IOP Publishing Journal of Physics A: Mathematical and Theoretical

J. Phys. A: Math. Theor. **53** (2020) 124001 (37pp) https://doi.org/10.1088/1751-8121/ab757f

Dense limit of the Dawid–Skene model for crowdsourcing and regions of sub-optimality of message passing algorithms

Christian Schmidt and Lenka Zdeborová

Institut de Physique Théorique, Université Paris Saclay, CEA Saclay and CNRS, 91191 Gif-sur-Yvette, France

E-mail: lenka.zdeborova@ipht.fr

Received 23 September 2019, revised 7 February 2020
Accepted for publication 12 February 2020
Published 3 March 2020

Abstract

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. Recently it was shown in the work of Ok *et al* that, in certain regimes, belief propagation is optimal for data generated from the Dawid–Skene model. This paper is motivated by this recent progress. We analyze a noisy dense limit of the Dawid–Skene model that has so long remained open. It is shown that it belongs to a larger class of low-rank matrix estimation problems for which it is possible to express the Bayes-optimal performance for large system sizes 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 further refines the results of Ok *et al* about optimality of message passing algorithms by characterizing regions of parameters where these algorithms do not match the Bayes-optimal performance. Besides, we 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.

Keywords: message passing algorithms, Bayesian statistics, signal reconstruction, approximation algorithms, disordered systems, spin glass

(Some figures may appear in colour only in the online journal)

1. 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 for 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 $0 \leqslant p_i^0 \leqslant 1$ for worker $i = 1, \ldots, N$. Further there are M tasks, each possessing a true label that we denote by $v_j^0 \in \{\pm 1\}$ for task $j = 1, \ldots, M$. The worker i is assigned a subset of tasks, denoted as $\partial i \subseteq \{1, \ldots, M\}$, to which it assigns an answer $L_{ij} \in \{\pm 1\}\ \forall j \in \partial i$. We denote $L_{ij} = 0$ if $j \notin \partial i$, i.e. for tasks that were not assigned to worker i. In the DS model labels provided by worker i for task j are modeled as

$$P(L_{ij}) = p_i^0 \delta(L_{ij} - v_j^0) + (1 - p_i^0)\delta(L_{ij} + v_j^0)\,. \tag{1}$$

Moreover it is assumed that the p_i^0s 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. [2].

While in general reconstructing the true labels and workers reliabilities from the observed answers, L_{ij}, is an NP-hard problem, a range of algorithmic approaches has been suggested in the literature, see e.g. [2–11]. The majority of the existing literature focuses on the regime where the probability of error per task goes to zero and studies the corresponding convergence rate under various conditions.

We are instead interested in a noisy high-dimensional regime where, in the limit of large system size, the probability of error per task stays bounded away from zero, and ranges between zero and as large as in the case of random guessing. Optimality results including exact constants in this regime are mathematically challenging. The authors of [12, 13] obtained a remarkable theorem stating that in certain regions of parameters in the noisy high-dimensional regime belief propagation (BP) reconstructs the true labels optimally in the sense that it minimizes the expected bitwise error. The BP algorithm for crowdsourcing was first suggested by [14].

The goal of the present paper is to carry out an analysis of the DS model in the noisy high-dimensional regime. Our theoretical results apply in a scaling where each worker is assigned randomly a constant fraction of the M tasks and M scales linearly with N. Otherwise we are in the same setting as [12, 13], i.e. with random worker reliabilities and on random graphs. From our analysis it is possible to characterize tightly the region of parameters for which BP 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. As we reveal later, such a first order phase transition is associated with a region of parameters in which BP does not match the Bayes-optimal performance. Our work can therefore also be seen as a follow-up on [12, 13] providing a refined analysis of the regions of parameters for which BP is or is not optimal.

1.1. Our contributions

In this paper we study a scaling limit of the DS model where the number of tasks and workers grow proportionally to each other, and the precision of each worker is poor in such a way that the limiting probability of error per task is bounded away from zero in the large size limit. We assume the observed data were generated from the DS model and we assume the parameters of the model to be known. We then study the performance of the so-called Bayes-optimal estimator that minimizes the *expected bitwise error* on the labels and the *expected mean square error* on the workers reliabilities.

Our main results are based on the realization that in this case the DS model is a special case of noisy high-dimensional low-rank estimation problems studied recently in [15–21]. We transfer results from those works to the present setting of the DS model and obtain the following contributions:

- We propose the approximate message passing (AMP) algorithm to approximate the Bayes-optimal estimator in the DS model.
- We provide a sharp (up to constants) characterization of the error achievable by the Bayes-optimal estimator.
- We analyze where the AMP algorithm achieves the Bayes-optimal error and provide a detailed phase diagram. This reveals a so-called hard-region where the AMP algorithm does not achieve the Bayes-optimal performance.
- We show numerically that the results obtained in the dense regime also translate into the sparse regime of the DS model. In turn this reveals the existence of regions where BP is sub-optimal.

1.2. Related work

Starting with the seminal paper of Dawid and Skene [1] many of the early works have focused on expectation-maximization (EM) algorithms and closely related approaches [3–8]. Besides its popularity for theoretical analysis, the DS model is sometimes too restrictive for practical applicability and several extensions have been proposed in the literature [8, 22–25]. Other works are based on spectral algorithmic methods, e.g. [10, 11] and [9] which combines a spectral method with the EM algorithm. Spectral methods can be advantageous when the underlying task-worker graph is not random [9, 11].

The authors of [2] were the first to propose a message passing scheme that turned out to be closely related to the BP algorithm. The derivation of the BP algorithm is based on the Bayesian analysis of a generative model and was first given in [14]. The authors also revealed how EM algorithms are related to mean-field methods and how majority voting and the iterative algorithm of [2] are related to BP. These efforts culminated in the works of [12, 13] that proved optimality of BP under certain assumptions on the parameters of the model. The present analysis of the dense DS 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.

We show that the dense DS model belongs to a class of low-rank matrix factorization problems, as studied in [15, 16] and analyzed recently by statistical physics techniques in [17, 18]. The authors of [17, 18] derived the AMP algorithm for low-rank matrix factorization and analyzed the Bayes-optimal performance in a closed form. We apply the results derived in those papers and identify the region of parameters for which the associated AMP algorithm is suboptimal.

One of the merits of AMP is that its asymptotic performance can be described via the so-called state evolution, as proven in [19, 20]. The performance of the Bayes-optimal estimator was also later put on fully rigorous bases in the work of [21] under assumptions that include the dense DS model as considered in this work.

From a physics point of view the model corresponds to a bipartite planted spin glass model that is closely related to the Sherrington–Kirkpatrick model [26]. The AMP equations correspond to the Thouless–Anderson–Palmer mean field equations [27] with correct time indices. Finally, the state evolution equations describe the stationary points of the replica-symmetric free energy that can be derived by means of the replica method from statistical physics.

1.3. Organization of the paper

In section 2 we first define a dense version of the DS model and outline the Bayesian inference setting considered in this work as well as the AMP algorithm for the dense DS model. In the following section 3 we apply the algorithm to a real-world dataset before we move to the theoretical analysis of the AMP algorithm for synthetic data in section 4. In the latter section we also draw a detailed phase diagrams for the dense DS model. Finally, in section 6 we investigate numerically how the results—valid in the dense regime—transfer into the sparse regime, as originally considered for BP in [13, 14].

2. Dense version of the Dawid–Skene model

In this section we introduce a dense version of the Dawid–Skene model (dDS) for crowdsourcing that is considered in this paper. We chose it in such a way that it can be mapped onto a low-rank matrix factorization problem studied previously with approximate message passing (AMP). The considered regime is dense in the sense that each of the N workers is assigned $\Theta(M)$ questions[1].

2.1. Definition of the large size limit of the DS model

Consider a crowd of N workers and a pool of M tasks. Assume that each task j comes with a true label $v_j^0 \in \{\pm 1\}$ and that the labels are independent and identically distributed (iid): $v_j^0 \sim P_{v^0}\ \forall j \in \{1, \ldots, M\}$. In the DS model (see (1)) the workers are assumed to be characterized by a single scalar parameter p_i^0. Similarly to the tasks we assume the probabilities of the workers to be iid: $p_i^0 \sim P_{p^0}\ \forall i \in \{1, \ldots, N\}$. Each worker, i, is assigned a subset $\partial i \subseteq \{1, \ldots, M\}$ of tasks from the pool. This subset is assumed to be drawn at random, such

[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.

that its expected size is $\mathbb{E}[|\partial i|] = (1-\rho)M$. Similarly, we denote by $\partial j \subseteq \{1,\dots,N\}$ the subset of workers that participate in task j. The rest of the paper is set in the high-dimensional dense regime where $M, N \to \infty$, while $\alpha \equiv M/N = \Theta(1)$ and $\rho = \Theta(1)$. Later, in section 6, we will discuss how to extrapolate the results into the sparse regime where each worker is only assigned to $O(1)$ tasks.

Under the above assumptions, the regime in which the error per task is bounded away from zero but (possibly) better that random guessing is such that

$$p_i^0 \equiv (1 + \sqrt{\nu/N}\,\theta_i^0)/2\,. \tag{2}$$

The parameter ν is an overall scale parameter, while θ_i^0 is the rescaled reliability of worker i, drawn iid from P_{θ^0}. The above scaling for the DS model is such that the noisy high-dimensional regime is amenable to closed form analysis in terms of both the Bayes-optimal performance and the corresponding message passing algorithm. Generalization to other scalings in which the resulting error remains strictly between zero and randomly bad is of interest, but is left for future work.

Workers with $\theta_i = 0$ give answers that are completely uninformative and will be called *spammers*. On the contrary if $\theta_i \gg 1$, then the answers are 'strongly' (yet only of order $1/\sqrt{N}$) aligned with the truth and we refer to such workers as *hammers*. Adversaries that willingly or unwillingly align against the ground true labels are characterized by $\theta_i < 0$. They may also be considered hammers if $\theta_i \ll -1$, because their answers are aligned against the truth, as opposed to the random alignment of the spammers.

We denote with L_{ij} the label assigned to question j by worker i and assume $L_{ij} \in \{0, \pm 1\}$. If question j was not in the set ∂j of answered questions we set $L_{ij} = 0$. We assume that for each ij the L_{ij} is generated independently of the others. Under the above assumptions the *likelihood* in the dDS model becomes

$$\begin{aligned} P\left(L_{ij} = \pm 1 \mid \theta_i,\, v_j\right) &= (1-\rho)\cdot\frac{1}{2}\cdot\left(1 \pm \sqrt{\frac{\nu}{N}}\,\theta_i v_j\right) \\ P\left(L_{ij} = 0 \mid \theta_i,\, v_j\right) &= \rho, \end{aligned} \tag{3}$$

where we assumed that the fraction of un-answered questions, ρ, is independent of ij.

2.2. Bayesian estimators

Given the matrix $\boldsymbol{L}$ the aim is to recover the true labels $\boldsymbol{v}^0$ and (rescaled) reliabilities $\boldsymbol{\theta}^0$[2]. In this work we assume a Bayesian inference setting. In Bayesian inference we aim to compute the estimators, $\hat{\boldsymbol{v}}$ and $\hat{\boldsymbol{\theta}}$, that minimize the expected bitwise error (ER) on the labels

$$\mathrm{ER}_v = \mathbb{E}_{\boldsymbol{v}^0|\boldsymbol{L}}\left[\frac{1}{M}\sum_j \mathbb{I}\left(\hat{v}_j \neq v_j^0\right)\right] = \mathbb{E}_{\boldsymbol{v}^0|\boldsymbol{L}}\left[\frac{1}{M}\sum_j \left(\frac{\hat{v}_j - v_j^0}{2}\right)^2\right] \tag{4}$$

$$= \mathbb{E}_{\boldsymbol{v}^0|\boldsymbol{L}}\left[\frac{1}{2}\frac{1}{M}\sum_j \left(1 - \hat{v}_j v_j^0\right)\right] \tag{5}$$

and the expected mean square error on the reliabilities

[2] Here and in the rest of the paper, upper case bold letter indicate matrices and lower case bold letters indicate column vectors.

$$\mathrm{MSE}_\theta = \mathbb{E}_{\boldsymbol{\theta}^0|\boldsymbol{L}}\left[\frac{1}{N}\sum_i\left(\hat{\theta}_i - \theta_i^0\right)^2\right]. \tag{6}$$

In order to compute the above objectives explicitly it is necessary to specify the posterior probability. This requires the specification of the likelihood $P(\boldsymbol{L} \mid \boldsymbol{\theta}, \boldsymbol{v})$ and the priors $P_\theta(\boldsymbol{\theta})$, $P_v(\boldsymbol{v})$. The first was specified above for the dDS model, while the latter remains unspecified for now. From Bayes' theorem we obtain

$$P(\boldsymbol{\theta}, \boldsymbol{v} \mid \boldsymbol{L}) = \frac{1}{Z(\boldsymbol{L})}\prod_{1\leqslant i\leqslant N} P_\theta(\theta_i)\prod_{1\leqslant j\leqslant M} P_v(v_j)\prod_{1\leqslant i\leqslant N, 1\leqslant j\leqslant M} P(L_{ij} \mid \theta_i, v_j)\,. \tag{7}$$

The estimators that minimize the above objectives are then computed using the posterior as

$$\hat{\theta}_i^{\mathrm{MMSE}}(\boldsymbol{L}) = \int \mathrm{d}\theta_i \theta_i P\left(\theta_i \mid \boldsymbol{L}\right) \tag{8}$$

$$\hat{v}_j^{\mathrm{MER}}(\boldsymbol{L}) = \mathrm{sign}\int \mathrm{d}v_j v_j P\left(v_j \mid \boldsymbol{L}\right), \tag{9}$$

where $P\left(x_k \mid \boldsymbol{L}\right)$, with $x_k \in \{\{\theta\}_{i=1,\ldots,N}, \{v_j\}_{j=1,\ldots,M}\}$ are the marginals of the posterior.

Inferring the reliabilities and labels in the crowdsourcing problem hence reduces to evaluating the marginal expectations of the posterior probability distribution. In general this is a difficult task. In the next section we show that the dDS model falls into a class of low-rank matrix estimation problems for which the posterior probability distribution can be evaluated in the above large size limit [18].

