta}_0\sim\tilde{\pi}$, $\mathbf{x}\sim\mathcal{N}(0,\frac{1}{N}I_N\otimes I_p)$ and $\mathbf{z}\sim\mathcal{N}(0,I_N)$. Let Π denote the projection onto the centered sphere of radius $\sqrt{p}$. Note that $\Pi(\boldsymbol{\beta}_0)$ thus follows the uniform distribution on the centered sphere of radius $\sqrt{p}$, which we denote by π. Set

$$\mathbf{y}':=\mathbf{x}\Pi(\boldsymbol{\beta}_0)+\sqrt{\Delta}\mathbf{z}. \tag{14}$$

Recall that $o_P(1)$ denotes a random variable vanishing in probability, and similarly for $o_P(\sqrt{p})$ etc.

Lemma 3. *Consider random $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{y}'$ defined in* (13) *and* (14)*. We have*

$$\begin{aligned}&\frac{1}{p}\ln\int e^{-\frac{1}{2\Delta}\|\mathbf{y}-\mathbf{x}\boldsymbol{\beta}\|^2}d\tilde{\pi}(\boldsymbol{\beta})\\ =&\frac{1}{p}\ln\int e^{-\frac{1}{2\Delta}\|\mathbf{y}'-\mathbf{x}\boldsymbol{\beta}\|^2}d\pi(\boldsymbol{\beta})+o_P(1).\end{aligned} \tag{15}$$

Note that as a consequence of this lemma and the result of Reeves and Pfister (2019), the right side of (12) (for spherical prior) can be calculated using (9) and (10) with the Gaussian prior.

Proof. Let $(u_p)_{p\geq 1}$ be a nonnegative vanishing sequence satisfying

$$\lim_{p\to\infty} \mathbb{P}\left[\left|\frac{\|\boldsymbol{\beta}_0\|}{\sqrt{p}} - 1\right| > u_p\right] = 0, \tag{16}$$

where $\boldsymbol{\beta}_0 \sim \tilde{\pi}$, which is possible if $u_p = \omega(\frac{1}{\sqrt{p}})$, as established by Vershynin (2010, Section 3.1).

We introduce an auxiliary probability measure $\tilde{P}$, which will serve as a convenient reference distribution in the subsequent analysis. Let $\tilde{P}_{\boldsymbol{\beta}_0\mathbf{xz}} = \tilde{\pi} \times \mathcal{N}(0, \frac{1}{N} I_{N\times p}) \times \mathcal{N}(0, I_N)$, let $\mathbf{y}$ be a function of $(\mathbf{x}, \mathbf{z}, \boldsymbol{\beta}_0)$ as define by (13), and then define $\tilde{P}_{\boldsymbol{\beta}_0\mathbf{xzy}\hat{\beta}} = \tilde{P}_{\boldsymbol{\beta}_0\mathbf{xzy}}\tilde{P}_{\hat{\boldsymbol{\beta}}|\mathbf{xy}}$ by setting $\tilde{P}_{\hat{\boldsymbol{\beta}}|\mathbf{xy}} = \tilde{P}_{\boldsymbol{\beta}_0|\mathbf{xy}}$. In other words, under $\tilde{P}$, the estimator $\hat{\boldsymbol{\beta}}$ is conditionally distributed as an independent posterior sample of the planted parameter $\boldsymbol{\beta}_0$ given the data $(\mathbf{x}, \mathbf{y})$. Hence $\hat{\boldsymbol{\beta}}$ is a posterior sample following the Bayesian rule. By symmetry, we have the equivalence of the marginal distributions $\tilde{P}_{\hat{\boldsymbol{\beta}}} = \tilde{P}_{\boldsymbol{\beta}_0}$. Consequently, if we define

$$\lambda_{\boldsymbol{\beta}_0\mathbf{xz}} := \mathbb{P}\left[\left|\frac{\|\hat{\boldsymbol{\beta}}\|}{\sqrt{p}} - 1\right| > u_p \,\middle|\, \boldsymbol{\beta}_0, \mathbf{x}, \mathbf{z}\right] \tag{17}$$

under $\tilde{P}$, then

$$\begin{aligned}\mathbb{E}[\lambda_{\boldsymbol{\beta}_0\mathbf{xz}}] &= \mathbb{P}\left[\left|\frac{\|\hat{\boldsymbol{\beta}}\|}{\sqrt{p}} - 1\right| > u_p\right] \\ &= \mathbb{P}\left[\left|\frac{\|\boldsymbol{\beta}_0\|}{\sqrt{p}} - 1\right| > u_p\right] = o(1).\end{aligned} \tag{18}$$

Now define

$$G := \left\{\boldsymbol{\beta}\colon \left|\frac{\|\boldsymbol{\beta}\|}{\sqrt{p}} - 1\right| \leq u_p\right\}. \tag{19}$$

Then

$$\begin{aligned}&\frac{1}{p}\ln\int e^{-\frac{1}{2\Delta}\|\mathbf{y}-\mathbf{x}\boldsymbol{\beta}\|^2} d\tilde{\pi}(\boldsymbol{\beta}) \\ =&\frac{1}{p}\ln\int_G e^{-\frac{1}{2\Delta}\|\mathbf{y}-\mathbf{x}\boldsymbol{\beta}\|^2} d\tilde{\pi}(\boldsymbol{\beta}) - \frac{1}{p}\ln \tilde{P}_{\boldsymbol{\beta}_0|\mathbf{xy}}[G|\mathbf{xy}] \\ =&\frac{1}{p}\ln\int_G e^{-\frac{1}{2\Delta}\|\mathbf{y}-\mathbf{x}\boldsymbol{\beta}\|^2} d\tilde{\pi}(\boldsymbol{\beta}) + o_P(1),\end{aligned} \tag{20}$$

where the last step follows by $\tilde{P}_{\boldsymbol{\beta}_0|\mathbf{xy}}[G|\mathbf{xy}] = 1-\lambda_{\boldsymbol{\beta}_0\mathbf{xz}}$ and (18). Now let μ be the distribution of $\boldsymbol{\beta}_0 \sim \tilde{\pi}$ conditioned on $\boldsymbol{\beta}_0 \in G$. We have from (16) that

$$\begin{aligned}&\frac{1}{p}\ln\int_G e^{-\frac{1}{2\Delta}\|\mathbf{y}-\mathbf{x}\boldsymbol{\beta}\|^2} d\tilde{\pi}(\boldsymbol{\beta}) + o_P(1) \\ =&\frac{1}{p}\ln\int e^{-\frac{1}{2\Delta}\|\mathbf{y}-\mathbf{x}\boldsymbol{\beta}\|^2} d\mu(\boldsymbol{\beta}) + o_P(1).\end{aligned} \tag{21}$$

