Gradient Descent for Sparse Rank-One Matrix Completion for Crowd-Sourced Aggregation of Sparsely Interacting Workers

Yao Ma [1]   Alex Olshevsky [1]   Venkatesh Saligrama [1]   Csaba Szepesvari [2]

Abstract

We consider worker skill estimation for the single-coin Dawid-Skene crowdsourcing model. In practice skill-estimation is challenging because worker assignments are sparse and irregular due to the arbitrary, and uncontrolled availability of workers. We formulate skill estimation as a rank-one correlation-matrix completion problem, where the observed components correspond to *observed* label correlation between workers. We show that the correlation matrix can be successfully recovered and skills identifiable if and only if the sampling matrix (observed components) is irreducible and aperiodic. We then propose an efficient gradient descent scheme and show that skill estimates converges to the desired global optima for such sampling matrices. Our proof is original and the results are surprising in light of the fact that even the weighted rank-one matrix factorization problem is NP hard in general. Next we derive sample complexity bounds for the noisy case in terms of spectral properties of the *signless* Laplacian of the sampling matrix. Our proposed scheme achieves state-of-art performance on a number of real-world datasets.

1. Introduction

We consider the problem of label estimation in crowdsourcing. The basis of our work is the single-coin model of Dawid & Skene (1979): In this model, the input comes in the form of a *sparsely filled* $W \times T$ worker-task label matrix. The workers possess unique unknown skills, and tasks assume unique unknown labels. The worker-task label matrix collects the random labels provided by the workers for the individual tasks. The skill level of a worker is the (scaled) probability of the worker's label matching the true unknown label for any of the tasks. The observed labels are independent of each other.

Given the workers' skill levels, the optimal way (Nitzan & Paroush, 1981; Shapley & Grofman, 1984) to reconstruct the unknown labels is to use weighted majority voting where the weights assigned to the label provided by a worker is equal to the log-odds underlying the worker's skill. In practice, the crowd is often highly heterogenous ranging from highly skilled to negatively skilled workers. Downweighting unskilled workers and upweighting skilled workers has a significant impact on performance. Since skill levels are unknown, we follow prior works (e.g., Dalvi et al., 2013; Berend & Kontorovich, 2014; Szepesvári, 2015; Bonald & Combes, 2016) and adopt a two-step approach, whereby worker skills are first estimated and then these skills are used with the optimal weighting method to recover labels.

Sparse & Irregular Assignments: In practice, skill estimation is challenging because worker assignments are sparse and irregular due to the arbitrary and uncontrolled availability of workers (Karger et al., 2013; Dalvi et al., 2013). An additional subtle issue is the lack of diversity in terms of interactions between the workers: A worker is often grouped with a limited subset of workers across all tasks [1].

Our Contributions: *(i)* we formulate skill estimation as a weighted least-squares rank-one problem where the weights are proposed to compensate for the varying accuracy of the moment estimates which is expected to be seen in most practical applications; *(ii)* we propose to use projected gradient descent to minimize the resulting objective function; *(iii)* we provide a theoretical justification of this approach: we give natural and mild conditions on the weighting matrix under which we prove that gradient descent, despite the objective being nonconvex, is guaranteed to find the rank-

---

[1] Division of Systems Engineering, Boston University, USA [2] Department of Computing Science, University of Alberta, Canada. Correspondence to: Venkatesh Saligrama <srv@bu.edu>.

---

[1] This situation is remarkably evident on benchmark datasets: The 'Web' dataset has 177 workers, with 3 to 20 workers/task and each worker on average interacting with about another 2.7 workers only, while the standard deviation of how many workers a worker is interacting with is 15. The RTE dataset has 164 workers, has only 10 workers/task on the average and each worker interacts with fewer than 2.5 other workers, while the standard deviation of the interaction degree is 20.

one decomposition of the true moment matrix despite the presence of sparse weighting; *(iv)* we provide experimental evidence for the strength of the proposed method both on synthetic and standard benchmark data. For the numerical illustration, as most datasets are multiclass (the labels take on more than two values), we also provide a naive extension of the method to the multiclass case, essentially following Karger et al. (2013; 2014), which works well in practice.

Technical Novelty: We derive a fundamental result for rank-one matrix completion: the unobserved entries can be recovered if and only if the sampling matrix is irreducible and aperiodic. Our results for convergence of the proposed gradient descent scheme should be especially surprising given that the weighted low-rank factorization problem is known to be NP-hard even for the rank-one case (Gillis & Glineur, 2011). The apparent contradiction is resolved by noting that we constrain both the data (rank-one) and the weighting matrices (irreducible and aperiodic). We present an entirely original proof that exploits combinatorial properties of bipartite graphs, which could be of independent interest. In contrast to our approach, existing results in low-rank matrix completion require strong assumptions on the weighting matrix. Typically, the weighting matrix is binary (i.e., an entry is either present or missing), and the assumptions require either incoherence or a random weighting matrix (e.g., Ge et al., 2016).

2. Related Work

Discriminative Approach: In contrast to our two-step approach, several works adopt a discriminative method for label prediction. These methods (Li & Yu, 2014; Tian & Zhu, 2015) directly identify true labels by various aggregation rules that incorporate worker reliability.

Skill Estimation: As mentioned earlier, we work in the problem of estimating skills under single-coin model. Past approaches to skill estimation are based on *maximum likelihood/maximum aposteriori* (ML/MAP) estimation, or *moment matching*, or a combination of these. In particular, various versions of the EM algorithm have been proposed to implement ML/MAP estimation, starting with the work of (Dawid & Skene, 1979). Variants and extensions of this method, tested in various problems, include (Hui & Walter, 1980; Smyth et al., 1995; Albert & Dodd, 2004; Raykar et al., 2010; Liu et al., 2012). A number of recent works were concerned with performance guarantees for EM and some of its variants (Gao & Zhou, 2013; Zhang et al., 2014; Gao et al., 2016). Another popular direction is to add priors over worker skills, labels or worker-task assignments. To properly deal with the extra information, various Bayesian methods (belief propagation, mean-field and variational methods) have been considered (Raykar et al., 2010; Karger et al., 2011; Liu et al., 2012; Karger et al., 2013;

