



HAL
open science

Dense Limit of the Dawid-Skene Model for Crowdsourcing and Regions of Sub-optimality of Message Passing Algorithms

Christian Schmidt, Lenka Zdeborová

► **To cite this version:**

Christian Schmidt, Lenka Zdeborová. Dense Limit of the Dawid-Skene Model for Crowdsourcing and Regions of Sub-optimality of Message Passing Algorithms. 2018. cea-01930667

HAL Id: cea-01930667

<https://cea.hal.science/cea-01930667>

Preprint submitted on 22 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Dense Limit of the Dawid-Skene Model for Crowdsourcing and Regions of Sub-optimality of Message Passing Algorithms

Christian Schmidt¹ and Lenka Zdeborová¹

¹ *Institut de Physique Théorique, CEA Saclay and CNRS, 91191, Gif-sur-Yvette, France.*

Crowdsourcing is a strategy to categorize data through the contribution of many individuals. A wide range of theoretical and algorithmic contributions are based on the model of Dawid and Skene [1]. Recently it was shown in [2, 3] that, in certain regimes, belief propagation is asymptotically optimal for data generated from the Dawid-Skene model. This paper is motivated by this recent progress. We analyze the dense limit of the Dawid-Skene model. It is shown that it belongs to a larger class of low-rank matrix estimation problems for which it is possible to express the asymptotic, Bayes-optimal, performance in a simple closed form. In the dense limit the mapping to a low-rank matrix estimation problem provides an approximate message passing algorithm that solves the problem algorithmically. We identify the regions where the algorithm efficiently computes the Bayes-optimal estimates. Our analysis refines the results of [2, 3] about optimality of message passing algorithms by characterizing regions of parameters where these algorithms do not match the Bayes-optimal performance. We further study numerically the performance of approximate message passing, derived in the dense limit, on sparse instances and carry out experiments on a real world dataset.

I. INTRODUCTION

The development of large-scale crowdsourcing platforms, such as Amazon’s MTurk, has popularized crowdsourcing as a simple approach to solve various problems that remain difficult for computers but require little effort to human workers. The overall strategy is simple: the requester poses a set of tasks that are allocated to several individuals from a pool of workers (the crowd). The workers answer according to their abilities and their will. Importantly, the set of answers is typically not unambiguous and post-processing has to be performed in order to infer the true information (typically labels) from the noisy observations (answers). With the crowds answers at hand the objective becomes to infer the true labels with as few mistakes as possible. The outcome of such a strategy strongly depends on the competences of the individuals; which makes it necessary to infer not only the true labels, but also the competences of the individuals.

A large fraction of the theoretical work on crowdsourcing focuses on the so-called Dawid-Skene (DS) model, after the authors of the seminal paper [1]. In the DS model we consider N workers, each of them of a certain reliability that denotes the probability that a worker gives the correct answer, represented by $0 \leq p_i^0 \leq 1$ for worker $i = 1, \dots, N$. Further there are M tasks, each having a true label that we denote by $v_j^0 \in \{\pm 1\}$ for task $j = 1, \dots, M$. The worker i is assigned a subset of tasks $j \in \partial i$ to which it assigns an answer $Y_{ij} \in \{\pm 1\}$. We denote $Y_{ij} = 0$ if $j \notin \partial i$, that is for tasks j that were not assigned to worker i . In the DS model labels provided by worker i for task j are modeled as

$$P(Y_{ij}) = p_i^0 \delta(Y_{ij} - v_j^0) + (1 - p_i^0) \delta(Y_{ij} + v_j^0). \quad (1)$$

Moreover it is assumed that the p_i^0 s are drawn independently from some probability distribution P_{p^0} .

The task allocation design (which tasks gets assigned to which worker) is in general part of the crowdsourcing problem and various strategies have been described and studied in the literature. It has been argued that designing the graph of assignments at random has practical and optimality advantages, among others it enables a sharp theoretical analysis of the problem, see e.g. [4]. On bipartite random regular graphs, where every worker is assigned r tasks and every task is assigned to l workers, the DS model has been studied in detail by [2, 4].

While in general reconstructing the true labels and workers reliabilities from the observed answers Y_{ij} is an NP-hard problem, the authors of [2, 3] obtained a remarkable theorem stating that in certain regions of parameters belief propagation reconstructs the true labels optimally in the limit of large system sizes. The belief propagation algorithm for crowdsourcing was first suggested by [5]. Many other algorithms for crowdsourcing exist in the literature, but as far as we know none of them reaches optimal performance for large random instances of the DS model for a regime where the probability of error per task stays bounded away from zero.

The goal of the present paper is to carry out an asymptotic analysis of the DS model in the dense regime where each worker is assigned a constant fraction of the M tasks. Otherwise we are in the same setting as [2, 3], i.e. with random worker reliabilities and on random graphs. From our analysis it is possible to characterize more tightly the region of parameters for which belief propagation is optimal and for which it is not. We find cases where a first order phase transition appears in the error of reconstruction of the true labels. Such a first order phase transition is associated with a region of parameters in which belief propagation does not match the asymptotically optimal performance. Our

work can thus be seen as a follow-up on [2, 3] providing a refined analysis of the regions of parameters for which belief propagation is or is not asymptotically optimal.

In section II we first define a dense version of the DS model. In the dense DS model workers reliabilities are close to 1/2 as otherwise inference becomes trivially easy. The dense DS model belongs to a class of low-rank matrix factorization problems, as studied recently by statistical physics techniques in [6, 7]. The authors derived the approximate message passing (AMP) algorithm that efficiently computes the Bayes-optimal estimator and analyzed the Bayes-optimal performance in a closed form. One of the merits of AMP is that its asymptotic performance can be described via the so-called state evolution, as proven in [8, 9]. The performance of the Bayes-optimal estimator was also later put on fully rigorous bases in the work of [10] under assumptions that include the dense DS model as considered in section II. We apply the results derived in those papers and identify the region of parameters for which the associated approximate message passing algorithm is suboptimal. In section IV we then investigate numerically how the results – valid in the dense regime – transfer into the sparse regime as originally considered in [2, 3]. We finally carry out some experiments on a real world dataset to show that AMP reaches similar performance to competing state-of-the-art algorithms.

II. DENSE LIMIT OF THE DAWID-SKENE MODEL

A. Definition of the dense limit

In this section the dense Dawid-Skene (dDS) model for crowdsourcing is introduced where each of the N workers is assigned a constant fraction of the M questions. It is shown that it can be modeled as a low-rank matrix factorization problem as studied in [6, 7, 10–12].

Let the probability that worker i provides a correct answer, p_i^0 , be close to 1/2 and introduce the parameters ν and θ_i^0 such that $p_i^0 = (1 + \sqrt{\nu/N} \theta_i^0)/2$. The parameter ν is an overall scale parameter, while θ_i^0 is the rescaled reliability of worker i taken from some probability distribution P_{θ^0} . In the dense limit the $1/\sqrt{N}$ scaling causes all the reliabilities to be close to 1/2 which is the interesting regime because a total of $\Theta(N)$ answers is received for each question.¹ If we had another scaling, the problem would become either trivially hard or easy in the thermodynamic limit, $N \rightarrow \infty$.

The true label of task j is v_j^0 , with $v_j^0 \in \{\pm 1\}$ distributed as P_{v^0} . We denote with Y_{ij} the label assigned to question j by worker i and assume $Y_{ij} \in \{0, \pm 1\}$. If $Y_{ij} = 0$ question j was left out by worker i . Set $\alpha := \frac{M}{N}$ and consider a system in which each worker is posed $(1 - \rho)M$ questions in the limit where $M, N \rightarrow \infty$, while $\alpha = \Theta(1)$ and $\rho = \Theta(1)$. In this limit the likelihood in the DS model becomes

$$\begin{aligned} P(Y_{ij} = \pm 1 \mid \theta_i, v_j) &= (1 - \rho) \cdot \frac{1}{2} \cdot \left(1 \pm \sqrt{\frac{\nu}{N}} \theta_i v_j \right), \\ P(Y_{ij} = 0 \mid \theta_i, v_j) &= \rho \end{aligned} \quad (2)$$

where we assumed that the fraction of un-answered questions, ρ , is independent of (i, j) . The rest of the present section is set in the limit where $N \rightarrow \infty$ and all the other parameters $\theta_i, \nu, \alpha, \rho = \Theta(1)$. Later, in section IV, we will discuss how to extrapolate the results into the sparse regime where each worker is only assigned to $O(1)$ tasks.

Worker with $\theta_i = 0$ give answers that are completely uninformative and will be called *spammers*. On the contrary, if $\theta_i \gg 1$ the answers are “strongly” aligned with the truth and we refer to such workers as *hammers*. Adversaries are characterized by $\theta_i < 0$. They are also considered hammers if $\theta_i \ll -1$ because their answers are aligned against the truth, as opposed to the random alignment of the spammers.