Note that the above estimators may be computed with either the true or a mismatched model. We therefore distinguish the Bayes-optimal estimators, that assume the true underlying model with which the data was generated, from the *mismatched* estimators that assume some model that is not matching the true one. In the theoretical part of this work (i.e. in all but section 3) *we assume that the distributions from which the ground truth, $\{\theta_i^0\}$ and labels $\{v_j^0\}$ are drawn, P_{θ^0} and P_{v^0} respectively, are known, as well as all other parameters*. Under these assumptions we aim to (a) compute efficiently the Bayes-optimal estimators of θ_i^0 and v_j^0, given the answers L_{ij} and (b) to evaluate the large size performance of the AMP algorithm. Note that in practice when some of the parameters are not known, they could be learned, but this is not the focus of our paper.

2.3. 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 [18], and whose results were proven rigorously in [21]. In the rest of this section we follow closely these two papers and review the results that will be applied to the dDS model.

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

$$\boldsymbol{w} \equiv \frac{\boldsymbol{\theta}\,\boldsymbol{v}^\intercal}{\sqrt{N}} \tag{10}$$

and re-express (3) as

$$P\left(L_{ij} \mid w_{ij}\right) = \exp\left(g(L_{ij}, w_{ij})\right) ,$$
$$g(L_{ij}, w_{ij}) = \begin{cases} \log\left(\frac{(1-\rho)}{2}\right) + \log\left(1 \pm \sqrt{\nu} w_{ij}\right) & \text{if } L_{ij} = \pm 1 \\ \log\left(\rho\right) & \text{if } L_{ij} = 0. \end{cases} \tag{11}$$

It is now not difficult to show [18, 28] that it suffices to expand $g(L_{ij}, w_{ij})$ w.r.t. w_{ij} up to second order, due to the $\Theta(1/\sqrt{N})$ scaling of $\boldsymbol{w}$. This allows to re-express the likelihood $P(L_{ij} \mid w_{ij})$ as a Gaussian with (inverse) effective noise Δ

$$\Delta^{-1} = \mathbb{E}_{P(L_{ij}|w_{ij}=0)} \left[\left(\frac{\partial g(L_{ij}, w_{ij})}{\partial w_{ij}} \bigg|_{w_{ij}=0} \right)^2 \right] = (1-\rho)\nu \,. \tag{12}$$

In turn this means that the dDS model as introduced before is asymptotically equivalent to a rank-1 matrix factorization problem under Gaussian white noise.

2.4. Background on algorithmic approaches

One of the basic algorithms used for low-rank estimation are the spectral methods that are based on the spectral properties of the labeling matrix $\boldsymbol{L}$, e.g. [9–11]. The spectral methods bring the advantage that they are typically algorithmically simple, robust and need no specification of an underlying generative model. On the other hand they are typically not Bayes-optimal (unless in some limits).

The widely deployed EM algorithm requires a specification of the likelihood function of the data, $\boldsymbol{L}$. The EM algorithm then computes the maximum likelihood estimators by alternating between a maximization step to estimate the reliabilities and an estimation step to compute the expected labels. Unlike the spectral algorithms it requires the specification of a model and much effort has been put into improving the model used for crowdsourcing [4–8].

BP also requires the specification of a generative model (as in (7)). Unlike EM, it directly approximates the Bayes-optimal estimators (9) via message passing. The messages are the conditional probabilities that live on the edges of the associated graphical model (as outlined in the appendix). The BP equations are self-consistent equations for these messages that can be solved iteratively. The main shortcoming of BP is that it requires the computation of a continuous distribution over the reliabilities for each message.

One possible simplification is obtained by absorbing all $\theta_{\partial j}$ into a common factor node in the graphical model. This is done by integration over $\boldsymbol{\theta}$ and is the strategy followed by [12–14]. Owing to the binary labels the resulting BP equations can then be solved more efficiently. However, this approach becomes unfeasible in the dense regime because too many messages need to be computed. Another approach to circumvent these issues, closely related to both BP and the spectral methods, is the iterative algorithm of [2] that can be interpreted as a linearization of BP. It can also be interpreted as a spectral method applied to the non-backtracking operator [29]. However, such a linearization is only justified in particular regimes and otherwise results in sub-optimal performance.

AMP bypasses the issues of BP while maintaining the performance: in the dense regime it is possible to derive a simplification of the BP equations (without requiring an integration over $\boldsymbol{\theta}$) that only requires the computation of $\Theta(N)$ beliefs. This AMP approach relies on the fact that the messages are effectively Gaussian as the underlying factor graph is densely connected.

Algorithm 1. Approximate message passing for crowd sourcing.

Data: $\boldsymbol{S}$, Δ, δ; // $\boldsymbol{S}$ and Δ according to (14) and (12) respectively.
Result: MMSE estimates $\hat{\boldsymbol{v}}$ and $\hat{\boldsymbol{\theta}}$
Initialize: $\hat{\boldsymbol{v}} \leftarrow \hat{\boldsymbol{v}}^{\text{init}} \sim P_v(\boldsymbol{v})$, $\hat{\boldsymbol{\theta}} \leftarrow \hat{\boldsymbol{\theta}}^{\text{init}} \sim P_\theta(\boldsymbol{\theta})$; $\sigma_v \leftarrow 1$, $\sigma_\theta \leftarrow 1$; $\hat{\boldsymbol{v}}^{\text{old}} \leftarrow \boldsymbol{0}$, $\hat{\boldsymbol{\theta}}^{\text{old}} \leftarrow \boldsymbol{0}$;
while $\left\|\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^{\text{old}}\right\|_2^2 + \left\|\hat{\boldsymbol{v}} - \hat{\boldsymbol{v}}^{\text{old}}\right\|_2^2 > \delta$ **do**
  $\boldsymbol{B}_\theta \leftarrow \frac{1}{\sqrt{N}} \boldsymbol{S}\hat{\boldsymbol{v}} - \frac{1}{\Delta} \hat{\boldsymbol{\theta}}^{\text{old}} \sigma_v$;
  $A_\theta \leftarrow \frac{1}{N\Delta} \hat{\boldsymbol{v}}^\intercal \hat{\boldsymbol{v}}$;
  $\boldsymbol{B}_v \leftarrow \frac{1}{\sqrt{N}} \boldsymbol{S}^\intercal \hat{\boldsymbol{\theta}} - \frac{\alpha}{\Delta} \hat{\boldsymbol{v}}^{\text{old}} \sigma_\theta$;
  $A_v \leftarrow \frac{1}{N\Delta} \hat{\boldsymbol{\theta}}^\intercal \hat{\boldsymbol{\theta}}$;
  $\hat{\boldsymbol{\theta}}^{\text{old}} \leftarrow \hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{v}}^{\text{old}} \leftarrow \hat{\boldsymbol{v}}$;
  $\hat{\boldsymbol{\theta}} \leftarrow f_\theta(A_\theta, \boldsymbol{B}_\theta)$, $\sigma_\theta \leftarrow \frac{1}{N} \sum_{1 \leqslant i \leqslant N} \partial_{B_{\theta_i}} f_\theta(A_\theta, B_{\theta_i})$;
  $\hat{\boldsymbol{v}} \leftarrow f_v(A_v, \boldsymbol{B}_v)$, $\sigma_v \leftarrow \frac{1}{M} \sum_{1 \leqslant j \leqslant M} \partial_{B_{v_j}} f_v(A_v, B_{v_j})$;
end

2.5. Approximate message passing

The approximate message passing (AMP) algorithm for low-rank matrix estimation is a simplification of BP in the limit of dense graphical models. In this limit both, BP 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 [29, 30] to the Thouless–Anderson–Palmer equations [27] from the theory of spin glasses. AMP for low-rank matrix factorization was first derived for special cases in [16, 19] and in its general form in [17, 18].

For the readers convenience, we give a derivation of the equations in the appendix. The AMP derivation procedure can be summarized as follows. One starts from BP and performs the following two simplifications. First, the BP messages are replaced by their means and variances which eradicate the necessity of tracking a whole function for each message. Secondly, each (mean and variance) *message* is replaced by its *marginal* version, reducing the necessary variables from $O(N^2)$ messages to $O(N)$ marginals. The AMP equations compute the estimates (9) in terms of a set of self-consistent equations. For details we refer the reader to the appendix.

We briefly recall the essential elements necessary to state the AMP algorithm for the dDS model, the details of which can be found in the appendix. To state the AMP algorithm for the dDS 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 $x_k \in \{\{\theta_i\}_{i=1,\ldots,N}, \{v_j\}_{j=1,\ldots,M}\}$) 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.

$$\hat{x} \equiv f_x(A_x, B_x) = \frac{1}{Z_x(A_x, B_x)} \int \mathrm{d}x\, x\, P_x(x)\, \mathrm{e}^{-\frac{1}{2}A_x x^2 + B_x x}, \qquad \sigma_x = \partial_{B_x} f_x(A_x, B_x). \tag{13}$$

To state the final equations we further need to define the Fisher score matrix

$$S_{ij} \equiv \left.\frac{\partial g(L_{ij}, w_{ij})}{\partial w_{ij}}\right|_{w_{ij}=0} = L_{ij} \cdot \sqrt{\nu}\,, \tag{14}$$

where $g(L_{ij}, w_{ij})$ is defined in equation (11).

Given these definitions AMP is the iterative scheme that we outline in algorithm 1. Unlike BP, the AMP equations close directly on marginal quantities. Very much like naive mean field equations, no message passing is necessary. Instead the equations are directly expressed in terms of the local fields, A_θ, A_v and B_θ, B_v, acting on the variables and from which the estimators are computed via (13). The major difference as compared to a naive mean field approach are the second terms that appear in the right hand side for B_θ and B_v. These are the so called *Onsager reaction terms*. They correct the naive mean field contribution of the first terms and add momentum to the equations, which is crucial for the convergence. The Onsager reaction term corrects for the fact that the equations were closed on the nodes of the graphical model (see figure A1) directly, instead of the messages, and prohibits a self-interaction of the estimators on the nodes.

The algorithm requires the specification of the priors, P_v and P_θ, as well as the parameters Δ and ν. In practice the true priors are typically not known (or they might simply not exist), but nevertheless the algorithm may be used, if *some* prior is specified, as is standard in any case of Bayesian analysis. Once a distribution is specified its remaining parameters can be learnt[3]. The algorithm further requires the specification of initial values for $\hat{\boldsymbol{\theta}}, \hat{v}$ that we draw from the prior distribution[4]. Finally, let us note that the numerical implementation of the algorithm might profit from an adequate damping scheme in order to enhance convergence on small instances or when the model assumptions are not satisfied.

The great advantage of the algorithm lays in its low computational complexity, simplicity and in its amenability to an exact analysis for synthetic data (as we shall see later in section 4).

3. Approximate message passing on real data

Before moving towards the theoretical analysis of the AMP algorithm (algorithm 1) we present a concrete application of the algorithm to a real-world dataset. We tested the algorithm on the bluebird dataset of Welinder *et al* [8]. The dataset contains labels for $M = 108$ tasks from $N = 39$ workers. This dataset is fully connected, minimizing effects introduced by poorly designed task-worker-graphs.

We stress that comparisons between BP and existing algorithms were already performed in [14], 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 and use the same priors and parameters as in [14] which also puts AMP in perspective to the other algorithms tested there (approximate mean field and EM).

Following [14] 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

[3] One approach in the setting of this work, closely related to expectation maximization, is to minimize the 'Bethe free energy' with respect to the parameters that are left open. The Bethe free energy is the objective that the AMP equations minimize and therefore also a function of all the parameters, appearing in the algorithm. The Bethe free energy can be found in [18].

[4] However this choice is not unique and different strategies can be employed in practice. In the numerical experiments we have also tried initializations and observed no visible difference in the fixed point.

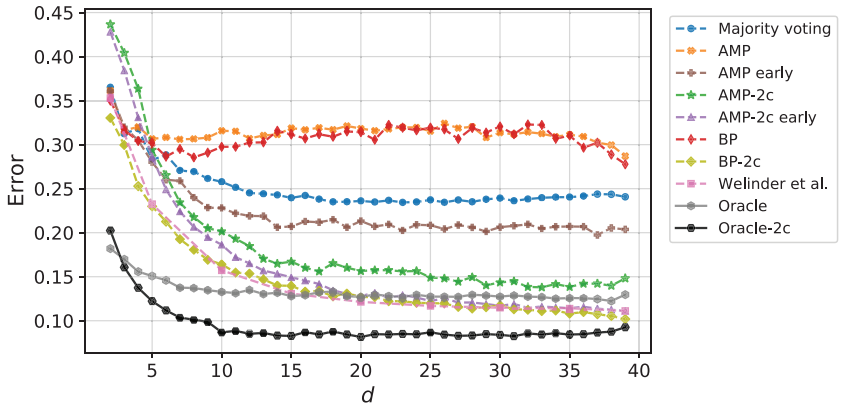

Figure 1. The bitwise error on the labels against the number of workers per task, d. The following different algorithms are compared to AMP on the bluebird dataset: BP, majority voting and the algorithm proposed by Welinder *et al* [8]. As explained in the text, we implemented two different version of 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, indicated by the 'early' suffix. The oracle lower-bound results from maximizing the posterior probability for known worker reliabilities (of course we do not have access to the true reliabilities of each worker, but we can estimate them as the fraction of correctly labeled data points from the knowledge of the ground truth). For BP and AMP the priors are set to independent Beta$(2, 1)$ distributions on $\boldsymbol{\theta}$. We averaged over 100 samples for each d.

$$
\begin{aligned}
P\left(L_{ij} = \pm 1 \mid \vec{\theta}_i,\ v_j = +1\right) &= (1-\rho)\cdot\frac{1}{2}\cdot\left(1 \pm \sqrt{\frac{\nu}{N}}\, s_i\right)\\
P\left(L_{ij} = \pm 1 \mid \vec{\theta}_i,\ v_j = -1\right) &= (1-\rho)\cdot\frac{1}{2}\cdot\left(1 \mp \sqrt{\frac{\nu}{N}}\, t_i\right)\\
P\left(L_{ij} = 0 \mid \vec{\theta}_i,\ v_j\right) &= \rho\,.
\end{aligned}
\tag{15}
$$

As in section 2.3 we cast the above model into a rank-2 matrix factorization problem by setting

$$
\vec{v}_j = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \text{ if question } j \text{ is true and } \vec{v}_j = \begin{pmatrix} 0 \\ -1 \end{pmatrix} \text{ if question } j \text{ is false.} \tag{16}
$$

The only difference is that the former rank-1 matrix $\boldsymbol{w}$, see (10), now becomes a rank-2 matrix with $\boldsymbol{\theta} \in \mathbb{R}^{M\times 2}$ and $\boldsymbol{v} \in \mathbb{R}^{N\times 2}$. The equations for a general rank are derived and given in detail in [18].