2014). Moment matching is also widely used (Ghosh et al., 2011; Dalvi et al., 2013; Zhang et al., 2014; Gao et al., 2016; Bonald & Combes, 2016; Zhang et al., 2016). With the exception of Bonald & Combes (2016), who propose an ad-hoc method, the algorithms in these works use matrix or tensor factorization.[2]

In theory, an ML/MAP method which is *guaranteed* to maximize the likelihood/posterior, is the ideal method to accommodate irregular worker-task assignments. However, as far as we know, none of the existing algorithms, unless initialized with a moment-matching-based spectral method, is proven to indeed find a satisfactory approximate maximizer of the objective that it is maximizing (Zhang et al., 2016). At the same time, moment matching methods that use spectral (and in general algebraic) algorithms implicitly assume the regularity of worker-task assignments, too. Indeed, the approach of Ghosh et al. (2011) crucially relies on the regularity of the worker-task assignment (as the method proposed uses unnormalized statistics). In particular, this method is not expected to work at all on non-regular data. Other spectral methods, being purely algebraic, implicitly treat all entries in the estimated matrices and tensors as if they had the same accuracy, which, in the case of irregular worker-task assignments, is far from the truth. In particular, the need to explicitly deal with data with unequal accuracy is a widely recognized issue that has a long history in the low-rank factorization community, going back to the work of Gabriel & Zamir (1979). Starting with this work, the standard recommendation is to reformulate the low-rank estimation problem as a weighted least-squares problem (e.g., Gabriel & Zamir, 1979; Srebro & Jaakkola, 2003). In this paper we will also follow this recommendation.

While Dalvi et al. (2013) also use a weighted least-squares objective, this is not by choice, but rather as a consequence of the need to normalize the data rather than to correct for the inaccuracy of the data. Furthermore, rather than considering the direct minimization of the resulting objective, they use two heuristic approaches that also use an unweighted spectral method.

In this light, our goal is to make spectral methods suitable for non-regular worker-task data often seen in practice.

Matrix Factorization/Completion: Unlike the general matrix factorization problem arising in recommender systems (Koren et al., 2009), we are primarily concerned with rank-one estimation of square symmetric matrices. Existing results on matrix completion (Ge et al., 2016) for square symmetric matrices are more general but require stronger assumptions on the matrix such as incoherence and random sampling.

---

[2]While Ghosh et al. (2011) pioneered the matrix factorization approach, their work is less relevant to this discussion as they estimate the labels directly.

Notation and conventions: The set of reals is denoted by $\mathbb{R}$, the set of natural numbers which does not include zero is denoted by $\mathbb{N}$. For $k \in \mathbb{N}$, $[k] \doteq \{1, \ldots, k\}$. Empty sums are defined as zero. We will use $\mathbb{P}$ to denote the probability measure over the measure space holding our random variables, while $\mathbb{E}$ will be used to denote the corresponding expectation operator. For $p \in [1, \infty]$, we use $\|v\|_p$ to denote the p-norm of vectors. Further, $\|\cdot\|$ stands for the 2-norm, $\|\cdot\|_F$ is the Frobenius-norm. The cardinality of a set S is denoted by $|S|$. For a real-valued vector x, $|x|$ denotes the vector whose ith component is $|x_i|$. Proofs of new results, missing from the main text are given in the appendix.

3. Problem Setup

We consider binary crowdsourcing tasks where a set of workers provide binary labels for large number of items. Let $W \in \mathbb{N}$ be a fixed positive integer denoting the number of workers. A problem instance $\theta \doteq (s, A, g)$ is given by a skill vector $s = (s_1, \ldots, s_W) \in [-1, 1]^W$, the worker-task assignment set $A \subset [W] \times \mathbb{N}$ and the vector of "ground truth labels" $g \in \{\pm 1\}^{\mathbb{N}}$.

When $A \subset [W] \times [T]$ for some $T \in \mathbb{N}$, we say that θ is a finite instance with T tasks, otherwise θ is an infinite instance. We allow infinite tasks to be able to discuss asymptotic identifiability. Θ_W denotes the set of all instances.

Definition 1 (Interaction Graph). *Let A be a worker-task assignment set. The (worker) interaction graph underlying A is a graph $G = G_A$ with vertex set $[W]$ such that $G = ([W], E)$ with $i, j \in [W]$ connected $((i, j) \in E)$ in G if there exists some task $t \in \mathbb{N}$ such that both (i, t) and (j, t) are elements of A.*

The problem in label recovery with crowdsourcing is to recover the ground truth labels $(g_t)_t$ given observations $(Y_{w,t})_{(w,t) \in A}$, a collection of ± 1-valued random variables such that $Y_{w,t} = Z_{w,t} g_t$ for $(w, t) \in A$, where $(Z_{w,t})_{(w,t) \in A}$ is a collection of mutually independent random variables that satisfy $\mathbb{E}[Z_{w,t}] = s_w$.

A (deterministic) *inference method* underlying an assignment set A takes the observations $(Y_{w,t})_{(w,t) \in A}$ and returns a real-valued score for each task in A; the signs of the scores give the label-estimates. Formally, we define an inference method as a map $\gamma : \{\pm 1\}^A \to \mathbb{R}^{\mathbb{N}}$, where given $Y \in \{\pm 1\}^A$, $\gamma_t(Y)$, the tth component of $\gamma(Y) \in \mathbb{R}$, is the score inferred for task t. Inference methods are aimed at working with finite assignment sets. To process an infinite assignment set, we define the notion of *inference schema*. In particular, an *inference schema* underlying an infinite assignment set A is defined as the infinite sequence of inference methods $\gamma^{(1)}, \gamma^{(2)}, \ldots$ such that $\gamma^{(t)}$ is an inference method for $A \cap [W] \times [T]$.