B. Equivalence to low-rank matrix estimation

The dDS model is a special case of bipartite low-rank (rank one in the present case) matrix factorization as formulated in a much more general setting in [7]. In the rest of this section we follow closely that paper and review the results that will be applied to the present model.

In the theoretical part of this work (i.e. in all but section IV C) we assume that the distributions from which the ground truth reliabilities θ_i^0 and labels v_j^0 are drawn, P_{θ^0} and P_{v^0} respectively, are known. Under these assumptions

¹ We make use of the standard big-theta and big-O notation. We refer to a function as $\Theta(N)$ if its dominant asymptotic growth rate is proportional to N . While $O(N)$ refers to an asymptotic growth rate that is bounded by some constant times N .

we aim to (a) compute efficiently the Bayes-optimal estimators of θ_i^0 and v_j^0 , given the answers Y_{ij} and (b) to evaluate the asymptotic inference performance.

Denoting by $\boldsymbol{\theta} \in \mathbb{R}^N$ the vector of rescaled reliabilities for all N workers, and $\mathbf{v} \in \mathbb{R}^M$ the vector of labels, we set

$$\mathbf{w} := \frac{\boldsymbol{\theta} \mathbf{v}^T}{\sqrt{N}} \quad (3)$$

and re-express (2) as

$$P(Y_{ij} | w_{ij}) = \exp(g(Y_{ij}, w_{ij})) \quad , \quad g(Y_{ij}, w_{ij}) = \begin{cases} \log\left(\frac{1-\rho}{2}\right) + \log(1 \pm \sqrt{\nu} w_{ij}) & \text{if } Y_{ij} = \pm 1 \\ \log(\rho) & \text{if } Y_{ij} = 0 \end{cases} . \quad (4)$$

From Bayes' theorem we obtain the corresponding posterior probability distribution

$$P(\boldsymbol{\theta}, \mathbf{v} | \mathbf{Y}) = \frac{1}{Z(\mathbf{Y})} \prod_{1 \leq i \leq N} P_\theta(\theta_i) \prod_{1 \leq j \leq M} P_v(v_j) \prod_{1 \leq i \leq N, 1 \leq j \leq M} e^{g(Y_{ij}, w_{ij})} . \quad (5)$$

The Bayes-optimal estimates, $\hat{\boldsymbol{\theta}}$, that minimize the mean-squared-error (MSE) on $\boldsymbol{\theta}$

$$\text{MSE}_\theta = \frac{1}{N} \sum_i \left(\hat{\theta}_i - \theta_i^0 \right)^2 \quad (6)$$

and the bitwise error-rate (ER) on \mathbf{v}

$$\text{ER}_v = \frac{1}{M} \sum_j \mathbb{I}[\hat{v}_j \neq v_j^0] = \frac{1}{M} \sum_j \left(\frac{\hat{v}_j - v_j^0}{2} \right)^2 = \frac{1}{2} \frac{1}{M} \sum_j (1 - \hat{v}_j v_j^0) \quad (7)$$

read

$$\hat{\theta}_i^{\text{MMSE}}(\mathbf{Y}) = \int d\theta_i \theta_i P(\theta_i | \mathbf{Y}) \quad \text{and} \quad \hat{v}_j^{\text{MER}}(\mathbf{Y}) = \text{sign} \int dv_j v_j P(v_j | \mathbf{Y}) , \quad (8)$$

where $P(x_k | \mathbf{Y})$, with $x_k \in \{\{\theta\}_{i=1, \dots, N}, \{v_j\}_{j=1, \dots, M}\}$, is the posterior marginal of (5) after integrating out all variables except x_k . Hence inferring the reliabilities and labels in the crowdsourcing problem reduces to evaluating the marginal expectations of the posterior probability distribution. In general this is a difficult task. The contribution of the present work is to realize that the dDS model falls into a class of low-rank matrix estimation problems for which the posterior probability distribution can be evaluated, as shown in [7]. Using these results, the phase diagram can be evaluated in great detail.

C. Approximate message passing

The approximate message passing (AMP) algorithm for low-rank matrix estimation is a simplification of belief propagation in the limit of dense graphical models. In this limit both, belief propagation and AMP have the same asymptotic performance. However, AMP is much simpler to implement and has a favorable scaling w.r.t the problem size. It is closely related to the Thouless-Anderson-Palmer equations [13] from the theory of spin glasses, with correct time indices [14, 15]. AMP for low-rank matrix factorization was first derived for special cases in [8, 12] and in its general form in [6, 7].

AMP can be derived starting from belief propagation for the graphical model where both the reliabilities and labels are variable nodes, and there are pair-wise factor nodes corresponding to the answers Y_{ij} . The following two simplifications of BP are then made. First, the BP messages are replaced by their means and variances which eradicates the necessity of tracking a whole function for each message. Secondly, each (mean and variance) *message* is replaced by its *marginal* version, reducing the complexity from $O(N^2)$ messages to $O(N)$ marginals. For details we refer the reader to [7].

To state the AMP algorithm for the dense DS model it is necessary to specify the denoising functions $f_\theta(A_\theta, B_\theta)$ and $f_v(A_v, B_v)$ that depend on the priors P_θ and P_v , respectively. A and B are estimates for the parameters of a Gaussian distribution that are computed self-consistently. The estimate \hat{x}_k – with $x_k \in \{\{\theta_i\}_{i=1, \dots, N}, \{v_j\}_{j=1, \dots, M}\}$

– are then computed as the mean of the prior weighted with this effective Gaussian. The estimates for their variance are obtained from the derivative w.r.t. B .

$$\hat{x} := f_x(A_x, B_x) = \frac{1}{Z_x(A_x, B_x)} \int dx x P_x(x) e^{-\frac{1}{2}A_x x^2 + B_x x}, \quad \sigma_x = \partial_{B_x} f_x(A_x, B_x), \quad (9)$$

where x can stay either for θ or v .

To state AMP we need to define the Fisher score matrix

$$S_{ij} := \left. \frac{\partial g(Y_{ij}, w_{ij})}{\partial w_{ij}} \right|_{w_{ij}=0} = Y_{ij} \cdot \sqrt{\nu}, \quad (10)$$

where $g(Y_{ij}, w_{ij})$ is defined in eq. (4). Further we define the Fisher information (inverse effective noise) of the noisy observation channel

$$\Delta^{-1} = \mathbb{E}_{P(Y_{ij}|w_{ij}=0)} \left[\left(\left. \frac{\partial g(Y_{ij}, w_{ij})}{\partial w_{ij}} \right|_{w_{ij}=0} \right)^2 \right] = (1 - \rho)\nu. \quad (11)$$

Given these definitions AMP is an iterative scheme that we outline in Algorithm 1. The numerical implementation might profit from an adequate damping scheme in order to enhance convergence even on small instances or when the model assumptions are not satisfied.

Algorithm 1: Approximate message passing for crowd sourcing.

Data: $\mathbf{S}, \Delta, \delta$; // \mathbf{S} and Δ according to (10) and (11) respectively.
Result: MMSE estimates $\hat{\mathbf{v}}$ and $\hat{\boldsymbol{\theta}}$
Initialize: $\hat{\mathbf{v}} \leftarrow \hat{\mathbf{v}}^{\text{init}} \sim P_v(\mathbf{v}), \hat{\boldsymbol{\theta}} \leftarrow \hat{\boldsymbol{\theta}}^{\text{init}} \sim P_\theta(\boldsymbol{\theta}); \sigma_v \leftarrow 1, \sigma_\theta \leftarrow 1; \hat{\mathbf{v}}^{\text{old}} \leftarrow \mathbf{0}, \hat{\boldsymbol{\theta}}^{\text{old}} \leftarrow \mathbf{0}$;
while $\|\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^{\text{old}}\|_2^2 + \|\hat{\mathbf{v}} - \hat{\mathbf{v}}^{\text{old}}\|_2^2 > \delta$ **do**
 $\mathbf{B}_\theta \leftarrow \frac{1}{\sqrt{N}} \mathbf{S} \hat{\mathbf{v}} - \frac{1}{\Delta} \hat{\boldsymbol{\theta}}^{\text{old}} \sigma_v$;
 $\mathbf{A}_\theta \leftarrow \frac{1}{N\Delta} \hat{\mathbf{v}}^T \hat{\mathbf{v}}$;
 $\mathbf{B}_v \leftarrow \frac{1}{\sqrt{N}} \mathbf{S}^T \hat{\boldsymbol{\theta}} - \frac{\alpha}{\Delta} \hat{\mathbf{v}}^{\text{old}} \sigma_\theta$;
 $\mathbf{A}_v \leftarrow \frac{1}{N\Delta} \hat{\boldsymbol{\theta}}^T \hat{\boldsymbol{\theta}}$;
 $\hat{\boldsymbol{\theta}}^{\text{old}} \leftarrow \hat{\boldsymbol{\theta}}, \hat{\mathbf{v}}^{\text{old}} \leftarrow \hat{\mathbf{v}}$;
 $\hat{\boldsymbol{\theta}} \leftarrow f_\theta(\mathbf{A}_\theta, \mathbf{B}_\theta), \sigma_\theta \leftarrow \frac{1}{N} \sum_{1 \leq i \leq N} \partial_{B_{\theta_i}} f_\theta(A_{\theta_i}, B_{\theta_i})$;
 $\hat{\mathbf{v}} \leftarrow f_v(\mathbf{A}_v, \mathbf{B}_v), \sigma_v \leftarrow \frac{1}{M} \sum_{1 \leq j \leq M} \partial_{B_{v_j}} f_v(A_{v_j}, B_{v_j})$;
end

D. State Evolution

The AMP algorithm depends on the realization of the disorder \mathbf{Y} and consequently so do the AMP estimates $\hat{\boldsymbol{\theta}}, \hat{\mathbf{v}}$ for the reliabilities and task labels. Quite remarkably, in the large size limit $N \rightarrow \infty$, the performance of the algorithm can be tracked with high probability by the so-called state evolution (SE) equations. This has been proven rigorously in [8, 9].

