myjournal manuscript No.
(will be inserted by the editor)

Distilling the Wisdom of Crowds: Weighted Aggregation of Decisions on Multiple Issues

Eyal Baharad[1] · **Jacob Goldberger**[2] · **Moshe Koppel**[3] · **Shmuel Nitzan**[4]

the date of receipt and acceptance should be inserted later

Abstract Given the judgments of multiple voters regarding some issue, it is generally assumed that the best way to arrive at some collective judgment is by following the majority. We consider here the now common case in which each voter expresses some (binary) judgment regarding each of a multiplicity of independent issues and assume that each voter has some fixed (unknown) probability of making a correct judgment for any given issue. We leverage the fact that multiple votes by each voter are known in order to demonstrate, both analytically and empirically, that a method based on maximum likelihood estimation is superior to the simple majority rule for arriving at true collective judgments.

Keywords judgment aggregation, dichotomous choice, expectation-maximization, Condorcet's Jury Theorem

Corresponding author:
Moshe Koppel
Department of Computer Sciences, Bar-Ilan University
Ramat-Gan 52900, Israel
email: koppel@cs.biu.ac.il
phone no. 972-3-5318075, fax no. 972-3-7360498

(1) Department of Economics, University of Haifa, Haifa 31905, Israel. E-mail: baharad@econ.haifa.ac.il, phone no. 972-4-8249585, fax no. 972-4-8240059 ·
(2) School of Engineering, Bar Ilan University, Ramat Gan 52900, Israel. E-mail: goldbej@eng.biu.ac.il, phone no. 972-3-5317053, fax no. 972-3-7384050 ·
(3) Department of Computer Sciences, Bar Ilan University, Ramat Gan 52900, Israel. E-mail: koppel@cs.biu.ac.il, phone no. 972-3-5318075, fax no. 972-3-7360498 ·
(4) Department of Economics, Bar Ilan University, Ramat Gan 52900, Israel. E-mail: nitzans@mail.biu.ac.il, phone no. 972-3-5318930, fax no. 972-3-7384034

1 Introduction

One of the key innovations spawned by the Internet is the compilation of judgments of non-experts regarding multiple questions of a particular type. Some web tools (for example, Amazon's Mechanical Turk [1] and social bookmarking tools such as Digg and de.licio.us) leverage collective judgment to obtain answers to simple questions (e.g., is this keyword an appropriate tag for this web page?) at low cost. Others attempt to tap into "the wisdom of crowds" to determine answers to questions with potentially significant political or commercial consequences, such as who will win an election or a sporting event or whether a given product idea is likely to succeed. In either case, the respective independent judgments of all the voters with regard to a given issue are then aggregated into a single collective judgment. Typically, this is done by straightforward averaging or, in the case of binary judgments, by following the simple majority rule (SMR). It has been often noted, however, that voter decisional skills are uneven; some voters offer judgments that are arbitrary, skewed or otherwise misguided. Thus, a fundamental question is how, in cases where we lack access to any ground truth against which to compare judgments, we can estimate each voter's decisional skills and accordingly reach a collective decision that is most likely to be correct.

Indeed where nothing at all is known about individual voter's relative decisional skills (except that, on average, they are better than random), we cannot improve upon SMR (Karotkin [8], Ben-Yashar and Paroush [1]). We will show, however, that in many common scenarios, although we have no direct information about individual voters' skills, we can still outperform SMR by exploiting the fact that we have each voter's judgments regarding each of a multiplicity of independent issues. That such tracking is possible is plainly true in the Internet setting mentioned above, but is also quite common in standing expert committees - such as courts, medical diagnostic bodies, investment committees, central bank committees - that periodically invoke voting to reach collective decisions.

Note that we consider here judgment scenarios where one of the alternatives is correct and the other incorrect. In principle, however, the method can be applied just as well to parliaments and other decision-making bodies where choices reflect preferences. In that context, we need to employ the fiction that there exists some "right" answer and that preferences reflect noisy judgments (Conitzer and Sandholm [3]).

Our basic insight is that, even without ever knowing who is right and who is wrong, voters whose judgments regarding many issues are different from those of other voters ought to be given less weight than other voters. While this simplistic insight is debatable, we will show in this paper that it can be leveraged into a highly accurate algorithm for vote aggregation whenever voter judgments across issues are available. The algorithm's objective is to find maximum likelihood values for the voters' competency. In particular, we will show how the output of this algorithm enables to aggregate votes in a manner that is vastly superior to SMR.