When important, we will use the subindex θ in $\mathbb{P}_\theta$ to denote the dependence of the probability distribution over the probability space holding our random variables. We will use $\mathbb{E}_\theta$ to denote the corresponding expectation operator. With this notation, the average loss suffered by an inference schema $\gamma = (\gamma^{(1)}, \gamma^{(2)}, \ldots)$ on the first T tasks of an instance θ is

$$\mathcal{L}_T(\gamma; \theta) = \frac{1}{T} \mathbb{E}_\theta \left[\sum_{t=1}^T \mathbb{I}\left\{ \gamma_t^{(T)}(Y) g_t \leq 0 \right\} \right].$$

The optimal inference schema for an assignment set A *given* the knowledge of the skill vector $s \in [-1, 1]^W$ is denoted by $\gamma_{s,A}^*$. The next section gives a simple explicit form for this optimal schema. The *average regret* of an inference schema $\gamma = (\gamma^{(1)}, \gamma^{(2)}, \ldots)$ for an instance $\theta \in \Theta$ is its excess loss on the instance as compared to the loss of the optimal schema:

$$\overline{R}_T(\gamma; \theta) = \mathcal{L}_T(\gamma; \theta) - \mathcal{L}_T(\gamma_{s,A}^*; \theta).$$

We define asymptotic consistency and learnability:

Definition 2 (Consistency and Learnability). *An inference schema is said to be (asymptotically) consistent for an instance set $\Theta \subset \Theta_W$ if for any $\theta \in \Theta$, $\limsup_{T \to \infty} \overline{R}_T(\gamma) = 0$. An instance set $\Theta \subset \Theta_W$ is (asymptotically) learnable, if there is a consistent inference schema for it.*

3.1. Two-Step Plug-in Approach

We propose a two-step approach based on first estimating the skills and then utilizing a plug-in classifier to predict the ground-truth labels. The motivation for a two-step approach stems from existing results that characterize accuracy in terms of skill estimation errors. For the sake of exposition, we recall some of these results.

For future reference, define the log-odds weighted majority vote parameterized by parameter vector $\alpha \in (-1, 1)^W$ as

$$\gamma_{t,\alpha}(Y) = \sum_{i : (i,t) \in A} v(\alpha_i) Y_{i,t}, \text{ where } v(\alpha) = \log \frac{1 + \alpha}{1 - \alpha}.$$

(Nitzan & Paroush, 1981) showed that the optimal decision rule $\gamma_{s,A}^*$, which in fact minimizes the probability of the error $\mathbb{P}(\gamma_{t,s}(Y) g_t \leq 0)$ individually for every $t \in \mathbb{N}$, takes this form with parameter $\alpha = s$, with weights $v_i^* = v(s_i)$.

When skills are known, (Berend & Kontorovich, 2014) provide an upper error bound, as well as an asymptotically matching lower error bound in terms of the so called *committee potential*. When skills are only approximately known, (Szepesvári, 2015; Berend & Kontorovich, 2014) also show that similar results can be obtained:

Lemma 1. *For any $\epsilon > 0$, the loss with estimated weights*

$\hat{v}_i = v(\hat{s}_i)$ *satisfies*

$$\frac{1}{T} \mathbb{E}_\theta \left[\sum_{t=1}^{T} \mathbb{I} \left\{ \gamma_{t,\hat{s}}(Y) g_t \leq 0 \right\} \right]$$

$$\leq \frac{1}{T} \mathbb{E}_\theta \left[\sum_{t=1}^{T} \mathbb{I} \left\{ \gamma^*(Y) g_t \leq \epsilon \right\} \right] + \mathbb{P}_\theta (\|v^* - \hat{v}\|_1 \geq \epsilon).$$

The error $\|v^* - \hat{v}\|_1$ can be bounded in terms of the multiplicative norm-differences in the skill estimates (see (Berend & Kontorovich, 2014)):

Lemma 2. *Suppose* $\frac{1+\hat{s}_i}{1+s_i}, \frac{1-\hat{s}_i}{1-s_i} \in [1 - \delta_i, 1 + \delta_i]$ *then* $|v(s_i) - v(\hat{s}_i)| \leq 2|\delta_i|$.

These results together imply that a plug-in estimator with a guaranteed accuracy on the skill levels in turn leads to a bound on the error probability of predicting ground-truth labels. This motivates the skill estimation problem.

4. Weighted Least-Squares Estimation

In this section, we propose an asymptotically consistent skill estimator for irregular worker-task assignments. By this we not only mean that only a subset of workers provide labels for a given task, but more importantly we mean that the interaction graph is not a clique and there is considerable variability in how often workers work on identical tasks.

Recall that given an instance $\theta = (s, A, g)$, the data of the learner is given in the sparse matrix $(Y_{i,t})_{(i,t) \in A}$ which is a collection of independent binary random variables such that $Y_{i,t} = g_t Z_{i,t}$ and $s_i = \mathbb{E}(Z_{i,t})$. Define $N \in \mathbb{N}^{W \times W}$ to be the matrix whose (i,j)th entry with $i \neq j$ gives the number of times the workers i and j labeled the same task:

$$N_{ij} = |\{t \in \mathbb{N} : (i,t), (j,t) \in A\}|.$$

We also let $N_{ii} = 0$. Note that the there is an edge between workers i and j in the interaction graph, denoted by $G = ([W], E)$, exactly when $N_{ij} > 0$. That is, $(i,j) \in E$ if and only if $N_{ij} > 0$. When A is infinite, N_{ij} may be infinite. In this case, for $i \neq j$ we also define $N_{ij}(T) = |\{t \in [T] : (i,t), (j,t) \in A\}|$ to denote the number of times workers i and j provide a label for the same task and let $N_{ii}(T) = 0$.

Let θ be a finite instance. When $(i,t), (j,t) \in A$, since $g_t^2 = 1$, by our independence assumptions, $\mathbb{E}[Y_{i,t}, Y_{j,t}] = s_i s_j$. This motivates estimating the skills using

$$\tilde{s} = \operatorname{argmin}_{x \in [-1,+1]^W} \frac{1}{2} \sum_{(i,t),(j,t) \in A} (Y_{i,t} Y_{j,t} - x_i x_j)^2 \quad (1)$$

Note that the number of terms constraining the skill estimate of particular worker in this objective scales with how many other workers this worker works with. Intuitively, this should feel "right": the more a worker works with others, the more information we should have about its skill level.

