Expert Voting and Error Correction

Kumail Jaffer, Sidhanth Mohanty, Phillip Wang

December 27, 2017

We would like to thank **Prof. Ariel Procaccia** for advising us on this project.

Abstract

We consider the model where there exists a ground truth, and the average distance, according to some metric, of votes from the ground truth is upper bounded. Under this setting, we derive a way to aggregate votes to return an alternative that is ranked fairly high in the ground truth in expectation: a rank that depends on the distribution of noise among the votes, with the worst case being achieved when all the votes have the same amount of uniform noise. Further, we explore how the model changes, when we introduce a bounded number of queries that allows the algorithm to learn the true rank of a picked alternative in a single query.

1 Introduction and related work

In the standard setting for voting theory problems, we are given a set of voters and candidates, and for each voter an ordering on the candidates representing their reported preferences. We are tasked with picking a candidate that reflects in some way the will of the group, with possible additional requirements on the selection mechanism so that it handles, for example, strategic voters. Notice that we do not try to find truth in such a situation. We simply try to find a candidate that satisfies everyone as much as possible.

Instead of finding a selection algorithm that is just to the voters, as in the above scenario, we find an algorithm that does justice to the truth. That is, we are interested in the setting where there is a ground truth, and every voter's preference list is some approximation of that. Concretely, this models situations such as that of a group of expert rocket scientists trying to reach consensus on the best design for a rocket. There is an objective "best" design, but each expert may have a subtly flawed perspective. Together, however, they may be able to discover the ground truth. This setting has been studied by Procaccia et. al, who found fairly tight bounds on how closely the ground truth can be approximated (in terms of some metric) if we know a bound on how well the experts approximate the ground truth [PSZ15].

We consider the related problem of picking a candidate that has some guarantee of being close to the best, in terms of the number of candidates, and the error of the experts. Formally, the problem looks like ths: We are a given a set of candidates N with $|N| = n$, a set of voters V, and for each $v \in V$ a ranking over N, $r_v \in S_N$ (the symmetric group on N). We also know that there is some true ranking r^*, part of which we are trying to recover. Further, we know that for some metric d and constant t, $\frac{1}{|V|} \sum_{v \in V} d(r_v, r^*) \leq t$. That is, the average distance of the expert rankings is bounded by t. The metric we consider is the Footrule distance, which is the sum of absolute values of difference in positions between profiles taken over all players.

Our algorithm is successful in finding a candidate $c \in N$ such that c is among the top k candidates in r^* in expectation, and k is always in $O(\sqrt{t})$.

Other related works include [PSZ15], which finds a ranking close to the ground truth. In [CPS13], the setting of votes being an approximation of the ground truth was studied. It differs from this work in that it assumes probabilistic noise, and looks at voting rules as Maximum Likelihood Estimators. [BM10] studies the setting where messages are permutations that have some added noise when sent, and one wants to recover a permutation close to the original.

2 Algorithms and Upper Bounds

2.1 Background and Motivation

Say there are m voters, whose preferences are the rankings $r_1, r_2, \ldots, r_m$, with average distance t from the true ranking r^*. This tells us that

$$\sum_{i=1}^{m} d(r_i, r^*) \leq mt$$

Now, for some alternative j, denote by q_{ij} its position in the ranking given by the ith voter and denote by r_j^* it's true rank according to r^*.

If we go by the Footrule distance metric, the total noise is

$$\sum_{j=1}^{n} \sum_{i=1}^{m} |r_j^* - q_{ij}| \leq mt$$

We know the value of m and t and hence know an upper bound on the total noise. Knowing how the noise is distributed along with information about the structure of the noise can help our algorithm pick a good alternative. Thus, as a starting point, we try to lower bound the total noise in terms of parameters we know. Fix an alternative j and consider the median value of the set $\{q_{ij} : i \in [m]\}$, denoted s_j. By the definition of the median, the expression $\sum_{i=1}^{m} |x - q_{ij}|$ is minimized when $x = s_j$.