In figure 1 we compare AMP with BP, majority voting and the algorithm developed by Welinder *et al* in [8]. We also compute the oracle lower bound of [2] for the two versions of AMP and BP. Note that the latter estimator has full information of the workers reliabilities. To evaluate the oracle one first estimates the probability that worker i provides the correct response: $\hat{p}_i^* \equiv \sum_{j\in\partial i} \mathbb{I}(L_{ij} = v_j^0)/|\partial i|$ and then compute the resulting estimator that

maximizes the posterior probability: denoting the log-ratio $r_i^* \equiv \log(\hat{p}_i^*/(1-\hat{p}_i^*))$ the oracle estimator can be expressed as $\hat{\boldsymbol{v}}^* = \text{sign}(\boldsymbol{r}^{*\intercal}\boldsymbol{L})$[5].

The experiments were run with identical beta-priors for BP and AMP for comparability with the results in [13, 14]. Concerning the prior on $\boldsymbol{v}$ one must specify the bias, β, of negative labels. We have implemented different strategies, but found that they all perform essentially the same. The results reported here are for a bias estimated from the ground truth, but we have tried different strategies. We found that setting the bias to arbitrary values by hand or setting it to the true value (estimated from the ground truth) led to comparable results as when it was learned. In our AMP implementation we initialize $\hat{\boldsymbol{v}}$ in the estimates obtained by majority voting and we have set $\rho = d/M$ and $\nu = N$ (from which Δ and $\boldsymbol{S}$ then follow). The BP and AMP algorithms stop either when the absolute values of the messages/estimators changed less than 10^{-6} from one iteration to the next, or after a maximum of 100 iterations.

Both, BP and AMP perform badly when the original model with $s_i = t_i$ is used as can be seen from figure 1 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 algorithm of [8].

The vanilla implementations of BP and AMP (with $s_i = t_i$) 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 figure 1. 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 it reaches much better performance in the rank-1 case. Here we used ten iterations, to illustrate the effect, but an exact study of the effects are beyond the scope of this paper. 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 running time complexity and scalability.

4. State evolution (approximate message passing on synthetic data)

The AMP algorithm depends on the data, $\boldsymbol{L}$, through $\boldsymbol{S}$ and consequently so do the AMP estimates for the reliabilities, $\hat{\boldsymbol{\theta}}$, and task labels, $\hat{\boldsymbol{v}}$. Quite remarkably, in the large size limit $N \to \infty$, the performance of the algorithm can be tracked by the so-called state evolution (SE) equations, which has been proven rigorously for the low-rank estimation in [19, 20]. In the appendix we provide a detailed derivation of the SE equations for the dDS model. Next we summarize the outcome of that derivation and outline how to make use of it in practice.

Looking at algorithm 1 one might expect that, under the assumptions made in our model, the vectors $\boldsymbol{B}_\theta, \boldsymbol{B}_v$ behave like vectors of iid Gaussian entries, as each of their components is a sum of many (nearly) independent terms. Similarly, A_θ, A_v, as a sum of $\Theta(N)$ terms, is expected to concentrate around its expectation in the limit where $M, N \to \infty$. It turns out that

[5] For the two-coin model this estimator must be adapted. Denoting $r_j^* = z_j^+/z_j^-$ with $z_j^+ = \prod_{i=1}^N s_j^{(1+L_{ij})/2} \cdot (1-s_i)^{(1-L_{ij})/2} \cdot (1-\hat{\beta})$ and $z_j^- = \prod_{i=1}^N t_j^{(1-L_{ij})/2} \cdot (1-t_i)^{(1+L_{ij})/2} \cdot \hat{\beta}$, where $\hat{\beta} = \sum_{j=1}^M \mathbb{I}(v_j = -1)/M$, the estimator can be expressed as $\hat{\boldsymbol{v}}^* = \text{sign}(\log \boldsymbol{r}^*)$.

this intuition holds (see [19, 20] for rigorous proves) and that one can trace the algorithmic performance of AMP exactly in the limit where $M, N \to \infty$.

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, this performance can be effectively expressed in terms of the two scalar *order parameters* below, thus reducing the high-dimensional problem to a scalar problem. The order parameters are the overlap of the AMP estimates with the ground true parameters:

$$\begin{aligned} M_\theta^t &= \frac{1}{N} \sum_{1 \leqslant i \leqslant N} \hat{\theta}_i^t \, \theta_i^0 \,, \\ M_v^t &= \frac{1}{M} \sum_{1 \leqslant j \leqslant M} \hat{v}_j^t \, v_j^0 \,, \end{aligned} \tag{17}$$

where x^0 indicates the true value of x, and t the iteration step of the AMP equations. The SE states that these order parameters concentrate and evolve as (see appendix)

$$\begin{aligned} M_v^{t+1} &= \mathbb{E}_{v^0,W} \left[f_v \left(\frac{M_\theta^t}{\Delta} \,, \, \frac{M_\theta^t}{\Delta} \, v^0 + \sqrt{\frac{M_\theta^t}{\Delta}} \, W \right) v^0 \right] \,, \\ M_\theta^t &= \mathbb{E}_{\theta^0,W} \left[f_\theta \left(\frac{\alpha \, M_v^t}{\Delta} \,, \, \frac{\alpha \, M_v^t}{\Delta} \, \theta^0 + \sqrt{\frac{\alpha \, M_v^t}{\Delta}} \, W \right) \theta^0 \right] \end{aligned} \tag{18}$$

where W is a standard Gaussian random variable, $v^0 \sim P_v$, $\theta^0 \sim P_\theta$, the functions f_v and f_θ are defined in (13), $\alpha = M/N$ and Δ is the effective noise (12).

Let us call M_θ^{SE} and M_v^{SE} the fixed points of the SE equations (18). These fixed points are then associated to the MSE (6) and ER (5) as reached by the AMP algorithm through

$$\mathrm{MSE}_\theta^{\mathrm{AMP}} = \mathbb{E}_\theta(\theta^2) - M_\theta^{\mathrm{SE}} \,, \tag{19}$$

$$\mathrm{ER}_v^{\mathrm{AMP}} = (1 - R_v^{\mathrm{SE}})/2 \,, \tag{20}$$

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

$$R_v^{\mathrm{SE}} = \mathbb{E}_{v^0,W} \left\{ \mathrm{sign} \left[f_v \left(\frac{M_\theta^{\mathrm{SE}}}{\Delta} \,, \, \frac{M_\theta^{\mathrm{SE}}}{\Delta} \, v^0 + \sqrt{\frac{M_\theta^{\mathrm{SE}}}{\Delta}} \, W \right) \right] v^0 \right\} \,. \tag{21}$$

At the same time, it is straightforward to observe that the SE equations are in fact stationarity conditions of the so-called replica-symmetric free energy (for short just 'free energy' in the following):

$$\begin{aligned} \phi(M_\theta, M_v) = \alpha \frac{M_\theta M_v}{2\Delta} &- \alpha \mathbb{E}_{v^0,W} \left[\log Z_v \left(\frac{M_\theta}{\Delta} \,, \, \frac{M_\theta}{\Delta} \, v^0 + \sqrt{\frac{M_\theta}{\Delta}} \, W \right) \right] \\ &- \mathbb{E}_{\theta^0,W} \left[\log Z_\theta \left(\frac{\alpha \, M_v}{\Delta} \,, \, \frac{\alpha \, M_v}{\Delta} \, \theta^0 + \sqrt{\frac{\alpha \, M_v}{\Delta}} \, W \right) \right] \end{aligned} \tag{22}$$

where the functions Z_θ and Z_v are defined in (13) and the rest of the variables are defined in the same way as in the SE. Hence the fixed points of the SE are critical points of the free energy.

As conjectured in [18] and proven rigorously in [21] the performance of the Bayes-optimal estimator (9) can be evaluated in the large size limit from the *global minimizer* of the free

energy. Denote by M_θ^* and M_v^* the global minimizers of the above free energy. The minimum-mean-squared-error (MMSE) and the minimum-error-rate (MER) are simply

$$\mathrm{MMSE}_\theta = \mathbb{E}_\theta(\theta^2) - M_\theta^* \,, \tag{23}$$

$$\mathrm{MER}_v = \frac{1}{2}(1 - R_v^*) \,, \tag{24}$$

where R_v^* is obtained from M_θ^* via (21).

We summarize: the performance of the AMP algorithm can be measured in terms of the scalar order parameters M_θ^t and M_v^t. Their evolution is tracked by the SE equations and therefore the SE characterizes the AMP algorithm. The final performance of the AMP algorithm ($t \to \infty$) is thus given by the fixed points of the SE initialized correspondingly (see below). The fixed points of the SE correspond to the stationary points of the free energy (22). At the same time, the optimal estimators (in the Bayesian sense as outlined in section 2.2) correspond to the global minimizer of the free energy. We conclude that if the SE converges to the fixed point that globally minimizes the free energy AMP is optimal (in the Bayesian sense outlined in the beginning of this paper).

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

Whether or not the SE reaches the global minimizer M_θ^*, M_v^* depends on the shape of the free energy and the initialization of the SE equations at $t = 0$. 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 SE (18) converges to the global minimum of the free energy (22). The way we check this in practice is that we initialize the SE in two different ways:

- Uninformative initialization, where $M_v^{t=0} = (\mathbb{E}_v(v))^2 + \delta_v$ and $M_\theta^{t=0} = (\mathbb{E}_\theta(\theta))^2 + \delta_\theta$. Where δ is some small perturbation. This setting corresponds to an infinitesimal alignment of the algorithm towards the direction of the signal. The error achieved by the AMP algorithm is then given by iteration of (18) from this uninformative initialization.
- Informative initialization, where $M_v^{t=0} = \mathbb{E}_v(v^2) + \delta_v$ and $M_\theta^{t=0} = \mathbb{E}_\theta(\theta^2) + \delta_\theta$ 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 (18) 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 smaller one corresponds to the Bayes-optimal performance [21]. 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 the trivial fixed point

To start analyzing the behaviour of the studied DS model, we first consider that both the priors P_θ and P_v have zero mean. In that case $M_\theta^{t=0} = M_v^{t=0} = 0$ is a fixed point of the SE. We refer to $M_\theta = M_v = 0$ as the trivial fixed point. The equations (18) can be expanded around this fixed point. In first order one obtains

$$M_\theta^t = \frac{\alpha}{\Delta}\left(\mathbb{E}_\theta\left[\theta^2\right]\right)^2 M_v^t \tag{25}$$

$$M_v^{t+1} = \frac{1}{\Delta}\left(\mathbb{E}_v\left[v^2\right]\right)^2 M_\theta^t\,, \tag{26}$$

implying that the uninformative fixed point is numerically stable for $\Delta^2 > \alpha\left(\mathbb{E}_v\left[v^2\right]\right)^2\left(\mathbb{E}_\theta\left[\theta^2\right]\right)^2$ and unstable otherwise. We define the critical effective noise, $\Delta_{\rm c}$, therefore as

$$\Delta_{\rm c} = \sqrt{\alpha}\cdot\mathbb{E}_\theta\left[\theta^2\right]\mathbb{E}_v\left[v^2\right]\,. \tag{27}$$

For $\Delta < \Delta_{\rm c}$ the uninformative initialization becomes numerically unstable. The algorithmic consequence of Δ_c is that while for $\Delta < \Delta_c$ the AMP algorithm reaches positive overlap with the ground truth, for $\Delta > \Delta_c$ the AMP algorithm does not reach a correlation with the ground truth. The threshold $\Delta_{\rm c}$ correspond to a *second order phase transition* in the behavior of the AMP algorithm, meaning that the overlap reached by the algorithm is non-analytic and continuous at $\Delta_{\rm c}$. This is in analogy to the critical temperature in ferromagnetic systems, below which aligned configurations can form.

At this point it is appropriate to point out that the symmetry in the priors leads to an invariance of the estimators w.r.t. a simultaneous sign-flip of $\boldsymbol{\theta} \to -\boldsymbol{\theta}$ and $\boldsymbol{v} \to -\boldsymbol{v}$. In that sense, our analysis only yields the possible achievable error up to this invariance. I.e. it is possible to detect the sets of tasks that belong to the same group ($v_j = +1$ versus $v_j = -1$), but it is per se impossible to tell to which of the two groups a particular task belongs. In practice one would require an infinitesimally small amount of supervision in order to break this invariance.

In the case where both the priors, P_θ and P_v, have zero mean, we can divide the region of parameters into the three phases outlined below. The same p
hase structure appears in many other inference problems; for a review we refer the reader to [29]. The three different phases are defined as follows:

- **Easy phase:** The fixed point of the SE (18) (initialized in the uninformative way) corresponds to the global minimum of the free energy (22), and at the same time this fixed point is associated with a *positive overlap* with the ground-truth configuration. AMP matches the Bayes-optimal performance in the large size limit.
- **Hard phase:** In this phase two or more minima of the free energy (22) coexist; at least one local minimum of small overlap and a global minimum of larger overlap. The outcome of iterating the SE 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 AMP algorithm do not reach the information-theoretically optimal performance and *AMP is sub-optimal*. We note that the fixed points reached by AMP in the hard phase can either have zero overlap with the ground truth, or can have positive (but not optimal) overlap with the ground truth.
- **Impossible phase:** When the global minimum of (22) is associated to the trivial, non-informative, fixed point corresponding to zero overlap, we talk about a phase of impossible inference. Algorithmically 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,

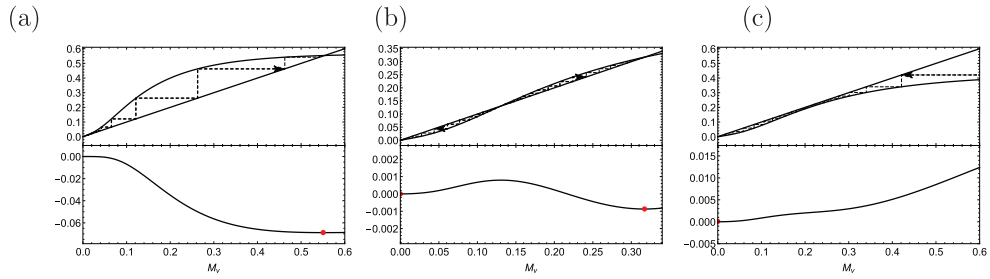

Figure 2. Exemplary exit charts and free energies for the state evolution equations in the (a) EASY (b) HARD and (c) impossible phases ($\Delta \in \{0.019, 0.028, 0.03\}$ respectively). The upper panel shows an exit chart for the SE equations and the lower depicts the free energy. The red dots in the lower plots mark the minima of the free energy. The straight line in the upper panels mark the identity mapping $M_v^{t+1} = M_v$ and the other solid line the SE mapping $M_v^{t+1} = G\left(\frac{1}{\Delta}T\left(\frac{\alpha}{\Delta}M_v^t\right)\right)$.

i.e. a discontinuity in the asymptotic value of the overlap reached by the Bayes-optimal estimator.

