Efficient Vote Elicitation under Candidate Uncertainty

Joel Oren and **Yuval Filmus** and **Craig Boutilier**
Department of Computer Science, University of Toronto
{oren,yuvalf,cebly}@cs.toronto.edu

Abstract

Top-k voting is an especially natural form of partial vote elicitation in which only length k prefixes of rankings are elicited. We analyze the ability of top-k vote elicitation to correctly determine true winners, with high probability, given probabilistic models of voter preferences and candidate availability. We provide bounds on the minimal value of k required to determine the correct winner under the plurality and Borda voting rules, considering both worst-case preference profiles and profiles drawn from the impartial culture and Mallows probabilistic models. We also derive conditions under which the special case of *zero-elicitation* (i.e., $k = 0$) produces the correct winner. We provide empirical results that confirm the value of top-k voting.

1 Introduction

Social choice has provided valuable foundations for the development of computational approaches to preference aggregation, group decision making and a variety of other problems in recent years. As algorithmic advances and data accessibility make the methods of social choice more broadly applicable, relaxing the assumptions of classical models to fit a richer class of practical problems becomes imperative. To this end, research has begun to address the *informational demands* of preference aggregation. For example, recent work has considered models in which information about the set of available candidates is imperfect [Lu and Boutilier, 2010; Baldiga and Green, 2011; Boutilier *et al.*, 2012]. Similarly, knowledge of voter preferences may be incomplete [Konczak and Lang, 2005; Xia and Conitzer, 2008; Lu and Boutilier, 2011b].

In this work, we bring together these two lines of research to investigate the feasibility and value of *top-k voting*. Our first motivation is to use intelligent *vote elicitation* techniques to minimize the amount of voter preference information required to determine the winner in an election (or more broadly, the desired outcome of a group decision). Vote elicitation has received considerable attention recently [Conitzer and Sandholm, 2005; Conitzer, 2009; Kalech *et al.*, 2011; Lu and Boutilier, 2011b; 2011c; Ding and Lin, 2012], and has proven to be effective in reducing the amount of information—and corresponding cognitive and communication burden—needed to determine winners in practice. Our second motivation is to develop methods that handle uncertainty in the set of available candidates. In many settings, voters may need to specify their preferences over a range of potential candidates prior to knowing which are in fact available or viable for selection [Lu and Boutilier, 2010; Baldiga and Green, 2011; Boutilier *et al.*, 2012]. Examples include ranking job candidates, public projects, or even restaurants. The potential impact of candidate unavailability on vote elicitation is clear: since certain desirable alternatives may turn out to be unavailable, one may need to elicit more preference information than is typical in the case of fully known candidates in order to ensure the correct winner is chosen.

We address the problems of efficient preference elicitation in this context in the form of *top-k elicitation*. In top-k voting, agents are asked to provide the length k prefix of their preference ranking instead of their full ranking. In the standard "known candidates" model, top-k voting has been used heuristically [Kalech *et al.*, 2011] and the optimal choice of k has been analyzed from a sample-complexity-theoretic perspective [Lu and Boutilier, 2011c]. However, bounds on the required values of k for specific preference distributions and voting rules have remained unaddressed, as has the impact of unavailable candidates on top-k voting.

In this work, we examine two common voting rules, plurality and Borda—these serve as a useful starting point for the investigation of our model, representing rather different extremes in space of so-called *scoring rules* for voting. Given a prior distribution on the preference profile, and a distribution over the set of available candidates (for which the standard "known candidates" model is a special case), we ask: what is the minimal value of k for which top-k voting determines the true winner (with high probability), with respect to the underlying preference profile? We provide theoretical results, in the form of upper and lower bounds on k, for both worst-case preferences and certain preference distributions (including impartial culture and Mallows distributions). As a special case, we consider *zero-elicitation protocols*, where $k = 0$. We show when, as a function of the election parameters, the true winner can be determined with high probability without eliciting *any* information from voters. We also provide empirical results demonstrating the extent to which top-k voting determines true winners as a function of k.