As it turns out, there is an alternative form for this objective, which is also very instrumental and which will form the basis of our algorithm and also of our analysis. To introduce this form, define $C_{ij} \doteq s_i s_j$ and let

$$\tilde{C}_{ij} = \frac{1}{N_{ij}} \sum_{(i,t),(j,t) \in A} Y_{i,t} Y_{j,t} . \quad (2)$$

The alternative form of the objective in Eq. (1) is given by the following result:

Lemma 3. *Let $L : [-1,1]^W \to [0, \infty)$ be defined by*

$$L(x) = \frac{1}{2} \sum_{(i,j) \in E} N_{ij} (\tilde{C}_{ij} - x_i x_j)^2.$$

The optimization problem of Eq. (1) is equivalent to the optimization problem $\operatorname{argmin}_{x \in [-1,+1]^W} L(x)$.

The proof, which is based on simple algebra, is given in Appendix A. In fact, the proof shows that the two objective functions are equal up to a shift by a constant.

The objective function from Lemma 3 can be seen as a weighted low-rank objective, first proposed by Gabriel & Zamir (1979). Clearly, the objective prescribes to approximate $\tilde{C}$ using $xx^\top$, with the error in the (i,j)th entry scaled by N_{ij}. Note that this weighting is reasonable as the variance of $\tilde{C}_{ij}$ is $1/N_{ij}$ and we expect from the theory of least-squares that an objective combining multiple terms where the data is heteroscedastic (has unequal variance), the terms should be weighted with the inverse of the data variances. Since, $N_{ii} = 0$, the weighting function N can in general be full-rank, and in this case the general weighted rank-one optimization is known to be NP-hard (Gillis & Glineur, 2011). However, our data has special structure, which may allow one to avoid the existing hardness results: On the one hand, as the number of data points increases, $\tilde{C}_{ij}$ will be near rank-one itself. On the other hand, we will put natural restrictions on the weighting matrix which are in fact necessary for identifiability. This restriction will essentially say that the limiting interaction graph, in which two workers are connected if and only if $N_{ij} = \infty$, should be irreducible and non-bipartite.

4.1. Plug-in Projected Gradient Descent

To solve the weighted least-squares objective, we propose a *Projected Gradient Descent* (PGD) algorithm (cf. Section 4). At each step we sequentially update the skill level based on following the negative gradient of the loss L:

$$\tilde{s}_i^{t+1} = s_i^t + \gamma \sum_{(i,j) \in E} N_{ij} (\hat{C}_{ij} - s_i^t s_j^t) s_j^t$$

$$s_i^{t+1} = \mathcal{P}(\tilde{s}_i^{t+1}),$$

where $\mathcal{P}(\cdot) : \mathbb{R} \to \left[-1 + \frac{\tau}{\sqrt{N_i}}, +1 - \frac{\tau}{\sqrt{N_i}} \right]$ is a projection function (i.e., $\mathcal{P}(x)$ truncates its input so that it belongs

Algorithm 1 Plug-in Projected Gradient Scheme

   Input: $N, Y = (Y_{i,t})_{(i,t) \in A}, \eta, \tau > 0$.
   $x_i \sim U[-1, 1], N_i = \sum_j N_{ij}$.
   $\tilde{C}_{ij} \leftarrow \frac{1}{N_{ij}} \sum_{(i,t),(j,t) \in A} Y_{i,t} Y_{j,t}, \forall (i,j)$ s.t. $N_{ij} > 0$.
   repeat
     for $i = 1, \ldots, W$ **do**
       $x_i \leftarrow x_i + \eta \sum_{j \in [W]} N_{ij} \tilde{C}_{ij} x_j$
             $- \eta \sum_{j \in [W]} N_{ij} x_i x_j^2$
       $x_i \leftarrow \min\{x_i, 1 - \frac{\tau}{\sqrt{N_i}}\}$
       $x_i \leftarrow \max\{x_i, -1 + \frac{\tau}{\sqrt{N_i}}\}$
     end for
   until x converges
   $\hat{s} \leftarrow \text{sgn}(\sum_{i \in [W]} x_i) x$
   for $t = 1, \ldots, T$ **do**
     $\hat{Y}_t \leftarrow \sum_{i \in [W]} Y_{i,t} \log \frac{1+\hat{s}_i}{1-\hat{s}_i}$
   end for
   return $(\hat{Y}_t)_{t \in [T]}$

to the interval it is projecting to), $\gamma > 0$ is the step size; $N_i = |\{t : (i, t) \in A\}|$ is the number of tasks labeled by worker i and $\tau > 0$ is a tuning parameter.

The purpose of the projection is to stay away from the boundary of the hypercube, where the log-odds function is changing very rapidly. The justification is that skills close to one have overwhelming impact on the plug-in rule and since the skill estimates are expected to have an uncertainty proportional to $\tau/\sqrt{N_i}$ with probability $\text{const} \times e^{-\tau^2}$, there is little loss in accuracy in confining the parameter estimates to the appropriately reduced hypercube. While in principle one could tune this parameter, we use $\tau = 1$ in this paper. As noted earlier, the skill vector can only be identified up to sign. To break this symmetry, in the paper we assume that the true unknown skill vector satisfies $\sum s_i > 0$. Thus, the final step of the algorithm reverses the sign of the skill vector estimate found if necessary to ensure that the estimate also has the property that the total skill level is positive.

5. Theoretical Results

In this section we derive theoretical results to shed light on the fundamental structural properties required of the interaction graph induced by an assignment set to ensure learnability with missing data. Subsequently, we analyze convergence properties of the PGD algorithm.

5.1. Learnability

There are different ways to let the number of tasks approach infinity while keeping an interaction graph fixed.
Case A: For a fixed interaction graph $G = ([W], E)$ we can consider assignment sets such that the minimum number of shared tasks, $T_{\min}(T) = \min_{(i,j) \in E} N_{ij}(T)$ approaches

infinity. Learnability in this context is a property of the interaction graph.
Case B: We can also start from an infinite assignment set A and define $G_A^\infty = ([W], E)$ as the graph where two workers are connected by an edge if $N_{ij} = \infty$. In other words define connectivity based on whether two workers interact finitely or infinitely many times.

We will follow the second approach as it is slightly more general than the first (the second approach allows assignment sets A where some workers interact only finitely many times, while the first approach does not allow such assignment sets). Thus, we fix an assignment set A and will consider a set of instances Θ sharing this assignment set.