Recall that Π is the projection of $\boldsymbol{\beta}$ onto the centered sphere of radius $\sqrt{p}$. For any $\boldsymbol{\beta}_0 \in G$, we have $\|\boldsymbol{\beta} - \Pi(\boldsymbol{\beta})\| = o(\sqrt{p})$, and hence

$$\begin{aligned}&\left|\|\mathbf{y}-\mathbf{x}\boldsymbol{\beta}\|^2 - \|\mathbf{y}-\mathbf{x}\Pi(\boldsymbol{\beta})\|^2\right| \\ =&\left|\langle \mathbf{x}(\boldsymbol{\beta}-\Pi(\boldsymbol{\beta})), 2\mathbf{y}-\mathbf{x}(\boldsymbol{\beta}+\Pi(\boldsymbol{\beta}))\rangle\right| &(22)\\ =&o(\sqrt{p}(\sigma_{\max}(\mathbf{x})\|y\| + \sigma_{\max}(\mathbf{x})^2\sqrt{p})) &(23)\\ =&o_P(p), &(24)\end{aligned}$$

where we used Lemma 4. Therefore, we obtain

$$\begin{aligned}&\frac{1}{p}\ln\int e^{-\frac{1}{2\Delta}\|\mathbf{y}-\mathbf{x}\boldsymbol{\beta}\|^2} d\mu(\boldsymbol{\beta}) + o_P(1) \\ =&\frac{1}{p}\ln\int e^{-\frac{1}{2\Delta}\|\mathbf{y}-\mathbf{x}\Pi(\boldsymbol{\beta})\|^2} d\mu(\boldsymbol{\beta}) + o_P(1) \\ =&\frac{1}{p}\ln\int e^{-\frac{1}{2\Delta}\|\mathbf{y}-\mathbf{x}\boldsymbol{\beta}\|^2} d\pi(\boldsymbol{\beta}) + o_P(1).\end{aligned} \tag{25}$$

Then

$$\begin{aligned}&\|\mathbf{y}-\mathbf{x}\boldsymbol{\beta}\|^2 - \|\mathbf{y}'-\mathbf{x}\boldsymbol{\beta}\|^2 \\ =&\langle \mathbf{y}-\mathbf{y}', \mathbf{y}+\mathbf{y}'-\mathbf{2x}\boldsymbol{\beta}\rangle &(26)\\ =&\langle \mathbf{x}(\boldsymbol{\beta}_0-\Pi(\boldsymbol{\beta}_0)), \mathbf{y}+\mathbf{y}'-\mathbf{2x}\boldsymbol{\beta}\rangle &(27)\\ =&o_P(p), &(28)\end{aligned}$$

since $\|\boldsymbol{\beta}_0 - \Pi(\boldsymbol{\beta}_0)\| = o_P(\sqrt{p})$ when $\boldsymbol{\beta}_0 \sim \tilde{\pi}$, and $\|\mathbf{y}+\mathbf{y}'-\mathbf{2x}\boldsymbol{\beta}\| = O_P(\sqrt{p})$. In turn,

$$\begin{aligned}&\frac{1}{p}\ln\int e^{-\frac{1}{2\Delta}\|\mathbf{y}-\mathbf{x}\boldsymbol{\beta}\|^2} d\pi(\boldsymbol{\beta}) + o_P(1) \\ =&\frac{1}{p}\ln\int e^{-\frac{1}{2\Delta}\|\mathbf{y}'-\mathbf{x}\boldsymbol{\beta}\|^2} d\pi(\boldsymbol{\beta}) + o_P(1).\end{aligned} \tag{29}$$

□

Lemma 4. (Vershynin, 2010) *Let A be a $N\times p$ matrix with i.i.d. standard normal entries. For every $k > 0$, the largest eigenvalue of A satisfies:*

$$\sigma_{max}(A) \leq \sqrt{N} + \sqrt{p} + k, \tag{30}$$

with probability at least $1 - 2\exp(-k^2/2)$.

4.2 TAP Free Energy at a Ridge Solution

Our proof relies on analyzing a ridge regression solution:

$$a_\Delta = A_\Delta^{-1}\mathbf{x}^\top\mathbf{x}\boldsymbol{\beta} + A_\Delta^{-1}\mathbf{x}^\top\epsilon, \tag{31}$$

where $A_\Delta := \mathbf{x}^\top\mathbf{x} + \Delta I$. To analyze the asymptotic behavior of the ridge regression solution, we consider

the normalized squared norm:

$$\lim_{p\to\infty}\frac{\|a_\Delta\|^2}{p}$$
$$=\lim_{p\to\infty}\frac{1}{p}(\|A_\Delta^{-1}\mathbf{x}^\top\mathbf{x}\boldsymbol{\beta}\|^2+\|A_\Delta^{-1}\mathbf{x}^\top\epsilon\|^2) \quad (32)$$
$$=\lim_{p\to\infty}\frac{1}{p}(\mathrm{tr}(\mathbf{x}^\top\mathbf{x}A_\Delta^{-2}\mathbf{x}^\top\mathbf{x})+\Delta\mathrm{tr}(\mathbf{x}A_\Delta^{-2}\mathbf{x}^\top)) \quad (33)$$
$$=1+\lim_{p\to\infty}-\frac{\Delta}{p}\mathrm{tr}(A_\Delta^{-1}), \quad (34)$$

where the second equation can be derived using the independence of ϵ and $\boldsymbol{\beta}$ as well as LLN.

The trace of A_Δ^{-1} in the preceding equations can be determined using the Marchenko–Pastur distribution, which provides the asymptotic behavior of these quantities for large matrices. As the the eigenvalues of $\mathbf{x}^\top\mathbf{x}$ follow Marchenko–Pastur law, the Stieltjes transform yields the following:

$$\lim_{p\to\infty}\frac{1}{p}\mathrm{tr}(A_t^{-1})=\lim_{p\to\infty}\frac{1}{t}T(t), \quad (35)$$
$$T(t):=\frac{-\alpha+1-t\alpha+\sqrt{(-\alpha+1-t\alpha)^2+4\alpha t}}{2}. \quad (36)$$

The complete statements and proof can be found in the work of Bai and Silverstein (2010).

