A Learning Theory of Ranking Aggregation

Anna Korba
LTCI, Télécom ParisTech,
Université Paris-Saclay

Stephan Clémençon
LTCI, Télécom ParisTech,
Université Paris-Saclay

Eric Sibony
Shift Technology

Abstract

Originally formulated in *Social Choice theory*, *Ranking Aggregation*, also referred to as *Consensus Ranking*, has motivated the development of numerous statistical models since the middle of the 20th century. Recently, the analysis of ranking/preference data has been the subject of a renewed interest in machine-learning, boosted by modern applications such as meta-search engines, giving rise to the design of various scalable algorithmic approaches for approximately computing *ranking medians*, viewed as solutions of a discrete (generally NP-hard) minimization problem. This paper develops a statistical learning theory for ranking aggregation in a general probabilistic setting (avoiding any rigid ranking model assumptions), assessing the generalization ability of empirical ranking medians. Universal rate bounds are established and the situations where convergence occurs at an exponential rate are fully characterized. Minimax lower bounds are also proved, showing that the rate bounds we obtain are optimal.

1 INTRODUCTION

In *ranking aggregation*, the goal is to summarize a collection of rankings over a set of alternatives by a single (consensus) ranking. This problem has been the subject of a good deal of attention in various fields: starting from elections in social choice theory (see Borda, 1781; Condorcet, 1785), it has been applied to meta-search engines (see for instance Desarkar et al., 2016; Dwork et al., 2001; Renda and Straccia, 2003), competitions ranking (see for instance Davenport and Lovell, 2005; Deng et al., 2014) or bioinformatics (see for instance Kolde et al., 2012; Patel et al., 2013) among others.

Two main approaches have emerged in the literature to state the rank aggregation problem. The first one, originating from the seminal work of Condorcet in the 18th century (Condorcet, 1785), considers a generative probabilistic model on the rankings and the problem then consists in maximizing the likelihood of a candidate aggregate ranking. This MLE approach has been widely used in machine-learning and computational social choice, see *e.g.* Conitzer et al. (2009); Conitzer and Sandholm (2005); Truchon (2008). Alternatively, the metric approach consists in choosing a (pseudo-) distance on the set of rankings and then finding a barycentric/median ranking, *i.e.* a ranking at minimum distance from the observed ones. It encompasses numerous methods, including the popular *Kemeny aggregation*, which the present paper focuses on. These two approaches can be related in certain situations however. Indeed, Kemeny aggregation can be given a statistical interpretation: it is equivalent to the MLE approach under the noise model intuited by Condorcet (see Young, 1988) then formalized as the *Mallows model* (see definition in Remark 7).

Consensus ranking has given rise to a wide variety of results, much too numerous to be listed exhaustively. In particular, Kemeny aggregation has been shown to satisfy many desirable properties (see for instance Young and Levenglick, 1978) but also to be NP-hard to compute (Bartholdi et al., 1989), even for four votes (Dwork et al., 2001). This has led to different approaches to apprehend the complexity of this problem: bounds on the cost of approximation procedures have been obtained by Diaconis and Graham (1977), Coppersmith et al. (2006) or Sibony (2014), whereas Conitzer et al. (2006) and Davenport and Kalagnanam (2004) obtained approximation bounds that can be computed based on the ranking data (from the pairwise majority graph namely) and developed greedy procedures. Approximation of the Kemeny aggregation under polynomial complexity cost was also

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54. Copyright 2017 by the author(s).

considered in several papers, among which one finds Ailon et al. (2008), Van Zuylen and Williamson (2007) or Betzler et al. (2008). In Saari and Merlin (2000) consistency relationships between Kemeny aggregation and the Borda Count have been exhibited, while Procaccia et al. (2012) considered the recovery of the top-k alternatives.

Concerning the metric approach, much effort has been devoted to developing efficient algorithms for the computation of a median permutation related to a given collection of rankings, whereas statistical issues about the generalization properties of such empirical medians have been largely ignored as far as we know. The sole statistical analyses of ranking aggregation have been carried out in the restrictive setting of parametric models. Hence, in spite of this uninterrupted research activity, the generalization ability of ranking aggregation rules has not been investigated in a formal probabilistic setup, with the notable exception of Soufiani et al. (2014), where a decision-theoretic framework is introduced and the properties of Bayesian estimators for parametric models are discussed (as popular axioms in social choice). In this paper, we develop a general statistical framework for Kemeny aggregation, on the model of the probabilistic results developed for pattern recognition (see Devroye et al., 1996), the flagship problem in statistical learning theory. Precisely, conditions under which optimal elements can be characterized are exhibited, universal rate bounds for empirical Kemeny medians are stated and shown to be minimax. A low noise property is also introduced that allows to establish exponentially fast rates of convergence, following in the footsteps of the results obtained in Koltchinskii and Beznosova (2005) for binary classification.

The paper is organized as follows. In section 2, key notions of consensus ranking are briefly recalled and the statistical framework considered through the paper is introduced at length, together with the main notations. Section 3 is devoted to the characterization of optimal solutions for the Kemeny aggregation problem, while section 4 provides statistical guarantees for the generalization capacity of empirically barycentric rankings in the form of rate bounds in expectation/probability. The sketch of technical proofs is deferred to the Appendix section, see the Supplementary Material for further details.

2 BACKGROUND

We start with a rigorous formulation of (the metric approach of) consensus ranking and describe next the probabilistic framework for ranking aggregation we consider in this paper. Here and throughout, the indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{I}\{\mathcal{E}\}$, the Dirac mass at any point a by δ_a, and we set $sgn(x) = 2\mathbb{I}\{x \geq 0\} - 1$ for all $x \in \mathbb{R}$. At last, the set of permutations of the ensemble $[\![n]\!] = \{1, \ldots, n\}$, $n \geq 1$ is denoted by $\mathfrak{S}_n$.

2.1 Consensus Ranking

In the simplest formulation, a (full) ranking on a set of items $[\![n]\!]$ is seen as the permutation $\sigma \in \mathfrak{S}_n$ that maps an item i to its rank $\sigma(i)$. Given a collection of $N \geq 1$ permutations $\sigma_1, \ldots, \sigma_N$, the goal of ranking aggregation is to find $\sigma^* \in \mathfrak{S}_n$ that best summarizes it. A popular approach consists in solving the following optimization problem:

$$\min_{\sigma \in \mathfrak{S}_n} \sum_{i=1}^{N} d(\sigma, \sigma_i), \tag{1}$$

where $d(.,\ .)$ is a given metric on $\mathfrak{S}_n$. Such a barycentric permutation, referred to as a *consensus/median ranking* sometimes, always exists, since $\mathfrak{S}_n$ is finite, but is not necessarily unique. In the most studied version of this problem, termed Kemeny ranking aggregation, the metric considered is equal to the Kendall's τ distance (see Kemeny, 1959): $\forall (\sigma, \sigma') \in \mathfrak{S}_n^2$,

$$d_\tau(\sigma, \sigma') = \sum_{i<j} \mathbb{I}\{(\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0\},$$

i.e. the number of pairwise disagreements between σ and σ'. Such a consensus has many interesting properties, but is NP-hard to compute. Various algorithms have been proposed in the literature to compute acceptably good solutions in a reasonable amount of time, their description is beyond the scope of the paper, see for example Ali and Meila (2012) and the references therein.