We note that the use of maximum likelihood estimators for vote aggregation has a rich history. Condorcet [4] already observed that in the case of dichotomous choice, if individual voter reliabilities are better than random, SMR yields a maximum likelihood estimator of the "correct" answer. Subsequently, such maximum likelihood estimators were computed explicitly given voter reliabilities (Nitzan and Paroush [11], Shapley and Grofman [12]). More recently Conitzer and Sandholm [3] showed that certain methods

[1] www.mturk.com/mturk

for vote (ranking) aggregation implicitly compute maximum likelihood estimators of the "correct" ranking using some implied underlying model of voter competency. Our problem differs from all this earlier work in that we wish to exploit multiple votes by the voters to find a maximum likelihood estimator of the voter reliabilities, which can then be used, as by Nitzan and Paroush [11] and Shapley and Grofman [12], to find a maximum likelihood estimator of the "correct" answer for each issue.

2 The Q Procedure

Let $N = \{1, ..., n\}$, $n \geq 3$, denote a finite set of voters and let M denote a set of m distinct binary issues, $m \geq 2$. The judgment of voter $i \in N$ on issue $j \in M$ is denoted by $a_{ij} \in \{0, 1\}$. The symbol a denotes the entire matrix of judgments (a_{ij}). The j column in the matrix a, denoted by a_j, is the judgment set on issue j. We assume that each issue has some (unknown) "correct" resolution, denoted by $t_j \in \{0, 1\}$, and that every voter i is associated with an unknown probability p_i of making the correct decision. The vector of individual probabilities $(p_1, ...p_n)$ is denoted by θ. For simplicity, we assume that, in the absence of any information, the two possible resolutions of an issue are equally likely; that is, for every j, the prior probability $p(t_j = 1) = 0.5$. (We will see below that relaxation of this assumption requires only minor adjustments to our basic algorithm.) We also assume that the issues are all independent of each other so that all outcomes over the set of issues are equiprobable. For this reason, we do not need to deal with questions of consistency that arise when issues are logically dependent (List and Pettit [10], Dokow and Holzman [7]).

A judgment aggregation rule V is a mapping from the judgments matrix $a = (a_{ij})$ to a set of binary decisions in $\{0, 1\}^m$. Our objective is to find an optimal judgment aggregation rule, given no information other than the matrix of judgments a. The suggested framework does not assume that the individual skills, $p_1, ..., p_n$, and the correct resolution for each issue are (ex-ante) known; hence, one might wonder in what sense a decision method could be optimal. In principle, given $\theta = (p_i)$ and assuming that, for all j, the prior probability $p(t_j = 1) = 0.5$, we can explicitly write the probability of the observed judgement matrix a:

$$p(a;\theta) = \prod_j p(a_j;\theta) = \prod_j (p(t_j = 0)p(a_j|t_j = 0) + p(t_j = 1)p(a_j|t_j = 1)) \quad (1)$$

$$= \prod_j (\frac{1}{2}\prod_i (a_{ij}p_i + (1 - a_{ij})(1 - p_i)) + \frac{1}{2}\prod_i (p_i(1 - a_{ij}) + (1 - p_i)a_{ij}))$$

Note that in this probabilistic modeling the individual skills $p_1, ..., p_n$ are viewed as (unknown) parameters and the "correct" resolutions $\{t_j\}$ are treated as hidden binary random variables. Thus, given some matrix of judgments a, optimality is obtained by the values of θ that maximize the probability of the observed data $p(a;\theta)$. That is, we wish to find a maximum likelihood estimator of θ:

$$\hat{\theta} = \arg \max_{\theta \in [0,1]^n} p(a;\theta) \quad (2)$$

The suggested iterative approach for finding this maximum is based on some initial estimate of θ. These values are re-used to compute, for each issue j, the probability that $t_j = 1$. Moreover, once all the conditional resolution probabilities $p(t_j = 1|a)$ are

given, one is able to compute, for each decision maker i, a more likely value of p_i, which we refer to as $p_i^{,}$. The iterative procedure is incomplete so long as $\theta \neq \theta^{,}$, i.e. $p_i \neq p_i^{,}$ for at least one decision maker; such $\theta = (p_i)$ is considered *inconsistent*. Inconsistency implies sub-optimality, in the sense that θ is not a maximal-likelihood estimate. The procedure is complete when $\theta = \theta^{,}$, i.e. for all i, $p_i = p_i^{,}$; at this stage a skill-evaluation equilibrium is obtained. Let us first show how $p(t_j = 1|a)$ is obtained given p_i (List [9], Nitzan and Paroush [11] and Grofman et al. [12]), and how an updated value of p_i is computed, given the probabilistic truth values $p(t_j = 1|a)$.