To express complete ignorance towards the true unknown labels assigned to tasks, we state our result for *truth-complete* instance sets: For any $\theta = (s, A, g) \in \Theta$, we require $\Theta_{s,A} \subset \Theta$ where $\Theta_{s,A} = \{(s, A, g) : g \in \{-1, +1\}^{\mathbb{N}}\}$.

As mentioned before, the inference problem is inherently symmetric: The likelihood assigned to some observed data Y under an instance $\theta = (s, A, g)$ is the same as under the instance $(-s, A, -g)$. Thus, an instance set cannot be learnable unless somehow these symmetric solutions are ruled out. To express the condition on this we need a few more definitions. In particular, given $\Theta \subset \Theta_W$, we let $S(\Theta) = \{s \in [-1, 1]^W : (s, A, g) \in \Theta\}$ be the set of skill vectors underlying Θ. For a skill vector $s \in [-1, 1]^W$ we let $P(s) = \{i \in [W] : s_i > 0\}$ be the set of workers whose skills are positive and we let $\mathcal{P}(s) = \{P(s), P(-s)\}$ be the (incomplete) partitioning of workers into workers with positive and negative skills. Note that workers with zero skill are left out. Finally, we say that Θ is *rich* if there exists $s \in [-1, 1]^W$ and $\alpha > 1$ such that $\times_{i \in [W]}\{\alpha s_i, s_i/\alpha\} \subset S(\Theta)$.

With this, we are ready to state our first main result:

Theorem 1 (Characterization of learnability). *Fix an infinite assignment set A and assume that $G = G_A^\infty$ is connected. Then, a rich, truth-complete set of instances $\Theta \subset \Theta_A$ over A is learnable if and only if the following hold:*

(i) For any $s, s' \in S(\Theta)$ such that $|s| = |s'|$ and $\mathcal{P}(s) = \mathcal{P}(s')$, it follows that $s = s'$;

(ii) The graph is non-bipartite, i.e, it has an odd-cycle.

Richness is required so that there is sufficient ambiguity about skills. Condition (i) requires that any $s \in \Theta$ should be uniquely identified by $|s|$ and knowing which components of s have the same sign and which components are zero. For example, this condition will be met if Θ is restricted so that it only contains skill vectors that have a positive sum (which is the condition we will make in the rest of the paper).

The forward direction of the theorem statement hinges upon the following result:

Lemma 4. *For any $g \in \{\pm 1\}$, $s \in [-1, 1]^W$ and an assignment set with a connected, non-bipartite interaction graph G_A^∞, there exists a method to estimate $|s|$ and $\mathcal{P}(s)$.*

The reverse implication in the theorem statement follows from the following result:

Lemma 5. *Assume that the lengths of all cycles in G are even. Then there exists $s, s' \in [-1, 1]^W$, $s \notin \{-s', s'\}$ such that $C_{ij} = s_i s_j = s_i' s_j'$.*

Learnability for Finite Tasks: We mention in passing that asymptotic learnability is a fundamental requirement, which if not met precludes any reasonable finite time result.

5.2. Convergence of the PGD Algorithm

The previous section established that for learnability the limiting interaction graph G_A^∞ must be a non-bipartite connected graph. We will now show that PGD under these assumptions converges to a unique minimum for both the noisy and noiseless cases; by the latter we mean that in the loss L of Lemma 3, we set $\tilde{C}_{ij} = C_{ij} = s_i s_j$ for $(i, j) \in E$. Note that the (non-bipartite) odd-cycle condition together with that G is connected gives that the worker-interaction count matrix N is irreducible and aperiodic (and vice versa). We show that in this case the loss has a unique minima and the PGD algorithm recovers the skill-vector.

Theorem 2. *Suppose the components of the skill vector are non-zero. The PGD Algorithm of Sec 4 for the noiseless and positive case ($\tilde{C}_{ij} = C_{ij} > 0$) converges to the global minima under the conditions (i) and (ii) of Theorem 1 for positive skills. Furthermore, skill vectors can be recovered uniquely by means of a post-processing step for arbitrary C_{ij} under conditions (i) and (ii) of Theorem 1.*

The proof of the result is based on analyzing the critical points of the loss L underlying the PGD algorithm. Specifically, we wish to verify whether or not there exists a vector $x \neq s$ such that, for each $i = 1, \ldots, W$, we have

$$\sum_{j=1}^W N_{ij}(x_i x_j - s_i s_j) x_j = 0. \qquad (3)$$

We argue that when worker-interaction matrix $N = [N_{ij}]$ is irreducible and is aperiodic, the only two points that satisfy this equation are $x = s$ and $x = -s$. We then rule out the incorrect equilibrium point by invoking our prior assumption that $\sum_i s_i > 0$. Figure 1 illustrates the key insight of our proof. Note that for the noiseless case, the theorem imposes few restrictions on the interactions in terms of number of tasks per worker, the total number of tasks, or whether task assignments can be asynchronous. Indeed,
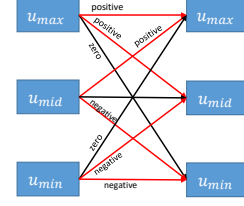

Figure 1. Illustration of the proof of Thm 2. The equilibrium points of PGD correspond to $\sum_j \tilde{N}_{ij}(u_i u_j - 1) = 0$ where $u_i = \frac{x_i}{s_i}$ and $\tilde{N}_{ij} = N_{ij} s_i s_j$, which is irreducible and aperiodic (for $s_i > 0$). We breakup nodes into three groups: (i) nodes i with $u_i = u_{\max} = \max_i \frac{x_i}{s_i}$ (ii) those nodes i with $u_i = u_{\min} = \min_i \frac{x_i}{s_i}$ (iii) all other nodes i and consider the case with $u_{\max} u_{\min} = 1$. The label on the edge going from node j to k is the sign of $u_j u_k - 1$. We show that the components $\tilde{N}_{jk}$ corresponding to $u_j u_k > 1$ (positive edges) and $u_j u_k < 1$ (negative edges) must be zero leaving behind a bipartite graph (black edges), which is a contradiction.

interactions could involve only two workers for each task and yet PGD converges to the skill-vector.