2 The Model

Let $C = \{c_1, \ldots, c_m\}$ be the set of (potential) candidates from which a winner is to be selected using some voting rule. Let $N = \{1, \ldots, n\}$ be the set of voters, and let voter i's *preference* π_i be a permutation of C: intuitively, for $1 \leq j < j' \leq m$, $\pi_i(j)$ is preferred by i to $\pi_i(j')$. Let $\mathcal{L}$ denote the set of all preferences over C. A *preference profile* $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n) \in \mathcal{L}^n$ represents the collection of voter preferences. A *voting rule* $v \colon \mathcal{L}^n \times 2^C \rightarrow C$ selects a winner from C given a vote profile and a set of available candidates.

We consider two voting rules: plurality and Borda. In *plurality voting*, given a profile $\boldsymbol{\pi}$, the *plurality score* of candidate c is the number of times that c is ranked first: $sc^P(c, \boldsymbol{\pi}) = |\{i \in N : \pi_i(1) = c\}|$. The plurality winner is the candidate with maximal plurality score (ties can be handled arbitrarily; the tie-breaking rule used does not impact our results). In *Borda* voting, the *Borda score* of c is the number of candidates ranked below it, summed over all preferences π_i: $sc^B(c, \boldsymbol{\pi}) = \sum_{i \in N}[m - \pi_i^{-1}(c)]$. The Borda winner is the candidate with maximal Borda score.

Unavailable Candidates. Recent attention has been paid to the possibility of voting over a slate of *potential candidates* C, prior to determining the availability of the *actual set* of candidates $A \subseteq C$. When determining the availability of candidates is costly or risky (e.g., making job offers, determining feasibility of public projects, calling restaurants for reservations), it often makes sense to elicit voter preferences prior to determining availability. Once preferences are known, one can focus availability determination on candidates most likely to be winners relative to the true available set A [Lu and Boutilier, 2010; Baldiga and Green, 2011; Boutilier *et al.*, 2012]. Following these recent models, we assume that each candidate $c \in C$ is available i.i.d. with some fixed probability $p \in (0, 1]$. The use of a fixed p simplifies our presentation; but using distinct probabilities p_c for different candidates c does not change the nature of our results—all can be adapted accordingly.

Given a set $A \subseteq C$ of available candidates, a *reduced preference* $\pi_i|_A$ is obtained by restricting π_i to the candidates in A; we denote by $\boldsymbol{\pi}|_A$ the *reduced preference profile* obtained in this way. Plurality and Borda voting in the unavailable candidate model are defined in the obvious way, using the scores obtained relative to the reduced profile. Notice that in the unavailable candidates model, it is no longer sufficient to run plurality voting by eliciting just the top-ranked candidate from each voter: in general the entire ranking may be needed.

Top-k voting. Recent research has focused on the use of intelligent preference elicitation schemes to minimize the burden on voters and obviate the need to provide full preference rankings. One especially natural approach is *top-k* voting, in which voters are asked to list only their k most preferred candidates (or the k-th prefix of their ranking) [Kalech *et al.*, 2011; Lu and Boutilier, 2011b; 2011c]. We discuss below alternative ways in which such votes can be used to determine winners; but here we adopt an especially simple approach.

Given a voting rule v and some $k \in [m]$, we denote by $(\boldsymbol{\pi}^{(k)}) = (\pi_1^{(k)}, \ldots, \pi_n^{(k)})$ the k-truncated preference profile. We use this truncated profile to determine plurality scores in the obvious fashion, by counting the number of first place rankings. We compute Borda scores by assigning a score of $\tilde{m} - \pi_i^{-1}(c)$ to any candidate c in voter i's k-truncated vote, where $\tilde{m}$ is the number of available candidates, and a score of zero otherwise. In the unavailable candidates model, we employ the same technique, restricting the *truncated* vote to the available set A. Our goal is to determine values of k that suffice to determine the true winner (with high probability) relative to the true (untruncated) preference profile.

If candidates are always available (i.e., $p = 1$) then $k = 1$ is sufficient to determine the correct plurality winner, and general top-k voting is of no value. By contrast, the possibility of unavailable candidates intuitively requires that one use larger values of k for plurality and other voting rules.