Next we shall establish the following result:

Lemma 5. *Define* $q_\Delta := 1-T(\Delta)$. *Then* $\lim_{p\to\infty}\mathbb{P}[\|a_\Delta\|<\sqrt{p}]=1$. *Moreover, under the convention that* $f_{\mathrm{TAP}}(a)=-\infty$ *for any* a *with* $\|a\|>\sqrt{p}$ *(i.e., outside the domain of the TAP free energy functional), we have convergence in probability:*

$$\lim_{p\to\infty}f_{\mathrm{TAP}}(a_\Delta)$$
$$=-\frac{1}{2}\alpha+\frac{1}{2}q_\Delta-\frac{\alpha}{2}\ln(1+\frac{1-q_\Delta}{\Delta\alpha})+\frac{1}{2}\ln(1-q_\Delta). \quad (37)$$

Proof. From (34) and (35), it remains to analyze the asymptotic behavior of the quantity $\frac{1}{p}\|\mathbf{y}-\mathbf{x}a_\Delta\|^2$. Exploiting the independence between ϵ and β, we obtain:

$$\lim_{p\to\infty}\frac{1}{p}\|\mathbf{y}-\mathbf{x}a_\Delta\|^2$$
$$=\lim_{p\to\infty}\frac{1}{p}\|\mathbf{x}(I-A_\Delta^{-1}\mathbf{x}^\top\mathbf{x})\boldsymbol{\beta}\|^2+\|(I-\mathbf{x}A_\Delta^{-1}\mathbf{x}^\top)\epsilon\|^2 \quad (38)$$
$$=\lim_{p\to\infty}\frac{1}{p}[\Delta^2\mathrm{tr}(A_\Delta^{-1})+\Delta(N-p)] \quad (39)$$
$$=\Delta T(\Delta)+\Delta(\alpha-1). \quad (40)$$

Substituting this result into the TAP functional defined in equation (5) concludes the proof. □

4.3 Asymptotics of the Free Energy

Similar to the work of Barbier et al. (2020), we expect that $1-q=\mathrm{MSE}$, and that the true free energy can be determined by fixed-point equations. This formulation allows for a direct validation of its equivalence to the TAP representation.

For the Gaussian prior $P_0\sim\mathcal{N}(0,1)$, the fourth moment is finite and the Single-Crossing Property holds (see Appendix B for details). Thus, P_0 satisfies Assumptions II and III as proposed by Reeves and Pfister (2019). The mutual information for the denoising model $i(\tilde{S};\tilde{Y})$ in (8) can then be expressed in a simplified form as:

$$i(\tilde{S};\tilde{Y})=-\frac{1}{2}\ln(\frac{\Sigma^2}{\Sigma^2+1}). \quad (41)$$

This result facilitates the calculation of RS potential in (9), which can be expressed as:

$$-i^{\mathrm{RS}}(E;\Delta)$$
$$=\frac{E}{2\Sigma^2}-\frac{1}{2}\alpha\ln(1+E/(\alpha\Delta))+\frac{1}{2}\ln(\frac{\Sigma^2}{\Sigma^2+1}). \quad (42)$$

By setting its derivative to zero, the above equation can be further simplified to derive the fixed-point condition for E. Specifically, we obtain:

$$E=\frac{\alpha\Delta+E}{\alpha\Delta+E+\alpha}=\frac{\Sigma^2}{\Sigma^2+1}, \quad (43)$$

which has a unique solution for $E\geq 0$. Alternatively, this can be expressed as:

$$E_\Delta=\frac{1-\alpha-\alpha\Delta+\sqrt{(\alpha\Delta+\alpha-1)^2+4\alpha\Delta}}{2}=T(\Delta). \quad (44)$$

This result also suggests that the assumptions formulated by Barbier et al. (2020), specifically that $i^{\mathrm{RS}}(E;\Delta)$ has at most three stationary points, as well as Assumption III proposed by Reeves and Pfister (2019) are valid. Consequently, the RS formula is correctly established.

The R.H.S. of (12) can then be written as:

$$-\min_{E\geq 0}i^{\mathrm{RS}}(E;\Delta)-\frac{\alpha}{2}=-\frac{\alpha}{2}+\frac{1}{2}(1-E_\Delta)$$
$$-\frac{1}{2}\alpha\ln(1+E_\Delta/(\alpha\Delta))+\frac{1}{2}\ln(E_\Delta). \quad (45)$$

Observing the equivalence between E_Δ and $1-q_\Delta$, we can readily derive the following conclusion.

Lemma 6. *Let* f_{TAP} *and* $i^{\mathrm{RS}}(E;\Delta)$ *be defined as* (5) *and* (42). *Then the TAP free energy provides a lower*

bound matching the replica prediction with convergence in probability:

$$\lim_{p\to\infty}\max_{\|a\|^2\le p} f_{\mathrm{TAP}}(a) \ge -\min_{E\ge 0} i^{\mathrm{RS}}(E;\Delta) - \frac{\alpha}{2}. \quad (46)$$

Proof. With a_Δ defined in (31), by lemma 5 and (45), we have:

$$\lim_{p\to\infty}\max_{\|a\|^2\le p} f_{\mathrm{TAP}}(a) \quad (47)$$

$$\ge \lim_{p\to\infty} f_{\mathrm{TAP}}(a_\Delta) = -\min_{E\ge 0} i^{\mathrm{RS}}(E;\Delta) - \frac{\alpha}{2}, \quad (48)$$

with the convergence in probability. □

4.4 Global Maximum

The preceding analysis shows that $f = f_{\mathrm{TAP}}(1-E_\Delta) \le \max_q f_{\mathrm{TAP}}(q)$, and $q_\Delta = 1 - E_\Delta$ is a stationary point of f_{TAP}.

Define

$$g_{\mathrm{TAP}}(q) := \max_{\|a\|^2=pq}\{-\frac{1}{2\Delta p}\|y - Xa\|^2\} - \frac{\alpha}{2}\ln(1+\frac{1-q}{\Delta\alpha}) + \frac{1}{2}\ln(1-q). \quad (49)$$

The goal is to show that $q_\Delta = 1 - E_\Delta$ is (asymptotically) the global maximizer of g_{TAP}. Define

$$g(q) := \max_{\|a\|^2=pq}\{-\frac{1}{2\Delta p}\|y - Xa\|^2\} - \frac{q}{2} + C, \quad (50)$$

where C is a constant chosen such that $g(1-E_\Delta) = g_{\mathrm{TAP}}(1-E_\Delta)$, that is,

$$C := \frac{1-E_\Delta}{2} - \frac{\alpha}{2}\ln(1+\frac{E_\Delta}{\Delta\alpha}) + \frac{1}{2}\ln E_\Delta. \quad (51)$$

Lemma 7. *Let g and g_{TAP} be defined as* (50) *and* (49)*. Then, for any $q \in [0,1]$, $g_{\mathrm{TAP}}(q) \le g(q)$.*