- The *algorithmic threshold* $\Delta_{\rm alg}$ is the largest value of effective noise, Δ, below which the AMP algorithm asymptotically *always* matches the Bayes-optimal performance.
- The *spinodal threshold*, $\Delta_{\rm sp}$, is the smallest values of effective noise above which the informative initialization converges to a different fixed point than the (perturbed) uninformative initialization. This threshold does not have any particularly interesting algorithmic consequences.
- The *information theoretic transition*, $\Delta_{\rm alg} < \Delta_{\rm IT} < \Delta_{\rm sp}$, is where the value of the free energy of the uninformative fixed point crosses with the free energy of the informative fixed point. The algorithmic consequences of this threshold is that at $\Delta_{\rm IT}$ the overlap between the ground truth and the Bayes-optimal estimator has a discontinuity. The performance of the Bayes-optimal estimator abruptly improves at $\Delta_{\rm IT}$.

A *first order phase transition* takes place when the free energy has two competing minima that *co-exist*. Such a situation can be found in figure 2(b). Referring to the figure, this happens when the value of the left minimum of the free energy becomes lower than the one on the right. Since the Bayes-optimal error corresponds to the order parameter, M_v^*, with lower free energy, a sudden discontinuous change in M_v^* takes place when the free energy of the two minima cross and the left one becomes the Bayes-optimal one instead of the right one. As opposed to a second-order phase transition, where the Bayes-optimal error varies continuously.

We summarize: when $\Delta < \Delta_{\rm alg}$ one is in the *easy* regime where AMP achieves Bayes-optimal performance. For $\Delta_{\rm alg} \leqslant \Delta \leqslant \Delta_{\rm IT}$ one is in the *hard* regime where AMP exhibits a gap to the Bayes-optimal performance. It is conjectured that no polynomial time algorithm succeeds in this region [29]. For $\Delta > \Delta_{\rm IT}$ AMP is once again Bayes-optimal, either reaching zero or positive overlap with the ground truth. In the regime $\Delta_{\rm IT} < \Delta < \Delta_{\rm sp}$, the free energy has still has a second minimum with positive overlap, but it can only be reached by initializing AMP in some informative initial state. The discontinuity in the Bayes-optimal overlap happens at $\Delta_{\rm IT}$. Note that while in some models, such as the stochastic block model [17], one finds $\Delta_{\rm c} = \Delta_{\rm alg}$ in general and in the present model $\Delta_{\rm c} \neq \Delta_{\rm alg}$. Note also that we do not refer to an algorithmic threshold if no hard region is present, because in that case AMP is always Bayes-optimal, with the mere distinction that for $\Delta < \Delta_{\rm c}$ the achieved error is better than that of random guessing whereas above it is not.

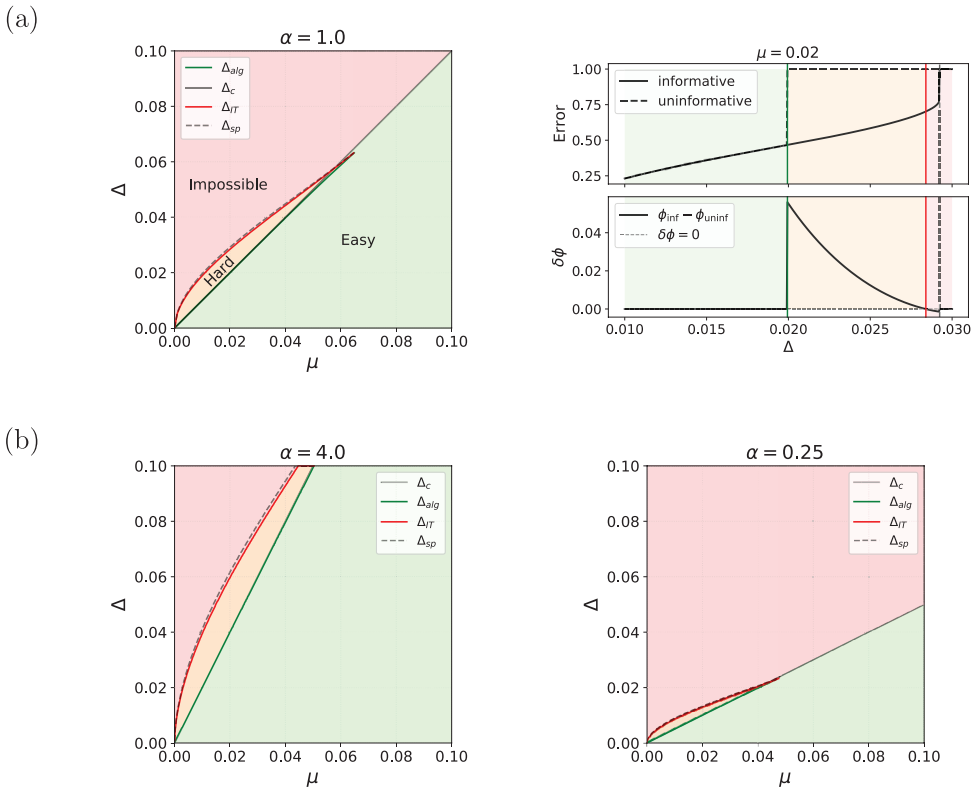

Figure 3. (a) left panel: the phase diagram for a Rademacher–Bernoulli prior on θ with $\lambda = 1/2$ and no bias in the distribution of the labels ($\beta = 1/2$). When the fraction of spammers is very large (small μ) a hard phase appears where the AMP algorithm is not able to reach the information-theoretically 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 $\Delta_c > \Delta_{\rm alg}$ but both are very close and not distinguishable in the figure. In the hard region (orange) the AMP algorithm reaches MSE $= 1$ for $\Delta > \Delta_c$, and MSE < 1 but not optimal in the tiny region of $\Delta_{\rm alg} < \Delta < \Delta_c$, while the Bayes-optimal estimator reaches smaller 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 $\mu \approx 0.077$.

4.2. An example

In order to illustrate the concepts, we provide a brief example that summarizes the key-points of this section graphically in figure 2.

Let us anticipate the model considered in section 5 with a simple parameter setting, such that the labels are ± 1 with equal probability and $P_\theta = (1-\mu)\delta(\theta) + \mu[\delta(\theta-1) + \delta(\theta+1)]$. The SE equations (18) can be brought into the following form

$$M_v^{t+1} = G\left(\frac{1}{\Delta} T\left(\frac{\alpha}{\Delta} M_v^t\right)\right), \tag{28}$$

with

$$G(x) = \frac{1}{2}\mathbb{E}_W\left[\tanh(x+\sqrt{x}W) - \tanh(-x+\sqrt{x}W)\right]$$
$$T(x) = \frac{\mu}{2}\mathbb{E}_W\left[\frac{\sinh(x+\sqrt{x}W)}{\frac{1-\mu}{\mu}e^{x/2}+\cosh(x+\sqrt{x}W))} - \frac{\sinh(-x+\sqrt{x}W)}{\frac{1-\mu}{\mu}e^{x/2}+\cosh(-x+\sqrt{x}W)}\right]. \tag{29}$$

We depict in figure 2 the resulting exit charts and free energies for three exemplary values of $\Delta \in \{0.019, 0.028, 0.03\}$ that are respectively in the easy, hard and impossible phase.

5. Phase diagrams for the dense David–Skene model

A key property of the results we described so far is that the asymptotic behavior 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. Since the model assumes that the ground truth task labels are iid and binary, we have

$$P_v(v) = (1-\beta)\delta(v-1) + \beta\delta(v+1)\,. \tag{30}$$

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

We start by considering worker reliabilities, θ_i, that were drawn from a skewed Rademacher–Bernoulli (RB) prior

$$P_\theta(\theta) = (1-\mu)\,\delta(\theta) + \mu\,[(1-\lambda)\delta(\theta-1) + \lambda\delta(\theta+1)]\,. \tag{31}$$

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 [2]: 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 $\Theta(1/\sqrt{N})$, see (3). Sending $\nu \to \infty$ and thus $\Delta \to 0$ approximates the hammers. The limit $\nu \to N$ will be considered in section 6.

5.1. 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 (18) have a trivial fixed point at $M_v = M_\theta = 0$. Expansion around this uninformative fixed point yields

$$M_v^{t+1} = \alpha\frac{\mu^2}{\Delta^2}\cdot M_v^t - \alpha^2\frac{\mu^2}{\Delta^2}\left[\frac{\mu}{\Delta} + \frac{\mu^2}{\Delta^2}\right]\cdot (M_v^t)^2 + O((M_v^t)^3)\,. \tag{32}$$

The linear term gives the stability criterion of the trivial fixed point that we had already derived in (27)

$$\Delta_c = \sqrt{\alpha}\cdot\mu\,. \tag{33}$$

In figure 3 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, as explained in section 4.1. We mark the phases where inference is algorithmically easy, hard and impossible.

We find that if there are not too many spammers in the crowd, then there are only two phases: the easy phase in which AMP achieves Bayes-optimal performance and positive overlap with the ground truth and the impossible phase where it does not. The easy phase is separated from the impossible phase by a second order phase transition at $\Delta = \Delta_c$. In the impossible phase AMP achieves Bayes-optimal performance, but the resulting estimators have zero overlap with the ground truth and inference is therefore asymptotically impossible. In this region where the hard phase is absent (33) provides the right criterion to locate the phase transition from the easy to the impossible phase.

If, however, there are too many spammers in the crowd, a hard region opens up in which AMP has a gap w.r.t. the Bayes-optimal performance. As explained in section 4.1, the hard phase is characterized by the coexistence of more than one minimum of the free energy and the fact that the one achieved by the SE equations from an uninformative initialization does not coincide with the lowest (Bayes-optimal) one.

The information theoretic transition line (Δ_{IT}) is the line where the lowest minimum of the free energy (i.e. the Bayes-optimal one) gets associated to a discontinuously different overlap. The free energy continues to possess multiple minima up until the spinodal transition line (Δ_{sp}). For the choice of priors made here the free energy only has two minima, one is achieved by the uninformative initialization of the SE equations and the other by the informative initialization (see figure A1). Thus, in order to reveal the information theoretic transition line we can compare the free energies of the uninformative and informative fixed points achieved by the SE equations. We do so in the right panel of figure 3(a). We plot the difference between the free energy of the uninformative and informative fixed point. As can be seen in the lower panel, the two free energies cross at some point and the uninformative fixed point starts to have lower free energy and becomes Bayes-optimal. In the upper panel we plot the achieved error by the SE equations from the different initializations. We see that the information theoretic transition does not show up in this plot and one can only see the spinodal transition. The lower end of the hard phase is separated from the easy phase by the algorithmic transition. As indicated in the upper right panel of figure 3(a) this is the point of largest effective noise, below which the uninformative fixed point of the SE always leads to the Bayes-optimal error.

5.2. The impact of α

Recall that α is the ratio of tasks to workers in our model. Increasing the number of workers decreases α which shrinks the hard region as we show in figure 3(b). In the other direction, i.e. when the total number of workers decreases and thus α grows, we show that the hard phase grows further.

By virtue of the $\sqrt{\nu/N}$ scaling of the signal, see (3), 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 (33) tells us that we should expect inference to become easier when α increases. Indeed, if we fix Δ and consider how the performance varies 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 $\boldsymbol{\theta}$, which in turn improves the estimate of $\boldsymbol{v}$. We depict this by plotting the error against α for two different values in figure 4.

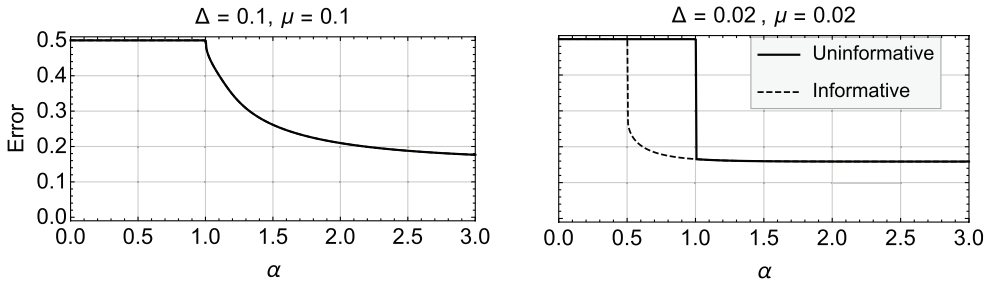

Figure 4. The behaviour of the error versus α for the RB prior (31) with $\lambda = 1/2$ and bias is set to $\beta = 1/2$.

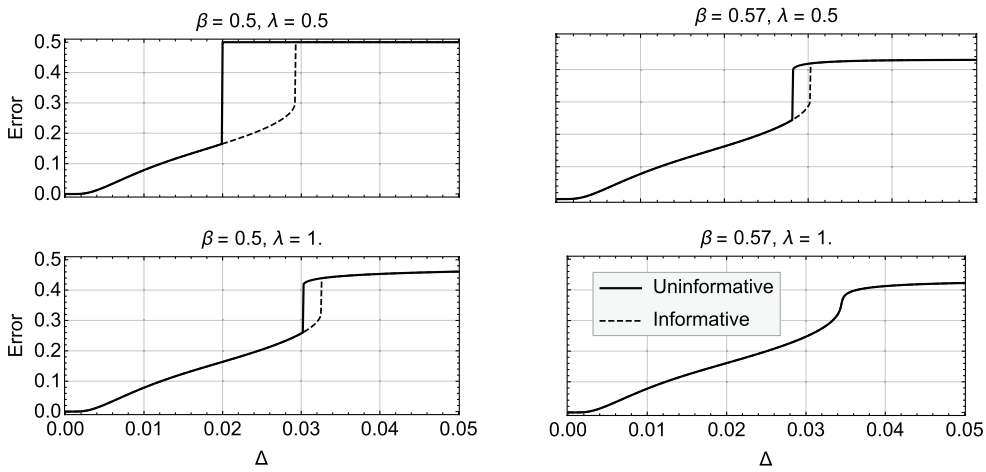

Figure 5. Influence of bias in the distribution of labels and worker reliabilities on the performance. Here we plot the resulting error (20) for $\alpha = 1$ and $\mu = 0.02$ as reached from the uninformative (bold) and informative (dashed) initialization. For bias in the labels ($\beta \neq 0.5$) or 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.