2.2 Statistical Framework

In the probabilistic setting we consider here, the collection of rankings to be aggregated is supposed to be composed of $N \geq 1$ i.i.d. copies $\Sigma_1, \ldots, \Sigma_N$ of a generic random variable Σ defined on a probability space $(\Omega,\ \mathcal{F},\ \mathbb{P})$ drawn from an unknown probability distribution P on $\mathfrak{S}_n$ (*i.e.* $P(\sigma) = \mathbb{P}\{\Sigma = \sigma\}$ for any $\sigma \in \mathfrak{S}_n$). With respect to a certain metric $d(.,\ .)$ on $\mathfrak{S}_n$ (*e.g.* the Kendall τ distance), a (true) median of distribution P w.r.t. d is any solution of the minimization problem:

$$\min_{\sigma \in \mathfrak{S}_n} L(\sigma), \tag{2}$$

where $L(\sigma) = \mathbb{E}_{\Sigma \sim P}[d(\Sigma, \sigma)]$ denotes the expected distance between any permutation σ and Σ and shall be referred to as the *risk* of the median candidate

σ throughout the paper. The objective pursued is to recover approximately a solution σ^* of this minimization problem, plus an estimate of this minimum $L^* = L(\sigma^*)$, as accurate as possible, based on the observations $\Sigma_1, \ldots, \Sigma_N$. The minimization problem (2) always has a solution since the cardinality of $\mathfrak{S}_n$ is finite (however exploding with n) but can be multimodal, see Section 3. A median permutation σ^* can be interpreted as a central value for P, a crucial *location parameter*, whereas the quantity L^* can be viewed as a dispersion measure. However, the functional $L(.)$ is unknown in practice, just like distribution P (in order to avoid any ambiguity, we write $L_P(.)$ when needed). We only have access to the dataset $\{\Sigma_1, \ldots, \Sigma_N\}$ to find a reasonable approximant of a median and would like to avoid rigid assumptions on P such as those stipulated by the Mallows model, see Mallows (1957) and Remark 7. Following the Empirical Risk Minimization (ERM) paradigm (see *e.g.* Vapnik, 2000), one replaces the quantity $L(\sigma)$ by a statistical version based on the sampling data, typically the unbiased estimator

$$\widehat{L}_N(\sigma) = \frac{1}{N}\sum_{i=1}^{N} d(\Sigma_i, \sigma). \tag{3}$$

It is the goal of the subsequent analysis to assess the performance of solutions $\widehat{\sigma}_N$ of

$$\min_{\sigma\in\mathfrak{S}_n} \widehat{L}_N(\sigma), \tag{4}$$

by establishing (minimax) bounds for the excess of risk $L(\widehat{\sigma}_N) - L^*$ in probability/expectation, when d is the Kendall's τ distance. In this case, any solution of problem (2) (resp., of problem (4)) is called a *Kemeny median* (resp., an *empirical Kemeny median*) throughout the paper.

Remark 1 (Alternative dispersion measure) *An alternative measure of dispersion which can be more easily estimated than $L^* = L(\sigma^*)$ is given by*

$$\gamma(P) = \frac{1}{2}\mathbb{E}[d(\Sigma, \Sigma')], \tag{5}$$

where Σ' is an independent copy of Σ. One may easily show that $\gamma(P) \leq L^* \leq 2\gamma(P)$. The estimator of (5) *with minimum variance among all unbiased estimators is given by the U-statistic*

$$\widehat{\gamma}_N = \frac{2}{N(N-1)}\sum_{i<j} d(\Sigma_i, \Sigma_j). \tag{6}$$

In addition, we point out that confidence intervals for the parameter $\gamma(P)$ can be constructed by means of Hoeffding/Bernstein type deviation inequalities for U-statistics and a direct (smoothed) bootstrap procedure can be applied for this purpose, see Lahiri (1993). In contrast, a bootstrap technique for building CI's for L^* would require to solve several times an empirical version of (2) *based on bootstrap samples.*

Remark 2 (Alternative framework) *Since the computation of Kendall's τ distance involves pairwise comparisons only, one could compute empirical versions of the risk functional L in a statistical framework stipulating that the observations are less complete than $\{\Sigma_1, \ldots, \Sigma_N\}$ and formed by i.i.d. pairs $\{(\mathbf{e}_k, \epsilon_k),\ k = 1, \ldots, N\}$, where the $\mathbf{e}_k = (\mathbf{i}_k, \mathbf{j}_k)$'s are independent from the Σ_k's and drawn from an unknown distribution ν on the set $\mathcal{E}_n$ such that $\nu(e) > 0$ for all $e \in \mathcal{E}_n$ and $\epsilon_k = sgn(\Sigma_k(\mathbf{j}_k) - \Sigma_k(\mathbf{i}_k))$ with $\mathbf{e}_k = (\mathbf{i}_k, \mathbf{j}_k)$ for $1 \leq k \leq N$. Based on these observations, an estimate of the risk $\mathbb{E}_\nu\mathbb{E}_{\Sigma\sim P}[\mathbb{I}\{\mathbf{e} = (i,j),\ \epsilon(\sigma(j) - \sigma(i)) < 0\}]$ of any median candidate $\sigma \in \mathfrak{S}_n$ is given by:*

$$\sum_{i<j}\frac{1}{N_{i,j}}\sum_{k=1}^{N}\mathbb{I}\{\mathbf{e}_k = (i,j),\ \epsilon_k(\sigma(j) - \sigma(i)) < 0\},$$

where $N_{i,j} = \sum_{k=1}^{N}\mathbb{I}\{\mathbf{e}_k = (i,j)\}$, see for instance Lu and Boutilier (2014) or Rajkumar and Agarwal (2014) for ranking aggregation results in this setting.