Proof. Define

$$h(q) := -q/2 + C + \frac{\alpha}{2}\ln(1+\frac{1-q}{\Delta\alpha}) - \frac{1}{2}\ln(1-q), \quad (52)$$

with C defined in (51), so $h(1-E_0) = 0$ by construction. The equation $h'(q) = 0$ is equivalent to (43) which has a unique solution for $q \in (0,1)$. Hence $h'(1-E_0) = 0$ by the definition of E_0. Furthermore, we observe that $h'(0) < 0$ and $\lim_{q\uparrow 1} h'(q) = +\infty$. Therefore $h'(q) > 0$ for $q \in (1-E_0, 1)$ and $h'(q) < 0$ for $q \in (0, 1-E_0)$. Consequently, $g(q) - g_{\mathrm{TAP}}(q) = h(q) \ge 0$ for $q \in (0,1)$. □

We are now ready to show the following:

Lemma 8. *We have the matching upper bound:*

$$\lim_{p\to\infty}\max_{\|a\|^2\le p} f_{\mathrm{TAP}}(a) \le -\min_{E\ge 0} i^{\mathrm{RS}}(E;\Delta) - \frac{\alpha}{2}. \quad (53)$$

Proof. We have

$$\max_{\|a\|^2\le p} f_{\mathrm{TAP}}(a) \le \sup_{q\in[0,1]} g(q) \quad (54)$$

$$\le -\frac{1}{2\Delta p}\|\mathbf{y} - \mathbf{x}a_\Delta\|^2 - \frac{1}{2p}\|a_\Delta\|^2 + C. \quad (55)$$

Using the expressions for $\|a_\Delta\|^2/p$ from (34) and $\|y - \mathbf{x}a_\Delta\|^2/p$ from (38), we can evaluate the asymptotic value of the upper bound. Specifically, we have:

$$\lim_{p\to\infty}\{-\frac{1}{2\Delta p}\|\mathbf{y} - \mathbf{x}a_\Delta\|^2 - \frac{1}{2p}\|a_\Delta\|^2\} = -\frac{1}{2}(T(\Delta) + \alpha - 1) - \frac{1}{2}(1 - T(\Delta)) = -\frac{\alpha}{2}. \quad (56)$$

This matches $C - \frac{\alpha}{2} = -i^{\mathrm{RS}}(E_\Delta;\Delta) - \frac{\alpha}{2} = -\min_{E\ge 0} i^{\mathrm{RS}}(E;\Delta) - \frac{\alpha}{2}$. Therefore, the upper bound converges to the desired quantity, completing the proof. □

Combining the lower bound in Lemma 6 with the upper bound in Lemma 8, we can conclude that

$$\lim_{p\to\infty}\max_{\|a\|^2\le p} f_{\mathrm{TAP}}(a) = -\min_{E\ge 0} i^{\mathrm{RS}}(E;\Delta) - \frac{\alpha}{2}, \quad (57)$$

which establishes the asymptotic equivalence between the TAP free energy and the RS prediction for all $\Delta > 0$ under the uniform spherical prior. To conclude the proof of Theorem 2 , it remains to control the fluctuations of F_p and show that convergence in expectation implies convergence in probability. We address this in the following subsection.

4.5 Concentration of the Free Energy

In the preceding analysis we established the convergence of the expected free energy. To upgrade this to convergence of the random free energy itself, it suffices to show that the fluctuations of F_p vanish in the high-dimensional limit.

Lemma 9 (Concentration of free energy)**.** *Under the assumptions of Theorem 2, we have*

$$\mathrm{Var}(F_p) = O\left(\frac{1}{p}\right). \quad (58)$$

In particular,

$$F_p - \mathbb{E}[F_p] \xrightarrow{P} 0. \quad (59)$$

The proof follows from Gaussian concentration applied to the Gaussian design matrix X and noise vector w. The key observation is that $\log Z_p$ is Lipschitz in these random inputs with a Lipschitz constant of order $O(\sqrt{p})$ (See Appendix C for details). Combining Lemma 6, Lemma 8, and Lemma 9, we establish that

$$F_p \xrightarrow{P} \max_{\|a\| \le \sqrt{p}} f_{\mathrm{TAP}}(a). \tag{60}$$

This completes the proof of Theorem 2.

5 NUMERICAL SIMULATION

We complement our theoretical results with numerical experiments that examine the finite-sample behavior of the TAP free energy in Bayesian linear regression. The purpose of this experiment is to provide empirical evidence for Corollary 1, which asserts that any near-maximizer of the TAP free energy lies close to the ridge regression solution. Accordingly, our simulations compare the maximizers of the TAP functional with the ridge estimator and examine how their distance behaves as the dimension grows.

We considered the model shown in the beginning of section 2. For each configuration (α, Δ, p), we generate $\mathbf{x}$, β, and $\mathbf{y} = \mathbf{x}\beta + \sqrt{\Delta}\mathbf{z}$. We then numerically optimize the TAP free energy functional $f_{\mathrm{TAP}}(a)$ with respect to $a \in \mathbb{R}^p$ under the constraint $\|a\| \le \sqrt{p}$. The optimization is performed by projected gradient ascent with a fixed step size and stopping tolerance. Because the TAP functional is non-convex, we use multiple restarts. Among all runs, we retain the maximum TAP value as our empirical estimate of the global maximizer.

To account for randomness, each experiment is repeated $R = 30$ times with independent seeds. For each repetition we record (i) the TAP free energy at the estimated maximizer, (ii) the ridge regression free energy, and (iii) the squared distance:

$$D = \frac{1}{p}\|a_{\mathrm{TAP}} - a_{\mathrm{ridge}}\|^2, \tag{61}$$

which quantifies the discrepancy between the TAP maximizer and the ridge solution. By Corollary 1, D is predicted to vanish asymptotically; our simulations examine this prediction across system sizes ranging from $p = 200$ to $p = 1600$.