Lemma 1: Given the judgments $a = (a_{ij})$ and the probabilities $\theta = (p_1, ..., p_n)$, the conditional "correct" resolution probabilities are given by:

$$p(t_j = 1|a) = \frac{H}{H+1} \qquad \text{where} \qquad H = \prod_i (\frac{p_i}{1-p_i})^{(2a_{ij}-1)}$$

Proof: Recalling that the prior probability $p(t_j = 1) = 1/2$, Bayes' rule implies that:

$$p(t_j = 1|a) = p(t_j = 1|a_j) = \frac{p(a_j|t_j = 1)}{p(a_j|t_j = 0) + p(a_j|t_j = 1)} \tag{3}$$

where a_j is the judgment set on issue j. From our independence assumption, it follows that

$$p(a_j|t_j = 1) = \prod_i (a_{ij}p_i + (1 - a_{ij})(1 - p_i))$$

$$p(a_j|t_j = 0) = \prod_i (p_i(1 - a_{ij}) + (1 - p_i)a_{ij})$$

Observing that $H = \frac{p(a_j|t_j=1)}{p(a_j|t_j=0)}$, we obtain that $p(t_j = 1|a) = \frac{H}{1+H}$. Q.E.D

In the event that the prior probability $p(t_j = 1) = \alpha_j$ other than $1/2$, we only need to modify Eq. (3) as follows:

$$p(t_j = 1|a_j) = \frac{\alpha_j p(a_j|t_j = 1)}{(1 - \alpha_j)p(a_j|t_j = 0) + \alpha_j p(a_j|t_j = 1)} = \frac{\alpha_j H}{(1 - \alpha_j) + \alpha_j H} \tag{4}$$

Now assume that the correct resolutions $\{t_j\}$ are known. They can be compared to the judgments of individual i, in order to compute the maximum-likelihood values of p_i:

$$\hat{p}_i = \frac{1}{m}|\{j|a_{ij} = t_j\}| \tag{5}$$

Assume that we are given only stochastic information on the correct resolution set $\{t_j\}$. Denote the probability that $t_j = 1$ by w_{j1} and the probability that $t_j = 0$ by w_{j0}. Then we can still apply a modified version of Eq. (5):

$$p_i^{,} = \frac{1}{m}E(|\{j|a_{ij} = t_j\}|) = \frac{1}{m}\sum_{j=1}^{m} p(a_{ij} = t_j) = \frac{1}{m}\sum_j (a_{ij}w_{j1} + (1 - a_{ij})w_{j0}) \tag{6}$$

Finally, we define the hill-climbing procedure, Q, for approximating a skill-evaluation equilibrium by iterating the two steps described above:

1. Choose some initial θ.

2. For each $j = 1, ..., m$, compute $p(t_j = 1|a; \theta)$ (as computed in Lemma 1).
3. Let $w_{j0} = p(t_j = 0|a_j; \theta)$ and $w_{j1} = p(t_j = 1|a_j; \theta)$. Replace $\theta = (p_i)$ with the induced $\theta' = (p_i')$ using Eq. (6).
4. Repeat until convergence (termination).

3 Analytic Results

The procedure Q is a special case of the EM algorithm (Dempster et al.[5]), but with special properties not implied by the general theory. Hence, though our proofs follow the general outline of Dempster et al. [5], we will prove the below theorems from scratch. Our two main theorems are as follows:

Theorem 1: For any voting matrix $a \in \{0, 1\}^{nm}$ and any initial choice of $\theta \in [0, 1]^n$, procedure Q converges to a skill-evaluation equilibrium θ_* (i.e., applying Q procedure iteration on θ_* yields the same parameter vector θ_*).

Theorem 2: For any voting matrix $a \in \{0, 1\}^{nm}$ and almost any initial choice of $\theta \in [0, 1]^n$, procedure Q converges to a probability vector θ_* that is a local maximum of the probability function $p(a; \theta)$.

The proofs of the two theorems will invoke the following three lemmas:

Lemma A1: Let the cross-entropy between two binary distributions $p = (p_0, p_1)$ and $q = (q_0, q_1)$ be $CE(p, q) = -p_0 \log q_0 - p_1 \log q_1$. Every pair of binary distributions p and q satisfies $CE(p, p) \leq CE(p, q)$ and there is equality if and only if $p = q$.

Proof of Lemma A1: $CE(p, p) - CE(p, q) = p_0 \log \frac{q_0}{p_0} + p_1 \log \frac{q_1}{p_1} \leq \log(p_0 \frac{q_0}{p_0} + p_1 \frac{q_1}{p_1}) = \log(1) = 0$. The inequality is obtained because log is a concave function. The log function is, in fact, strictly concave (the second derivative is always negative), hence there is equality if and only if $p = q$. Q.E.D.