2.3 Connection to Voting Rules

In Social Choice, we have a collection of votes under the form of rankings $P_N = (\sigma_1, \ldots, \sigma_N)$. Such a collection of votes $P_N \in \mathfrak{S}_n^N$ is called a *profile* and a voting rule, which outputs a consensus ranking on this profile, is classically defined as follows:

$$\sigma_{P_N} = \operatorname*{argmin}_{\sigma\in\mathfrak{S}_n} g(\sigma, P_N)$$

where $g : \mathfrak{S}_n \times \bigcup_{t=1}^{\infty}\mathfrak{S}_n^t \to \mathbb{R}$. This definition can be easily translated in order to be applied to any given distribution P instead of a profile. Indeed, the authors of Prasad et al. (2015) define a *distributional rank aggregation procedure* as follows:

$$\sigma_P = \operatorname*{argmin}_{\sigma\in\mathfrak{S}_n} g(\sigma, P)$$

where $g : \mathfrak{S}_n \times \mathcal{P}_n \to \mathbb{R}$ where $\mathcal{P}_n$ is the set of all distributions on $\mathfrak{S}_n$. Many classic aggregation procedures are naturally extended through this definition and thus to our statistical framework, as we have seen for Kemeny ranking aggregation previously. To detail some examples, we denote by $p_{i,j} = \mathbb{P}\{\Sigma(i) < \Sigma(j)\} = 1 - p_{j,i}$ for $1 \leq i \neq j \leq n$ and define the associated empirical estimator by $\widehat{p}_{i,j} = (1/N)\sum_{m=1}^{N}\mathbb{I}\{\Sigma_m(i) < \Sigma_m(j)\}$. The Copeland method (Copeland, 1951) consists on P_N in ranking the items by decreasing order of their Copeland score, calculated for each one as the number of items it beats in pairwise duels minus the number of items it looses against: $s_N(i) = \sum_{k\neq i}\mathbb{I}\{\widehat{p}_{i,k} \leq 1/2\} - \mathbb{I}\{\widehat{p}_{i,k} > 1/2\}$. It thus naturally applies to a distribution P using the scores $s(i) = \sum_{k\neq i}\mathbb{I}\{p_{i,k} \leq 1/2\} - \mathbb{I}\{p_{i,k} > 1/2\}$. Similarly, Borda aggregation (Borda, 1781) which consists in ranking items in increasing order of their score $s_N(i) = \sum_{t=1}^{N}\sigma_t(i)$ when applied on P_N, naturally extends to P using the scores $s(i) = \mathbb{E}_P[\Sigma(i)]$.

3 OPTIMALITY

As recalled above, the discrete optimization problem (2) always has a solution, whatever the metric d chosen. In the case of the Kendall's τ distance however, the optimal elements can be explicitly characterized in certain situations. It is the goal of this section to describe the set of Kemeny medians under specific conditions. As a first go, observe that the risk of a permutation candidate $\sigma \in \mathfrak{S}_n$ can be then written as

$$L(\sigma) = \sum_{i<j} p_{i,j}\mathbb{I}\{\sigma(i) > \sigma(j)\} + \sum_{i<j}(1-p_{i,j})\mathbb{I}\{\sigma(i) < \sigma(j)\}. \quad (7)$$

Remark 3 (Connection to binary classification) *Let $(\mathbf{i},\mathbf{j})$ be a random pair defined on $(\Omega, \mathcal{F}, \mathbb{P})$, uniformly distributed on the set $\{(i,j) : 1 \leq i < j \leq n\}$ and independent from Σ. Up to the factor $n(n-1)/2$, the risk (7) can be rewritten as the expectation of the error made when predicting the sign variable $sgn(\Sigma(\mathbf{j}) - \Sigma(\mathbf{i}))$ by the specific classifier $sgn(\sigma(\mathbf{j}) - \sigma(\mathbf{i}))$:*

$$L(\sigma) = \frac{n(n-1)}{2}\mathbb{E}\left[l_{\mathbf{i},\mathbf{j}}(\Sigma, \sigma)\right], \quad (8)$$

where we set $l_{i,j}(\sigma,\sigma') = \mathbb{I}\{(\sigma(i)-\sigma(j))\cdot(\sigma'(i)-\sigma'(j)) < 0\}$ for all $i<j$, $(\sigma,\sigma') \in \mathfrak{S}_n^2$. The r.v. $p_{\mathbf{i},\mathbf{j}}$ can be viewed as the posterior related to this classification problem.

We deduce from (7) that $L^* \geq \sum_{i<j}\min\{p_{i,j},\ 1-p_{i,j}\}$. In addition, if there exists a permutation σ with the property that $\forall i<j$ s.t. $p_{i,j} \neq 1/2$,

$$(\sigma(j) - \sigma(i)) \cdot (p_{i,j} - 1/2) > 0, \quad (9)$$

it would be necessarily a median for P (notice incidentally that $L^* = \sum_{i<j}\min\{p_{i,j},\ 1-p_{i,j}\}$ in this case).

Definition 4 *The probability distribution P on $\mathfrak{S}_n$ is said to be stochastically transitive if it fulfills the condition: $\forall(i,j,k) \in [\![n]\!]^3$,*

$$p_{i,j} \geq 1/2 \text{ and } p_{j,k} \geq 1/2 \Rightarrow p_{i,k} \geq 1/2$$

In addition, if $p_{i,j} \neq 1/2$ for all $i<j$, P is said to be strictly stochastically transitive.

Let $s^* : [\![n]\!] \rightarrow [\![n]\!]$ be the mapping defined by:

$$s^*(i) = 1 + \sum_{k\neq i}\mathbb{I}\{p_{i,k} < \frac{1}{2}\} \quad (10)$$

for all $i \in [\![n]\!]$, which induces the same ordering as the Copeland method (see Subsection 2.3). Observe that, if the *stochastic transitivity* is fulfilled, then: $p_{i,j} < 1/2 \Leftrightarrow s^*(i) < s^*(j)$. Equipped with this notation, property (9) can be also formulated as follows: $\forall i<j$ s.t. $s^*(i) \neq s^*(j)$,

$$(\sigma(j) - \sigma(i)) \cdot (s^*(j) - s^*(i)) > 0. \quad (11)$$

The result stated below describes the set of Kemeny median rankings under the conditions introduced above, and states the equivalence between the Copeland method and Kemeny aggregation in this setting.

Theorem 5 *If the distribution P is stochastically transitive, there exists $\sigma^* \in \mathfrak{S}_n$ such that (9) holds true. In this case, we have*

$$\begin{aligned} L^* &= \sum_{i<j}\min\{p_{i,j},\ 1-p_{i,j}\} \quad (12)\\ &= \sum_{i<j}\left\{\frac{1}{2} - \left|p_{i,j} - \frac{1}{2}\right|\right\}, \end{aligned}$$

the excess of risk of any $\sigma \in \mathfrak{S}_n$ is given by

$$L(\sigma) - L^* = 2\sum_{i<j}|p_{i,j} - 1/2| \cdot \mathbb{I}\{(\sigma(j)-\sigma(i))(p_{i,j}-1/2) < 0\}$$

and the set of medians of P is the class of equivalence of σ^* w.r.t. the equivalence relationship:

$$\sigma\mathcal{R}_P\sigma' \Leftrightarrow (\sigma(j)-\sigma(i))(\sigma'(j)-\sigma'(i)) > 0 \quad \text{for all } i<j \text{ such that } p_{i,j} \neq 1/2. \quad (13)$$

In addition, the mapping s^* belongs to $\mathfrak{S}_n$ iff P is strictly stochastically positive. In this case, s^* is the unique median of P.