Figure 1 reports the squared distance $D = \|a_{\mathrm{TAP}} - a_{\mathrm{ridge}}\|^2/p$ between the TAP maximizer and the ridge regression estimator for a representative configuration $(\alpha = 0.75, \Delta = 0.1)$. The results clearly demonstrate that D decreases as p grows, in agreement with Corollary 1. Although finite-sample oscillations are visible

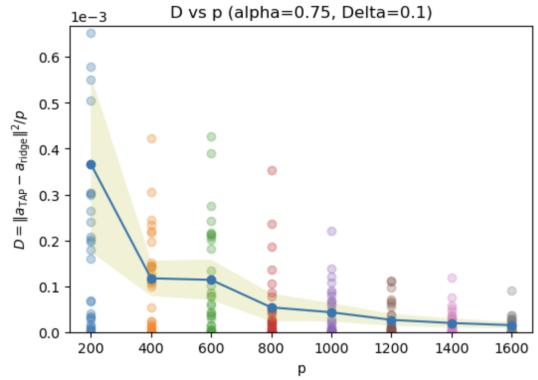

Figure 1: Numerical Results Under $\Delta = 0.1$, $\alpha = 0.75$.

for small p, the overall trend supports the theoretical prediction that maximizer of TAP concentrate around the ridge solution in the proportional high-dimensional regime. (see Appendix D for additional experiments).

6 CONCLUSION

Our results establish that under suitable conditions, the TAP free energy admits a maximizer whose value converges to the true free energy. This work advances the theoretical understanding of the TAP free energy approximation in high-dimensional Bayesian inference. By establishing its equivalence to the RS formula, this study enhances the applicability of the TAP framework for high-dimensional Bayesian inference, offering a more accurate alternative to the widely used mean-field approximation.

Although we have proven that the TAP free energy converges in probability to the true free energy, the behavior of its maximizer under the uniform spherical prior remains to be explored. In particular, while the TAP maximizer coincides with the posterior mean in the case of a Gaussian prior, Corollary 1 suggests that the same holds under the spherical prior, though a rigorous justification requires additional work. Moreover, while our result shows that any near-maximizer of the TAP functional must be close to the ridge solution, this does not in itself provide the full geometric control established in the high-temperature analysis of Qiu and Sen (2023). Our theorem resolves the question about the global maximizer of the TAP formula, but does not imply band-wise free energy control or pure-state structure. Establishing such a geometric characterization remains open, and we view it as a promising direction for future work.

Future directions include extending the TAP framework to more general prior distributions, including non-spherical and non-Gaussian settings, and assessing its applicability in more complex models. A nat-

ural next question is how to extend our results for spherical priors to the case of general i.i.d. priors, as conjectured by Krzakala et al. (2014). To this end, the method-of-types approach from Liu (2022, 2023) for the free energy could provide an analogue to the geometric arguments used in spherical models (Qiu and Sen, 2023). Another important future direction is to develop and analyze algorithms for optimizing the TAP free energy in this non-convex regime. Understanding the geometry of the TAP landscape and designing robust optimization methods that can navigate potential spurious local maxima are essential steps toward practical inference in high-dimensional settings.

Acknowledgments

This research was supported in part by NSF Grant DMS-2515510. The authors also thank Kevin Luo and Qiang Wu for helpful discussions.

References

Bai, Z. and Silverstein, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*. Springer, New York.

Barbier, J., Krzakala, F., Macris, N., Miolane, L., and Zdeborová, L. (2019). Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460.

Barbier, J., Macris, N., Dia, M. G., and Krzakala, F. (2020). Mutual information and optimality of approximate message-passing in random linear estimation. *IEEE Transactions on Information Theory*, 66(7):4270–4303.

Belius, D. (2022). High temperature TAP upper bound for the free energy of mean field spin glasses. *arXiv preprint arXiv:2204.00681*.

Bethe, H. A. (1935). Statistical theory of superlattices. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 150(871):552–575.

Brennecke, C., Xu, C., and Yau, H.-T. (2023). Operator norm bounds on the correlation matrix of the SK model at high temperature. *arXiv preprint arXiv:2307.12535*.

Celentano, M., Fan, Z., Lin, L., and Mei, S. (2023). Mean-field variational inference with the TAP free energy: Geometric and statistical properties in linear models.

Chen, W.-K., Panchenko, D., and Subag, E. (2023). Generalized TAP free energy. *Communications on Pure and Applied Mathematics*, 76(7):1329–1415.

Cornelius, R., Heberling, D., Koep, N., Behboodi, A., and Mathar, R. (2016). Compressed sensing applied to spherical near-field to far-field transformation. In *2016 10th European Conference on Antennas and Propagation (EuCAP)*, pages 1–4. IEEE.

Cui, H., Yu, Z., and Liu, J. (2024). Sampling from the random linear model via stochastic localization up to the amp threshold. *arXiv preprint arXiv:2407.10763*.

El Alaoui, A., Montanari, A., and Sellke, M. (2022). Sampling from the sherrington-kirkpatrick gibbs measure via algorithmic stochastic localization. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 323–334. IEEE.

Fan, Z., Mei, S., and Montanari, A. (2021). TAP free energy, spin glasses and variational inference. *The Annals of Probability*, 49(1):1–45.

Huang, B. (2024). Capacity threshold for the ising perceptron. In *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1126–1136.

Koehler, F. (2019). Fast convergence of belief propagation to global optima: Beyond correlation decay. *Advances in Neural Information Processing Systems*, 32.

Krzakala, F., Manoel, A., Tramel, E. W., and Zdeborová, L. (2014). Variational free energies for compressed sensing. In *Proceedings of the 2014 IEEE International Symposium on Information Theory (ISIT)*, pages 1499–1503, Honolulu, HI, USA.

Liu, J. (2022). Minoration via mixed volumes and cover's problem for general channels. *Probability Theory and Related Fields*, 183(1):315–357.

Liu, J. (2023). From soft-minoration to information-constrained optimal transport and spiked tensor models. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pages 666–671.

Mézard, M., Parisi, G., and Virasoro, M. A. (1987). *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company.

Montanari, A. and Wu, Y. (2023). Posterior sampling in high dimension via diffusion processes. *arXiv preprint arXiv:2304.11449*.

Mukherjee, S. and Sen, S. (2022). Variational inference in high-dimensional linear regression. *Journal of Machine Learning Research*, 23(304):1–56.

Qiu, J. and Sen, S. (2023). The TAP free energy for high-dimensional linear regression. *The Annals of Applied Probability*, 33(4):2643–2680.

Reeves, G. and Pfister, H. D. (2019). The replica-symmetric prediction for random linear estimation

with gaussian matrices is exact. *IEEE Transactions on Information Theory*, 65(4):2252–2283.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2021). Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*.

Subag, E. (2017). The geometry of the Gibbs measure of pure spherical spin glasses. *Inventiones mathematicae*, 210:135–209.

Subag, E. (2023). TAP approach for multispecies spherical spin glasses II: The free energy of the pure models. *The Annals of Probability*, 51(3):1004–1024.

Tanaka, T. (2002). A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors. *IEEE Transactions on Information Theory*, 48(11):2888–2910.