In the Bayes-optimal setting, where the true distributions P_{θ^0} and P_{v^0} are known and equal to P_θ and P_v respectively, the overlap of the AMP estimates with the true solution can be quantified in terms of the two order parameters

$$M_\theta^t = \frac{1}{N} \sum_{1 \leq i \leq N} \hat{\theta}_i^t \theta_i^0, \quad (12)$$

$$M_v^t = \frac{1}{M} \sum_{1 \leq j \leq M} \hat{v}_j^t v_j^0.$$

Where x^0 indicates the true value of x , and t the iteration step of the AMP equations (Alg. 1).

The SE equations imply that these order parameters evolve with high probability as

$$\begin{aligned} M_v^{t+1} &= \mathbb{E}_{v^0, W} \left[f_v \left(\frac{M_\theta^t}{\Delta}, \frac{M_\theta^t}{\Delta} v^0 + \sqrt{\frac{M_\theta^t}{\Delta}} W \right) v^0 \right], \\ M_\theta^t &= \mathbb{E}_{\theta^0, W} \left[f_\theta \left(\frac{\alpha M_v^t}{\Delta}, \frac{\alpha M_v^t}{\Delta} \theta^0 + \sqrt{\frac{\alpha M_v^t}{\Delta}} W \right) \theta^0 \right]. \end{aligned} \quad (13)$$

where W is an effective Gaussian random variable of zero mean and unit variance, $v^0 \sim P_v$, $\theta^0 \sim P_\theta$, the functions f_v and f_θ are defined in (9), $\alpha = M/N$ and Δ is the effective noise (11).

Let us call M_θ^{SE} and M_v^{SE} the fixed points of the SE equations (13). These fixed points are then associated to the MSE (6) and ER (7) as reached by the AMP algorithm through

$$\text{MSE}_\theta^{\text{AMP}} = \mathbb{E}_\theta(\theta^2) - M_\theta^{\text{SE}}, \quad (14)$$

$$\text{ER}_v^{\text{AMP}} = (1 - R_v^{\text{SE}})/2, \quad (15)$$

where we introduced the order parameter $R_v^t = 1/M \sum_i \text{sign}(\hat{v}_i^t) v_i^0$

$$R_v^{\text{SE}} = \mathbb{E}_{v^0, W} \left\{ \text{sign} \left[f_v \left(\frac{M_\theta^{\text{SE}}}{\Delta}, \frac{M_\theta^{\text{SE}}}{\Delta} v^0 + \sqrt{\frac{M_\theta^{\text{SE}}}{\Delta}} W \right) v^0 \right] \right\}. \quad (16)$$

E. Bayes-optimal error and sub-optimality of message passing algorithms

As conjectured in [7] and proven rigorously in [10] the performance of the Bayes-optimal estimator (8) can be evaluated in the large size limit $N \rightarrow \infty$ with $\alpha = \Theta(1)$ from the global minimizer of the so-called replica symmetric Bethe free energy, which reads

$$\begin{aligned} \phi_{\text{Bethe}}(M_\theta, M_v) &= \alpha \frac{M_\theta M_v}{2\Delta} - \alpha \mathbb{E}_{v^0, W} \left[\log Z_v \left(\frac{M_\theta}{\Delta}, \frac{M_\theta}{\Delta} v^0 + \sqrt{\frac{M_\theta}{\Delta}} W \right) \right] - \\ &\quad - \mathbb{E}_{\theta^0, W} \left[\log Z_\theta \left(\frac{\alpha M_v}{\Delta}, \frac{\alpha M_v}{\Delta} \theta^0 + \sqrt{\frac{\alpha M_v}{\Delta}} W \right) \right]. \end{aligned} \quad (17)$$

where the functions Z_θ and Z_v are defined in (9) and the rest of the variables are defined in the same way as in the SE. Assume M_θ^* and M_v^* are the global minimizers of the above Bethe free energy. Then the minimum-mean-squared-error (MMSE) and the minimum-error-rate (MER) are expressed as

$$\text{MMSE}_\theta = \mathbb{E}_\theta(\theta^2) - M_\theta^*, \quad (18)$$

$$\text{MER}_v = \frac{1}{2}(1 - R_v^*), \quad (19)$$

where R_v^* is obtained from M_θ^* via (16).

It is straightforward to observe that the SE equations are in fact stationarity conditions of the Bethe free energy. Hence the fixed points of the state evolution are critical points of the Bethe free energy. Whether or not the SE reaches the global minimizer M_θ^* , M_v^* depends on the shape of the Bethe free energy and the initialization of the SE equations at $t = 0$. Canonically the SE is initialized in such a way that the initial estimators are simply taken from the prior distributions.

We can now explain the key point of the present paper. Previous work [2, 3] proved asymptotic optimality of belief propagation under certain assumptions on the parameters of the model. The present analysis of the dDS model is able to determine sharply in what regions of parameters AMP matches the Bayes-optimal estimator and when it does not, thus refining the previous picture in the limit, where AMP and BP are asymptotically equivalent.

Previously we reduced the high-dimensional model into the investigation of the two-variable free energy function (17). In particular, the phases in which AMP does not match the Bayes-optimal estimator can be characterized in terms of the critical points of the free energy and whether or not the state evolution (13) converges to the global minimum of the free energy (17). The way we check this in practice is that we initialize the state evolution in two different ways:

- Uninformative initialization, where $M_v^{t=0} = (\mathbb{E}_v(v))^2$ and $M_\theta^{t=0} = (\mathbb{E}_\theta(\theta))^2$. This corresponds to the uninformative initialization of the algorithm where the initial values of the estimators are simply taken equal to the mean of the prior distributions P_θ and P_v . The error achieved by the AMP algorithm is then given by iteration of (13) from this uninformative initialization.
- Informative initialization, where $M_v^{t=0} = \mathbb{E}_v(v^2)$ and $M_\theta^{t=0} = \mathbb{E}_\theta(\theta^2)$ so that the initial mean-squared-errors are zero. This is not possible within the algorithm without the knowledge of the ground truth and it is purely used for the purpose of the analysis. If the iteration of the SE equations (13) from this informative initialization leads to a different fixed point than from the uninformative initialization, then the free energies of the two fixed points need to be compared and the larger one surely does not correspond to the Bayes-optimal performance.

This procedure is sufficient provided there are no other fixed points. If there are, the free energy of all of them needs to be compared.

Zero-mean priors and uninformative fixed point If both prior distributions P_θ and P_v have zero mean, the uninformative initialization $M_\theta = M_v = 0$ is a fixed point of the SE and equations (13) can be expanded around this fixed point. In first order we obtain

$$M_\theta^t = \frac{\alpha}{\Delta} (\mathbb{E}_\theta [\theta^2])^2 M_v^t \quad (20)$$

$$M_v^{t+1} = \frac{1}{\Delta} (\mathbb{E}_v [v^2])^2 M_\theta^t, \quad (21)$$

implying that the uninformative fixed point is numerically stable for $\Delta^2 > \alpha (\mathbb{E}_v [v^2])^2 (\mathbb{E}_\theta [\theta^2])^2$ and unstable otherwise. Therefore we define the critical effective noise, Δ_c , as

$$\Delta_c = \sqrt{\alpha} \cdot \mathbb{E}_\theta [\theta^2] \mathbb{E}_v [v^2]. \quad (22)$$

For $\Delta < \Delta_c$ the uninformative initialization becomes numerically unstable. The threshold Δ_c correspond to the 2nd order phase transition in the behaviour of the AMP algorithm, meaning that the overlap reached by the algorithm is non-analytic and continuous at Δ_c .