We will now extend these results to the noisy case. First notice that we can obtain asymptotic consistency as a direct corollary of Theorem 2:

Corollary 1. *Skill estimates are asymptotically consistent if the limiting interaction graph with weights $\rho_{ij} = \lim_{T \to \infty} N_{ij}(T)/T > 0$ is irreducible and aperiodic.*

The proof of this result follows from $\tilde{C}_{ij} \to C_{ij}$.

NOISY OBSERVATIONS: We next consider the noisy case, namely, $\tilde{C}_{ij} \neq C_{ij}$. In particular, let $\Delta = [\Delta_{ij}]$ be a $W \times W$ matrix with $C_{ij} = \tilde{C}_{ij} + \Delta_{ij}$. We leverage local strong convexity of the gradient to bound the skill-estimation error in terms of Δ. To this end, we consider the equilibrium points of the PGD with perturbation Δ:

$$\sum_{j=1}^W N_{ij}(x_i x_j - s_i s_j + \Delta_{ij}) x_j = 0$$

$$\implies \sum_{j=1}^W N_{ij}(x_i x_j - s_i s_j) x_j = \sum_{j=1}^W N_{ij} \Delta_{ij}. \qquad (4)$$

Let f denote L in the case when $\tilde{C}_{ij} = s_i s_j$ ("noise-free" objective). Note that the LHS of the last equality is $\nabla f(x)$. Hence, it follows that if x is a stationary point of L and x is bounded away from zero, then ∇f at x is small. Now, $\nabla f(s) = 0$ since s is a minimum of f. If we knew that f is strongly convex in a neighborhood of s, it would follow, at least, for small enough Δ that x is close to s. To show that f is indeed strongly convex, write $\nabla^2 f(x) = D_s P(x/s) D_s$ where $P_{ii}(x) = \sum_j 2N_{ij} x_i^2$, $P_{ij}(x) = 4N_{ij} x_i x_j - 2N_{ij}$, for $i \neq j$. Positive definiteness of $P(1)$ follows from the fact that $P(1)$ is a so-called unsigned Laplacian matrix (cf.

Proposition 2.1 of (Desai & Rao, 1994)). The argument is finished by resorting to continuity. This gives the following theorem (the detailed proof is in the appendix):

Theorem 3. *Suppose the worker-interaction matrix satisfies the assumptions in Theorem 2, Then for each $\epsilon \in (0,1)$ there exists a constant $c_\epsilon > 0$ with the following property: for any $\Delta \in \mathbb{R}^{W \times W}, x \in \mathbb{R}^W$ such that $\|\Delta\| \leq c_\epsilon$ and x is the solution of Eq. 4 and $\min_i |x_i| \geq \epsilon$,*

$$\|x - s\| \leq \frac{\|N\|_F \|\Delta\|}{s_{\min}^2 \sigma_{\min}(P(1))}, \quad where \quad s_{\min} = \min_i |s_i|.$$

Note that $\|N\|_F$ term in numerator and $\sigma_{\min}(P(1))$ (the smallest eigenvalue of $P(1)$) in the denominator both scale linearly with the number of tasks. Thus, the theorem states that for small enough perturbations, the error of stationary solutions scales proportionally to $\|\Delta\|_2$, with the proportional constant governed by the squared inverse minimum skill level and the minimum eigenvalue of $P(1)$, which is known to characterize how far the weighted graph with weights N is from being "bipartite" (Desai & Rao, 1994).

FINITE-TASK BOUND: Note that we can directly apply this result to obtain a finite task characterization as well. In particular consider a connected and non-bipartite interaction graph. Define, $T_{\min} = \min_{(i,j) \in E} N_{ij}$ as the minimum number of shared tasks; $d_{\max}$ as the maximum degree and D sum of the degrees. It follows by standard Hoeffding bounds that with probability greater than $(1 - \delta)$ we have $\max_{(i,j) \in E} |C_{ij} - \hat{C}_{ij}| \leq \frac{\log(D/\delta)}{\sqrt{N_{\min}}}$. By setting $\Delta = C_{ij} - \hat{C}_{ij}$ and invoking the Gershgorin circle theorem we conclude that with $\|\Delta\| \leq \frac{d_{\max} \log(D/\delta)}{\sqrt{N_{\min}}}$ with probability greater than $1 - \delta$. Substituting this expression in Theorem 3 yields with probability greater than $1 - \delta$ that

$$\|x - s\| \leq \frac{d_{\max} \log(D/\delta) \|N\|_F}{s_{\min}^2 \sigma_{\min}(P(1))\sqrt{T_{\min}}}.$$

6. Experimental Results

SYNTHETIC EXPERIMENTS: We will experiment with different graph types, increasing levels of label noise, graph-size, skill distribution, and different weighting functions on synthetic data. Results for graph size, skill distribution and weightings appear in supplementary material. Here we describe results for different graph types and noise.

Impact of Graph Type: We consider three 11-node (# workers) irreducible, non-bipartite graphs, namely, a Clique (G_1), Star with augmented odd cycle (G_2), and a Ring (G_3) to illustrate the impact of sparsity (Clique has dense worker interactions while Star/Ring have fewer than 3 worker interactions) and graph-type (Ring vs. Star). These graphs satisfy condition (ii) of Thm 1.

Noise Robustness: To see the impact of noise, we vary the noise level by increasing the number of tasks, which in turn reduces the error in the correlation matrix. Tasks are randomly assigned to binary classes ± 1 with total number of tasks ranging from 11 to 330. Skills are randomly assigned on a uniform grid between 0.8 and -0.3[3]

We compare the average prediction error $PE = \frac{1}{T} \sum_{t=1,\dots,T} 1\{\hat{Y}_t \neq g_t\}$ with the Majority Voting (MV) algorithm, the KOS algorithm (Karger et al., 2013), Opt-D&S algorithm (Zhang et al., 2014), the ER algorithm (Dalvi et al., 2013), the IWMV algorithm (Li & Yu, 2014), and the M3V algorithm (Tian & Zhu, 2015). The KOS algorithm is based on belief propagation, Opt-D&S uses a spectral method to initialize EM, the ER algorithm is the more successful spectral method of the paper defining it, the IWMV algorithm is an EM-style algorithm. Each algorithm is averaged over 15 trials on each dataset. The average prediction errors are presented in Figure 2. As the number of tasks grows, the average prediction error of PGD algorithm decreases. PGD is evidently robust to missing data/sparsity and graph-type. OPT-DS, which is close to PGD performance suffers significant performance degradation on sparse graphs such as rings. We can attribute this to the fact that a tensor-based method requires at least 3 worker annotations for each task (Zhang et al., 2014).