Therefore, we know that the following chain of inequalities is true:

$$\sum_{j=1}^{n} \sum_{i=1}^{m} |s_j - q_{ij}| \leq \sum_{j=1}^{n} \sum_{i=1}^{m} |r_j^* - q_{ij}| \leq mt$$

The reason this inequality is useful is that each s_j can be computed by us, which leads to the following insight: If we consider a subset of alternatives, S, let $F(S)$ be the total contribution of the alternatives in S to the total noise, written as

$$F(S) = \sum_{j \in S} \sum_{i=1}^{m} |r_j^* - q_{ij}| \geq \sum_{j \in S} \sum_{i=1}^{m} |s_j - q_{ij}|$$

In particular, for any S, we know that the total noise is equal to $F(S) + F(S^c)$, which leads to another chain of inequalities:

$$mt \geq F(S) + F(S^c) \geq \sum_{j \in S} \sum_{i=1}^{m} |s_j - q_{ij}| + F(S^c)$$

And it follows that:

$$F(S^c) \leq mt - \sum_{j \in S} \sum_{i=1}^{m} |s_j - q_{ij}|$$

Further, the above inequality provides a bound on the average noise in S^c.

$$\text{Average noise in } S^c = \frac{F(S^c)}{m} \leq t - \frac{\sum_{j \in S} \sum_{i=1}^{m} |s_j - q_{ij}|}{m}$$

The takeaway from the above inequality is that if we can find a set S of lower ranked alternatives with a high amount of noise, we could restrict our attention to S^c, a set of mostly high ranked players with noise less than t.

Before we describe our algorithm, we consider an algorithm in section 3 of [PSZ15] to obtain a ranking σ satisfying

1. The Footrule distance between σ and the ground truth is $\leq 2t$.
2. The average Footrule distance between the preference profiles of the voters and σ is $\leq t$.

The algorithm isn't specific for the Footrule distance but works for any metric.
Algorithm 1. Described in detail in [PSZ15]. At a high level, this algorithm considers the set of all permutations that have average distance at most t from the votes – this set is nonempty because the ground truth is in this set – and then picks a permutation that minimizes the maximum distance to this set.

2.2 The algorithm and it's analysis

The background we developed in the last section motivates the following algorithm:
Algorithm 2. Compute the corresponding median s_j for each candidate and do the following.

1. Use Algorithm 1 to find a ranking σ, then sort the alternatives according to σ and place them in a sorted sequence L.

2. Calculate the lower bound on the noise of each suffix of L. That is, for suffix S, compute

$$\gamma_P = \sum_{j \in S} \sum_{i=1}^{m} |s_j - q_{ij}|$$

   And also calculate an upper bound on the the noise of the corresponding prefix P with the expression

$$\alpha_P = mt - \gamma_P$$

   Note that γ_P also lower bounds $\sum_{j \in S} \sum_{i=1}^{m} |L_j - q_{ij}|$ where L_j is the rank of person j according to L.

3. Pick the shortest prefix P^* for which the following inequality is satisfied:

$$|P^*| \geq \sqrt{\frac{2\alpha_{P^*}}{m}}$$

4. Pick a candidate from P^* uniformly at random.

The following lemma motivates step 3 of the algorithm and will also be useful in the lower bound analysis.
Lemma 2.1. *The Maximum Footrule distance of a permutation of length n from the identity permutation is* $\lceil \frac{n^2}{2} \rceil$.

Proof. The main idea of this proof is from [SFD].

The Footrule distance of a permutation σ is given by

$$\sum_{i=1}^{n} |\sigma(i) - i|$$

Let S be the set on which $\sigma_i > i$. Then the above sum can be written as

$$\sum_{i \in S} (\sigma(i) - i) + \sum_{i \in [n] \setminus S} (i - \sigma(i))$$

Since there are exactly n positive terms and n negative terms in the sum and each integer between 1 and n occurs twice, the above quantity is maximized when the positive terms are all occurences of numbers $\lfloor \frac{n}{2} \rfloor + 1, \ldots, n$ and negative numbers are all occurences of $1, \ldots, \lfloor \frac{n}{2} \rfloor$, yielding a result of $\lceil \frac{n^2}{2} \rceil$.

□

Since every permutation of numbers $1, 2, \ldots, l$ has Footrule distance bounded by $\lceil \frac{l^2}{2} \rceil$, having $\lceil \frac{l^2}{2} \rceil$ amount of noise in l elements admits the players to rank any permutation of these l elements, which is the rationale for the cutoff defined in step 3. We pick the shortest such prefix because it narrows down the set with high ranked players. Also note that it is possible to perform step 3 since the full sequence is a prefix that satisfies the required inequality.

Theorem 2.2. *Algorithm 2 finds a candidate with expected rank* $3\sqrt{\frac{\alpha_{P^*}}{m}}$*.*

Proof. For simplicity of exposition, let $\beta = \sqrt{\frac{2\alpha_{P^*}}{m}}$. First, we argue that $|P^*| = \lceil \beta \rceil$.

Let β' be the corresponding value for the prefix of length $|P^*| - 1$. By choice of P^*, it is true that $|P^*| - 1 < \beta'$ and $\beta \leq |P^*|$. Since the noise can only decrease on shortening the prefix, $\beta' \leq \beta$. It follows that $|P^*| - 1 < \beta \leq |P^*|$, which establishes our claim.