The proof is detailed in the Appendix Section. Before investigating the accuracy of empirical Kemeny medians, a few remarks are in order.

Remark 6 (Borda consensus) *We say that the distribution P is strongly stochastically transitive if $\forall(i,j,k) \in [\![n]\!]^3$:*

$$p_{i,j} \geq 1/2 \text{ and } p_{j,k} \geq 1/2 \Rightarrow p_{i,k} \geq \max(p_{i,j}, p_{j,k}).$$

Then under this condition, and for $i < j$, $p_{i,j} \neq \frac{1}{2}$, there exists a unique $\sigma^* \in \mathfrak{S}_n$ such that (9) *holds true, corresponding to the Kemeny and Borda consensus both at the same time (see the Supplementary Material for the proof).*

Remark 7 (Mallows model) *The Mallows model introduced in the seminal contribution Mallows (1957) is a probability distribution P_θ on $\mathfrak{S}_n$ parametrized by $\theta = (\sigma_0, \phi) \in \mathfrak{S}_n \times [0,1]$: $\forall \sigma \in \mathfrak{S}_n$,*

$$P_{\theta_0}(\sigma) = \frac{1}{Z}\phi^{d_\tau(\sigma_0,\sigma)}, \tag{14}$$

where $Z = \sum_{\sigma\in\mathfrak{S}_n} \phi^{d_\tau(\sigma_0,\sigma)}$ is a normalization constant. One may easily show that Z is independent from σ and that $Z = \prod_{i=1}^{n-1}\sum_{j=0}^{i}\phi^j$. Observe firstly that the smallest the parameter ϕ, the spikiest the distribution P_θ (equal to a Dirac distribution for $\phi = 0$). In contrast, P_θ is the uniform distribution on $\mathfrak{S}_n$ when $\phi = 1$. Observe in addition that, as soon as $\phi < 1$, the Mallows model P_θ fulfills the strict stochastic transitivity property. Indeed, it follows in this case from Corollary 3 in Busa-Fekete et al. (2014) that for any $i < j$, we have:

(i) $\sigma_0(i) < \sigma_0(j) \Leftarrow p_{i,j} \geq \frac{1}{1+\phi} > \frac{1}{2}$ *with equality holding iff* $\sigma_0(i) = \sigma_0(j) - 1$,

(ii) $\sigma_0(i) > \sigma_0(j) \Leftarrow p_{i,j} \leq \frac{\phi}{1+\phi} < \frac{1}{2}$ *with equality holding iff* $\sigma_0(i) = \sigma_0(j) + 1$,

(iii) $p_{i,j} > \frac{1}{2}$ *iff* $\sigma_0(i) < \sigma_0(j)$ *and* $p_{i,j} < \frac{1}{2}$ *iff* $\sigma_0(i) > \sigma_0(j)$.

This directly implies that for any $i < j$:

$$|p_{i,j} - \frac{1}{2}| \geq \frac{|\phi - 1|}{2(1+\phi)}$$

Therefore, according to (12), *we have in this setting:*

$$L^*_{P_\theta} \leq \frac{n(n-1)}{2}\frac{\phi}{1+\phi}. \tag{15}$$

The permutation σ_0 of reference is then the unique mode of distribution P_θ, as well as its unique median.

Remark 8 (Bradley-Terry-Luce-Plackett model) *The Bradley-Terry-Luce-Plackett model (Bradley and Terry, 1952; Luce, 1959; Plackett, 1975) assumes the existence of some hidden preference vector $w = [w_i]_{1\leq i \leq n}$, where w_i represents the underlying preference score of item i. For all $i < j$, $p_{ij} = \frac{w_i}{w_i + w_j}$. If $w_1 \leq \cdots \leq w_n$, we have in this case $L^*_{P_\theta} = \sum_{i<j} w_i/(w_i + w_j)$. Observe in addition that as soon as for all $i < j$, $w_i \neq w_j$, the model fulfills the strict stochastic transitivity property. The permutation σ_0 of reference is then the one which sorts the vector w in decreasing order.*

4 EMPIRICAL CONSENSUS

Here, our goal is to establish sharp bounds for the excess of risk of *empirical Kemeny medians*, of solutions $\widehat{\sigma}_N$ of (4) in the Kendall's τ distance case namely. Beyond the study of universal rates for the convergence of the expected distance $L(\widehat{\sigma}_N)$ to L^*, we prove that, under the stochastic transitivity condition, exponentially fast convergence occurs, if the $p_{i,j}$'s are bounded away from 1/2, similarly to the phenomenon exhibited in Koltchinskii and Beznosova (2005) for binary classification under extremely low noise assumption.

4.1 Universal Rates

Such rate bounds are classically based on the fact that any minimizer $\widehat{\sigma}_n$ of (4) fulfills

$$L(\widehat{\sigma}_N) - L^* \leq 2 \max_{\sigma\in\mathfrak{S}_n} |\widehat{L}_N(\sigma) - L(\sigma)|. \tag{16}$$

As the cardinality of the set $\mathfrak{S}_n$ of median candidates is finite, they can be directly derived from bounds (tail probabilities or expectations) for the absolute deviations of i.i.d. sample means $\widehat{L}_N(\sigma)$ from their expectations, $|\widehat{L}_N(\sigma) - L(\sigma)|$. Let $\widehat{p}_{\mathbf{i},\mathbf{j}} = (1/N)\sum_{m=1}^{N}\mathbb{I}\{\Sigma_m(i) < \Sigma_m(j)\}$ and $p_{\mathbf{i},\mathbf{j}}$ the r.v. defined in Remark 3. First notice that same as in (7) one has for any $\sigma \in \mathfrak{S}_n$:

$$\begin{aligned}\widehat{L}_N(\sigma) = \frac{n(n-1)}{2}\mathbb{E}\,[\widehat{p}_{\mathbf{i},\mathbf{j}}\mathbb{I}\{\sigma(\mathbf{i}) > \sigma(\mathbf{j})\} \\ + (1 - \widehat{p}_{\mathbf{i},\mathbf{j}})\mathbb{I}\{\sigma(\mathbf{i}) < \sigma(\mathbf{j})\}]\end{aligned} \tag{17}$$

which, combined with (16), gives

$$|\widehat{L}_N(\sigma) - L(\sigma)| \leq \frac{n(n-1)}{2}\mathbb{E}_{\mathbf{i},\mathbf{j}}\left[|p_{\mathbf{i},\mathbf{j}} - \widehat{p}_{\mathbf{i},\mathbf{j}}|\right]. \tag{18}$$

This leads to the bounds in expectation and probability for ERM in the context of Kemeny ranking aggregation stated below, unsurprisingly of order $O(1/\sqrt{N})$.