In the case where both the priors, P_θ and P_v , have zero mean, we can divide the region of parameters into the following three phases:

- **Easy phase:** The free energy (17) has a unique minimum and this minimum is associated with a positive overlap with the ground-truth configuration. Consequently iterating the state evolution (13) yields an informative fixed point from both, the informative, as well as the (perturbed) uninformative initializations. AMP is Bayes-optimal.
- **Hard phase:** In this phase at least two minima of the free energy (17) coexist; at least one local minimum of small overlap and a global minimum of larger overlap. The outcome of iterating the state evolution equations now depends on the initialization: while the informative initialization yields a fixed point with large overlap, the uninformative initialization leads to a fixed point of low overlap. This is precisely the region of parameters where the approximate message passing algorithms do not reach the information-theoretically optimal performance and *AMP is not Bayes-optimal*.
- **Impossible phase:** When the global minimum of (17) is associated to the trivial, non-informative, fixed point corresponding to zero overlap, we talk about a phase of impossible inference. Otherwise this region is indeed similar to the easy phase in the sense that AMP is Bayes-optimal.

If at least one of the priors has non-zero mean, then the distinction of an impossible phase is not meaningful and one would only have an easy and a hard phase, the later is defined by asymptotic sub-optimality of the AMP algorithm.

Let us further define the following three thresholds that are associated with the existence of a hard phase. The hard phase is always linked to the presence of a first order phase transition, i.e. a discontinuity in the asymptotic value of the overlap reached by the Bayes-optimal estimator. The *algorithmic threshold* Δ_{alg} is the largest value of effective noise, Δ , below which the AMP algorithm asymptotically matches the Bayes-optimal performance. The *spinodal threshold*, Δ_{sp} , is the smallest values of effective noise above which the informative initialization converges to a different fixed point than the (perturbed) uninformative initialization. The *information theoretic transition*, $\Delta_{\text{alg}} < \Delta_{\text{IT}} < \Delta_{\text{sp}}$, is where the value of the Bethe free energy of the fixed point reached from the uninformative initialization crosses with the free energy of the fixed point reached from the informative initialization. The discontinuity in overlap happens at Δ_{IT} . Remark that while in some models, such as the stochastic block model [6], we find $\Delta_c = \Delta_{\text{alg}}$ in general and in the present model $\Delta_c \neq \Delta_{\text{alg}}$.

III. PHASE DIAGRAMS FOR THE DENSE DAVID-SKENE MODEL

A key property of the results we described so far is that the asymptotic behaviour of the AMP algorithm and of the Bayes-optimal estimator depend only on the priors P_v , P_θ and the effective noise $\Delta = 1/[(1-\rho)\nu]$. In what follows concrete priors will be considered.

It is assumed that the ground truth task labels are generated from

$$P_v(v_j) = (1 - \beta)\delta(v_j - 1) + \beta\delta(v_j + 1). \quad (23)$$

With the parameter $\beta \in [0, 1]$ accounting for a bias in the dataset.

We start our discussion with worker reliabilities θ_i that were drawn from a skewed Rademacher-Bernoulli (RB) prior

$$P_\theta(\theta) = (1 - \mu)\delta(\theta) + \mu[(1 - \lambda)\delta(\theta - 1) + \lambda\delta(\theta + 1)]. \quad (24)$$

Besides its simplicity the phase diagram for this case comprises the essential features. Tuning μ from zero to one interpolates between an uninformative crowd of mere spammers and an informative crowd. The fraction of adversaries is controlled by λ . In physics terms the workers with $\theta = -1$ are spins that are coupled to the questions by an anti-ferromagnetic interaction, whereas the workers with $\theta = 1$ are ferromagnetically coupled. Consequently also the adversaries enhance our ability to recover the correct labels, if they can be identified, as they align anti-parallel to the truth.

The RB prior is the dense version of what is sometimes referred to as the “spammer-hammer” model in the literature [4]: Workers are either spammers that provide random answers or hammers that align very strongly with (or opposed to) the truth. Here the situation is slightly different as we assume a very weak alignment of $\Theta(1/\sqrt{N})$, cf. (2). In the dDS sending $\nu \rightarrow \infty$ and thus $\Delta \rightarrow 0$ approximates the hammers. The limit $\nu \rightarrow N$ will be considered in section IV.

A. The case of symmetric priors

If $\lambda = 1/2$ and $\beta = 1/2$ both the priors P_v and P_θ have zero mean and the SE equations in (13) have a trivial fixed point at $M_v^* = M_\theta^* = 0$. Expansion around this fixed point yields

$$M_v^{t+1} = \alpha \frac{\mu^2}{\Delta^2} \cdot M_v^t - \alpha^2 \frac{\mu^2}{\Delta^2} \left[\frac{\mu}{\Delta} + \frac{\mu^2}{\Delta^2} \right] \cdot (M_v^t)^2 + O((M_v^t)^3). \quad (25)$$

The linear term gives the stability criterion of the trivial fixed point that we had already derived in (22)

$$\Delta_c = \sqrt{\alpha} \cdot \mu. \quad (26)$$

In Fig. 1 we present the phase diagram for several values of $\alpha = M/N$. We plot the stability threshold Δ_c as well as the three phase transitions associated with the existence of the hard phase. We mark the phases where inference is algorithmically easy, hard and impossible. In particular, we find that a *hard phase appears* for small enough μ as depicted in the figure. Regions with small μ correspond to crowds that contain mostly spammers. For $\alpha = 1$ the hard phase appears only if the vast majority of the workers are spammers. When α grows (shrinks) the hard region grows (shrinks) as well. In the region where the hard phase is absent (26) provides the right criterion to locate the phase transition from the easy to the impossible phase.

B. Biased labels and worker reliabilities

If $\lambda \neq 1/2$ or $\beta \neq 1/2$ the trivial fixed point $M_v = M_\theta = 0$ does not exist anymore. We illustrate in Fig. 2 how this changes the phase diagram and the achievable MSE. For the case $\alpha = 1$ and $\mu = 0.02$ we plot the MSE reached by the state evolution from the informative and the uninformative initialization.

First (left top panel), we consider the unbiased case with $\beta = 1/2$, but $\lambda \neq 1/2$ as already plotted in Fig. 1. In the bottom-left panel we consider the case where λ changes. Due to the present symmetry it suffices to restrict the attention to $\lambda > 1/2$. When more hammers than adversaries are present, i.e. for $\lambda > 1/2$ the trivial fixed point at $M_v = 0$ disappears and instead another fixed point with low but positive overlap (i.e. error smaller than 1) appears. The hard phase shrinks as shown in the bottom-left panel of Fig. 2.

If the dataset is biased, i.e. $\beta \neq 1/2$, the change is quantitatively more dramatic, but phenomenologically very similar, cf. top-right panel in Fig. 2. Upon slight change in β the hard phase shrinks considerably. For a large range of values of β and λ the hard phase entirely disappears as in the bottom-right panel in Fig. 2.