Thus, we have a prefix of $\lceil \beta \rceil$ alternatives whose total noise is bounded by α_{P^*}. Note that α_{P^*} is an upper bound to both

$$mt - \sum_{j \in [n] \setminus P^*} \sum_{i=1}^{m} |r_j^* - q_{ij}| \geq \sum_{j \in P^*} \sum_{i=1}^{m} |r_j^* - q_{ij}|$$

and

$$mt - \sum_{j \in [n] \setminus P^*} \sum_{i=1}^{m} |q_{ij} - L_j| \geq \sum_{j \in P^*} \sum_{i=1}^{m} |L_j - q_{ij}|$$

It follows from the triangle inequality that

$$2\alpha_{P^*} \geq \sum_{j \in P^*} \sum_{i=1}^{m} |r_j^* - q_{ij}| + |L_j - q_{ij}| \geq \sum_{i=1}^{m} \sum_{j \in P^*} |r_j^* - L_j|$$

$$\frac{2\alpha_{P^*}}{m} \geq \sum_{j \in P^*} |r_j^* - L_j|$$

In other words, we have just shown that the Footrule distance restricted to elements of P^* between the ranking returned by the [PSZ15] algorithm and the true ranking is upper bounded by $\frac{2\alpha_{P^*}}{m}$.

Denote by $T(i)$ the true rank of the alternative assigned in position i in L. Denote by $a(i)$ the value of $|i - T(i)|$, how much L differs on the alternative it ranks i from its true rank. Picking a uniformly random alternative gives the expected rank as

$$\sum_{i=1}^{\lceil \beta \rceil} \frac{T(i)}{\lceil \beta \rceil} \leq \sum_{i=1}^{\lceil \beta \rceil} \frac{i + a(i)}{\lceil \beta \rceil} = \sum_{i=1}^{\lceil \beta \rceil} \frac{i}{\lceil \beta \rceil} + \sum_{i=1}^{\lceil \beta \rceil} \frac{a(i)}{\lceil \beta \rceil}$$

$$\leq \frac{\lceil \beta \rceil + 1}{2} + \frac{2\alpha_{P^*}}{m \lceil \beta \rceil} \leq \sqrt{\frac{\alpha_{P^*}}{2m}} + \sqrt{\frac{2\alpha_{P^*}}{m}} \leq 3\sqrt{\frac{\alpha_{P^*}}{m}}$$

□

Corollary 2.3. *For any set of input votes, algorithm 2 finds a candidate with expected rank* $\leq 3\sqrt{t}$*.*

Proof. Since $\frac{\alpha_{P^*}}{m} \leq t$, the result follows from theorem 2.2. □

3 Lower bounds and Tightness

First, note that our algorithm is asymptotically optimal in the worst case.

Theorem 3.1. *For any randomized algorithm, there exists a profile of voters, p, such that the algorithm has guarantee $\Omega(\sqrt{t})$ in expectation on p, where t is the bound on the average footrule distance of the voters to the ground truth.*

Proof. Take any profile with a single voter. Let his ranking be r. Let r' be a ranking with the first $\lfloor\sqrt{2t}\rfloor$ ranked candidates reversed. By lemma 2.1, this ranking is also a possible ground truth for the profile. Let $p_1 \dots p_n$ be the proabilities assigned to candidates $1 \dots n$ of the original ranking by a randomized algorithm. Let R be the rank of the candidate output by the algorithm if the ground truth were r, and R' be the rank if the ground truth were r'. We have (considering just the first $\lfloor\sqrt{2t}\rfloor$) the following:

$$\begin{aligned}
\mathbf{E}[R] + \mathbf{E}[R'] &\geq \sum_{i=1}^{\lfloor\sqrt{2t}\rfloor} p_i i + \sum_{i=1}^{\lfloor\sqrt{2t}\rfloor} p_i(\lfloor\sqrt{2t}\rfloor - i) \\
&\geq \sum_{i=1}^{\lfloor\sqrt{2t}\rfloor} p_i(\lfloor\sqrt{2t}\rfloor) = \lfloor\sqrt{2t}\rfloor \\
\max(\mathbf{E}[R], \mathbf{E}[R']) &\geq \frac{\lfloor\sqrt{2t}\rfloor}{2}
\end{aligned}$$

So for any randomized algorithm, on one of these situations, it has at best a $O(\sqrt{t})$ guarantee in expectation. □

This is not entirely satisfying, however, since there may be profiles on which significantly better guarantees are possible. So next we'll try to develop a lower bound that applies generally to any profile. Note that the bound we developed for our algorithm had a guarantee incorporating characteristics of the noise of the profile. We'd like to be able do something similar on the lower bound side.