Lemma A2: The probability function $p(a_j; \theta)$ (see Eq. (1)) satisfies:

$$\log p(a_j; \theta) = L_j(\theta, \theta_0) + CE(p(t|a_j; \theta_0), p(t|a_j; \theta)) \tag{7}$$

where

$$L_j(\theta, \theta_0) = \sum_{t=0,1} p(t|a_j; \theta_0) \log p(t, a_j; \theta)$$

and θ_0 can be any other possible parameter value.

Proof of Lemma A2: Taking the log of $p(t_j, a_j; \theta) = p(a_j; \theta)p(t_j|a_j; \theta)$ we obtain:

$$\log p(a_j; \theta) = \log p(t_j, a_j; \theta) - \log p(t_j|a_j; \theta) \tag{8}$$

Multiplying each term in (8) by $p(t_j|a_j; \theta_0)$ and summing over $t_j = 0, 1$, we obtain:

$$\begin{aligned} \log p(a_j; \theta) &= \sum_t p(t|a_j; \theta_0) \log p(t, a_j; \theta) - \sum_t p(t|a_j; \theta_0) \log p(t|a_j; \theta) \qquad (9) \\ &= L_j(\theta, \theta_0) + CE(p(t|a_j; \theta_0), p(t|a_j; \theta)) \end{aligned}$$

Q.E.D.

Lemma A3: The Q procedure satisfies $p(a;\theta_{t+1}) \geq p(a;\theta_t)$ where θ_t is the value of θ obtained in the t-th iteration of the Q procedure.
Proof of Lemma A3: Using the notation $L(\theta,\theta_0) = \sum_j L_j(\theta,\theta_0)$, Lemma 2 implies that for each parameter value θ:

$$\log p(a;\theta) = L(\theta,\theta_t) + \sum_j CE(p(t|a_j;\theta_t), p(t|a_j;\theta)) \tag{10}$$

$$\log p(a;\theta_t) = L(\theta_t,\theta_t) + \sum_j CE(p(t|a_j;\theta_t), p(t|a_j;\theta_t)) \tag{11}$$

Lemma A1 implies that for every issue j

$$CE(p(t|a_j;\theta_t), p(t|a_j;\theta)) \geq CE(p(t|a_j;\theta_t), p(t|a_j;\theta_t))$$

Subtracting Eq. (11) from Eq. (10) we obtain:

$$\log p(a;\theta) - \log p(a;\theta_t) \geq L(\theta,\theta_t) - L(\theta_t,\theta_t)$$

Hence, to prove that $p(a;\theta_{t+1}) \geq p(a;\theta_t)$, it is enough to show that $L(\theta_{t+1},\theta_t) \geq L(\theta_t,\theta_t)$. We shall show that $\theta_{t+1} = \arg\max_\theta L(\theta,\theta_t)$. Denote $w_{j0} = p(t=0|a_j;\theta_t)$ and $w_{j1} = p(t=0|a_j;\theta_t)$.

$$L(\theta,\theta_t) = \sum_j w_{j1} \sum_i (a_{ij}\log p_i + (1-a_{ij})\log(1-p_i)) + \tag{12}$$
$$\sum_j w_{j0} \sum_i ((1-a_{ij})\log p_i + a_{ij}\log(1-p_i))$$
$$= \sum_i \log p_i (\sum_j (a_{ij}w_{j1} + (1-a_{ij})w_{j0})) + \log(1-p_i)(\sum_j (a_{ij}w_{j0} + (1-a_{ij})w_{j1}))$$

Using the notation $p_i^{\cdot} = \frac{1}{m}\sum_j (a_{ij}w_{j1} + (1-a_{ij})w_{j0})$, we obtain:

$$L(\theta,\theta_t) = -m \sum_i CE((p_i^{\cdot}, 1-p_i^{\cdot}), (p_i, 1-p_i)) \tag{13}$$

Eq. (13) implies that we can maximize $L(\theta,\theta_t)$ separably for each p_i. Lemma A1 implies that for every voter i

$$p_i^{\cdot} = \arg \max_{p_i \in [0,1]} -CE((p_i^{\cdot}, 1-p_i^{\cdot}), (p_i, 1-p_i)) \tag{14}$$

and this maximum is unique. Therefore, the unique maximum of $L(\theta,\theta_t)$ is obtained at the parameter vector $\theta_{t+1} = (p_i^{\cdot})$ obtained by applying a single iteration of the Q procedure on θ_t. Hence, $\theta_{t+1} = \arg\max L(\theta,\theta_t)$ and therefore $p(a;\theta_{t+1}) \geq p(a;\theta_t)$. Q.E.D