Probabilistic preference models. It has become increasingly common to analyze voting rules under the assumption that agent preferences are drawn from a prior distribution over permutations. One important class of distributions, widely used in psychometrics, statistics, and machine learning, is the *Mallows φ-distribution* [Mallows, 1957; Marden, 1995]. It is described by two parameters: a *reference ranking* $\hat{\pi} \in \mathcal{L}$, and a *dispersion parameter* φ (controlling variance). The probability of a permutation π under this model is $\Pr(\pi) = \varphi^{\tau(\pi, \hat{\pi})}/Z_m$, where $\tau(\pi, \hat{\pi})$ is the Kendall-tau distance,

$$\tau(\pi_1, \pi_2) = |\{c, c' : \pi_1^{-1}(c) < \pi_1^{-1}(c') \text{ and } \pi_2^{-1}(c) > \pi_2^{-1}(c')\}|,$$

and Z_m is a normalization constant. Importantly, when $\varphi = 1$, one obtains the uniform distribution over $\mathcal{L}$, the so-called *impartial culture (IC)* assumption, a modeling assumption widely used in social choice.

Related Work. As mentioned, vote elicitation has attracted considerable recent attention, usually in the context of standard "known available candidate" models. Of particular relevance is work on top-k voting. Unlike our model, in which we "zero out" the scores of unavailable candidates, other work has treated the uncertainty in the missing candidates more cautiously. Kalech et al. [2011] use top-k ballots to determine *possible and necessary winners* [Konczak and Lang, 2005] and develop heuristic elicitation schemes to extend these ballots to quickly identify true winners for several different voting rules. Lu and Boutilier [2011b] use *minimax regret* to measure error in winner determination and to guide elicitation heuristically as well. Both methods show good empirical performance (and handle general partial votes) but provide no theoretical guarantees on the required values of k. The optimal choice of k has been analyzed from a sample-complexity-theoretic perspective by Lu and Boutilier [2011c], who provide bounds on the *required number of sampled profiles* needed to estimate the required value of k for arbitrary distributions; but this does not provide direct bounds on k itself. Theoretical communication complexity results show that Borda (and other rank-based rules) cannot benefit from the use of top-k voting in the worst-case [Conitzer and Sandholm, 2005], a point to which we return below.

None of the models above consider candidate unavailability. The idea of voting with unavailable candidates was considered by Lu and Boutilier [2010] and Baldiga and Green

Voting rule	Adversarial	IC
Plurality, $n = \mathrm{poly}(m)$	$k = O(\log m)$	$k = O(\log m)$
Plurality, $n = \exp(m)$	$k = \Omega(m)$	$k = \Theta(\log m)$
Borda, $n = \Omega(m^3 \log m)$	$k = \Omega(m)$	$k = \Omega(m/\log m)$

Table 1: Top-k voting: bounds on k

[2011], who study the impact of missing candidates on the fidelity of a winner using voting rules such as Borda, and how close *ranking policies* for selecting winners approximate the true winner. More general querying policies, assuming costly availability tests, were studied by Boutilier *et al.* [2012]. Unavailable candidate models also bear a strong connection to the study of manipulation by candidate addition and deletion [Hemaspaandra *et al.*, 2007; Bartholdi III *et al.*, 1992]. These models do not consider partial preferences. Chevalyre *et al.* [2010] analyze the possible and necessary winner problem under (general) partial preferences, when new candidates are added to an election, for several voting rules, but do not consider elicitation or quantifying the amount of information needed to determine a necessary winner.

Our results. In most of our theoretical bounds, we say that a value of k *produces a correct winner with high probability (w.h.p.)* if the probability that top-k voting returns the true (full profile) winner is $1 - o(1)$, where $o(1) \to 0$ as $m \to \infty$. For plurality, we provide an upper bound of $O(\log m)$ on the k that produces the correct winner w.h.p., if n is polynomial in m, even if the preference profile is selected by an adversary. If n is exponentially larger than m, we show that under impartial culture we require $k = \Theta(\log m)$, while $k = \Omega(m)$ is needed in the worst case. For Borda, we show that for a sufficiently large n (polynomial in m), k is $\Omega(m/\log m)$ under impartial culture, even if $p = 1$; and it has a lower-bound of $k = \Omega(m)$ in the worst case. Our top-k results are summarized in Table 1.

We also consider the case where preferences are distributed according to a Mallows model with reference ranking $\hat{\pi}$. In this model, we provide theoretical results for the special case of $k = 0$, in other words *zero elicitation protocols*. We provide lower bounds on the required number of voters n needed to find winners w.h.p., as a function of φ and m. For plurality, we show that if $n = \Omega(\log m/(1-\varphi)^3)$, then the top candidate in $\hat{\pi}$ is the winner w.h.p. For Borda, we derive a lower bound of $\ln m \cdot \Gamma(\varphi)$ on n, where $\Gamma(\varphi) = (8(1+\varphi)^2(1-\varphi)^3 + (1+\varphi))/(1-\varphi)^7$.