Theorem 3.2. *Let q be, as before, a vector encoding the rankings of each of the voters, let n be the number of candidates, and m the number of voters. Then for any ranking σ consistent with the voters, and any suffix S of the voters ordered by σ and corresponding prefix P_S such that*

$$|P_S| \leq \frac{t - \frac{1}{m}\sum_{j\in S}\sum_{i=1}^{m} |\sigma(j) - q_{ij}|}{n}$$

no randomized algorithm can do better than $\frac{\sqrt{2|P_S|}}{2}$ in expectation.

Proof. Let $\gamma_S = \frac{1}{m}\sum_{j\in S}\sum_{i=1}^{m} |\sigma(j) - q_{ij}|$. This is the average footrule distance between the ranking σ and the voters on S. If we create a new profile σ' whose average footrule distance from the voters on the corresponding prefix P_S is up to $t - \gamma_S$, σ' will also be a consistent ranking. In particular, if we have that $t - \gamma_S \geq |P_S|n$, then both σ, and σ with the first $|P_S|$ elements moved anywhere are both consistent rankings for the profile. Even more particularly, if we take σ' to be σ with the first

$|P_S|$ elements reversed, by reasoning analagous to that of Theorem 3.1, this would give us a lower bound of $\frac{\sqrt{2|P_S|}}{2}$. □

Pick σ to be the permutation consistent with the rankings that minimizes its distance to the medians, $\sum_j \sum_{i=1}^{m} |\sigma(j) - s_j|$, where s_j denotes the median of the rankings of candidate j across the voters, as before. Heuristically, this is a good σ to choose since, by, the triangle inequality we get something along the lines of

$$\gamma_S \leq \frac{1}{m} \sum_{j \in S} \sum_{i=1}^{m} (|\sigma(j) - s_j| + |s_j - q_{ij}|) \approx 2\frac{1}{m} \sum_{j \in S} \sum_{i=1}^{m} |s_j - q_{ij}|$$

and then we have a way to relate the bound to the median ranking.

Also note that this result is not particularly interesting when $t \in O(n)$, but could become interesting if it is not (note that t could be up to $\frac{n^2}{2}$). In general this is not at all a satisfying result and can almost definitely be heavily improved.

4 Conclusion and further directions

Regardless of the voter's profiles, our algorithm finds a candidate with expected rank $\leq 3\sqrt{t}$ and performs better on distributions where more of the noise is concentrated among worse alternatives. It is reasonable to assume that more of the noise is distributed among worse alternatives, since in real life people judge more appealing alternatives more acutely and correctly than less appealing ones.

The next natural step would be to find ways to close the gap between the upper and lower bound.

On the upper bounds side, a possible way to improve the algorithm would be to find an alternate way to lower bound the suffix noise that performs better than the median method, especially when the noise in a suffix is $O(t)$. For example, consider the scenario when the suffix noise of the permutation returned by [PSZ15] is actually lower bounded by t, that is for some set that isn't the full set of alternatives, S, any permutation π satisfies

$$\sum_{i \in [m]} \sum_{j \in S} |\pi(j) - q_{ij}| \geq t$$

and some superior algorithm knows this, but the median method gives a lower bound of $0.99t$. The superior algorithm could then upper bound the noise of the corresponding prefix by 0 and select the top person but our algorithm would only upper bound it by $0.01t$, and give a much worse guarantee.

On the lower bounds side, the aim would be to have an improved version of Theorem 3.2, in particular, a version that improves the bound on $|P_S|$. Additionally, it would help to come up with a better or more easily analyzable choice of the ranking σ used in the lower bound, and to then relate this directly to the upper bound we found for our algorithm.

Another task would be to compare how well the algorithm we came up with performs in the setting of probabilistic noise and draw a comparison to the methods in [CPS13].

References

[BM10] Alexander Barg and Arya Mazumdar. Codes in permutations and error correction for rank modulation. *Information Theory, IEEE Transactions on*, 56(7):3158–3165, 2010.

[CPS13] Ioannis Caragiannis, Ariel D Procaccia, and Nisarg Shah. When do noisy votes reveal the truth? In *Proceedings of the fourteenth ACM conference on Electronic commerce*, pages 143–160. ACM, 2013.

[PSZ15] Ariel D Procaccia, Nisarg Shah, and Yair Zick. Voting rules as error-correcting codes. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[SFD] https://mikespivey.wordpress.com/2014/01/20/the-maximum-value-of-spearmans-footrule-distance/.