Proof of Theorem 1: Lemma A3 asserts that the sequence $\{p(a;\theta_t)\}$ is monotonically increasing. It is also a bounded sequence since $\{p(a;\theta)\}$ viewed as a function of θ is a polynomial function defined on the compact n-dimensional set $[0,1]^n$. Therefore $p(a;\theta_t)$ converges to a limit point $p(a;\theta_*)$. It remains only to show that the Q procedure does not oscillate between points with the same likelihood. Denote the result of

applying the Q procedure on θ_* by θ_{**}. We need to show that $\theta_* = \theta_{**}$. Since $p(a, \theta_*)$ is a limit point, it follows that $p(a; \theta_*) = p(a; \theta_{**})$. From Lemma A3 we have that, by definition of the Q procedure, $\theta_{**} = \arg\max L(\theta, \theta_*)$. It follows from the proof of Lemma 3 that $\max_\theta L(\theta, \theta_*)$ is obtained at a single point. Hence if $\theta_* \neq \theta_{**}$ then $L(\theta_{**}, \theta_*) > L(\theta_*, \theta_*)$ and therefore $p(a; \theta_{**}) > p(a; \theta)$. This, however, contradicts our assumption that $p(a, \theta_*)$ is a limit point and therefore $p(a; \theta_*) = p(a; \theta_{**})$. Hence θ_* is the unique maximum point of $L(\theta, \theta_*)$, i.e $\theta_{**} = \theta_*$. Q.E.D.

Proof of Theorem 2: Let θ_* be an equilibrium point of the Q procedure. By Lemma A2 we obtain that for every θ

$$\log p(a; \theta) = L(\theta, \theta_*) + \sum_j CE(p(t|a_j; \theta_*)), p(t|a_j; \theta)) \tag{15}$$

Hence,

$$\frac{d}{d\theta} \log p(a; \theta) = \frac{d}{d\theta} L(\theta, \theta_*) + \frac{d}{d\theta} \sum_j CE(p(t|a_j; \theta_*)), p(t|a_j; \theta)) \tag{16}$$

Note that as long as θ_* falls in the interior of the parameter space $[0, 1]^n$ the functions above are all differentiable. By Theorem 1, θ_* is a fixed point of the Q procedure, i.e., $\theta_* = \arg\max_\theta L(\theta, \theta_*)$. Hence

$$\frac{d}{d\theta} L(\theta, \theta_*)|_{\theta=\theta_*} = 0 \tag{17}$$

Lemma A1 implies that $\theta_* = \arg\min_\theta \sum_j CE(p(t|a_j; \theta_*), p(t|a_j; \theta))$. Thus

$$\frac{d}{d\theta} (\sum_j CE(p(t|a_j; \theta_*), p(t|a_j; \theta)))|_{\theta=\theta_*} = 0 \tag{18}$$

By substituting (15) and (16) into (18) we obtain:

$$\frac{d}{d\theta} \log p(a; \theta_*) = 0$$

Hence the Q-procedure converges to a stationary point of $p(a; \theta)$. A stationary point is a local maximum, a local minimum or a saddle point. In fact, though, it can be shown that, since local minima and saddle points are not attractors, the probability of selecting an initial point that leads to convergence to a local minimum or to a saddle point is 0. Q.E.D.

To conclude this section, we note that in our setting two properties hold which are not implied by the general EM theory. Let θ_t be the value of θ obtained in the t-th iteration of the Q procedure. The EM theory [5] guarantees that the likelihood sequence $p(a; \theta_t)$ is monotonically increasing and therefore, if it is bounded, it necessarily converges to a limit number. This is exactly what is shown in Lemma A3. In our case, however, there are two additional properties that do not necessary hold in the general EM theory. First, in the general EM theory, the fact that the sequence $\{p(\theta_t)\}$ converges does not guarantee that the parameter sequence $\{\theta_t\}$ converges to a limit point in the parameter domain. It can happen that the sequence $\{\theta_t\}$ oscillates between different points whose likelihoods are equal. Second, although EM converges

to a limit point of the likelihood function $p(a, \theta)$, there is no theoretical guarantee that this limit point is indeed a local maximum (see, e.g., Wu [13]). The convergence of the sequence θ_t (proved in Theorem 1) follows in our case from the fact that the parameter re-estimation step (corresponding to the M-step in EM) in each iteration is applied to the cross-entropy, which is strictly concave (as a function of its second argument) and thus has a unique maximum point. The proof (in Theorem 2) that this limit point is a local maximum follows from the same fact. In the general EM theory, there is no guarantee of convexity on the function maximized in the M-step.