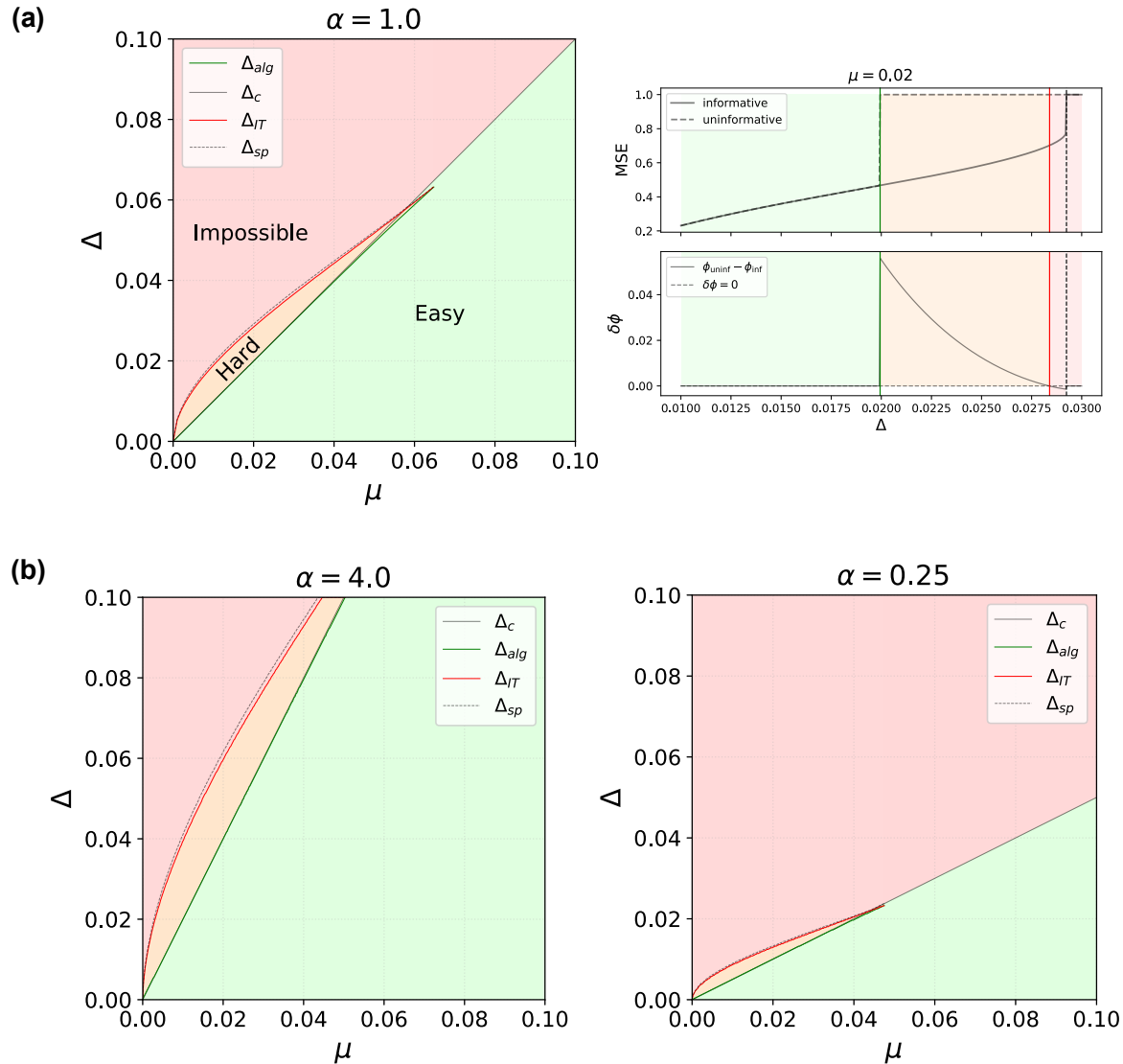


Figure 1. (a) left panel: The phase diagram for a Rademacher-Bernoulli prior on θ with $\lambda = 1/2$ and no bias in the distribution of the labels ($\beta = 1/2$). When the fraction of spammers is very large (small μ) a hard phase appears where the AMP algorithm is not able to reach the information-theoretically optimal performance. (a) right panel: cut of the phase diagram corresponding to $\mu = 0.02$, i.e. only 2% of non-spammers. We plot the MSE (top) and the difference in the free energy (bottom) of the two fixed points as a function of Δ . Note that in this case we still have $\Delta_c > \Delta_{alg}$ but both are indistinguishably close. In the hard region (orange) the AMP algorithm reaches $MSE = 1$ while the Bayes-optimal estimator reaches the depicted MSE. (b) Phase diagrams with all parameters set to the same values, but α different. When α grows (shrinks) inference becomes easier (harder) and the hard region grows (shrinks). The tricritical point for $\alpha = 1/4$ is located around $\mu \approx 0.048$ whereas for $\alpha = 4$ it is around $\mu \approx 0.077$.

C. The impact of α

Recall that α is the ratio of tasks to workers in our model. By virtue of the $\sqrt{\nu/N}$ scaling of the signal, cf. (2), we have two competing mechanisms when N is increased: on the one hand the signal becomes weaker, on the other hand we obtain more answers per question. Equation (26) tells us that we should expect inference to become easier when α increases. Indeed, if we fix Δ and consider how the performance changes with α it follows from the SE that in order to achieve higher overlap it is necessary to increase the fraction of questions distributed to each worker, i.e. by increasing α . This improves the estimation of θ , which in turn improves the estimate of \mathbf{v} . We depict this by plotting the error rate against α for two different values in Fig. 3.

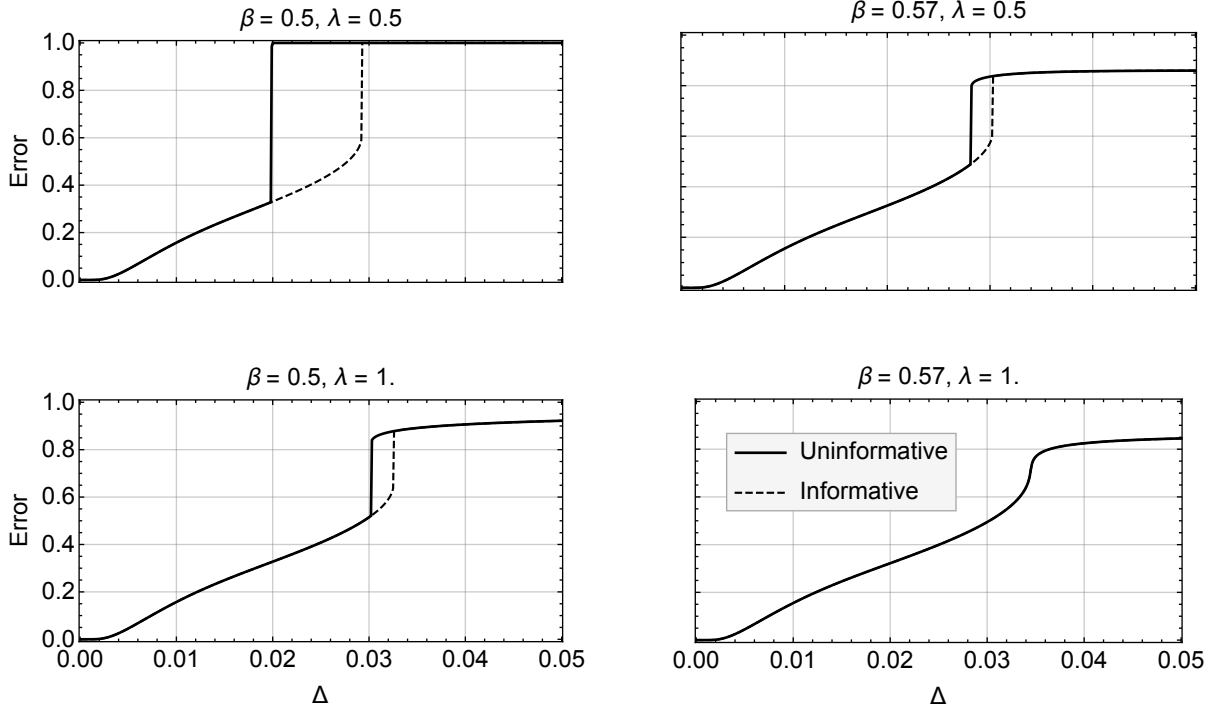


Figure 2. Influence of bias in the distribution of labels and worker reliabilities on the performance. Here we plot the resulting error rate (15) for $\alpha = 1$ and $\mu = 0.02$ as reached from the uninformative (bold) and informative (dashed) initialization. For bias in the labels ($\beta \neq 0.5$) or in the workers abilities ($\lambda \neq 0.5$) the trivial fixed point (error equal to one) is replaced by another fixed point with slightly lower error. The hard phase in these examples appears at larger noise and shrinks or might disappear as in the bottom right panel.

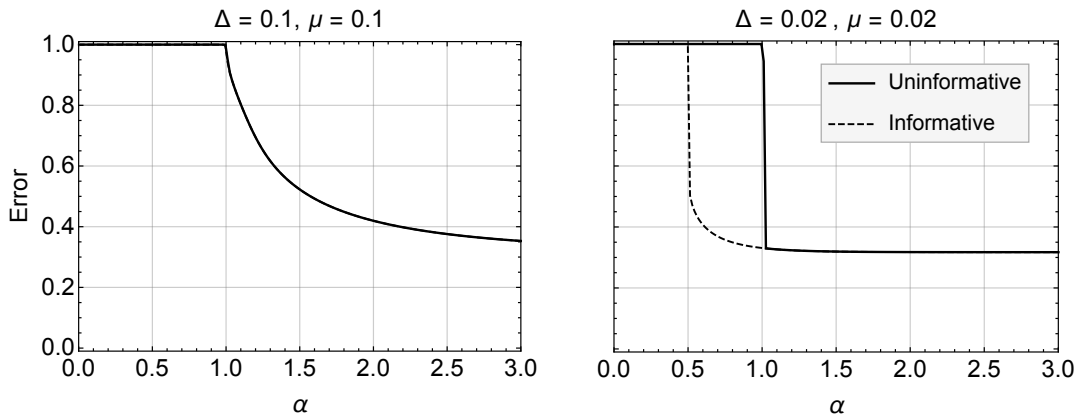


Figure 3. The behaviour of the error rate versus α for the RB prior (24) with $\lambda = 1/2$ and bias is set to $\beta = 1/2$.

How does the hard phase vary with α ? We answer this question in Fig. 1 (b) where we show that the hard phase grows further in the impossible phase when α is increased, while it shrinks when α is decreased.

D. Dealing with other priors

The derivation of section II applies to any prior as long as $\theta = O(1)$. Indeed, many features persist if we replace (24) by

$$P_\theta(\theta) = (1 - \mu)\delta(\theta) + \mu\phi(\theta).$$

with $\phi(\theta)$ some appropriate distribution (we have considered $\phi(\theta)$ being a beta distribution or a Gaussian). For instance (25) still holds when $\phi(\theta)$ is a standard Gaussian and as for the RB prior a first order transition is triggered by very noisy θ , i.e. only very few hammers and mostly spammers in the crowd.