We support our theoretical findings by testing the performance of top-k voting (including the special case of zero elicitation) under varying parameter values (k, n, m, φ). Our empirical results suggest that when the dispersion parameter is bounded away from 1, fairly low values of k are sufficient for correct winner determination.

Space limitations preclude the inclusion of proofs of certain results. Omitted proofs can be found in an extended version of this paper.[1]

3 Top-k Voting and Plurality Scoring

We start with a theoretical analysis of the performance of top-k voting with plurality scoring, assessing the values of k needed to determine the true plurality winner w.h.p. As noted above, if the candidate availability probability p is 1, setting $k = 1$ trivially guarantees correct winner selection. Therefore, in this section we assume that p is a *fixed* probability, bounded away from 1. We distinguish: (a) *worst-case results*, in which an adversarial preference profile is selected to minimize the odds of correct winner selection, and expectations are taken over available candidate sets A; and (b) *average-case results*, in which profiles are drawn from some distribution (e.g., impartial culture), and expectations are taken over both profiles and available sets.

We first show that, even in the worst case, when the number of voters n is "small" relative to the number of candidates m, a small value of k suffices for plurality:

Theorem 1 (Worst-case upper bound, poly. n)**.** *If $n = \mathrm{poly}(m)$, then top-k voting with $k = O(\log m)$ determines the correct plurality winner w.h.p. in the worst case.*

Proof. Consider a vote $\pi \in \mathcal{L}$. Set $k = 2\log n/\log(\frac{1}{1-p})$. The probability that all top-k candidates are unavailable is $1/n^2$. Taking a union bound over all votes, the probability that some vote has all top-k candidates unavailable is $1 - 1/n = 1 - o(1)$. □

Since this $O(\log m)$ upper bound applies in the worst case, it also applies to the average case for any profile distribution. However, in the worst case, having n sub-exponential in m is required if we want a small k.

Theorem 2 (Worst-case lower bound, exp. n)**.** *If $n = \exp(\mathrm{poly}(m))$, top-$k$ voting requires $k = \Omega(m)$ to determine the correct plurality winner w.h.p. in the worst-case.*

Proof. Let $C = \{c_1, \ldots, c_m\} \cup \{a, b\}$, and $p = 1/2$. A key observation is that the unavailable set has size at least $m/2$ with probability very close to $1/2$ (we assume for simplicity that m is even). We create a scenario in which a and b have very close plurality scores, requiring a large value of k to tell which has the higher score. Consider the set $\mathcal{H} = \{S \subseteq C : |S| = m/2\}$ containing all subsets of C of size $m/2$. We show that $k \geq m/2$ is required. Create two sets of votes:

1. V_1: This set ensures a and b have the two highest scores if they are available (which occurs with constant probability, so assume both are). Let $t = 2 \cdot |\mathcal{H}|$, and for a set $S \subseteq C$, let $\mathit{lin}(S)$ be an arbitrary ordering of S. Create $t+1$ copies of $a > \mathit{lin}(C \setminus \{a\})$, and t copies of $b \succ \mathit{lin}(C \setminus b)$. Note: a gets one more vote than b in V_1.
2. V_2: For every $S \in \mathcal{H}$, create two copies of the ranking $\mathit{lin}(S) \succ b \succ a \succ \mathit{lin}(C \setminus (S \cup \{a, b\}))$.

Now, suppose the unavailable set has size at least $m/2$. The plurality score of a is $t+1$, the score of b is at least $t+2$, and so b is the true winner. Otherwise, the score of a is $t+1$, that of b is t, and a is the winner. (All other candidates have score at most t.) If $k \leq m/2$ then the voting scheme doesn't see b in the set V_2, and so it gives incorrect results with probability roughly $p^2/2$. □

[1] See www.cs.toronto.edu/~cebly/papers.html.