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

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

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

Welling, M. and Teh, Y. W. (2013). Belief optimization for binary networks: A stable alternative to loopy belief propagation. *arXiv preprint arXiv:1301.2317*.

Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2001). Bethe free energy, kikuchi approximations, and belief propagation algorithms. *Advances in neural information processing systems*, 13(24).

Yedidia, J. S., Freeman, W. T., Weiss, Y., et al. (2003). Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8(236-239):0018–9448.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

The appendix is structured as follows: Appendix A collects technical preliminaries and auxiliary lemmas that are used throughout the proofs. Appendix B verifies the replica-symmetric (RS) assumptions in our setting. Appendix C provides the full proof of the concentration result for the free energy, completing the argument of Lemma 9. Appendix D presents additional numerical experiments, including different values of Δ and α, and error bars from repeated trials, which further support the theoretical predictions.

A PROOF OF NON-CONCAVITY

In this section, we present a rigorous proof to illustrate that the TAP free energy functional may not be globally concave under the spherical prior.

Let

$$h(q) := \frac{\alpha}{2}\log\left(1+\frac{1-q}{\Delta\alpha}\right) - \frac{1}{2}\log(1-q). \tag{62}$$

The first and second derivative of h can then be calculated as:

$$h'(q) = -\frac{\alpha}{2(\Delta\alpha+1-q)} + \frac{1}{2(1-q)}, \tag{63}$$

$$h''(q) = -\frac{\alpha}{2(\Delta\alpha+1-q)^2} + \frac{1}{2(1-q)^2}. \tag{64}$$

Thus, the Hessian of the TAP free energy satisfies:

$$-p\cdot\nabla^2 f_{\mathrm{TAP}}(a) = \frac{4}{p}h''(q)\cdot aa^\top + 2h'(q)I_p + \frac{1}{\Delta}\mathbf{x}^\top\mathbf{x}, \tag{65}$$

where $q := \|a\|^2/p$. Let $\lambda_{\min}$ denote the smallest eigenvalue of $X^\top X$, and $u_{\min}$ the corresponding eigenvector. From random matrix theory, the empirical spectral distribution of $X^\top X$ converges (as $p, n \to \infty$ with $n/p \to \alpha \in (1,\infty)$) to the Marchenko–Pastur law. In particular, the smallest eigenvalue $\lambda_{\min}$ converges almost surely to

$$\lambda_{\min} \longrightarrow \left(1-\sqrt{\alpha^{-1}}\right)^2. \tag{66}$$

Then, we consider a vector a aligned with $u_{\min}$ of the form

$$a = \frac{u_{\min}}{\|u_{\min}\|}\cdot\sqrt{pq}. \tag{67}$$

Then the second directional derivative becomes

$$-a^\top\nabla^2 f_{\mathrm{TAP}}(a)a = q\left(4qh''(q) + 2h'(q) + \frac{\lambda_{\min}}{\Delta}\right). \tag{68}$$

Dividing both sides by q, we obtain

$$\frac{1}{q}\cdot\left(-a^\top\nabla^2 f_{\mathrm{TAP}}(a)a\right) = -\frac{2\alpha q}{(\Delta\alpha+1-q)^2} + \frac{2q}{(1-q)^2} + \frac{1}{1-q} - \frac{\alpha}{\Delta\alpha+1-q} + \frac{\lambda_{\min}}{\Delta} \tag{69}$$

$$= -\frac{(\Delta\alpha+1+\ q)\alpha}{(\Delta\alpha+1-q)^2} + \frac{1+q}{(1-q)^2} + \frac{\lambda_{\min}}{\Delta} \tag{70}$$

Letting $\Delta\alpha = k$, and recalling the asymptotic limit of $\lambda_{\min}$, as $p \to \infty$, the expression becomes:

$$-\frac{(k+1+\ q)\alpha}{(k+1-q)^2} + \frac{1+q}{(1-q)^2} + \frac{\alpha+1-2\sqrt{\alpha}}{k}. \tag{71}$$

To ensure this quantity becomes negative as $\alpha \to \infty$, it suffices to find $k > 0$ and $q \in (0,1)$ such that:

$$\frac{1}{k} - \frac{k+1+\ q}{(k+1-q)^2} < 0. \tag{72}$$

As an example, one can choose $k = 1$, $q = 0.5$, and set $\alpha = 10$, $\Delta = 0.1$, which yields:

$$-\frac{(k+1+\ q)\alpha}{(k+1-q)^2} + \frac{1+q}{(1-q)^2} + \frac{\alpha+1-2\sqrt{\alpha}}{k} \approx -0.4357 < 0 \tag{73}$$

which implies that $-\nabla^2 f_{\mathrm{TAP}}(a)$ may have negative eigenvalues in the neighborhood of $a = \frac{u_{\min}}{\|u_{\min}\|} \cdot \sqrt{p/2}$. Therefore, the TAP free energy is not globally concave under the uniform spherical prior.

B VERIFICATION OF RS ASSUMPTIONS FOR THE GAUSSIAN PRIOR

In this section, we verify that the assumptions required for the validity of the RS formula, as formulated by Reeves and Pfister (2019), are satisfied under the standard Gaussian prior $P_0 = \mathcal{N}(0,1)$. We adopt the random linear estimation model in (11).

Define the normalized mutual information and MMSE functions:

$$I_p(\alpha) := \frac{1}{p} I(\boldsymbol{\beta}; \mathbf{y} \mid \mathbf{x}), \qquad M_p(\alpha) := \frac{1}{p}\mathrm{mmse}(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{x}). \tag{74}$$

Let $\beta \sim P_0$ and $z \sim \mathcal{N}(0,1)$. Consider the scalar Gaussian channel

$$y = \sqrt{s}\beta + z. \tag{75}$$

We define the scalar functions:

$$I_\beta(s) = I(\beta; \sqrt{s}\beta + z), \qquad \mathrm{mmse}_\beta(s) = \mathrm{mmse}(\beta \mid \sqrt{s}\beta + z). \tag{76}$$

The RS potential is then defined as:

$$\mathcal{R}(\alpha, u) = I_\beta\Big(\frac{\alpha}{1+u}\Big) + \frac{\alpha}{2}\Big(\log(1+u) - \frac{u}{1+u}\Big), \tag{77}$$

Using the I-MMSE identity and stationarity of $\mathcal{R}$, one obtains the RS fixed-point equation

$$u = \mathrm{mmse}_\beta\Big(\frac{\alpha}{1+u}\Big). \tag{78}$$

It is convenient to view (78) as the intersection of two curves in the (u, α)-plane. Define

$$\alpha_{\mathrm{FP}}(u) := (1+u)\mathrm{mmse}_\beta^{-1}(u), \qquad \alpha_{\mathrm{RS}}(u) := \inf\{\alpha \geq 0 :\ u \in \arg\min_{v \geq 0} \mathcal{R}(\alpha, v)\}. \tag{79}$$

Equivalently, $\alpha_{\mathrm{RS}}(u)$ is characterized by the vanishing first variation of $\mathcal{R}$ at $u = v$.