BENCHMARK DATASET EXPERIMENTS: We illustrate the performance of PGD algorithm against state-of-art algorithms. Each algorithm is executed on four data-sets, i.e. RTE1 (Snow et al.), Temp (Snow et al.), Dogs (Deng et al., 2009), and WebSearch (Zhou et al., 2012). Following convention we report errors between ground-truth and recovered labels. A summary of these data-sets is presented in Table 1. RTE1 and Temp data-sets have binary labels where our algorithm could be directly applied.

Multi-Class Datasets: For Dogs and Web (multiclass) we run our algorithm with one-vs-rest strategy for each class by assuming class-independent models determine the probability of the worker flipping the ground truth. A score function for class-conditional skill is calculated for each class k using $score(k) = \sum_{(i,t) \in A} \log \frac{1+s_i}{1-s_i} \mathbf{1}(Y_{i,t} = k)$, where $k \in \mathcal{K}$ is the class index and $\mathbf{1}(\cdot)$ is a ± 1 indicator. We predict the label by finding the class corresponding to the maximum of the score function. We also consider a closely related strategy (Li & Yu, 2014) (see also Supplementary) where the flipped ground-truth label is randomly assigned to one of the other classes. The skill estimation and label estimation for this scenario is a straightforward extension of our proposed scheme since the confusion matrix is characterized by a single skill parameter. We report

---

[3] The reason for this choice is to satisfy condition (i) in Theorem 1, i.e., requiring overall skills to be positive. Aggregate skill is about 0.25.

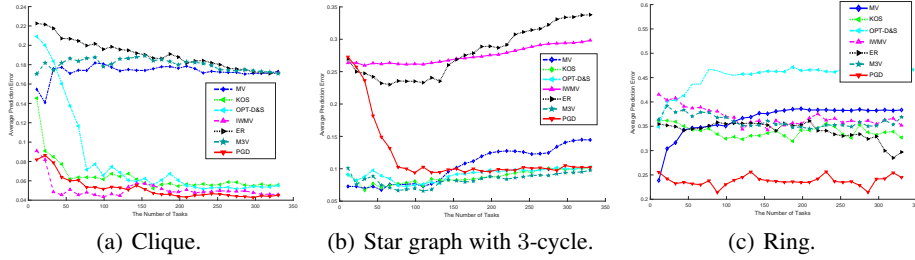| (a) Clique. | (b) Star graph with 3-cycle. | (c) Ring. |

Figure 2. Illustrative comparisons of prediction performance for three graph types. Only mean values are plotted for exposition. For the clique, standard deviation values with 11 tasks were 0.09, 0.10, 0.14, 0.14, 0.19, 0.09, and 0.09 for MV, KOS, OPT-D&S, PGD, ER, IWMV, and M3V respectively; and with 330 tasks they were 0.02, 0.01, 0.017, 0.014, 0.012, 0.013, and 0.018 respectively. For the star-graph the standard deviations for 11 tasks were 0.09, 0.13, 0.13, 0.06, 0.13 0.09, and 0.07 for MV, KOS, OPT-D&S, PGD, ER, IWMV, M3V respectively and for 330 tasks they were 0.016, 0.015, 0.013, 0.012, 0.04, 0.03, and 0.013. For the ring the standard deviation for 11 tasks were 0.096, 0.05, 0.08, 0.1, 0.11, 0.09, 0.08 and for 330 tasks they were 0.017, 0.05, 0.02, 0.043, 0.05, 0.086, 0.05. Standard deviations decrease with growing number of tasks.

Table 1. Benchmark Datasets with Prediction Errors for Different Methods.

| Datasets	Tasks	Workers	Instances	Classes	Sparsity level	Data	PGD	MV	Opt-D&S	KOS	ER	IWMV	M3W
RTE1	800	164	8000	2	0.0610	RTE1	**0.07**	0.1031	0.0712	0.3975	0.14	0.08	0.0813
Temp	462	76	4620	2	0.1316	Temp	**0.0512**	0.0639	0.0584	0.0628	0.052	0.06	0.0606
Dogs	807	109	8070	4	0.0917	Dogs	**0.1660**	0.1958	0.1689	0.3172	0.18	0.19	0.1822
Web	2665	177	15567	5	0.0033	Web	**0.1485**	0.2693	0.1586	0.4293	0.22	0.22	0.1847

| (a) RTE1. | (b) TEMP. |
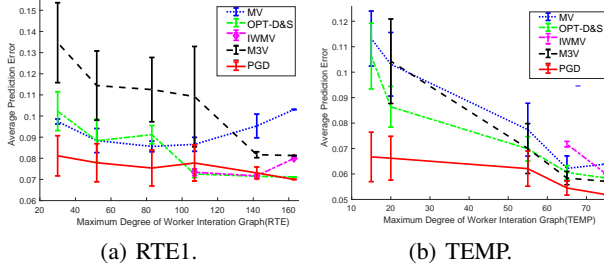
Figure 3. Impact of Graph Sparsification.

the best results among these two setups in Table 1. PGD algorithm uniformly outperforms the state-of-art algorithms.

Influence of Graph Sparsification: Here we consider the scenario where fewer workers label each task on the binary classification benchmark datasets. Binary classification tasks are aligned with our theoretical results. This experiment will highlight the performance of state-of-art algorithms under sparse task-assignments. We simulate this effect based on random sparsification. In particular, we sort the degree of each node on the interaction graph. To sparsify the graph we randomly delete edges starting with the highest degree node and continue this process for other nodes until we obtain an interaction graph with desired maximum degree. We also remove symmetrically remove corresponding edges of incident workers to maintain symmetry. This has the implicit effect of deleting some of the tasks as well (for instance, if a task is annotated by two workers). Higher levels of sparsification leads to fewer availability of tasks for training. We iteratively run PGD and the other algorithms for 50 Monte-Carlo trials with different desired maximum degrees. The average prediction errors are displayed in Figure 3. The reason IWMV performs poorly is that majority votes are no longer reliable, which IWMV relies on. Our PGD algorithm

is surprisingly robust to sparsification of interactions and degrades gracefully relative to other schemes. This highlights the fact that PGD is capable of leveraging sparse interactions among workers and obtain fairly robust estimates of skill-levels required for accurate prediction.