4 Empirical Results

While our analytic result only shows that Q finds a local maximum, we find empirically that in fact Q does considerably better than that. On simulated examples, where we know the true θ used to generate the voting data, Q nearly always converges to some θ_* that is at least as good as the true θ, in the sense that $p(a; \theta_*) \geq p(a; \theta)$. This strongly suggests that Q generally converges to some value close to the global maximum of $p(a; \theta)$. Thus, our main claim for practical purposes is that given a matrix of votes, the proper method of reaching a collective decision regarding each issue j is to assign the consensus judgment 1 if Q converges to a value of $p(t_j = 1|a)$ greater than (or equal to) 1/2 and 0 otherwise. (Of course, in principle, if in some context we regard Type 1 errors with different severity than Type 2 errors, we can use a threshold other than 1/2 (Nitzan and Paroush [11] and Dietrich [6]).) We will see that this method is vastly superior to SMR precisely in the sense that it is more likely to arrive at the correct answer for any given issue.

4.1 The Simulation Procedure

We use the following simulation procedure. Choose values of n and m, representing the number of voters and the number of issues, respectively. Assign some random binary correct answer t_j to each issue j and some random reliability level p_i to each voter i (subject to a single condition, as will be described below). For each $1 \leq i \leq n$ and each $1 \leq j \leq m$, generate $a_{ij} \in \{0, 1\}$, the vote of voter i on issue j, by tossing a coin with probability p_i of yielding t_j. The task is to ascertain t_j for each issue j, given only the voting matrix a.

We do need to make one assumption about the values of $\theta = (p_i)$. Note that for every vector $\theta_* = (p_i)$, there is a dual vector, $\bar{\theta}_* = (1 - p_i)$, such that $p(a; \theta) = p(a; \bar{\theta})$. Thus, for every "sensible" solution, θ, there is a counter-intuitive one, $\bar{\theta}$. In order to distinguish between them, we note that for at most one of them, $\prod p_i > \prod(1 - p_i)$, namely, that if voters are unanimous on some issue, their vote is correct with probability greater than 1/2. Thus, to break the symmetry between the sensible solution and its dual, we assume that $\prod p_i > \prod(1 - p_i)$.

4.2 A Simple Example

Now, to illustrate our simulation procedure, we first run through a single toy example with $n = 5$ and $m = 10$. We arbitrarily assign the correct answer 0 to the first

(p_i)						Votes									
.99	.94	.86	.82	.79	.9	0	0	0	1	1	1	1	1	1	1
.50	.53	.57	.63	.68	.9	1	0	0	0	1	0	0	0	1	1
.99	.94	.86	.82	.79	.9	0	0	0	1	1	1	1	1	1	1
.50	.55	.63	.67	.70	.9	0	0	1	1	0	0	0	0	1	1
.70	.75	.82	.86	.87	.9	0	0	1	1	1	0	0	1	1	1
						$\{p(t_j = 1\|a)\}$									
						.01	.01	.10	.99	.99	.10	.10	.90	.99	.99
						.01	.01	.34	.99	.99	.31	.31	.96	.99	.99
						.01	.01	.29	.99	.99	.50	.50	.98	.99	.99
						.01	.01	.13	.99	.99	.80	.80	.99	.99	.99
						.01	.01	.01	.99	.99	.98	.98	.99	.99	.99
						.01	.01	.01	.99	.99	.99	.99	.99	.99	.99

Fig. 1 A matrix of votes by five voters over ten issues. Values of (p_i) at respective iterations are shown in left columns progressing from right to left. Corresponding values of $p(t_j = 1|a)$ are shown at bottom progressing from top to bottom.

three issues and 1 to the other seven issues and assign the respective reliability values $\{0.82, 0.61, 0.83, 0.60, 0.76\}$ to the five voters. Using coin-tossing as described, we obtain the matrix shown in Figure 1. Note that SMR would return incorrect answers for issues 6 and 7. We now apply Q to the matrix in the hope of reconstructing the correct answers (which are, of course, unknown to us).

As in all our simulations below, we initialize all p_i to 0.9. (Other initial values return essentially the same overall results.) We terminate when $p(a; \theta_{t+1}) - p(a; \theta_t) < 0.0001$.

The columns to the left of the matrix represent reliability levels for the respective voters during successive iterations of the algorithm and the rows below the matrix represent probabilities that the correct answer for the respective issues is 1 during successive iterations of the algorithm. The algorithm converges after six iterations and, unlike SMR, reaches the correct answer for every issue. Moreover, $p(a, \theta)$ increases after each iteration and the value to which it converges ($\exp(-20.2)$) is in fact far greater than that obtained from the actual reliability values used to generate the votes ($\exp(-29.0)$).