One might also replace the delta distribution by some other sparsity inducing distribution. A case for which the corresponding integrals are tractable analytically is that of a mixture of two Gaussians, centered around $\bar{\theta}_L$ ($\bar{\theta}_R$) with variance σ_L^2 (σ_R^2).

$$P_\theta(\theta) = (1 - \mu)\mathcal{N}(\theta; \bar{\theta}_L, \sigma_L^2) + \mu\mathcal{N}(\theta; \bar{\theta}_R, \sigma_R^2).$$

Under this choice and with $\beta = 1/2$ in (23) the SE equations (13) can be expressed as

$$M_v^{t+1} = G\left(\frac{1}{\Delta}T\left(\frac{\alpha}{\Delta}M_v^t\right)\right) \quad (27)$$

with

$$\begin{aligned} G(x) &= \mathbb{E}_W \left\{ \tanh(x + \sqrt{x}W) - \tanh(-x + \sqrt{x}W) \right\} \\ T(q) &= \mu \cdot \mathbb{E}_W \left\{ \frac{\left[\left(\bar{\theta}_R + \sqrt{\frac{q}{1+q\sigma_R^2}}\sigma_L^2 W \right) + \frac{1-\mu}{\mu} \left(\frac{1+q\sigma_R^2}{1+q\sigma_L^2} \right)^{\frac{3}{2}} \left(\frac{\bar{\theta}_L + q\sigma_L^2 \bar{\theta}_R}{1+q\sigma_R^2} + \sqrt{\frac{q}{1+q\sigma_R^2}} W \right) \cdot \exp\left(-\frac{1}{2}Q(W)\right) \right]^2}{1 + \frac{1-\mu}{\mu} \sqrt{\frac{1+q\sigma_R^2}{1+q\sigma_L^2}} \exp\left(-\frac{1}{2}Q(W)\right)} \right\} \\ Q(W) &= \frac{1 + q\sigma_R^2}{1 + q\sigma_L^2} \left(W + \sqrt{\frac{q}{1+q\sigma_R^2}}(\bar{\theta}_R - \bar{\theta}_L) \right) - W^2, \end{aligned}$$

where \mathbb{E}_W indicates the average over the standard Gaussian measure on W . Varying the means ($\bar{\theta}_L$, $\bar{\theta}_R$) and variances (σ_L^2 , σ_R^2) then allows to interpolate between different scenarios.

IV. RELEVANCE OF THE RESULTS IN THE SPARSE REGIME

Our analysis of the dense DS model is based on the ground that the underlying graphical model (the bipartite question-worker-graph) is densely connected. This means that each task-node is connected to $\Theta(N)$ worker-nodes – and reversely each worker-node is connected to $\Theta(M)$ task-nodes. Allowing that some of the tasks remain unanswered introduces a sense of sparsity in the channel (cf. (2)). Our analysis assumes that $1 - \rho = \Theta(1)$. Existing mathematical literature on low-rank matrix estimation shows that the formulas we derived for the Bayes-optimal performance, hold true even when the degrees in the graph grow with N slower than linearly, i.e. when $(1 - \rho)N$ diverges with $N \rightarrow \infty$ [16, 17]. The regime where the above asymptotic results do not hold anymore is when $1 - \rho = O(1/N)$, which we refer to as the sparse regime. In this section we investigate numerically how the behaviour of the sparse DS model deviates from the predictions drawn from the dense DS model.

In the sparse regime considered here every worker is connected to d randomly chosen tasks, where $d = \Theta(1)$. Unless the quality of each answer is very high, the effective noise $\Delta = [(1 - \rho)\nu]^{-1}$ is overwhelming and inference impossible, unless $\nu = \Theta(N)$. Therefore we will consider the following “mapping”

$$\rho = 1 - \frac{d}{M} \quad \nu = n \cdot N, \quad (28)$$

with $n \in [0, 1]$ being a constant. Consequently in the sparse regime we are dealing with high quality workers as compared to the dense regime. This brings us close to the setting of previous literature on the DS model [2–5].

A. Approximate message passing on sparse graphs

We study numerically how the AMP algorithm behaves when the average degree of the nodes is small. In the following we will set $M = N$ such that the average degree of the task-nodes equals the average degree, d , of the worker-nodes.

Figure 4 (a) depicts results that are obtained by running AMP in the dense regime where $d = \Theta(N)$, for a system with 10^4 nodes. Except from finite size effects close to the phase transition the SE prediction agrees with the empirical results. For Fig. 4 (b) we fixed different values of Δ – by adjusting n so that $\Delta = \alpha/(nd)$ – and plotted the relative deviation from the SE when the degree d is varied. We also show the results obtained with the BP algorithm of [5] that are obtained by matching the prior and signal to noise ratio. In the limit of large N the BP results are exact even for finite d . We find as expected that when d is increased, the AMP performance approaches the prediction of the associated dense model and so does BP. While for very small d BP slightly outperforms AMP, the difference is not very significant (up to fluctuations).

We further quantify the difference in performance of BP and AMP in the sparse regime in Fig. 5. Now ν (i.e. n) is fixed and d (and hence Δ) varies. We compare AMP with its BP equivalent and find that BP always outperforms AMP, but again only slightly. The general trend is as expected: in the sparse regime BP is optimal and no other algorithm can outperform it. However, it is remarkable how quickly AMP becomes comparable to BP. In Fig. 5 (b) we fix d and vary ν (i.e. n), such that Δ varies in the same range as in Fig. 5 (a). We cannot explore the full range of Δ because we must restrict $n \leq 1$.

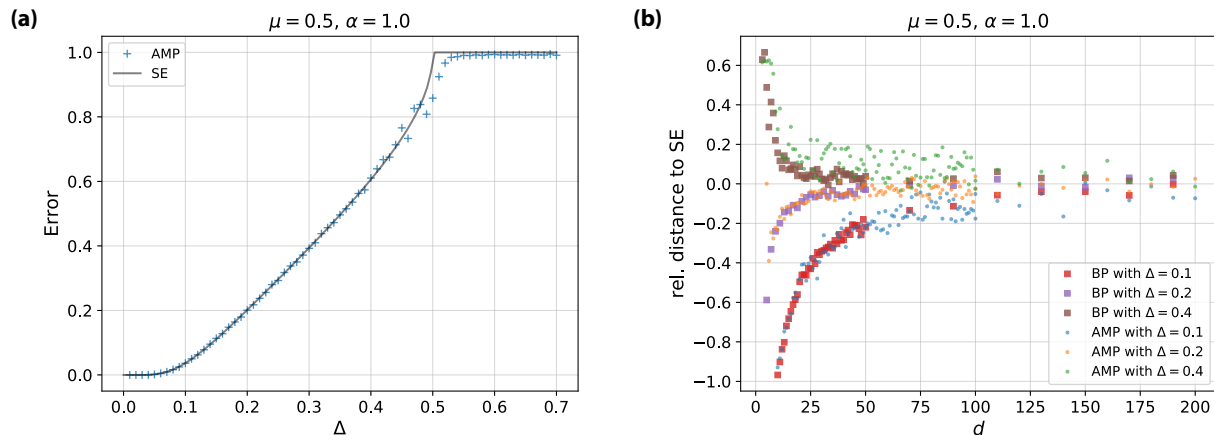


Figure 4. Evaluation of the AMP results in the sparse regime. (a) Numerical results for AMP in the dense regime for $N = M = 10^4$, averaged over 20 samples. (b) The relative distance of the AMP results to the SE prediction of the error when the average degree d and signal to noise ratio ν are varied such that Δ remains fixed. We also compare to the BP algorithm that is asymptotically exact in the sparse regime. We see that the SE gives an accurate description, already for d around 30 – 50. While AMP is suboptimal for low degrees d and BP still asymptotically optimal, we see that AMP and BP give comparable results down to average degrees around 10.