5.3. 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 no longer exist. We illustrate in figure 5 how this changes the phase diagram and the achievable error. For the case $\alpha = 1$ and $\mu = 0.02$ we plot the error reached by the SE from the informative and the uninformative initialization.

First (top left panel), we consider the unbiased case with $\beta = 1/2$, but $\lambda \neq 1/2$ as already plotted in figure 3. 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 1/2) appears. The hard phase shrinks as shown in the bottom-left panel of figure 5.

If the dataset is biased, i.e. $\beta \neq 1/2$, the change is quantitatively more dramatic, but phenomenologically very similar, see top-right panel in figure 5. 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 figure 5.

Not surprisingly, adding skew in the priors shrinks the hard region and has thus a similar effect as increasing the number of workers (decreasing α). The effect on the second order transition is even more dramatic: since the trivial fixed point disappears when any kind of skew is present in the priors, the second order transition (i.e. the transition at Δ_c) disappears. We depict this in a figure in the appendix.

Next we comment on the dependence w.r.t. λ for small Δ, i.e. in the easy phase. Note that once $\lambda \neq 1/2$ the prior has non-zero mean and the impossible phase is absent. Therefore it is also no longer sensible to distinguish the two phases. We find that the smaller Δ, the smaller is the dependence on λ. This is depicted in a figure in the appendix (for the case in which $\beta = 1/2$ and $\mu = 1/2$) and can also be observed in figure 5 that shows little change in the error to the left of the hard phase. For $\Delta \to 0$ all curves culminate and it is possible to show (derivation provided in the appendix) from an expansion of the SE equations that, *independent of* λ, they all approach zero error as

$$\mathrm{ER}_v \overset{\Delta\to 0}{\longrightarrow} \frac{1}{2}\mathrm{e}^{-\frac{1}{2}\frac{\mu}{\Delta}-\frac{1}{2}\log\pi\frac{\mu}{\Delta}}\,. \tag{34}$$

Indeed μ/Δ is the dominant quantity and plays the role of a signal-to-noise ratio. Apparently, in the zero-noise limit, λ plays no role. It is sensible that both the adversaries as well as the hammers carry the same information, but why this behavior only shows in the limit $\Delta \to 0$ is not clear to us.

Finally, another interesting limit that can be deduced from the SE equations is when $\mu \to 0$. It was previously observed that the error of majority voting and other inference algorithms seems to coincide when the number of spammers becomes overwhelming (and for Δ significantly small) [2]. An expansion of the SE equations in this limit (see the appendix) shows that this is indeed the case and the errors, achieved by AMP and majority voting, coincide for $\beta = 1/2$ and $\lambda \geqslant 1/2$ (and $\rho = 0$ for simplicity):

$$\mathrm{ER}_v \overset{\mu\to 0}{\longrightarrow} \frac{1}{2} - \sqrt{\frac{2\nu}{\pi}}\mu\left|\lambda - \frac{1}{2}\right|\,. \tag{35}$$

5.4. Dealing with other priors

The derivation of section 2 applies to any prior as long as $\theta = O(1)$. Indeed, many features persist if we replace (31) by

$$P_\theta(\theta) = (1-\mu)\delta(\theta) + \mu\phi(\theta)\,. \tag{36}$$

Where $\phi(\theta)$ is some appropriate distribution (we have considered $\phi(\theta)$ being a beta distribution or a Gaussian). For instance (32) 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).

$$P_\theta(\theta) = (1-\mu)\mathcal{N}(\theta;\bar{\theta}_L,\sigma_L^2) + \mu\mathcal{N}(\theta;\bar{\theta}_R,\sigma_R^2)\,. \tag{37}$$

Under this choice and with $\beta = 1/2$ in (30) the SE equations (18) can again be expressed as $M_v^{t+1} = G\left(\frac{1}{\Delta}T\left(\frac{\alpha}{\Delta}M_v^t\right)\right)$ with

$$G(x) = \frac{1}{2}\mathbb{E}_W\left\{\tanh(x+\sqrt{x}W) - \tanh(-x+\sqrt{x}W)\right\}$$

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

$$Q(W) = \frac{1+q\sigma_R^2}{1+q\sigma_L^2}\left(W + \sqrt{\frac{q}{1+q\sigma_R^2}}(\bar{\theta}_R - \bar{\theta}_L)\right) - W^2\,,$$

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

6. Relevance of the results in the sparse regime

Our analysis of the dDS model is based on the ground that the underlying graphical model (the bipartite question-worker-graph) is a densely connected random graph: each task-node connects to $\Theta(N)$ worker-nodes and reversely each worker-node is to $\Theta(M)$ task-nodes. A sense of sparsity was introduced in our model by allowing that some of the tasks remain unanswered (see the likelihood (3)). Our analysis assumes that the number of questions answered by each worker is extensive, i.e. $1-\rho = \Theta(1)$. 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 $N\to\infty$ [31, 32]. 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 the dense limit, the central limit theorem allows to reduce BP to AMP. However, this is no longer true in the sparse limit, where the messages are not sufficiently independent, which causes the arguments to break down and consequently also the equations can no longer be reduced to two scalar order parameters, thus rendering BP a much harder algorithm to analyze in the sparse regime. In this section we investigate numerically how the behavior of the sparse DS model deviates from the predictions drawn for the dDS model.

In the sparse regime 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'

$$\rho = 1 - \frac{d}{M} \qquad \nu = n\cdot N\,, \tag{38}$$

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, 12–14].

6.1. 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 $M = N$ and draw the bipartite worker-task graph at

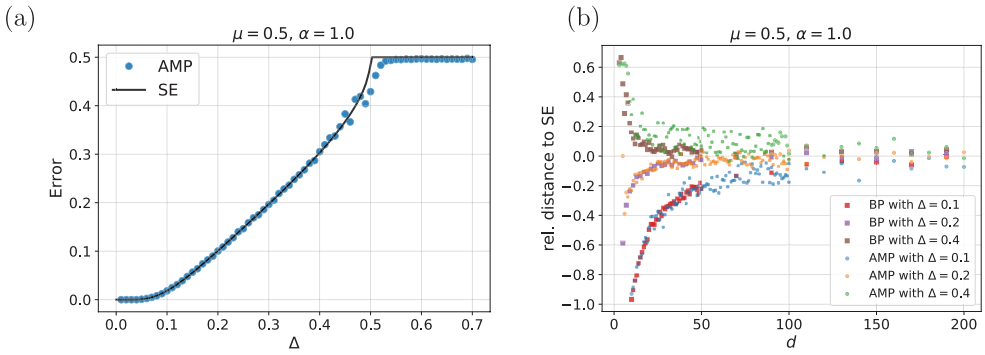

Figure 6. (a) Numerical results for AMP in the dense regime for $N = M = 10^4$, averaged over 20 samples. (b) The relative distance of the AMP results to the SE prediction of the error when the average degree d and signal-to-noise ratio ν are varied such that Δ remains fixed. We also compare to the BP algorithm that is asymptotically exact in the sparse regime. We see that the SE gives an accurate description, already for d around $30 - 50$. Although 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.

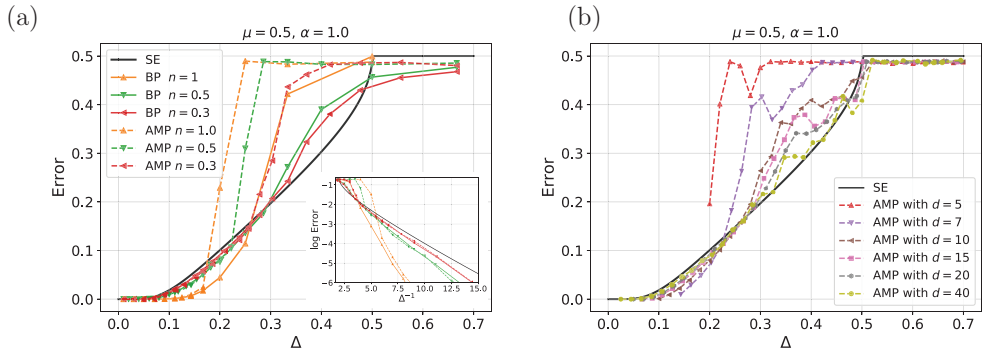

Figure 7. The effect of variation of either d or ν (i.e. n) on the performance of AMP. (a) AMP results for fixed ν. We also compared to the BP results that have the same prior and matching signal-to-noise ratio. (b) AMP results for fixed d. The fact that the error found in the experiments for large Δ 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.

random, such that the average degree of the task-nodes equals the average degree, d, of the worker-nodes.

Figure 6(a) depicts results that are obtained by running AMP in the dense regime where $d = \Theta(N)$ in order to have a reference (here 10^4 nodes were used). Except from finite size effects close to the phase transition the SE prediction agrees with the empirical results as expected. For figure 6(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 [14] 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 figure 7. This time 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 particular the two become very close for small error, while for larger error the gap between them tends to be larger around the transition. This behavior persists also when $\lambda \neq 1/2$, as can be examined for instance in figure 8. In figure 7(b) we fix d and vary ν (i.e. n), such that Δ varies in the same range as in figure 7(a). We cannot explore the full range of Δ because we must restrict $n \leqslant 1$. We see again that AMP quickly approaches its asymptotic performance when the graph becomes more and more connected.

The results clearly suggest that (for finite size systems) AMP can indeed be run even on moderately sparse instances. Compared to BP it is algorithmically less complex and more memory efficient, as fewer messages need to be stored. Further, the SE prediction seems to remain a good qualitative approximation to the algorithmic performance[6]. It suggests that the phenomenology found in the dense limit should be rather generic and also appear in sparse systems. In the following section it is shown that this is the case.

6.2. First order phase transition in belief propagation

We have already established in section 4.1 that the dDS 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 sub-optimal. We recall that the first order transition is characterized by the co-existence of multiple minima of the free energy. In this region of co-existing minima, the AMP algorithm is sub-optimal up until the information theoretic transition point. Below this point (i.e. for $\Delta < \Delta_{\mathrm{IT}}$) the fixed point reached by the SE, and therefore the estimates of the AMP algorithm, do not coincide with the Bayes-optimal minimum (i.e. the lowest) of the free energy. It is conjectured that no polynomial time algorithm exists that is more efficient then AMP [29].

In the previous section, we have established that even on sparse graphs BP and AMP behave very similar. On the one hand, the authors of [12, 13] established that BP is optimal in the sparse regime for sufficiently large signal-to-noise ratio. On the other hands, we have shown that a hard region, in which AMP is sub-optimal, exists in the dense regime. It hence remains to be tested whether we can observe a first order phase transition also in the sparse regime of the DS model? This question is answered in figure 8: it depicts numerical results obtained for BP with a Bernoulli-prior on $\boldsymbol{\theta}$ ($\lambda = 0$ in (31)) with very sparse signals ($\mu = 0.01$). In the same figure we also plot the AMP performance (in the same sparse regime) as well as the asymptotic prediction of the SE in the dense case (with the mapping (38)). Indeed, a clear first order transition appears which is associated with a region of parameters for which BP converges to different fixed points from the informative and from the uninformative initialization. This establishes the sub-optimality of BP by virtue of the dependency on the initialization.

[6] Note however the interesting deviations for small Δ that seem to deviate from (34) towards the expected $\exp(-\mu/\Delta)$ optimal error scaling in the sparse regime [2].

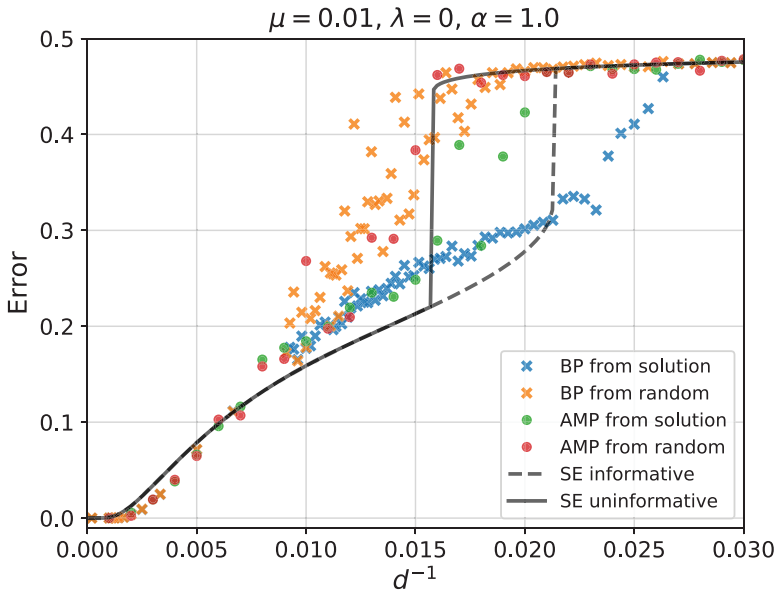

Figure 8. Numerical results obtained for the BP algorithm of [12, 14]. The experiments were carried out on graphs of size $N = M = 10^4$, and are reported as a function of the inverse average degree of the worker nodes d. A region of coexistence associated to a first order phase transition opens up and 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.

7. Conclusion

In this paper the dense limit of the Dawid–Skene model for crowdsourcing was considered. It was shown that in this regime the Dawid–Skene model 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 in terms of the so-called state evolution equations.

Although we did not provide proves in the present paper, the results can be considered rigorous. They fall into the class of problems considered in the works of [20, 21, 33] on the low-rank matrix factorization problem from which the proves can be deduced. While the theory only holds rigorously for the dense limit of the Dawid–Skene model, we have carried out numerical experiments that establish that the asymptotic analysis provides a good qualitative prediction even in the sparse regime. Further we have shown that approximate message passing still performs well and provides a comparable algorithm to belief propagation, with favorable time complexity and simplicity.

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 the

numerical experiments we also found instances of first order transitions in the belief propagation algorithm of [14]. This shows that there are regimes in the Dawid–Skene model where belief propagation is not optimal. This complements recent results on [12, 13] about regimes of optimality of belief propagation.

We also carried out experiments on real-world data and showed that AMP performs comparable to other state-of-the-art algorithms. The 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 do inference on the model.

Acknowledgment

We would like to thank T Lesieur and A Manoel for advice and guidance as well as F 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.

Appendix

A.1. Approximate message passing

In this appendix, for the convenience of the reader, we re-derive the AMP equations for the dDS model that were first derived in [18].