4.3 Simulation Results

We now systematically compare the performance of our algorithm Q on this task with that of the standard algorithm, namely, SMR. For our first experiment, we fix the number of voters at $n = 50$ and let the number of issues m vary. For each value of m, we run 10,000 trials, each representing a new randomized choice of (p_i) and $\{t_j\}$. The values of (p_i) are sampled uniformly in the range $[0, 1]$, subject to the single requirement that $\prod p_i > \prod(1 - p_i)$. For each trial, the accuracy of the algorithm is the proportion of j for which the algorithm correctly ascertains the value of t_j. Results are shown in Figure 2. As can be seen, SMR remains steady at accuracy of 0.7 and does not improve as the number of issues increases since it does not learn from one issue to the other. On the other hand, Q increases steadily to near-perfect accuracy as the number of issues

increases, since the greater the number of issues the more accurately we can estimate voter reliability.

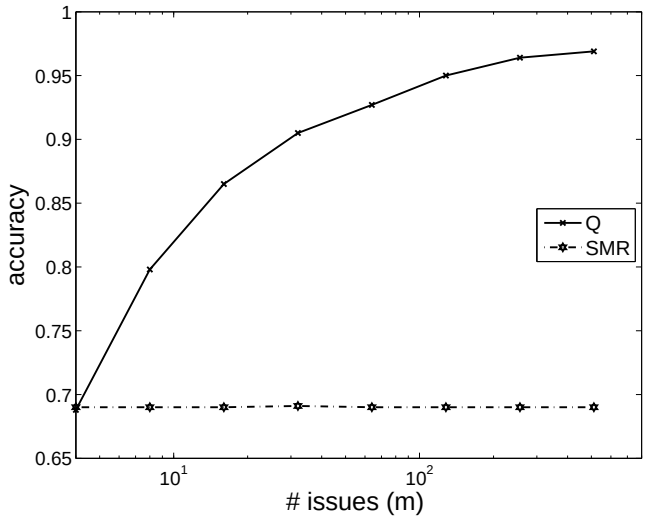


Fig. 2 Accuracy of SMR and Q for 50 voters and increasing number of issues. Each datapoint represents 10,000 trials in each of which values of p_i are sampled uniformly in the range [0,1] subject to the unanimity condition.

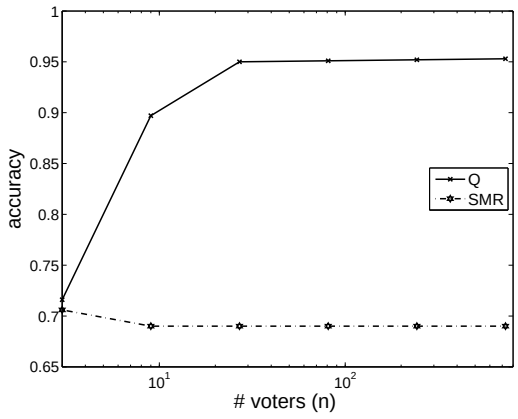
Our second experiment operates according to the same principles as the first, except the number of issues is held fixed at $m = 100$ and the number of voters n varies. Results are shown in Figure 3a. Remarkably, we find the same phenomenon in this experiment as in the previous one. SMR remains steady at accuracy of 0.7 and does not improve as the number of voters increases. On the other hand, Q increases steadily to accuracy of 0.95 as the number of voters increases.

Note that when voters' collective decision skills are sufficient, Condorcet's Jury Theorem [4] ensures convergence of SMR to perfect decision-making as the number of voters grows. However, in this experiment, our assumptions regarding voters' collective decision skills are too weak for that theorem to hold (Berend and Paroush [2]). We thus reran our second experiment, this time sampling values of (p_i) uniformly in the range [0.1,1], so that Condorcet's Jury theorem would hold. As can be seen in Figure 3b, even in this case, as the number of voters grows, the accuracy of Q converges to 1 much faster than that of SMR. In Figure 3c, we show results for the same experiment with values of (p_i) sampled uniformly in the range [0.5,1]. As is evident, in such cases SMR converges rapidly and the advantage of Q is diminished.