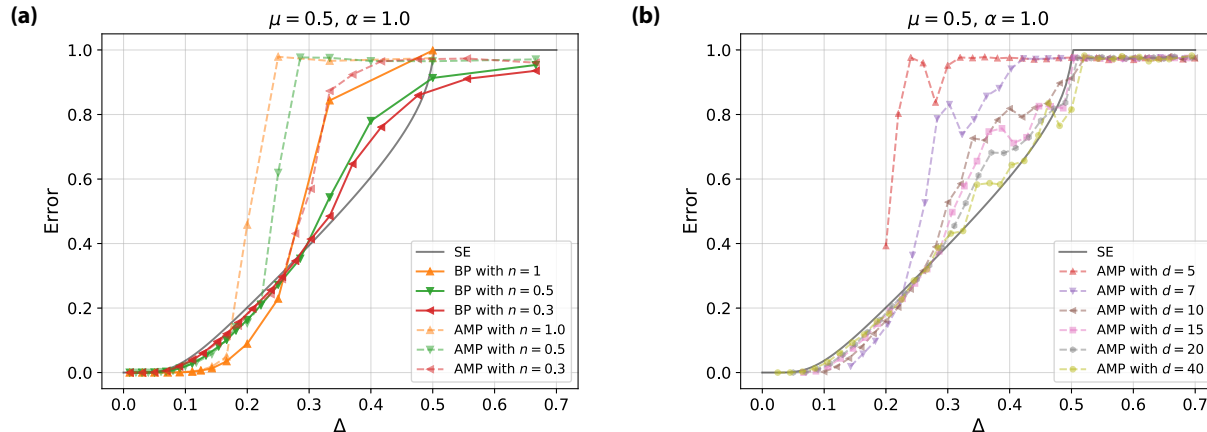


Figure 5. The effect of variation of either d or ν (i.e. n) on the performance of AMP. (a) AMP results for fixed ν . We also compared to the BP results that have the same prior and matching signal to noise ratio. (b) AMP results for fixed d . The fact that the error found in the experiments for large Δ is slightly below the SE is due to finite size effects. Increasing the average degree pushes the results closer to the SE prediction. The experiments were carried out with $N = 10^3$ and are averaged over 100 samples.

The results clearly suggests that (for finite size systems) AMP can indeed be run even on relatively sparse instances. Compared to BP it is algorithmically less complex and more memory efficient, as fewer messages need to be stored. Further, the state evolution prediction seems to remain a good qualitative approximation to the algorithmic performance. It suggests that the phenomenology found in the dense limit should be rather generic and also appear in sparse systems.

B. First order phase transition in belief propagation

So far we have shown that the dense DS model can exhibit both second and first order phase transitions. The first order transitions are more interesting algorithmically as they are associated with the presence of an algorithmically *hard* region where the corresponding message passing algorithm is suboptimal.

The authors of [2, 3] established that BP is optimal in the sparse DS model for sufficiently large signal-to-noise-ratio. It remains to be tested whether we can observe a first order phase transition also in the sparse version of the model. This shall be the aim of the present section.

Suboptimality of BP is associated with a region of parameters for which BP converges to different fixed points from the informative and from the uninformative initialization. We use our intuition based on the results of the dense case to show that there exist regimes where BP is sub-optimal. Figure 6 depicts numerical results obtained for BP with a Bernoulli-prior on θ ($\lambda = 0$ in (24)) with very sparse signals ($\mu = 0.01$). We also plot the AMP performance (in the same sparse regime) as well as the asymptotic prediction that would be expected in the dense case (28). Indeed, a clear first order transition appears. This establishes the suboptimality of BP by virtue of the dependency on the initialization.

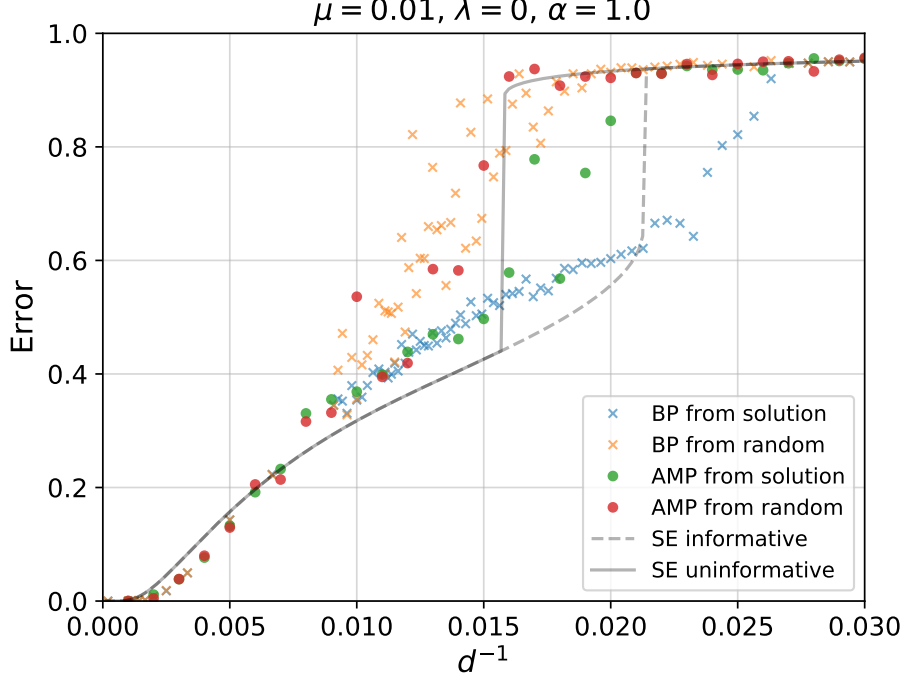


Figure 6. Numerical results obtained for the BP algorithm of references [2, 5]. The experiments were carried out on graphs of size $N = M = 10^4$, and are reported as a function of the inverse average degree of the worker nodes d . A region of coexistence associated to a first order phase transition opens up and an informative initialization leads to another fixed point than the uninformative one. This makes BP sub-optimal in the part of this region, where the free energy of the fixed point reached from the uninformative initialization is higher than the one of the fixed point reached from the informative initialization. We found in our experiments that the first order transition appears more pronounced the larger the system size, suggesting that the phenomenon persists asymptotically.

C. Approximate message passing on real data

We tested the AMP algorithm (Alg. 1) on the bluebird dataset of Welinder et al. [18]. This dataset is fully connected, minimizing effects introduced by poorly designed task-worker-graphs. We used the same priors and parameters as in [5] to compare AMP to other algorithms. Following [5] we also implemented a “two-coin” extension of AMP that assumes that the true positive and true negative rates are different. We define $\vec{\theta}_i = (s_i, t_i)$ with s_i the sensitivity of worker i and t_i indicating its specificity. We have

$$\begin{aligned}
 P\left(Y_{ij} = \pm 1 \mid \vec{\theta}_i, v_j = +1\right) &= (1 - \rho) \cdot \frac{1}{2} \cdot \left(1 \pm \sqrt{\frac{\nu}{N}} s_i\right) \\
 P\left(Y_{ij} = \pm 1 \mid \vec{\theta}_i, v_j = -1\right) &= (1 - \rho) \cdot \frac{1}{2} \cdot \left(1 \mp \sqrt{\frac{\nu}{N}} t_i\right) \\
 P\left(Y_{ij} = 0 \mid \vec{\theta}_i, v_j\right) &= \rho.
 \end{aligned}$$

As in section II B we cast the above model into a rank-2 matrix factorization problem by setting

$$\vec{v}_j = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \text{ if question } j \text{ is true and } \vec{v}_j = \begin{pmatrix} 0 \\ -1 \end{pmatrix} \text{ if question } j \text{ is false.}$$

The only difference is that the former rank-1 matrix \mathbf{w} , cf. (3), now becomes a rank-2 matrix with $\boldsymbol{\theta} \in \mathbb{R}^{M \times 2}$ and $\mathbf{v} \in \mathbb{R}^{N \times 2}$. The equations for a general rank are derived and given in detail in [7].

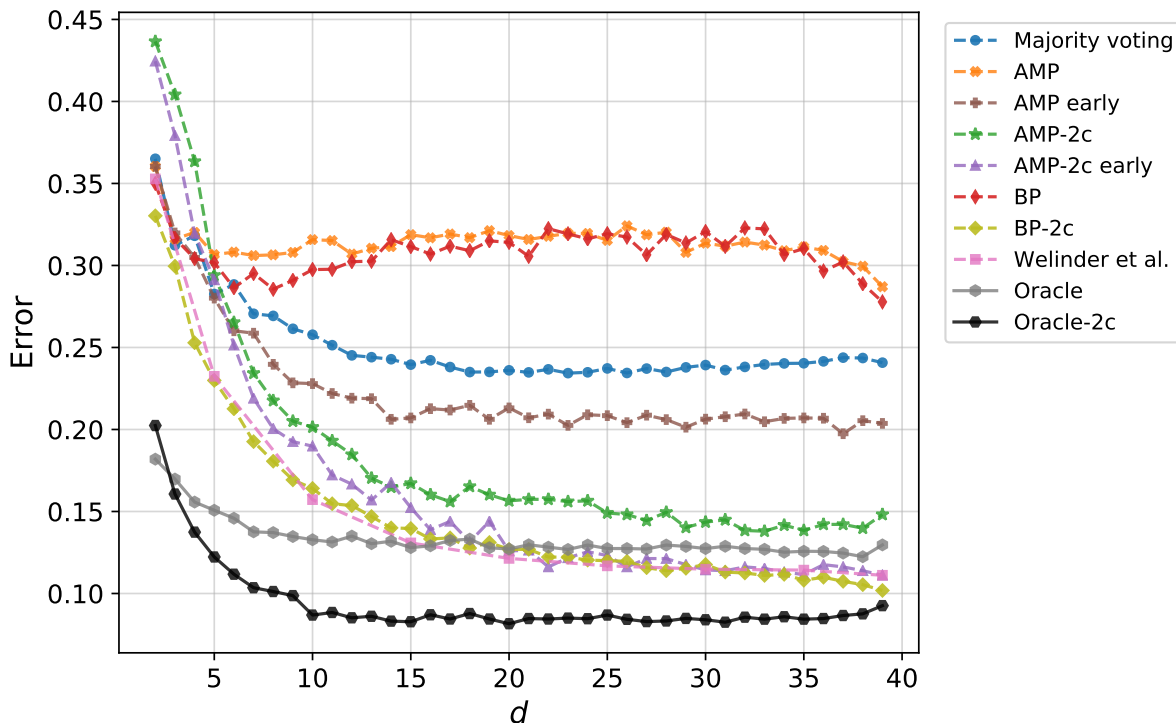


Figure 7. The error against the number of workers-per-task, d . Different algorithms are compared to AMP on the bluebird dataset. We compare to the results obtained with BP, majority voting and the algorithm proposed by Welinder et al. [18]. As explained in the text, we implemented two different version of AMP and BP: a symmetric one in which the sensitivity and specificity are equal and an asymmetric version (referred to as “2-c” in the legend). Finally we also plot results obtained when AMP is run with an early stopping criterion of 10 iterations. For BP and AMP the priors are set to independent Beta(2, 1) distributions on θ . We averaged over 100 samples for each d .