Proposition 9 *Let $N \geq 1$ and $\widehat{\sigma}_N$ be any Kemeny empirical median based on i.i.d. training data $\Sigma_1, \ldots, \Sigma_N$, i.e. a minimizer of* (3) *over $\mathfrak{S}_n$ with $d = d_\tau$. The excess risk of $\widehat{\sigma}_N$ is upper bounded:*

(i) In expectation by

$$\mathbb{E}\left[L(\widehat{\sigma}_N) - L^*\right] \leq \frac{n(n-1)}{2\sqrt{N}}$$

(ii) With probability higher than $1-\delta$ for any $\delta \in (0,1)$ by

$$L(\widehat{\sigma}_N) - L^* \leq \frac{n(n-1)}{2}\sqrt{\frac{2\log(n(n-1)/\delta)}{N}}.$$

The proof is given in the Appendix section.

Remark 10 *As the problem* (4) *is NP-hard in general, one uses in practice an optimization algorithm to produce an approximate solution $\widetilde{\sigma}_N$ of the original minimization problem, with a control of the form: $\widehat{L}_N(\widetilde{\sigma}_N) \leq \min_{\sigma \in \mathfrak{S}_n} \widehat{L}_N(\sigma) + \rho$, where $\rho > 0$ is a tolerance fixed in advance, see e.g. Jiao et al. (2016). As pointed out in Bottou and Bousquet (2008), a bound for the expected excess of risk of $\widetilde{\sigma}_N$ is then obtained by adding the quantity ρ to the estimation error given in Proposition 9.*

We now establish the tightness of the upper bound for empirical Kemeny aggregation stated in Proposition 9. Precisely, the next theorem provides a lower bound of order $O(1\sqrt{N})$ for the quantity below, referred to as the *minimax risk*,

$$\mathcal{R}_N \overset{def}{=} \inf_{\sigma_N} \sup_P \mathbb{E}_P \left[L_P(\sigma_N) - L_P^* \right], \tag{19}$$

where the supremum is taken over all probability distributions on $\mathfrak{S}_n$ and the infimum is taken over all mappings σ_N that maps a dataset $(\Sigma_1, \ldots, \Sigma_N)$ composed of N independent realizations of P to an empirical median candidate .

Proposition 11 *The minimax risk for Kemeny aggregation is lower bounded as follows:*

$$\mathcal{R}_N \geq \frac{1}{16e\sqrt{N}}.$$

The proof of Proposition 11 relies on the classical Le Cam's method, it is detailed in the Supplementary Material. The result shows that no matter the method used for picking a median candidate from $\mathfrak{S}_n$ based on the training data, one may find a distribution such that the expected excess of risk is larger than $1/(16e\sqrt{N})$. If the upper bound from Proposition 9 depends on n, it is also of order $O(1/\sqrt{N})$ when N goes to infinity. Empirical Kemeny aggregation is thus optimal in this sense.

Remark 12 (Dispersion estimates) *In the stochastically transitive case, one may get an estimator of L^* by plugging the empirical estimates $\widehat{p}_{i,j}$ into Formula* (12)*:*

$$\begin{aligned} \widehat{L}^* &= \sum_{i<j} \min\{\widehat{p}_{i,j},\ 1 - \widehat{p}_{i,j}\} \qquad (20) \\ &= \sum_{i<j} \left\{ \frac{1}{2} - \left| \widehat{p}_{i,j} - \frac{1}{2} \right| \right\}. \end{aligned}$$

One may easily show that the related MSE is of order $O(1/N)$: $\mathbb{E}[(\widehat{L}^* - L^*)^2] \leq n^2(n-1)^2/(16N)$, see the Supplementary Material. Notice also that, in the Kendall's τ case, the alternative dispersion measure (5) *can be expressed as $\gamma(P) = \sum_{i<j} p_{i,j}(1-p_{i,j})$ and that the plugin estimator of $\gamma(P)$ based on the $\widehat{p}_{i,j}$'s coincides with* (6)*.*

While Proposition 9 makes no assumption about the underlying distribution P, it is also desirable to understand the circumstances under which the excess risk of empirical Kemeny medians is small. Following in the footsteps of results obtained in binary classification, it is the purpose of the subsequent analysis to exhibit conditions guaranteeing exponential convergence rates in Kemeny aggregation.

4.2 Fast Rates in Low Noise

The result proved in this subsection shows that the bound stated in Proposition 9 can be significantly improved under specific conditions. In binary classification, it is now well-known that (super) fast rate bounds can be obtained for empirical risk minimizers, see Massart and Nédélec (2006), Tsybakov (2004), and for certain *plug-in* rules, see Audibert and Tsybakov (2007). As shown below, under the stochastic transitivity hypothesis and the following *low noise assumption* (then implying strict stochastic transitivity), the risk of empirical minimizers in Kemeny aggregation converges exponentially fast to L^* and remarkably, with overwhelming probability, empirical Kemeny aggregation has a unique solution that coincides with a natural *plug-in* estimator of the true median (namely s^* in this situation, see Theorem 5). For $h > 0$, we define condition:

$$\mathbf{NA}(h)\text{: } \min_{i<j} |p_{i,j} - 1/2| \geq h.$$

Remark 13 (Low noise for parametric models) *Condition $\mathbf{NA}(h)$ is fulfilled by many parametric models. For example, the Mallows model* (14) *parametrized by $\theta = (\sigma_0, \phi) \in \mathfrak{S}_n \times [0,1]$ satisfies $\mathbf{NA}(h)$ iff $\phi \leq (1-2h)/(1+2h)$. For the Bradley-Terry-Luce-Plackett model with preference vector $w = [w_i]_{1 \leq i \leq n}$, condition $\mathbf{NA}(h)$ is satisfied iff $\min_{1 \leq i \leq n} |w_i - w_{i+1}| \geq (4h)/(1-2h)$, see Chen and Suh (2015) where minimax bounds are obtained for the problem of identifying top-K items.*

This condition may be considered as analogous to that introduced in Koltchinskii and Beznosova (2005) in binary classification, and was used in Shah et al. (2015) to prove fast rates for the estimation of the matrix of pairwise probabilities.