The single-crossing property requires that the two curves $\alpha_{\mathrm{FP}}(u)$ and $\alpha_{\mathrm{RS}}(u)$ intersect at most once for $u \in (0,1)$. Under this property the following lemma holds:

Lemma 10. (Reeves and Pfister, 2019) *Consider the random linear model in* (11) *where* $\mathbf{x} \in \mathbb{R}^{n \times p}$ *has i.i.d. entries* $\mathbf{x}_{ij} \sim \mathcal{N}(0, 1/p)$*, the noise* $Z \sim \mathcal{N}(0, I_m)$ *is independent of* $\mathbf{x}$*, and the signal* $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ *has i.i.d. entries* $\beta_i \sim P_0$ *with* $\mathbb{E}[\beta^4] < \infty$*.*

Assume P_0 *satisfies the single-crossing property. Then, as* $p \to \infty$*,*

$$I_p(\alpha) \longrightarrow I_{\mathrm{RS}}(\alpha) := \min_{u \geq 0} \mathcal{R}(\alpha, u), \tag{80}$$

$$M_p(\alpha) \longrightarrow M_{\mathrm{RS}}(\alpha) \in \arg\min_{u \geq 0} \mathcal{R}(\alpha, u). \tag{81}$$

Now we consider the special case $P_0 = \mathcal{N}(0,1)$. We have $\mathbb{E}\beta^4 = 3 < \infty$. For the scalar channel, $\mathrm{mmse}_\beta(s) = \frac{1}{1+s}$ and $I'_\beta(s) = \frac{1}{2}\mathrm{mmse}_\beta(s)$. Since $\mathrm{mmse}_\beta(s) = 1/(1+s)$, its inverse on $(0,1)$ is $\mathrm{mmse}_\beta^{-1}(z) = (1-z)/z$. Hence

$$\alpha_{\mathrm{FP}}(u) = (1+u)\frac{1-u}{u} = \frac{1-u^2}{u}. \tag{82}$$

The RS stationarity condition yields the same parametric relation between α and u:

$$u = \frac{1}{1+\alpha/(1+u)} \quad \Longleftrightarrow \quad \alpha = \frac{1-u^2}{u}, \tag{83}$$

so $\alpha_{\mathrm{RS}}(u) \equiv \alpha_{\mathrm{FP}}(u)$ on $(0,1)$. Since $\alpha_{\mathrm{RS}} \equiv \alpha_{\mathrm{FP}}$ pointwise, the two curves coincide and in particular cross at most once. The single crossing property holds trivially.

All assumptions required by the framework of Reeves and Pfister (2019) are satisfied for $P_\beta = \mathcal{N}(0,1)$. Consequently, along $n, p \to \infty$ with $n/p \to \alpha$,

$$I_p(\alpha) \;\longrightarrow\; I_{\mathrm{RS}}(\alpha) = \min_{z\geq 0} \mathcal{R}(\alpha, u). \tag{84}$$

where $u_\star(\alpha)$ is the unique minimizer of $\mathcal{R}(\alpha,\cdot)$ and solves $u = \mathrm{mmse}_\beta(\alpha/(1+u))$. And from (84), we can easily obtain the convergence we want in (10).

C PROOF OF LEMMA 9

In this section we present the Concentration of F_p. Recall that, conditional on X,

$$\mathcal{Z}_p(X,y) = (2\pi)^{-n/2}\,\det(A_\Delta)^{-1/2}\,\exp\Big(-\tfrac{1}{2}\,y^\top A_\Delta^{-1} y\Big), \tag{85}$$

where $A_\Delta = \Delta I_n + XX^\top$. Define

$$f(X,y) := \frac{1}{p}\log \mathcal{Z}_p(X,y) = -\frac{1}{2p}\log\det A_\Delta - \frac{1}{2p}\,y^\top A_\Delta^{-1} y - \frac{n}{2p}\log(2\pi). \tag{86}$$

We use the law of total variance:

$$\mathrm{Var}\big(F_p\big) = \mathbb{E}\big[\mathrm{Var}(f \mid X)\big] + \mathrm{Var}\big(\mathbb{E}[f \mid X]\big). \tag{87}$$

Conditioning on X, we have $y \mid X \sim \mathcal{N}(0, A_\Delta)$ and

$$\nabla_y f(X,y) = -\frac{1}{p} A_\Delta^{-1} y. \tag{88}$$

The Poincaré inequality with covariance A_Δ states that for smooth $g : R^n \to R$ and $Y \sim \mathcal{N}(0, A_\Delta)$, $\mathrm{Var}(g(Y)) \leq \mathbb{E}\big[\nabla g(Y)^\top A_\Delta \nabla g(Y)\big]$. Applying this with $g(y) = f(X,y)$ gives

$$\mathrm{Var}(f \mid X) \leq \mathbb{E}\left[\frac{1}{p^2} y^\top A_\Delta^{-1} A_\Delta A_\Delta^{-1} y \Big| X\right] = \frac{1}{p^2}\mathbb{E}\left[y^\top A_\Delta^{-1} y \Big| X\right] = \frac{1}{p^2}\mathrm{tr}(A_\Delta^{-1} A_\Delta) = \frac{n}{p^2}. \tag{89}$$

Taking expectation over X yields

$$\mathbb{E}[\mathrm{Var}(f \mid X)] \leq \frac{n}{p^2} = \frac{\alpha}{p}. \tag{90}$$

Then, we compute $\mathbb{E}[f \mid X]$. Since $\mathbb{E}[y^\top A_\Delta^{-1} y \mid X] = \mathrm{tr}(A_\Delta^{-1}\mathbb{E}[yy^\top \mid X]) = \mathrm{tr}(A_\Delta^{-1} A_\Delta) = n$,

$$\mathbb{E}[f \mid X] = -\frac{1}{2p}\log\det A_\Delta - \frac{n}{2p} - \frac{n}{2p}\log(2\pi), \tag{91}$$

only the $\log\det A_\Delta$ term depends on X. Matrix calculus gives $\partial(\log\det A_\Delta)/\partial A_\Delta = A_\Delta^{-1}$ and

$$dA_\Delta = (dXX^\top + XdX^\top). \tag{92}$$

By the chain rule and cyclicity of trace,

$$d\mathbb{E}[f \mid X] = -\frac{1}{2p}\langle A_\Delta^{-1}, dA_\Delta\rangle = -\frac{\tau^2}{2p}\mathrm{tr}\big(A_\Delta^{-1}dXX^\top + A_\Delta^{-1}XdX^\top\big) = -\frac{1}{p}\mathrm{tr}\big(A_\Delta^{-1}XdX^\top\big). \tag{93}$$