Time Complexity: We also compare the time complexity of proposed algorithm against state-of-art algorithms. Our PGD algorithm requires fewer iterations in comparison to other iterative methods and each iteration scales linearly with W and the maximum degree, D_{max}, of the worker-interaction graph which is bounded by W. Time complexity of different algorithms is summarized in Table 2 [4].

Table 2. Time Complexity/Iteration of Different Methods.

Alg.	PGD	IWMV	M3W
Com.	$O(D_{max}W)$	$O(TW)$	$O(W^2T)$

7. Conclusions

Worker skill estimation is a critical subtask in crowdsourcing and the general problem is related to the problem of estimation with intermittent sensor communications (Saligrama & Castanon, 2006), which we plan to explore in future work. To estimate the skills of workers we propose a new moment-matching approach with weighted rank-one approximation and propose a gradient algorithm. In contrast to prior work, the weights are set up to correct for the spread of the measured worker-worker agreements accuracies which are typical in real-world problems where who works on the same task with whom is out of control. We show that various intuitive graph-theoretical quantities play a fundamental role in the success of this approach.

---

[4]Opt-D&S, KOS, and ER algorithms are omitted. They employ spectral factorization and have high time complexity.

Acknowledgements

References

Albert, P. S. and Dodd, L. E. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics*, 60(2):427–435, 2004.

Berend, D. and Kontorovich, A. Consistency of weighted majority votes. In *NIPS*, pp. 3446–3454, 2014.

Bonald, T. and Combes, R. Crowdsourcing: Low complexity, minimax optimal algorithms. arXiv preprint arXiv 1606.00226, 2016.

Dalvi, N., Dasgupta, A., Kumar, R., and Rastogi, V. Aggregating crowdsourced binary ratings. In *WWW*, pp. 285–294, 2013.

Dawid, A. P. and Skene, A. M. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, 28(1):20–28, 1979.

Deng, J., Dong, W., Socher, R., Li, L.-j., Li, K., and Fei-fei, L. Imagenet: A large-scale hierarchical image database. In *In CVPR*, 2009.

Desai, M. and Rao, V. A characterization of the smallest eigenvalue of a graph. *J. Graph Theory*, 1994.

Gabriel, K. R. and Zamir, S. Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics*, 21(4):489–498, November 1979. ISSN 0040-1706. doi: 10.1080/00401706.1979.10489819.

Gao, C. and Zhou, D. Minimax optimal convergence rates for estimating ground truth from crowdsourced labels. arXiv preprint arXiv:1310.5764, 2013.

Gao, C., Lu, Y., and Zhou, D. Exact exponent in optimal rates for crowdsourcing. In *ICML*, pp. 603–611, 2016.

Ge, R., Lee, J. D., and Ma, T. Matrix completion has no spurious local minimum. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 2973–2981. Curran Associates, Inc., 2016.

Ghosh, A., Kale, S., and McAfee, P. Who moderates the moderators? crowdsourcing abuse detection in user-generated content. In *Proc. of the 12th ACM conference on Electronic commerce*, pp. 167–176, 2011.

Gillis, N. and Glineur, F. Low-rank matrix approximation with weights or missing data is NP-hard. *SIAM J. Matrix Analysis Applications*, 32(4):1149–1165, 2011.

Hui, S. L. and Walter, S. D. Estimating the error rates of diagnostic tests. *Biometrics*, pp. 167–171, 1980.

Karger, D., Oh, S., and Shah, D. Efficient crowdsourcing for multi-class labeling. In *SIGMETRICS*, pp. 81–92, 2013.

Karger, D., Oh, S., and Shah, D. Budget-optimal task allocation for reliable crowdsourcing systems. *Operations Research*, 62(1):1–24, 2014.

Karger, D. R., Oh, S., and Shah, D. Iterative learning for reliable crowdsourcing systems. In *NIPS*, pp. 1953–1961, 2011.

Koren, Y., Bell, R., and Volinsky, C. Matrix factorization techniques for recommender systems. *Computer*, 2009.

Li, H. and Yu, B. Error rate bounds and iterative weighted majority voting for crowdsourcing. 11 2014.

Liu, Q., Peng, J., and Ihler, A. T. Variational inference for crowdsourcing. In *NIPS*, pp. 692–700, 2012.

Nesterov, Y. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, 2004.

Nitzan, S. and Paroush, J. The characterization of decisive weighted majority rules. *Economics Letters*, 7(2):119–124, 1981.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322, 2010.

Saligrama, V. and Castanon, D. A. Reliable distributed estimation with intermittent communications. In *Proceedings of the 45th IEEE Conference on Decision and Control*, Dec 2006.

Shapley, L. and Grofman, B. Optimizing group judgmental accuracy in the presence of interdependencies. *Public Choice*, 43(3):329–343, January 1984. ISSN 0048-5829, 1573-7101. doi: 10.1007/BF00118940.

Smyth, P., Fayyad, U., Burl, M., Perona, P., and Baldi, P. Inferring ground truth from subjective labelling of venus images. In *NIPS*, volume 7, pp. 1085–1092, 1995.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pp. 254–263, Stroudsburg, PA, USA. Association for Computational Linguistics.

Srebro, N. and Jaakkola, T. Weighted low-rank approximations. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 720–727, 2003.

Szepesvári, D. A statistical analysis of the aggregation of crowdsourced labels. Master's thesis, University of Waterloo, 2015.

Tian, T. and Zhu, J. Max-margin majority voting for learning from crowds. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28*, pp. 1621–1629. Curran Associates, Inc., 2015.

Zhang, Y., Chen, X., Zhou, D., and Jordan, M. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. In *NIPS*, pp. 1260–1268, 2014.

Zhang, Y., Chen, X., Zhou, D., and Jordan, M. I. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. *The Journal of Machine Learning Research*, 17(1):3537–3580, 2016.

Zhou, D., Basu, S., Mao, Y., and Platt, J. Learning from the wisdom of crowds by minimax entropy. In *NIPS*, pp. 2195–2203, 2012.