Proposition 14 *Assume that P is stochastically transitive and fulfills condition $\mathbf{NA}(h)$ for some $h > 0$. The following assertions hold true.*

(i) For any empirical Kemeny median $\widehat{\sigma}_N$, we have: $\forall N \geq 1$,

$$\mathbb{E}\left[L(\widehat{\sigma}_N) - L^*\right] \leq \frac{n^2(n-1)^2}{8} e^{-\frac{N}{2}\log\left(\frac{1}{1-4h^2}\right)}.$$

(ii) With probability at least $1 - (n(n-1)/4)e^{-\frac{N}{2}log\left(\frac{1}{1-4h^2}\right)}$, the mapping

$$\widehat{s}_N(i) = 1 + \sum_{k \neq i} \mathbb{I}\{\widehat{p}_{i,k} < \frac{1}{2}\}$$

for $1 \leq i \leq n$ belongs to $\mathfrak{S}_n$ and is the unique solution of the empirical Kemeny aggregation problem (4)*. It is then referred to as the plug-in Kemeny median.*

The technical proof is given in the Supplementary Material. The main argument consists in showing that, under the hypotheses stipulated, with very large probability, the empirical distribution $\widehat{P}_N = (1/N)\sum_{i=1}^{N} \delta_{\Sigma_i}$ is strictly stochastically transitive and Theorem 5 applies to it. Proposition 14 gives a rate in $O(e^{-\alpha_h N})$ with $\alpha_h = \frac{1}{2}log\left(1/(1-4h^2)\right)$. Notice that $\alpha_h \to +\infty$ as $h \to 1/2$, which corresponds to the situation where the distribution converges to a Dirac δ_σ since P is supposed to be stochastically transitive. Therefore the greatest h is, the easiest is the problem and the strongest is the rate. On the other hand, the rate decreases when h gets smaller. The next result proves that, in the low noise setting, the rate of Proposition 14 is almost sharp in the minimax sense.

Proposition 15 *Let $h > 0$ and define*

$$\widetilde{\mathcal{R}}_N(h) = \inf_{\sigma_N} \sup_{P} \mathbb{E}_P\left[L_P(\sigma_N) - L_P^*\right],$$

where the supremum is taken over all stochastically transitive probability distributions P on $\mathfrak{S}_n$ satisfying $\mathbf{NA}(h)$*. We have: $\forall N \geq 1$,*

$$\widetilde{\mathcal{R}}_N(h) \geq \frac{h}{4} e^{-N2h\log\left(\frac{1+2h}{1-2h}\right)}. \tag{21}$$

The proof of Proposition 15 is provided in the Supplementary Material. It shows that the minimax rate is lower bounded by a rate in $O(e^{-\beta_h N})$ with $\beta_h = 2h\log((1+2h)/(1-2h))$. Notice that $\alpha_h \sim \beta_h/2$ when $h \to 1/2$. The rate obtained for empirical Kemeny aggregation in Proposition 14 is thus almost optimal in this case. The bound from Proposition 15 is however too small when $h \to 0$ as it goes to 0. Improving the minimax lower bound in this situation is left for future work.

4.3 Computational Issues

As mentioned previously, the computation of an empirical Kemeny consensus is NP-hard and therefore usually not tractable in practice. Proposition 9 and 14 can therefore be seen as providing theoretical guarantees for the ideal estimator $\widehat{\sigma}_N$. Under the low noise assumption however, Proposition 14 also has a practical interest. Part (ii) says indeed that in this case, the Copeland method (ordering items by decreasing score $\widehat{s}_N$), which has complexity in $O(N\binom{n}{2})$, outputs the exact Kemeny consensus with high probability. Furthermore, part (i) actually applies to any empirical median $\tilde{\sigma}_N$ that is equal to $\widehat{\sigma}_N$ with probability at least $1-(n(n-1)/4)e^{-(N/2)\log(1/(1-4h^2))}$ thus in particular to the Copeland method. In summary, under assumption $\mathbf{NA}(h)$ with $h > 0$, the tractable Copeland method outputs the exact Kemeny consensus with high probability and has almost optimal excess risk convergence rate.

5 CONCLUSION

Whereas the issue of computing (approximately) ranking medians has received much attention in the literature, just like statistical modelling of the variability of ranking data, the generalization ability of practical ranking aggregation methods has not been studied in a general probabilistic setup. By describing optimal elements and establishing learning rate bounds for empirical Kemeny ranking medians, this paper provides a first statistical explanation for the success of these techniques.

APPENDIX - PROOFS

In this section, we denote by $\binom{n}{k}$ the binomial coefficient indexed by n and k, by $\mathcal{B}(N,p)$ the binomial distribution indexed by parameters N, p, by (i,j) the transposition that swaps item i and j, and by $K(P||Q)$ the Kullback-Leibler divergence between two probability distributions P and Q.

Proof of Theorem 5

Suppose that distribution P is stochastically transitive. The pairwise probabilities can be then represented as a directed graph on the n items with the following definition: each item i is represented by a vertex, and an edge is drawn from i to j whenever $p_{i,j} > \frac{1}{2}$ (no edge is drawn if $p_{i,j} = \frac{1}{2}$). This graph is thus a directed acyclic graph, since the stochastic transitivity prevents the occurrence of any cycle. Hence, there exists a partial order on the graph (also referred to as *topological ordering*, the vertices representing the items are sorted by their in-degree) and any permutation σ^* extending this partial order satisfies (9). In addition, under the stochastic transitivity condition, for $i < j$, $p_{i,j} < 1/2 \Leftrightarrow s^*(i) < s^*(j)$. So s^* belongs to $\mathfrak{S}_n$ iff P is stricly stochastically transitive and in this case, s^* is the unique median of P.

Proof of Remark 6

With simple calculations, the Borda score of any item $i \in [\![n]\!]$ can be written as $s(i) = 1 + \sum_{k \neq i} p_{k,i}$. Suppose that $p_{i,j} > 1/2$ ($\Leftrightarrow s^*(i) < s^*(j)$ under *stochastic transitivity*). We have $s(j) - s(i) = \sum_{k \neq i,j} p_{k,j} - p_{k,i} + (2p_{i,j} - 1)$ with $2p_{i,j} - 1 > 0$. We prove that for any $k \neq i, j$, we have $p_{k,j} - p_{k,i} \geq 0$ under the strong stochastic transitivity condition (by considering firstly the case where $p_{j,k} \geq 1/2$ and next the case where $p_{k,j} > 1/2$). We obtain that $(s(i) - s(j))(s^*(i) - s^*(j)) > 0$, the scoring functions s and s^* yield exactly the same ranking on the set of items.