Hence the gradient is

$$\nabla_X\mathbb{E}[f \mid X] = -\frac{1}{2p}A_\Delta^{-1}X. \tag{94}$$

Therefore

$$\|\nabla_X\mathbb{E}[f \mid X]\|_F^2 = \frac{1}{4p^2}\mathrm{tr}\big(X^\top A_\Delta^{-2}X\big) \le \frac{1}{4p^2}\|A_\Delta^{-2}\|\|X\|_F^2 \le \frac{1}{4p^2\Delta^2}\|X\|_F^2, \tag{95}$$

since $A_\Delta \succeq \Delta I_n$ implies $\|A_\Delta^{-1}\| \le 1/\Delta$ and thus $\|A_\Delta^{-2}\| \le 1/\Delta^2$.

Now apply the Poincaré inequality to X:

$$\mathrm{Var}(\mathbb{E}[f \mid X]) \le \frac{1}{p}\mathbb{E}\|\nabla_X\mathbb{E}[f \mid X]\|_F^2 \le \frac{1}{p}\cdot\frac{1}{4p^2\Delta^2}\mathbb{E}\|X\|_F^2 = \frac{1}{4p^3\Delta^2}\,\mathbb{E}\Big[\sum_{i,j}X_{ij}^2\Big] = \frac{1}{4p^3\Delta^2}n = \frac{1}{4\Delta^2}\frac{\alpha}{p^2}. \tag{96}$$

Combing the bounds in (90) and (96),

$$\mathrm{Var}(F_p) \le \frac{\alpha}{p} + \frac{1}{4\Delta^2}\frac{\alpha}{p^2}. \tag{97}$$

Chebyshev's inequality then implies $F_p - \mathbb{E}[F_p] \to 0$ in probability.

D ADDITIONAL NUMERICAL STUDIES

In this section, we provide extended experiments that go beyond those presented in Section 5. Specifically, we study a grid of parameters with sampling ratios $\alpha = 0.75, 1.0, 2.0$, and noise levels $\Delta = 0.1, 0.5, 1.0$. For each (α, Δ) pair, we evaluate the discrepancy

$$D = \frac{1}{p}\|a_{\mathrm{TAP}} - a_{\mathrm{ridge}}\|^2, \tag{98}$$

as a function of the system size p.

In all settings tested, D decreases with p, consistent with the theoretical prediction. The small oscillations observed at intermediate values of p are consistent with finite-size effects and are expected to vanish at rate $O(1/p)$.

The complete set of nine plots, covering all combinations of (α, Δ), is shown in Figure 2.

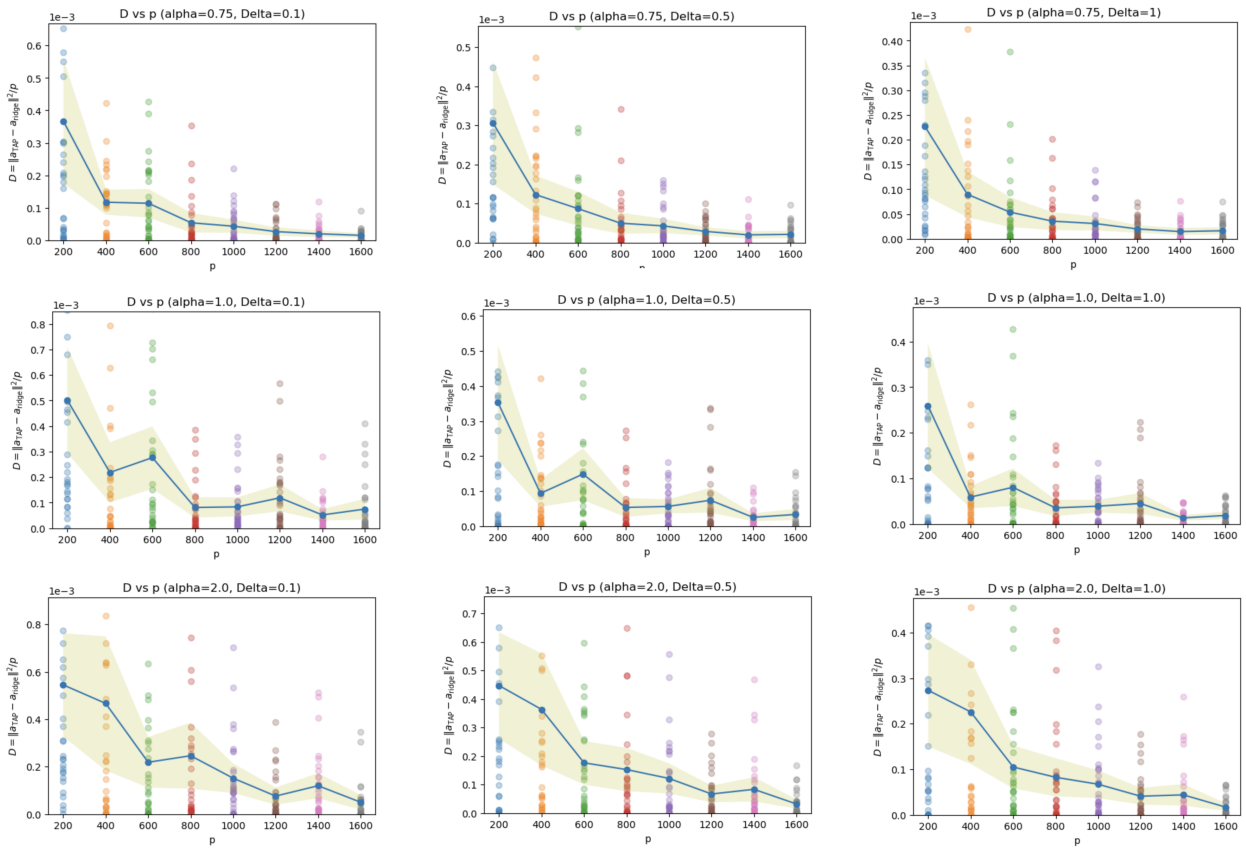

Figure 2: Complete Collection of Results, Displaying the Discrepancy D across All Combinations of (α, Δ).