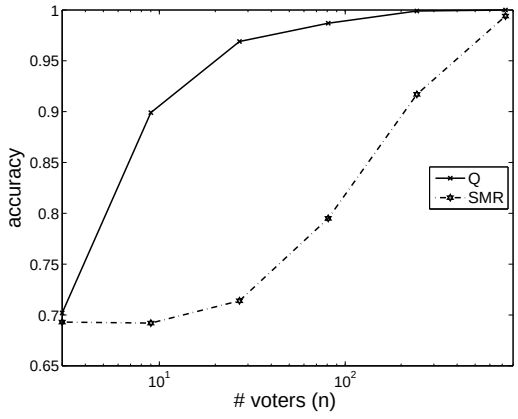
5 Conclusions

To summarize, we have considered scenarios in which the following conditions hold:

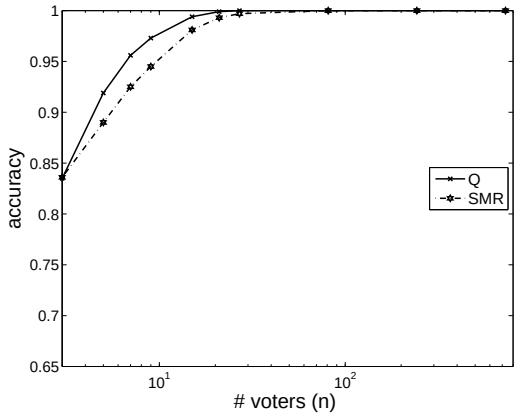
1. Voter records are available over a variety of issues.



(a)



(b)



(c)

Fig. 3 Accuracy of SMR and Q for 100 issues and increasing number of voters. Each datapoint represents 10,000 trials in each of which values of p_i are sampled uniformly in the range: (a) [0,1], (b) [0.1,1], (c) [0.5,1].

2. For each voter, there is some fixed (unknown) probability that that voter gives the correct answer for any given issue.
3. Voters' judgments are independent of each other.
4. Voters are collectively at least minimally competent in the sense that a unanimous vote is more likely to be right than wrong.

To be sure, some of these assumptions are quite restrictive, particularly the independence assumptions. Moreover, our method applies principally to voting with regard to the truth of propositions rather than voting with regard to personal preferences, and assumes that voter's competency is the same for all propositions. Nevertheless, such scenarios are quite common, especially in the context of standing expert committees and online judgment aggregation sites.

We have found that in such scenarios a judgment aggregation method in the expectation-maximization framework is far more likely to reach correct answers than the standard SMR. This result holds regardless of the number of voters or the number of issues each has voted on, but the advantage of the proposed method over SMR is especially pronounced when the track records of individual voters are sufficiently abundant and when voters' decisional skills are sufficiently heterogeneous.

Our approach is trivially generalizable to cases in which voting records are incomplete so that we have the votes of each voter on some subset of all issues. A common instance of this scenario is refereeing of conference papers; Q could be used to aggregate judgments of referees in a manner that optimally discounts referees whose judgments of a variety of papers suggest idiosyncratic views. Moreover, as our preliminary investigations indicate, the method is easily generalized to cases in which votes are not binary, but rather real numbers in the range $[0, 1]$. In such cases, we assume a different generative model, for example, that each voter's noise is (truncated) normally distributed with fixed bias and variance.

References

1. R. Ben-Yashar and J. Paroush. A non-asymptotic Condorcet jury theorem. *Social Choice and Welfare*, 17:189–99, 2000.
2. D. Berend and J. Paroush. When is Condorcet's jury theorem valid. *Social Choice and Welfare*, 15:481–88, 1998.
3. V. Conitzer and T. Sandholm. Common voting protocols as maximum likelihood estimators. *Proc. of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2005.
4. N.C. de Condorcet. Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. pages 27–32, 1785.
5. A. Dempster, N. Laird, and D. Rubin D. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, pages 1–38, 1977.
6. F. Dietrich. General representation of epistemically optimal procedures. *Social Choice and Welfare*, 26:263–283, 2006.
7. E. Dokow and R. Holzman. Aggregation of binary evaluations for truth-functional agendas. *Social Choice and Welfare*, 32:221–241, 2009.
8. D. Karotkin. Justification of the simple majority and chairman rules. *Social Choice and Welfare*, 13:479–86, 1996.
9. C. List. On the significance of the absolute margin. *The British Journal for the Philosophy of Science*, 55(3):521–544, 2004.
10. C. List and P. Pettit. Aggregating sets of judgments: An impossibility result. *Economics and Philosophy*, 18:89–110, 2002.
11. S. Nitzan and J. Paroush. A general theorem and eight corollaries in search of correct decision. *Theory and Decision*, 17:211–220, 1984.
12. L.S. Shapley and B. Grofman. Optimizing group judgmental accuracy in the presence of interdependencies. *Public Choice*, 43:329–343, 1984.

13. C. F. J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11(1):95–103, 1983.