Proof of Proposition 9

Apply first Cauchy-Schwarz inequality, so as to get

$$\mathbb{E}_{\mathbf{i},\mathbf{j}}\left[|p_{\mathbf{i},\mathbf{j}} - \widehat{p}_{\mathbf{i},\mathbf{j}}|\right] \leq \sqrt{\mathbb{E}_{\mathbf{i},\mathbf{j}}\left[(p_{\mathbf{i},\mathbf{j}} - \widehat{p}_{\mathbf{i},\mathbf{j}})^2\right]} = \sqrt{Var(\widehat{p}_{\mathbf{i},\mathbf{j}})},$$

since $\mathbb{E}_{\mathbf{i},\mathbf{j}}\left[p_{\mathbf{i},\mathbf{j}} - \widehat{p}_{\mathbf{i},\mathbf{j}}\right] = 0$. Then, for $i < j$, $N\widehat{p}_{i,j} \sim \mathcal{B}(N, p_{i,j}))$ and thus $Var(\widehat{p}_{\mathbf{i},\mathbf{j}}) \leq \frac{1}{4N}$. Combining (18) with the last upper bound on the variance finally gives the upper bound stated.

A related probability bound can also be established as follows. By (16) and (18), we have: for any $t > 0$,

$$\mathbb{P}\Big\{L(\widehat{\sigma}_N) - L^* > t\Big\} \leq \mathbb{P}\Big\{\sum_{i<j} |p_{i,j} - \widehat{p}_{i,j}| > \frac{t}{2}\Big\}.$$

On the other hand, we have:

$$\mathbb{P}\Big\{\sum_{i<j} |p_{i,j} - \widehat{p}_{i,j}| > \frac{t}{2}\Big\} \leq \sum_{i<j} \mathbb{P}\Big\{|p_{i,j} - \widehat{p}_{i,j}| > \frac{t}{n(n-1)}\Big\}.$$

The last step consists in applying Hoeffding's inequality to each $p_{i,j}$ for $i < j$

$$\mathbb{P}\Big\{|p_{i,j} - \widehat{p}_{i,j}| > \frac{t}{n(n-1)}\Big\} \leq 2e^{-2N(\frac{t}{n(n-1)})^2}.$$

Combining the three preceding inequalities gives the bound.

Proof of Proposition 11

The proof of the minimax lower bound is based on Le Cam's method, see section 2.3 in Tsybakov (2009). Consider two Mallows models P_{θ_0} and P_{θ_1} where $\theta_k = (\sigma_k^*, \phi) \in \mathfrak{S}_n \times (0,1)$ and $\sigma_0^* \neq \sigma_1^*$. We clearly have:

$$\begin{aligned}\mathcal{R}_N \geq &\inf_{\sigma_N} \frac{|\phi - 1|}{2(1+\phi)} \times \\ &\max_{k=0,\,1} \sum_{i<j} \mathbb{E}_{P_{\theta_k}}\left[\mathbb{I}\{(\sigma_N(i) - \sigma_N(j)(\sigma_k^*(i) - \sigma_k^*(j)) < 0\}\right] \\ \geq &\frac{|\phi - 1|}{4} \inf_{\sigma_N} \max_{k=0,\,1} \mathbb{E}_{P_{\theta_k}}\left[d_\tau(\sigma_N, \sigma_k^*)\right].\end{aligned}$$

Following line by line Le Cam's method, we obtain

$$\mathcal{R}_N \geq \frac{|\phi - 1|}{32} e^{-NK(P_{\theta_0}||P_{\theta_1})}.$$

Some calculations show that $K(P_{\theta_0}||P_{\theta_1}) = \log(\frac{1}{\phi})(1-\phi)/(1+\phi)$, for $\sigma_1 = (i,j)\sigma_0$ where $i < j$ verify $\sigma_0(i) = \sigma_0(j) - 1$. The desired lower bound is then obtained by taking $\phi = 1 - 1/\sqrt{N}$. More details can be found in the Supplementary Material.

Proposition 14

Let $\mathcal{A}_N = \bigcap_{i<j} \{(p_{i,j} - \frac{1}{2})(\widehat{p}_{i,j} - \frac{1}{2}) > 0\}$. On the event $\mathcal{A}_N$, both distributions p and $\widehat{p}$ satisfy the strong stochastic transitivity property, and agree on each pair: $\widehat{\sigma}_N = \sigma$ and $L(\widehat{\sigma}_N) - L^* = 0$. Therefore we only have to bound $\mathbb{P}\Big\{\mathcal{A}_N^c\Big\}$. Since for $i < j$, $N\widehat{p}_{i,j} \sim \mathcal{B}(N, p_{i,j}))$, we have

$$\mathbb{P}\Big\{\widehat{p}_{i,j} \leq \frac{1}{2}\Big\} = \sum_{k=0}^{\lfloor \frac{N}{2} \rfloor} \binom{N}{k} p_{i,j}^k (1 - p_{i,j})^{N-k}. \qquad (22)$$

Then, $\sum_{k=0}^{\lfloor \frac{N}{2} \rfloor} \binom{N}{k} \leq 2^{N-1}$ and since $p_{i,j} > 1/2$, for $k \leq \frac{N}{2}$, we have

$$p_{i,j}^k (1 - p_{i,j})^{N-k} \leq p_{i,j}^{\frac{N}{2}} (1 - p_{i,j})^{\frac{N}{2}} \leq \left(\frac{1}{4} - h^2\right)^{\frac{N}{2}}$$

Finally, we have $\mathbb{P}\Big\{\mathcal{A}_N^c\Big\} \leq \sum_{i<j} \mathbb{P}\Big\{\widehat{p}_{i,j} \leq \frac{1}{2}\Big\}$ and the desired bound is proved.

Proof of Proposition 15

Similarly to Proposition 11, we can bound by below the minimax risk as follows

$$\begin{aligned}\mathcal{R}_N \geq &\inf_{\sigma_N} \max_{k=0,\,1} h\mathbb{E}_{P_{\theta_k}}\left[d_\tau(\sigma_N, \sigma^*)\right] \\ \geq &\frac{h}{8} e^{-NK(P_{\theta_0}||P_{\theta_1})},\end{aligned}$$

with $K(P_{\theta_0}||P_{\theta_1}) = \log(\frac{1}{\phi})(1-\phi)/(1+\phi)$. Now we take $\phi = (1-2h)/(1+2h)$ so that both P_{θ_0} and P_{θ_1} satisfy $\mathbf{NA}(h)$, and we have

$$K(P_{\theta_0}||P_{\theta_1}) = 2h \log\left(\frac{1+2h}{1-2h}\right),$$

which finally gives us the bound.

References

Ailon, N., Charikar, M., and Newman, A. (2008). Aggregating inconsistent information: ranking and clustering. *Journal of the ACM (JACM)*, 55(5):23.