A.1.1. From belief propagation to relaxed belief propagation. Starting point to derive the AMP equations are the belief propagation equations for the graphical model in figure A1

$$\begin{aligned}
\tilde{m}_{ij\to j}(v_j) &= \frac{1}{Z_v^{ij\to j}} \int \mathrm{d}\theta_i \, n_{i\to ij}(\theta_i) P(L_{ij} \mid \theta_i, v_j) \\
n_{i\to ij}(\theta_i) &= \frac{1}{Z_\theta^{i\to ij}} P_\theta(\theta_i) \prod_{k\in\partial i\setminus j} \tilde{n}_{ik\to i}(\theta_i) \\
\tilde{n}_{ik\to i}(\theta_i) &= \frac{1}{Z_\theta^{ik\to i}} \sum_{v_k} m_{k\to ik}(v_k) P(L_{ik} \mid \theta_i, v_k) \\
m_{k\to ik}(v_k) &= \frac{1}{Z_v^{k\to ik}} P_v(v_k) \prod_{l\in\partial k\setminus i} \tilde{m}_{lk\to k}(v_k) \,.
\end{aligned} \tag{A.1}$$

One sees that, in the most general form written above, these equations are quite involved: (a) the variables θ_i are in general continuous, which requires the computation of an integral to obtain $\tilde{m}_{ij\to j}(v_j)$ and (b) one deals with $\Theta(NM)$ messages in the dense setting. However, the fully connected nature of the factor graph under the dense regime, together with the iid assumptions we have made, permit a simplification by application of the central limit theorem that leads to the *relaxed BP* equations that re-parametrizes each message in term of its mean and variance

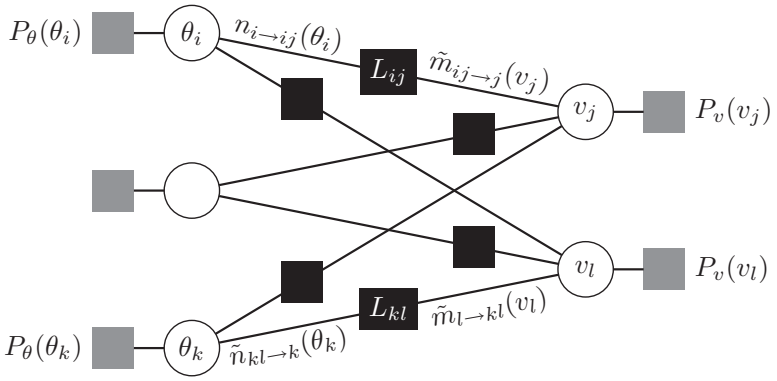

Figure A1. Factor graph representation of the dense Dawid–Skene model. Both the reliabilities and labels are associated to the circular variable nodes (left and right respectively). The pair-wise interactions between them (represented by the black factor nodes) corresponding to the collected answers, L_{ij}, that make the two side of the graphical model interact. The gray factor nodes are priors that act on the variable nodes.

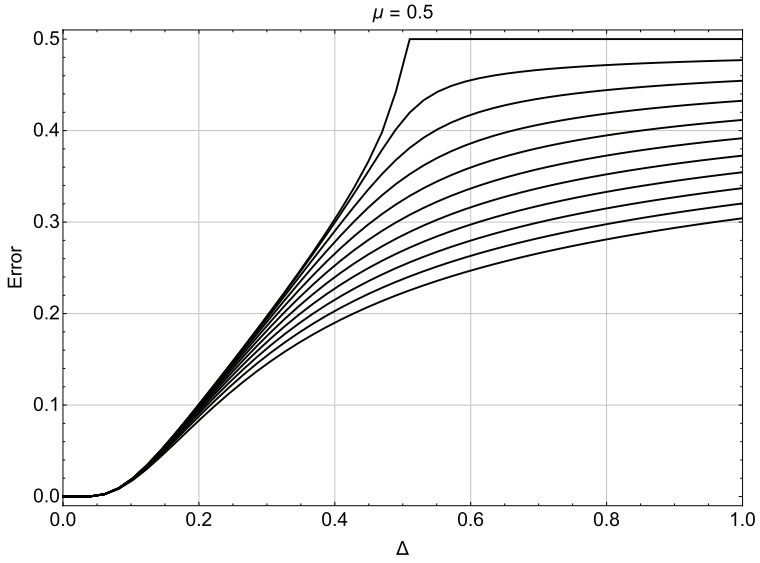

Figure A2. Exemplary plot of the impact of variation of λ in the skewed Rademacher–Bernoulli prior $P_\theta(\theta) = (1-\mu)\delta(\theta) + \mu[(1-\lambda)\delta(\theta-1) + \lambda\delta(\theta+1)]$. The prior on the labels is such that they are ± 1 with equal probability. From top to bottom we vary $\lambda \in \{0.5, 0.55, \ldots, 1.0\}$. As soon as $\lambda \neq 0.5$ the second order transition disappears and asymptotically ($\Delta \to \infty$) all the curves approach an error of 1/2 at different rates.

$$\begin{aligned}
\hat{\theta}_{i\to ij} &\equiv \int \mathrm{d}\theta_i n_{i\to ij}(\theta_i)\,\theta_i \\
\sigma_{\theta,\,i\to ij} &\equiv \int \mathrm{d}\theta_i n_{i\to ij}(\theta_i)\,(\theta_i)^2 - \hat{\theta}^2_{i\to ij} \\
\hat{v}_{k\to ik} &\equiv \int \mathrm{d}v_k m_{k\to ik}(v_k)\,v_k \\
\sigma_{v,\,k\to ik} &\equiv \int \mathrm{d}v_k m_{k\to ik}(v_k)\,v_k^2 - \hat{v}^2_{k\to ik}\,.
\end{aligned} \tag{A.2}$$

Owing to the $1/\sqrt{N}$ scaling we may expand $P(L_{ij} \mid \theta_i, v_j)$ in (A.1) up to second order. Denoting

$$S_{ij} \equiv \left.\frac{\partial g(L_{ij}, w)}{\partial w}\right|^{w=0}, \tag{A.3}$$

$$R_{ij} \equiv \left(\left.\frac{\partial g(L_{ij}, w)}{\partial w}\right|^{w=0}\right)^2 + \left.\frac{\partial^2 g(L_{ij}, w)}{\partial w^2}\right|^{w=0}, \tag{A.4}$$

we obtain

$$\begin{aligned}
\mathrm{e}^{g(L_{ij},w_{ij})} &= \mathrm{e}^{g(L_{ij},0)}\left[1 + S_{ij}w_{ij} + \frac{R_{ij}w_{ij}^2}{2} + O(N^{-\frac{3}{2}})\right] \\
&= \mathrm{e}^{g(L_{ij},0)+S_{ij}w_{ij}+\frac{1}{2}(R_{ij}-S_{ij}^2)w_{ij}^2} + O(N^{-\frac{3}{2}}).
\end{aligned} \tag{A.5}$$

The messages (A.1) can now be written in a simplified Gaussian form

$$\begin{aligned}
n_{i\to ij}(\theta_i) &= \frac{1}{Z_\theta^{i\to ij}}\,P_\theta(\theta_i)\exp\left(B_{\theta,\,i\to ij}\,\theta_i - \frac{1}{2}A_{\theta,\,i\to ij}\,\theta_i^2\right) \\
m_{k\to ik}(v_k) &= \frac{1}{Z_v^{k\to ik}}\,P_v(v_k)\exp\left(B_{v,\,k\to ik}\,v_k - \frac{1}{2}A_{v,\,k\to ik}\,v_k^2\right),
\end{aligned} \tag{A.6}$$

where the new variables A_θ, A_v and B_θ, B_v follow the equations

$$\begin{aligned}
B^t_{\theta,\,i\to ij} &= \frac{1}{\sqrt{N}}\sum_{k=1,\,k\neq j}^{M} S_{ik}\hat{v}^t_{k\to ik} \\
A^t_{\theta,\,i\to ij} &= \frac{1}{N}\sum_{k=1,\,k\neq j}^{M}\left[S_{ik}^2(\hat{v}^t_{k\to ik})^2 - R_{ik}\left((\hat{v}^t_{k\to ik})^2 + \sigma^t_{v,\,k\to ik}\right)\right] \\
B^t_{v,\,k\to ik} &= \frac{1}{\sqrt{N}}\sum_{l=1,\,l\neq i}^{N} S_{lk}\hat{\theta}^t_{l\to lk} \\
A^t_{\theta,\,k\to ik} &= \frac{1}{N}\sum_{l=1,\,l\neq k}^{N}\left[S_{lk}^2(\hat{\theta}^t_{l\to lk})^2 - R_{lk}\left((\hat{\theta}^t_{l\to lk})^2 + \sigma^t_{\theta,\,l\to lk}\right)\right].
\end{aligned} \tag{A.7}$$

The equations now close on the means and variances of the messages (A.2):

$$\begin{aligned}
\hat{\theta}^t_{i\to ij} &= f_\theta\left(A^t_{\theta,\,i\to ij},\ B^t_{\theta,\,i\to ij}\right)\\
\sigma^t_{\theta,\,i\to ij} &= \frac{\partial f_\theta}{\partial B}\left(A^t_{\theta,\,i\to ij},\ B^t_{\theta,\,i\to ij}\right)\\
\hat{v}^{t+1}_{k\to ik} &= f_v\left(A^t_{v,\,k\to ik},\ B^t_{v,\,k\to ik}\right)\\
\sigma^{t+1}_{v,\,k\to ik} &= \frac{\partial f_v}{\partial B}\left(A^t_{v,\,k\to ik},\ B^t_{v,\,k\to ik}\right)\,,
\end{aligned} \tag{A.8}$$

where we have introduced the input functions (with x indicating either θ or v)

$$f_x\left(A,\,B\right) \equiv \frac{1}{Z_x(A,B)}\int \mathrm{d}x P_x(x)\mathrm{e}^{-\frac{1}{2}Ax^2+Bx}\,x\,. \tag{A.9}$$

Equations (A.8) together with (A.9) are the relaxed BP (rBP) equations.

A.1.2. From relaxed belief propagation to approximate message passing. The rBP equations are a direct consequence of the central limit theorem. Each message is a random variable and since the $\Theta(N)$ incoming messages are only weakly correlated they result in an effective Gaussian field, acting on each variable node. This field is additionally weighted with the prior on each of the sides. The outgoing messages also only weakly depend on the target node:

$$\begin{aligned}
B^t_{\theta,\,i} &\equiv \frac{1}{\sqrt{N}}\sum_{k=1}^{M} S_{ik}\hat{v}^t_{k\to ik}\\
&= B^t_{\theta,\,i\to ij} + \frac{1}{\sqrt{N}}S_{ij}\hat{v}^t_{j\to ij}\\
A^t_{\theta,\,i} &\equiv \frac{1}{N}\sum_{k=1}^{M}\left[S^2_{ik}(\hat{v}^t_{k\to ik})^2 - R_{ik}\left((\hat{v}^t_{k\to ik})^2+\sigma^t_{v,\,k\to ik}\right)\right]\\
&= A^t_{\theta,\,i\to ij} + O(\frac{1}{N})\\
B^t_{v,\,k} &\equiv \frac{1}{\sqrt{N}}\sum_{l=1}^{N} S_{lk}\hat{\theta}^t_{l\to lk}\\
&= B^t_{v,\,k\to ik} + \frac{1}{\sqrt{N}}S_{ik}\hat{\theta}^t_{i\to ik}\\
A^t_{\theta,\,k} &\equiv \frac{1}{N}\sum_{l=1}^{N}\left[S^2_{lk}(\hat{\theta}^t_{l\to lk})^2 - R_{lk}\left((\hat{\theta}^t_{l\to lk})^2+\sigma^t_{\theta,\,l\to lk}\right)\right]\\
&= A^t_{\theta,\,k\to ik} + O(\frac{1}{N})\,.
\end{aligned} \tag{A.10}$$

The marginals can now be expressed in terms of the messages

$$\begin{aligned}
\hat{\theta}^t_i &= f_\theta\left(A^t_{\theta,\,i},\,B^t_{\theta,\,i}\right) = \hat{\theta}^t_{i\to ij} + \sigma^t_{v,\,i\to ij}\frac{1}{\sqrt{N}}S_{ij}\hat{v}^t_{j\to ij} + O(\frac{1}{N})\\
&= \hat{\theta}^t_{i\to ij} + \sigma^t_{v,\,i}\frac{1}{\sqrt{N}}S_{ij}\hat{v}^t_j + O(\frac{1}{N})\\
\hat{v}^t_k &= f_v\left(A^{t-1}_{v,\,k},\,B^{t-1}_{v,\,k}\right) = \hat{v}^t_{k\to ik} + \sigma^t_{\theta,\,k\to ik}\frac{1}{\sqrt{N}}S_{ik}\hat{\theta}^{t-1}_{i\to ik} + O(\frac{1}{N})\\
&= \hat{v}^t_{k\to ik} + \sigma^t_{\theta,\,k}\frac{1}{\sqrt{N}}S_{ik}\hat{\theta}^{t-1}_i + O(\frac{1}{N})\,.
\end{aligned} \tag{A.11}$$

Similarly $\sigma^t_{\theta,i} = \sigma^t_{\theta,i\to ij} + O(1/N)$ and $\sigma^t_{v,k} = \sigma^t_{v,k\to ik} + O(1/N)$. This process is sometimes referred to as 'TAPification'. We finally obtain a set of equations that is independent of the messages and only depends on the marginals, by plugging (A.11) back into (A.10):

$$\begin{aligned}
B^t_{\theta,i} &= \frac{1}{\sqrt{N}}\sum_{k=1}^{M} S_{ik}\hat{v}^t_k - \left(\frac{1}{N}\sum_{k=1}^{M} S^2_{ik}\sigma^t_{v,k}\right)\hat{\theta}^{t-1}_i \\
A^t_{\theta,i} &= \frac{1}{N}\sum_{k=1}^{M}\left[S^2_{ik}(\hat{v}^t_k)^2 - R_{ik}\left((\hat{v}^t_k)^2 + \sigma^t_{v,k}\right)\right] \\
\hat{\theta}^t_i &= f_\theta\left(A^t_{\theta,i}, B^t_{\theta,i}\right) \\
\sigma^t_{\theta,i} &= \frac{\partial f_\theta}{\partial B}\left(A^t_{\theta,i}, B^t_{\theta,i}\right) \\
B^t_{v,k} &= \frac{1}{\sqrt{N}}\sum_{l=1}^{N} S_{lk}\hat{\theta}^t_l - \left(\frac{1}{N}\sum_{l=1}^{N} S^2_{lk}\sigma^t_{\theta,l}\right)\hat{v}^t_k \\
A^t_{\theta,k} &= \frac{1}{N}\sum_{l=1}^{N}\left[S^2_{lk}(\hat{\theta}^t_l)^2 - R_{lk}\left((\hat{\theta}^t_l)^2 + \sigma^t_{\theta,l}\right)\right] \\
\hat{v}^{t+1}_k &= f_v\left(A^t_{v,k}, B^t_{v,k}\right) \\
\sigma^{t+1}_{v,k} &= \frac{\partial f_v}{\partial B}\left(A^t_{v,k}, B^t_{v,k}\right) .
\end{aligned} \tag{A.12}$$

These are the AMP equations. The additional terms that appear in B_θ and B_v after TAPification are *Onsager reaction terms* that correct the mean field contribution from the first sum. The equations found in section 2.5 are a further simplification of these equations under the assumption that (a) the model matched the one with which the data was generated and (b) that the terms in the brackets of the equations for the B_θ, B_v are self-averaging. The terms S^2_{ij} and R_{ij} can then be replaced by their averages

$$\begin{aligned}
\Delta^{-1} &= \mathbb{E}_{P(L_{ij}|w_{ij}=0)}\left[S^2_{ij}\right] \\
R &= \mathbb{E}_{P(L_{ij}|w_{ij}=0)}\left[R_{ij}\right] = 0 .
\end{aligned} \tag{A.13}$$

Where the last equality is a consequence of the normalization of the conditional probability: $\int \mathrm{d}L P(L \mid w) = 1 \Rightarrow \int \mathrm{d}L \partial_{w^2} P(L \mid w) = 0$. For the dDS model we have

$$\begin{aligned}
S_{ij} &= L_{ij}\sqrt{\nu} \\
R_{ij} &= L^2_{ij}\nu - L^2_{ij}\nu = 0.
\end{aligned} \tag{A.14}$$

In the Bayes optimal setting, where $P_x(x) = P_{x^0}(x)$ (for $x \in \{\theta, v\}$) and $P(L_{ij} \mid w_{ij}) = P_0(L_{ij} \mid w_{ij})$, we have

$$\begin{aligned}
\Delta^{-1} &= (1-\rho)\,\nu \\
R &= 0.
\end{aligned} \tag{A.15}$$

And finally we obtain the AMP equations, as outlined in the main text.