In Fig. 7 we compare AMP with BP, majority voting and the algorithm developed by Welinder et al. in [18]. We also compute the oracle lower bound of [4] for the two versions of AMP and BP. To evaluate the oracle we first estimate the true parameters θ from the ground truth and then compute the resulting Bayes-optimal estimator that maximizes the posterior probability. Note that the latter estimator has full information of the workers reliabilities.

We stress that analogous comparisons between existing algorithms and BP were already performed in [5], where BP was found to be superior. Our main point in this section is that AMP, which is simpler than BP, gives a comparable performance to BP even on real-world data. We therefore focus on the comparison between BP and AMP. Both, BP and AMP perform badly when the original model with $s_i = t_i$ is used as can be seen from Fig. 7 by comparing them to majority voting as a baseline algorithm. Running the same experiments with the two-coin version improves the results significantly. Indeed BP and AMP perform essentially as well as the much more involved algorithm of [18].

The experiments were run with identical beta-priors for BP and AMP ($a = 2$, $b = 1$) for comparability with the results in [3, 5]. For AMP different strategies were implemented for the prior on \mathbf{v} . Setting σ to the true value (estimated from the ground truth) or to 1/2 led to comparable results as when it was learned. In our AMP implementation we initialize $\hat{\mathbf{v}}$ in the estimates obtained by majority voting.

In the symmetric case BP and AMP are very close in performance. The difference for the two-coin models tends to be slightly larger, while the general trend persists. We also observe that it can be beneficial to implement AMP with an early stopping criterion as depicted in Fig. 7. Early stopping can be reasonable because the assumptions made in the derivation are likely to be imprecise, especially for small system sizes.

In summary, AMP performs quite well on real world datasets. The vanilla implementation yields slightly worse results, as compared to BP. However, when AMP is stopped after few iterations (we used 10) it reaches much better performance in the rank-1 case. A significant improvement is also obtained in the rank-2 version of AMP: for small d BP outperforms AMP, but they soon become quasi indistinguishable. Besides its good performance it has the great advantage of algorithmic simplicity, better time complexity and scalability.

V. CONCLUSION

In this paper the dense limit of the Dawid-Skene model for crowdsourcing was considered. It was shown that the problem can be mapped onto a larger class of low-rank matrix factorization problems. This leads to an approximate message passing algorithm for crowdsourcing and a closed-form asymptotic analysis of its performance. Due to the previous work of [10, 19] this analysis can be considered rigorous. While the theory only holds rigorously for the dense Dawid-Skene model, numerical experiments suggest that in the sparse regime AMP still performs well and also the asymptotic analysis provides a good qualitative prediction.

When the crowd consists mainly of spammers with only few workers that provide useful information, we found that a first order transition appears in the Bayes-optimal performance. Algorithmically this first order transition translates into the presence of a hard phase in which the AMP algorithm is sub-optimal. As a proof of concept we showed numerically that this feature persists even in the sparse regime where the rigor of our analysis breaks down. In experiments we also found instances of first order transitions in the belief propagation algorithm of [5]. This shows that there are regimes in the Dawid-Skene model where BP is not optimal. This complements recent results on [2, 3] about regimes of optimality of BP.

We also carried out experiments on real-world data and showed that AMP performs comparable to other state-of-the-art algorithms, while being of lower time complexity. Our experiments on the real-world dataset also show that having a model that described data accurately is more important than the precise algorithm that is used to do inference on the model.

ACKNOWLEDGEMENT

We would like to warmly thank T. Lesieur for advice and guidance as well as Florent Krzakala for suggesting to look at a real world dataset. LZ acknowledges funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement No 714608 - SMiLe). This work is supported by the “IDI 2015” project funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02.

-
- [1] A. P. Dawid and A. M. Skene, “Maximum likelihood estimation of observer error-rates using the EM algorithm,” *Journal of the Royal Statistical Society. Series C* **28**, 20–28 (1979).
 - [2] Jungseul Ok, Sewoong Oh, Jinwoo Shin, and Yung Yi, “Optimality of belief propagation for crowdsourced classification,” in *International Conference on Machine Learning* (2016) pp. 535–544.
 - [3] Jungseul Ok, Sewoong Oh, Jinwoo Shin, and Yung Yi, “Optimal inference in crowdsourced classification via belief propagation,” arXiv preprint arXiv:1602.03619 (2016).
 - [4] David R. Karger, Sewoong Oh, and Devavrat Shah, “Iterative learning for reliable crowdsourcing systems,” in *Advances in Neural Information Processing Systems 24* (2011) pp. 1953–1961.
 - [5] Qiang Liu, Jian Peng, and Alexander T Ihler, “Variational inference for crowdsourcing,” in *Advances in Neural Information Processing Systems 25* (2012) pp. 692–700.
 - [6] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová, “MMSE of probabilistic low-rank matrix estimation: Universality with respect to the output channel,” in *53rd Annual Allerton Conference on Communication, Control, and Computing* (2015) pp. 680–687.
 - [7] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová, “Constrained low-rank matrix estimation: phase transitions, approximate message passing and applications,” *Journal of Statistical Mechanics: Theory and Experiment*, 073403 (2017).
 - [8] Sundeep Rangan and Alyson K Fletcher, “Iterative estimation of constrained rank-one matrices in noise,” in *IEEE International Symposium on Information Theory Proceedings* (2012) pp. 1246–1250.
 - [9] Adel Javanmard and Andrea Montanari, “State evolution for general approximate message passing algorithms, with applications to spatial coupling,” *Information and Inference* **2**, 115–144 (2013).
 - [10] L. Miolane, “Fundamental limits of low-rank matrix estimation: the non-symmetric case,” arXiv preprint arXiv:1702.00473 (2017).
 - [11] Yash Deshpande and Andrea Montanari, “Information-theoretically optimal sparse PCA,” in *IEEE International Symposium on Information Theory* (2014) pp. 2197–2201.
 - [12] Ryosuke Matsushita and Toshiyuki Tanaka, “Low-rank matrix reconstruction and clustering via approximate message passing,” in *Advances in Neural Information Processing Systems 26* (2013) pp. 917–925.
 - [13] D. J. Thouless, P. W. Anderson, and R. G. Palmer, “Solution of ‘Solvable model of a spin glass’,” *The Philosophical Magazine: A Journal of Theoretical Experimental and Applied Physics* **35**, 593–601 (1977).
 - [14] Erwin Bolthausen, “An iterative construction of solutions of the TAP equations for the Sherrington–Kirkpatrick model,” *Communications in Mathematical Physics* **325**, 333–366 (2014).

- [15] Lenka Zdeborová and Florent Krzakala, “Statistical physics of inference: Thresholds and algorithms,” *Advances in Physics* **65**, 453–552 (2016).
- [16] Y. Deshpande, E. Abbe, and A. Montanari, “Asymptotic mutual information for the binary stochastic block model,” in *IEEE International Symposium on Information Theory* (2016) pp. 185–189.
- [17] Francesco Caltagirone, Marc Lelarge, and Léo Miolane, “Recovering asymmetric communities in the stochastic block model,” *IEEE Transactions on Network Science and Engineering* **PP**, 1–1 (2017).
- [18] Peter Welinder, Steve Branson, Pietro Perona, and Serge J. Belongie, “The multidimensional wisdom of crowds,” in *Advances in Neural Information Processing Systems 23* (2010) pp. 2424–2432.
- [19] Jean Barbier, Mohamad Dia, Nicolas Macris, Florent Krzakala, Thibault Lesieur, and Lenka Zdeborová, “Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula,” in *Advances in Neural Information Processing Systems 29* (2016) pp. 424–432.