Ali, A. and Meila, M. (2012). Experiments with kemeny ranking: What works when? *Mathematical Social Sciences*, 64(1):28 – 40.

Audibert, J. and Tsybakov, A. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.*, 32:608–633.

Bartholdi, J. J., Tovey, C. A., and Trick, M. A. (1989). The computational difficulty of manipulating an election. *Social Choice and Welfare*, 6:227–241.

Betzler, N., Fellows, M. R., Guo, J., Niedermeier, R., and Rosamond, F. A. (2008). Fixed-parameter algorithms for kemeny scores. In *International Conference on Algorithmic Applications in Management*, pages 60–71. Springer.

Borda, J. C. (1781). Mémoire sur les élections au scrutin.

Bottou, L. and Bousquet, O. (2008). The trade-offs of large-scale learning. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Proceedings of NIPS'07*, pages 161–168.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Busa-Fekete, R., Hüllermeier, E., and Szörényi, B. (2014). Preference-based rank elicitation using statistical models: The case of mallows. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1071–1079.

Chen, Y. and Suh, C. (2015). Spectral mle: Top-k rank aggregation from pairwise comparisons. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 371–380.

Condorcet, N. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. L'imprimerie royale, Paris.

Conitzer, V., Davenport, A., and Kalagnanam, J. (2006). Improved bounds for computing kemeny rankings. In *AAAI*, volume 6, pages 620–626.

Conitzer, V., Rognlie, M., and Xia, L. (2009). Preference functions that score rankings and maximum likelihood estimation. In *IJCAI*, volume 9, pages 109–115.

Conitzer, V. and Sandholm, T. (2005). Common voting rules as maximum likelihood estimators. In *Proceedings of the Twenty-First Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)*, pages 145–152, Arlington, Virginia. AUAI Press.

Copeland, A. H. (1951). A reasonable social welfare function. In *Seminar on applications of mathematics to social sciences, University of Michigan*.

Coppersmith, D., Fleischer, L., and Rudra, A. (2006). Ordering by weighted number of wins gives a good ranking for weighted tournaments. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithm*, SODA '06, pages 776–782.

Davenport, A. and Kalagnanam, J. (2004). A computational study of the kemeny rule for preference aggregation. In *AAAI*, volume 4, pages 697–702.

Davenport, A. and Lovell, D. (2005). Ranking pilots in aerobatic flight competitions. Technical report, IBM Research Report RC23631 (W0506-079), TJ Watson Research Center, NY.

Deng, K., Han, S., Li, K. J., and Liu, J. S. (2014). Bayesian aggregation of order-based rank data. *Journal of the American Statistical Association*, 109(507):1023–1039.

Desarkar, M. S., Sarkar, S., and Mitra, P. (2016). Preference relations based unsupervised rank aggregation for metasearch. *Expert Systems with Applications*, 49:86–98.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. Springer.

Diaconis, P. and Graham, R. L. (1977). Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 262–268.

Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. (2001). Rank aggregation methods for the Web. In *Proceedings of the 10th International WWW conference*, pages 613–622.

Jiao, Y., Korba, A., and Sibony, E. (2016). Controlling the distance to a kemeny consensus without computing it. In *Proceeding of ICML 2016*.

Kemeny, J. G. (1959). Mathematics without numbers. *Daedalus*, 88:571–591.

Kolde, R., Laur, S., Adler, P., and Vilo, J. (2012). Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, 28(4):573–580.

Koltchinskii, V. and Beznosova, O. (2005). Exponential convergence rates in classification. In *Proceedings of COLT 2005*.

Lahiri, M. (1993). Bootstrapping the studentized sample mean of lattice variables. *Journal of Multivariate Analysis.*, 45:247–256.

Lu, T. and Boutilier, C. (2014). Effective sampling and learning for mallows models with pairwise-preference data. volume 15, pages 3963–4009.

Luce, R. D. (1959). *Individual Choice Behavior*. Wiley.

Mallows, C. L. (1957). Non-null ranking models. *Biometrika*, 44(1-2):114–130.

Massart, P. and Nédélec, E. (2006). Risk bounds for statistical learning. *Annals of Statistics*, 34(5).

Patel, T., Telesca, D., Rallo, R., George, S., Xia, T., and Nel, A. E. (2013). Hierarchical rank aggregation with applications to nanotoxicology. *Journal of Agricultural, Biological, and Environmental Statistics*, 18(2):159–177.

Plackett, R. L. (1975). The analysis of permutations. *Applied Statistics*, 2(24):193–202.

Prasad, A., Pareek, H., and Ravikumar, P. (2015). Distributional rank aggregation, and an axiomatic analysis. In Blei, D. and Bach, F., editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2104–2112. JMLR Workshop and Conference Proceedings.

Procaccia, A. D., Reddi, S. J., and Shah, N. (2012). A maximum likelihood approach for selecting sets of alternatives. *arXiv preprint arXiv:1210.4882*.

Rajkumar, A. and Agarwal, S. (2014). A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *Proceedings of the 31st International Conference on Machine Learning*.

Renda, M. E. and Straccia, U. (2003). Web metasearch: rank vs. score based rank aggregation methods. In *Proceedings of the 2003 ACM symposium on Applied computing*, pages 841–846. ACM.

Saari, D. G. and Merlin, V. R. (2000). A geometric examination of kemeny's rule. *Social Choice and Welfare*, 17(3):403–438.

Shah, N. B., Balakrishnan, S., Guntuboyina, A., and Wainright, M. J. (2015). Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *arXiv preprint arXiv:1510.05610*.

Sibony, E. (2014). Borda count approximation of kemenys rule and pairwise voting inconsistencies. In *Proceedings of the NIPS 2014 Workshop on Analysis of Rank Data*.

Soufiani, H. A., Parkes, D. C., and Xia, L. (2014). A statistical decision-theoretic framework for social choice. In *Advances in Neural Information Processing Systems*, pages 3185–3193.

Truchon, M. (2008). Borda and the maximum likelihood approach to vote aggregation. *Mathematical Social Sciences*, 55(1):96–102.

Tsybakov, A. (2004). Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32(1):135–166.

Tsybakov, A. B. (2009). Introduction to nonparametric estimation. revised and extended from the 2004 french original. translated by vladimir zaiats.

Van Zuylen, A. and Williamson, D. P. (2007). Deterministic algorithms for rank aggregation and other ranking and clustering problems. In *International Workshop on Approximation and Online Algorithms*, pages 260–273. Springer.

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory*. Lecture Notes in Statistics. Springer.

Young, H. P. (1988). Condorcet's theory of voting. *American Political Science Review*, 82(4):1231–1244.

Young, H. P. and Levenglick, A. (1978). A consistent extension of condorcet's election principle. *SIAM Journal on applied Mathematics*, 35(2):285–300.