A.2. State evolution

The AMP equations depend explicitly on the realization of the data (for the crowdsourcing these are the labels L_{ij}). These enter through S_{ij} and possibly R_{ij}. Therefore the $B^t_{\theta_i}$, $B^t_{v_j}$ and $A^t_{\theta,i}$, $A^t_{v,j}$ are random variables in the equations (A.12). Let us consider equations (A.10) in

order to derive their distributions. Recalling that the different messages, incoming to one node are independent by BP assumption we can apply the CLT to the sums on the r.h.s. of the equations for $B^t_{\theta,i}$, $B^t_{v,j}$ in (A.10). The mentioned independence holds only approximately because the underlying graph is not a tree, but on account of the $O(1/\sqrt{N})$ scaling this suffices in the $N \to \infty$ limit. Thus we have

$$\begin{aligned} B^t_{\theta,i} &\sim \mathcal{N}\left(\mathbb{E}B^t_{\theta,i},\, \mathbb{E}(B^t_{\theta,i})^2 - (\mathbb{E}B^t_{\theta,i})^2\right) \\ B^t_{v,j} &\sim \mathcal{N}\left(\mathbb{E}B^t_{v,j},\, \mathbb{E}(B^t_{v,j})^2 - (\mathbb{E}B^t_{v,j})^2\right)\,. \end{aligned} \tag{A.16}$$

Further more, by the law of large numbers, the r.h.s. of the equation for $A^t_{\theta,i}$ and $A^t_{v,j}$ in (A.10) can be replaced by their averages to obtain

$$\begin{aligned} A^t_{\theta,i} &\overset{N\to\infty}{\to} \mathbb{E}A^t_{\theta,i} \\ A^t_{v,j} &\overset{N\to\infty}{\to} \mathbb{E}A^t_{v,j}. \end{aligned} \tag{A.17}$$

It remains to compute the first two moments of $B^t_{\theta,i}$, $B^t_{v,j}$ and the first moment of $A^t_{\theta,i}$, $A^t_{v,j}$. We introduce the following order parameters that will turn up naturally during the computation

$$\begin{aligned} M^t_\theta &= \frac{1}{N}\sum_{i=1}^N \hat{\theta}^t_i \theta^0_i\,, M^t_v = \frac{1}{M}\sum_{j=1}^M \hat{v}^t_j v^0_j\,, \\ Q^t_\theta &= \frac{1}{N}\sum_{i=1}^N (\hat{\theta}^t_i)^2\,, Q^t_v = \frac{1}{M}\sum_{j=1}^M (\hat{v}^t_j)^2\,, \\ \Sigma^t_\theta &= \frac{1}{N}\sum_{i=1}^N \sigma^t_{\theta,i}\,, \Sigma^t_v = \frac{1}{M}\sum_{j=1}^M \sigma^t_{v,j}\,. \end{aligned} \tag{A.18}$$

We start by consideration of the first moment of $B^t_{\theta,i}$

$$\mathbb{E}\left[B^t_{\theta,i}\right] = \frac{1}{\sqrt{N}}\sum_{k=1}^M \int \mathrm{d}L_{ik} P_0\left(L_{ik} \mid w^0_{ik}\right) S_{ik}\hat{v}^t_{k\to ik}\,. \tag{A.19}$$

Expansion w.r.t. w^0_{ik} leads to

$$\begin{aligned} \mathbb{E}\left[B^t_{\theta,i}\right] = \frac{1}{\sqrt{N}}\sum_{k=1}^M \int \mathrm{d}L_{ik} P_0\left(L_{ik} \mid 0\right) \\ \cdot \left[1 + \frac{\theta^0_i v^0_j}{\sqrt{N}} \left.\frac{\partial \log P_0(L_{ik} \mid w)}{\partial w}\right|_0 + O(\frac{1}{N})\right] S_{ik}\hat{v}^t_{k\to ik}\,, \end{aligned} \tag{A.20}$$

which can be further simplified because the first order term vanishes

$$\int \mathrm{d}L_{ik} P_0\left(L_{ik} \mid 0\right) S_{ik} = 0 \tag{A.21}$$

and therefore

$$\mathbb{E}\left[B^t_{\theta,i}\right] = \frac{\alpha}{\hat{\Delta}} M^t_v \theta^0_i + O(1/\sqrt{N})\,, \tag{A.22}$$

where

$$\hat{\Delta}^{-1} = \mathbb{E}_{P_0(L|w^0=0)}\left[\left(\frac{\partial \log P_0(L_{ik} \mid w)}{\partial w}\right)^2\right]\,, \tag{A.23}$$

which will be equal to Δ^{-1} from (12) in the Bayes optimal case.

In the Bayes optimal setting, where all the distributions and parameters are known, this follows from

$$\int \mathrm{d}L_{ik} P(L_{ik} \mid w) = 1 \Rightarrow \int \mathrm{d}L_{ik} \frac{\partial}{\partial w} P(L_{ik} \mid w) = 0\,, \tag{A.24}$$

because

$$\frac{\partial}{\partial w} P(L_{ik} \mid w)|_{w=0} = P(L_{ik} \mid 0) \frac{\partial}{\partial w} \log P(L_{ik} \mid w)|_{w=0} \tag{A.25}$$

and thus

$$\int \mathrm{d}L_{ik} P(L_{ik} \mid 0) S_{ik} = \int \mathrm{d}L_{ik} P(L_{ik} \mid 0) \log P(L_{ik} \mid 0)\,. \tag{A.26}$$

If we are not in the Bayes optimal setting, then this still holds, as long as S_{ik} has mean $o(1/\sqrt{N})$, which will be the case in the example of crowdsourcing. Finally, replacing

$$\hat{v}^t_{k\to ik} = \hat{v}^t_k + O(1/\sqrt{N})$$

yields the result.

The second moment can be computed straight forwardly. To leading order one finds

$$\mathbb{E}\left[(B^t_{\theta,\,i})^2\right] = \frac{\alpha}{\hat{\Delta}} Q^t_v + O(\frac{1}{N})\,. \tag{A.27}$$

An analogue computation can be carried out to obtain

$$\mathbb{E}\left[B^t_{v,\,j}\right] = \frac{1}{\hat{\Delta}} M^t_u v^0_j + O(\frac{1}{\sqrt{N}})\,, \tag{A.28}$$

and

$$\mathbb{E}\left[(B^t_{v,\,j})^2\right] = \frac{1}{\hat{\Delta}} Q^t_\theta + O(\frac{1}{N})\,. \tag{A.29}$$

After the introduction of

$$\hat{R} = \mathbb{E}_{P_0(L|w^0=0)} \left[\left(\frac{\partial \log P_0(L_{ik} \mid w)}{\partial w} \right)^2 + \frac{\partial^2 \log P_0(L_{ik} \mid w)}{\partial w^2} \right] \tag{A.30}$$

similar arguments lead to an expression for the averages of A^t_θ, A^t_v in terms of order parameters:

$$\begin{aligned} \mathbb{E}\left[A^t_{\theta,\,i}\right] &= \frac{\alpha}{\hat{\Delta}} Q^t_v - \alpha \hat{R}(Q^t_v + \Sigma^t_v) + O(\frac{1}{\sqrt{N}}) \\ \mathbb{E}\left[A^t_{v,\,j}\right] &= \frac{1}{\hat{\Delta}} Q^t_\theta - \hat{R}(Q^t_\theta + \Sigma^t_\theta) + O(\frac{1}{\sqrt{N}})\,. \end{aligned} \tag{A.31}$$

The estimators $\hat{\theta}^t_i$ and $\hat{v}^t_j$ are functions of the random variables $B^t_{\theta,\,i}$ and $B^t_{v,\,j}$ respectively, which distribution is now known in terms of the order parameters. Therefore we do now also know the distributions of $\hat{\theta}^t_i$ and $\hat{v}^t_j$

$$\hat{\theta}_i^t = f_\theta\left(\frac{\alpha Q_v^t}{\hat{\Delta}} - \alpha\hat{R}(Q_v^t + \Sigma_v^t),\ \frac{\alpha M_v^t}{\hat{\Delta}}\theta^0 + \sqrt{\frac{\alpha Q_v^t}{\hat{\Delta}}}W\right)$$
$$\hat{v}_j^{t+1} = f_v\left(\frac{Q_\theta^t}{\hat{\Delta}} - \hat{R}(Q_\theta^t + \Sigma_\theta^t),\ \frac{M_\theta^t}{\hat{\Delta}}v^0 + \sqrt{\frac{Q_\theta^t}{\hat{\Delta}}}W\right), \tag{A.32}$$

where W is a standard normal distributed random variable. The equations can now be closed on the order parameters (A.18) as summarized in the following equations

$$M_\theta^t = \mathbb{E}_{\theta^0,W}\left[f_\theta\left(\frac{\alpha Q_v^t}{\hat{\Delta}} - \alpha\hat{R}(Q_v^t + \Sigma_v^t),\ \frac{\alpha M_v^t}{\hat{\Delta}}\theta^0 + \sqrt{\frac{\alpha Q_v^t}{\hat{\Delta}}}W\right)\theta^0\right]$$
$$Q_\theta^t = \mathbb{E}_{\theta^0,W}\left[\left(f_\theta\left(\frac{\alpha Q_v^t}{\hat{\Delta}} - \alpha\hat{R}(Q_v^t + \Sigma_v^t),\ \frac{\alpha M_v^t}{\hat{\Delta}}\theta^0 + \sqrt{\frac{\alpha Q_v^t}{\hat{\Delta}}}W\right)\right)^2\right]$$
$$\Sigma_\theta^t = \mathbb{E}_{\theta^0,W}\left[\frac{\partial f_\theta}{\partial B}\left(\frac{\alpha Q_v^t}{\hat{\Delta}} - \alpha\hat{R}(Q_v^t + \Sigma_v^t),\ \frac{\alpha M_v^t}{\hat{\Delta}}\theta^0 + \sqrt{\frac{\alpha Q_v^t}{\hat{\Delta}}}W\right)\right]$$
$$M_v^{t+1} = \mathbb{E}_{v^0,W}\left[f_v\left(\frac{Q_\theta^t}{\hat{\Delta}} - \hat{R}(Q_\theta^t + \Sigma_\theta^t),\ \frac{M_\theta^t}{\hat{\Delta}}v^0 + \sqrt{\frac{Q_\theta^t}{\hat{\Delta}}}W\right)v^0\right]$$
$$Q_v^{t+1} = \mathbb{E}_{v^0,W}\left[\left(f_v\left(\frac{Q_\theta^t}{\hat{\Delta}} - \hat{R}(Q_\theta^t + \Sigma_\theta^t),\ \frac{M_\theta^t}{\hat{\Delta}}v^0 + \sqrt{\frac{Q_\theta^t}{\hat{\Delta}}}W\right)\right)^2\right]$$
$$\Sigma_v^{t+1} = \mathbb{E}_{v^0,W}\left[\frac{\partial f_\theta}{\partial B}\left(\frac{Q_\theta^t}{\hat{\Delta}} - \hat{R}(Q_\theta^t + \Sigma_\theta^t),\ \frac{M_\theta^t}{\hat{\Delta}}v^0 + \sqrt{\frac{Q_\theta^t}{\hat{\Delta}}}W\right)\right]. \tag{A.33}$$

These equations track the evolution of the AMP equations (A.12).

In the Bayes optimal setting $\hat{\Delta} = \Delta$ and $\hat{R} = R$. Further more, the set of order parameters can be reduced because $M_x^t = Q_x^t$ and $\Sigma_x^t = \mathbb{E}_{x^0}[(x^0)^2] - Q_x^t$, where x stands for either θ or v. The above equations simplify to

$$M_\theta^t = \mathbb{E}_{\theta^0,W}\left[f_\theta\left(\frac{\alpha M_v^t}{\Delta},\ \frac{\alpha M_v^t}{\Delta}\theta^0 + \sqrt{\frac{\alpha M_v^t}{\Delta}}W\right)\theta^0\right]$$
$$M_v^t = \mathbb{E}_{v^0,W}\left[f_v\left(\frac{M_\theta^t}{\Delta},\ \frac{M_\theta^t}{\Delta}v^0 + \sqrt{\frac{M_\theta^t}{\Delta}}W\right)v^0\right], \tag{A.34}$$

which are the equations found in the main body of the paper. For the reader, interested in a rigorous treatment, we refer to [19, 20].

A.3. Derivation of (34)

The starting point are the SE equations (18), that we recall here for convenience

$$M_v^{t+1} = G(\frac{M_\theta^t}{\Delta})$$
$$M_\theta^t = T(\frac{\alpha}{\Delta}M_v^t). \tag{A.35}$$

We will be assuming for simplicity that there is no bias in the labels, i.e. $\beta = 1/2$ in (30), and a generic skewed RB prior

$$P_\theta(\theta) = (1-\mu)\,\delta(\theta) + \mu\left[(1-\lambda)\delta(\theta-1) + \lambda\delta(\theta+1)\right], \tag{A.36}$$

which leads to

$$\begin{aligned} G(x) &= \frac{1}{2}\mathbb{E}_W\left[\tanh(x+\sqrt{x}W) - \tanh(-x+\sqrt{x}W)\right] \\ T(x) &= \mu\mathbb{E}_W\left[(1-\lambda)\frac{(1-\lambda)\mathrm{e}^{x+\sqrt{x}W} - \lambda\mathrm{e}^{-x-\sqrt{x}W}}{\frac{1-\mu}{\mu}\mathrm{e}^{x/2} + (1-\lambda)\mathrm{e}^{x+\sqrt{x}W} + \lambda\mathrm{e}^{-x-\sqrt{x}W}}\right. \\ &\qquad \left. - \lambda\frac{(1-\lambda)\mathrm{e}^{-x+\sqrt{x}W} - \lambda\mathrm{e}^{x-\sqrt{x}W}}{\frac{1-\mu}{\mu}\mathrm{e}^{x/2} + (1-\lambda)\mathrm{e}^{-x+\sqrt{x}W} + \lambda\mathrm{e}^{x-\sqrt{x}W}}\right]. \end{aligned} \tag{A.37}$$

We are interested in the $\Delta \to 0$ limit which can be deduced from the above equations via an asymptotic expansion that is valid as long as the order parameters stay well away from zero (which can be deduced self-consistently below). The strategy is therefore as follows: we first derive the asymptotic behavior of $G(x)$ and $T(x)$, which in turn leads to a simplified fixed-point equation for

$$M_v = G\left(\frac{1}{\Delta}T\left(\frac{\alpha}{\Delta}M_v\right)\right) \tag{A.38}$$

that can be solved for M_v, which finally leads to the resulting error via (21). For (30) with $\beta = 1/2$ we have $R_v^{\mathrm{SE}} \equiv R_v(\frac{M_\theta}{\Delta})$, with

$$R_v(x) = \mathrm{erf}\left(\sqrt{\frac{x}{2}}\right), \tag{A.39}$$

and $\mathrm{erf}(x)$ denoting the error function.

The $x \to \infty$ behavior of $T(x)$ is dominated by the value of x and the $\sqrt{x}W$-terms can be neglected similarly as the exponentials with negative x-arguments, such that

$$T(x) \overset{x\to\infty}{\longrightarrow} \mu\left[(1-\lambda)^2\frac{\mathrm{e}^x}{\frac{1-\mu}{\mu}\mathrm{e}^{x/2} + (1-\lambda)\mathrm{e}^x} + \lambda^2\frac{\mathrm{e}^x}{\frac{1-\mu}{\mu}\mathrm{e}^{x/2} + \lambda\mathrm{e}^x}\right]. \tag{A.40}$$

Re-writing leads to

$$T(x) \overset{x\to\infty}{\longrightarrow} \mu(1 - 2\frac{1-\mu}{\mu}\mathrm{e}^{-\frac{x}{2}}). \tag{A.41}$$

Similarly, one finds

$$G(x) \overset{x\to\infty}{\longrightarrow} 1 - 2\mathrm{e}^{-2x}. \tag{A.42}$$

The fixed point equation can be simplified in this limit

$$M_v = G\left(\frac{1}{\Delta}\underbrace{T\left(\frac{\alpha}{\Delta}M_v\right)}_{=M_\theta \overset{\Delta\to 0}{\longrightarrow}\mu}\right) \tag{A.43}$$

$$\xrightarrow{\Delta\to 0} 1 - 2\mathrm{e}^{-2\frac{\mu}{\Delta}}\,. \tag{A.44}$$

It remains to express the ER. Since (see (20))

$$\mathrm{ER}_v = \frac{1}{2}(1 - R_v(\frac{M_\theta}{\Delta}))$$

and with $M_\theta^* = \mu$ in leading order and (A.39) we have

$$\mathrm{ER}_v = \frac{1}{2}\mathrm{erfc}\left(\sqrt{\frac{\mu}{2\Delta}}\right)\,, \tag{A.45}$$

where $\mathrm{erfc}(x)$ is the complementary error function. Another asymptotic expansion of the error function then leads to the desired result:

$$\mathrm{ER}_v = \frac{1}{2}\mathrm{erfc}\left(\sqrt{\frac{\mu}{2\Delta}}\right) \xrightarrow{\Delta\to 0} \frac{1}{\sqrt{\frac{\mu}{2\Delta}\pi}}\mathrm{e}^{-\frac{\mu}{2\Delta}}\,. \tag{A.46}$$

The interesting fact that the error is independent of λ can also be seem from the visualization in the following figure A2.

A.4. Derivation of (35)

As in the previous derivation, the starting point are the SE equations (A.37), however this time we perform an expansion for small μ, i.e. for few valuable workers in the crowd. We find that

$$T(x) \xrightarrow{\mu\to 0} \mu^2((1-2\lambda)^2\cosh(x) + \sinh(x)) + O(\mu^3)\,. \tag{A.47}$$

Since $T(x) = O(\mu^2)$ and Δ is a μ independent parameter we expand $G(x)$ for small x in order to resolve the fixed point equation (A.38). Such an expansion leads to

$$G(x) \xrightarrow{x\to 0} x + O(x^2)\,. \tag{A.48}$$

We can expect (and will find self-consistently) that M_v is small in the limit we are interested in. Thus we expand (A.47) for small x:

$$T(x) \xrightarrow[x\to 0]{\mu\to 0} \mu^2((1-2\lambda)^2 + x)\,. \tag{A.49}$$

Employing the latter two equations in the fixed point equation (A.38) leads to

$$M_v = \alpha\frac{\mu^2}{\Delta^2}M_v + \frac{1}{\Delta}\mu^2(1-2\lambda)^2\,,$$

which can be resolved to yield the fixed point to leading order in μ

$$M_v = \frac{1}{\Delta}\mu^2(1-2\lambda)^2\,. \tag{A.50}$$

Consequently, from (A.33), we have $M_\theta = T(\frac{\alpha}{\Delta}M_v) = \mu^2(1-2\lambda)^2 + O(\mu^4)$ and using once again (A.45) we obtain

$$\mathrm{ER}_v = \frac{1}{2}\left[1 - \mathrm{erf}\left(\sqrt{\frac{1}{2\Delta}}\,\mu|1-2\lambda|\right)\right]\,. \tag{A.51}$$

For small μ this is equal in leading order to

$$\mathrm{ER}_v \xrightarrow{\mu\to 0} \frac{1}{2} - \frac{2}{\pi\Delta}\mu|1-2\lambda|\,, \tag{A.52}$$

where we have used that $\mathrm{erf}(x) = 2/\sqrt{\pi}(x - x^3/3 + O(x^5))$.

This should be compared with the error resulting from majority voting. We assume $\rho = 0$ for simplicity. In order to derive the error obtained by majority voting under our model one can consider the sum

$$S_N \equiv \sum_{i=1}^{N} \mathbb{I}(L_{ij} = v_j),$$

with

$$\mathbb{I}(L_{ij} = v_j) = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

and (see section 2.1)

$$p = \frac{1}{2}\left(1 + \sqrt{\frac{\nu}{N}}\mu(1-2\lambda)\right).$$

Majority voting succeeds when

$$S_N \geqslant \lfloor \frac{N}{2} \rfloor + 1$$

or in other words, when the normalized sum (introducing $\sigma^2 = Np(1-p)$) follows

$$\frac{S_N - Np}{\sqrt{N\sigma^2}} \geqslant \frac{-\frac{1}{2}\sqrt{\nu N}\mu(1-2\lambda)}{\sqrt{N\frac{1}{4}(1-\frac{\nu}{N}\mu^2(1-2\lambda^2))}} + O(\frac{1}{\sqrt{N}})\,.$$

Deploying the de Moivre–Laplace theorem, we can estimate

$$\lim_{N\to\infty} \Pr\left(-\sqrt{\nu}\mu(1-2\lambda) \leqslant \frac{S_N - N*p}{\sqrt{N\sigma^2}} \leqslant \infty\right) = \int_{-\sqrt{\nu}\mu(1-2\lambda)}^{\infty} \mathrm{D}x\,. \tag{A.53}$$

With Dx indicating the standard normal measure. The right hand side can be rewritten in terms of the error function as

$$\frac{1}{2}\left[1 + \mathrm{erf}\left(\sqrt{\frac{\nu}{2}}\mu(1-2\lambda)\right)\right]$$

and consequently we have

$$\mathrm{ER}_v^{\mathrm{MV}} = \frac{1}{2}\left[1 - \mathrm{erf}\left(\sqrt{\frac{\nu}{2}}\mu(1-2\lambda)\right)\right]. \tag{A.54}$$

Comparing to (A.51) and recalling that for $\rho = 0 \;\Rightarrow\; \Delta^{-1} = \nu$ we see that the two errors coincide asymptotically, for $N \to \infty$, $\mu \to 0$, $\beta = 1/2$ and $\lambda \geqslant 1/2$. This is the result claimed in (35).

ORCID iDs

Christian Schmidt https://orcid.org/0000-0001-7349-8119

References

[1] Dawid A P and Skene A M 1979 *J. R. Stat. Soc.* C **28** 20–8
[2] Karger D R, Oh S and Shah D 2011 Iterative learning for reliable crowdsourcing systems *Advances in Neural Information Processing Systems* vol **24** pp 1953–61
[3] Dempster A P, Laird N M and Rubin D B 1977 *J. R. Stat. Soc.* B **39** 1–38
[4] Smyth P, Fayyad U M, Burl M C, Perona P and Baldi P 1995 Inferring ground truth from subjective labelling of venus images *Advances in Neural Information Processing Systems* pp 1085–92
[5] Raykar V C, Yu S, Zhao L H, Valadez G H, Florin C, Bogoni L and Moy L 2010 *J. Mach. Learn. Res.* **11** 1297–322
[6] Jin R and Ghahramani Z 2003 Learning with multiple labels *Advances in Neural Information Processing Systems* pp 921–8
[7] Whitehill J, Wu T F, Bergsma J, Movellan J R and Ruvolo P L 2009 Whose vote should count more: optimal integration of labels from labelers of unknown expertise *Advances in Neural Information Processing Systems* pp 2035–43
[8] Welinder P, Branson S, Perona P and Belongie S J 2010 The multidimensional wisdom of crowds *Advances in Neural Information Processing Systems 23* pp 2424–32
[9] Zhang Y, Chen X, Zhou D and Jordan M I 2014 Spectral methods meet em: a provably optimal algorithm for crowdsourcing *Advances in Neural Information Processing Systems* pp 1260–8
[10] Ghosh A, Kale S and McAfee P 2011 Who moderates the moderators?: crowdsourcing abuse detection in user-generated content *Proc. 12th ACM Conf. on Electronic Commerce* (ACM) pp 167–76
[11] Dalvi N, Dasgupta A, Kumar R and Rastogi V 2013 Aggregating crowdsourced binary ratings *Proc. 22nd Int. Conf. on World Wide Web* (ACM) pp 285–94
[12] Ok J, Oh S, Shin J and Yi Y 2016 Optimality of belief propagation for crowdsourced classification *Int. Conf. on Machine Learning* pp 535–44
[13] Ok J, Oh S, Shin J and Yi Y 2018 Optimal inference in crowdsourced classification via belief propagation *IEEE Trans. Inf. Theory* **64** 6127–38
[14] Liu Q, Peng J and Ihler A T 2012 Variational inference for crowdsourcing *Advances in Neural Information Processing Systems 25* pp 692–700
[15] Deshpande Y and Montanari A 2014 Information-theoretically optimal sparse PCA *IEEE Int. Symp. on Information Theory* pp 2197–201
[16] Matsushita R and Tanaka T 2013 Low-rank matrix reconstruction and clustering via approximate message passing *Advances in Neural Information Processing Systems 26* pp 917–25
[17] Lesieur T, Krzakala F and Zdeborová L 2015 MMSE of probabilistic low-rank matrix estimation: universality with respect to the output channel *53rd Annual Allerton Conf. on Communication, Control, and Computing* pp 680–7
[18] Lesieur T, Krzakala F and Zdeborová L 2017 *J. Stat. Mech.* 073403
[19] Rangan S and Fletcher A K 2012 Iterative estimation of constrained rank-one matrices in noise *IEEE Int. Symp. on Information Theory Proceedings* pp 1246–50
[20] Javanmard A and Montanari A 2013 *Inf. Inference* **2** 115–44
[21] Miolane L 2017 (arXiv:1702.00473)
[22] Zhou D, Liu Q, Platt J C, Meek C and Shah N B 2015 (arXiv:1503.07240)
[23] Khetan A and Oh S 2016 Achieving budget-optimality with adaptive schemes in crowdsourcing *Advances in Neural Information Processing Systems* pp 4844–52
[24] Shah N B, Balakrishnan S and Wainwright M J 2016 (arXiv:1606.09632)
[25] Shah D and Lee C 2018 Reducing crowdsourcing to graphon estimation, statistically *Int. Conf. on Artificial Intelligence and Statistics* pp 1741–50
[26] Sherrington D and Kirkpatrick S 1975 *Phys. Rev. Lett.* **35** 1792
[27] Thouless D J, Anderson P W and Palmer R G 1977 *Phil. Mag.* A **35** 593–601

[28] Krzakala F, Xu J and Zdeborová L 2016 Mutual information in rank-one matrix estimation *IEEE Information Theory Workshop (ITW) pp* 71–5
[29] Zdeborová L and Krzakala F 2016 *Adv. Phys.* **65** 453–552
[30] Bolthausen E 2014 *Commun. Math. Phys.* **325** 333–66
[31] Deshpande Y, Abbe E and Montanari A 2016 Asymptotic mutual information for the binary stochastic block model *IEEE Int. Symp. on Information Theory* pp 185–9
[32] Caltagirone F, Lelarge M and Miolane L 2018 Recovering asymmetric communities in the stochastic block mode *IEEE Trans. Netw. Sci. Eng.* **5** 237–46
[33] Barbier J, Dia M, Macris N, Krzakala F, Lesieur T and Zdeborová L 2016 Mutual information for symmetric rank-one matrix estimation: a proof of the replica formula *Advances in Neural Information Processing Systems 29* pp 424–32