Thus, for large n, we must set $k \geq m/2$ in the worst-case. However, under impartial culture, a small value of $k = O(\log m)$ again suffices for the average case:

Theorem 3 (Avg. case upper bound, exp. n)**.** *If $n = \exp(\Omega(m))$, then top-k voting with $k = O(\log m)$ determines the correct plurality winner w.h.p. under impartial culture.*

Proof. Let V be an arbitrary vote profile. Partition V into two sets: $V_1 = \{\pi_i \in V : \text{one of } \pi_i(1), \dots, \pi_i(k) \text{ is available}\}, V_2 = V \setminus V_1$. Let $A \subseteq C$ be the available set, let $\tilde{m} = |A|, n_1 = |V_1|, n_2 = |V_2|$. For $c \in C$, let $sc_1^P(c)$ and $sc_2^P(c)$ be its plurality scores in elections $(V_1, A), (V_2, A)$, respectively. W.l.o.g., order candidates based on $sc_1^P(\cdot)$: $sc_1^P(c_1) \geq sc_1^P(c_2) \geq \cdots \geq sc_1^P(c_{\tilde{m}})$. We prove that c_1 is the true winner w.h.p.

By a simple Chernoff-bound argument, $\frac{m \cdot p}{2} \leq \tilde{m} \leq 2m \cdot p$, w.h.p. Similarly, a simple calculation shows that $\mathbb{E}[n_2] = n \cdot (1-p)^k$, and using a Chernoff bound we obtain $n_2 \leq 2n \cdot (1-p)^k$ w.h.p. Hence, $n_1 \geq n - 2n \cdot (1-p)^k$ w.h.p.

We now give an anti-concentration argument about the difference between the scores according to V_1. We let $D^1_{i,j} = sc_1^P(c_i) - sc_1^P(c_j)$ (we define $D^2_{i,j}$ similarly).

Lemma 4. *$D^1_{1,2} = \Omega(n_1/m^{3.5})$ with high probability.*

Proof. After conditioning on A, consider the votes V_1 sequentially. By a simple balls and bins argument, the difference between the scores of c_i and c_j increases by 1 due to vote π_t $(t = 1, \dots, n_1)$ with probability $1/\tilde{m}$, decreases by 1 with probability $1/\tilde{m}$, and does not change with probability $1 - 2/\tilde{m}$. We can thus treat this change as a random variable X_t, rewriting $D^1_{i,j} = \sum_{t=1}^{n_1} X_t$, where $X_t = 1, X_t = -1$ each with probability $1/\tilde{m}$, and $X_t = 0$ with probability $1 - 2/\tilde{m}$. Then $Var(X_t) = \mathbb{E}[X_i^2] = \frac{2}{\tilde{m}}, \mathbb{E}[D^1_{i,j}] = \frac{2n_1}{\tilde{m}}$, and $\rho = \mathbb{E}[|X_t|^3] = \frac{2}{\tilde{m}}$. The Berry-Esseen Theorem allows us to prove that $D^1_{1,2}$ (and hence $D^1_{1,j}$ for every j s.t. $c_j \in A$) is "large enough."

Lemma 5 (Berry-Esseen [?])**.** *Let $X = X_1 + \cdots + X_n$ be the sum of i.i.d. zero-mean random variables s.t. $\mathbb{E}[X_i^2] = \sigma^2 > 0, \mathbb{E}[|X_i|^3] = \rho < \infty$. Let $F_n(\cdot)$ be the cdf of X, and let $\Phi(\cdot)$ be the cdf of the normal distribution. Then:*

$$\sup_x |F_n(x) - \Phi(x)| < \frac{C\rho}{\sigma^3 \sqrt{n}} \tag{1}$$

where $0 < C \leq 0.4784$.

In our case: $\frac{C\rho}{\sigma^3\sqrt{n}} = \frac{C}{\sqrt{n_1}} \cdot \frac{2}{\tilde{m}} \cdot \left(\frac{\tilde{m}}{2}\right)^{3/2} = C' \cdot \sqrt{\frac{\tilde{m}}{n_1}}$. Hence, we may assume that $D^1_{i,j}$ is effectively given by the normal distribution $\mathcal{N}(0, \sigma^2 = \frac{2n_1}{\tilde{m}})$, which gives us:

$$\begin{aligned} \Pr[|D^1_{i,j}| < t] &\leq \frac{1}{\sqrt{2\pi}} \int_{-t/\sigma}^{t/\sigma} e^{-x^2/2} \mathrm{d}x < \frac{1}{\sqrt{2\pi}} \cdot (2t/\sigma) \\ &= t \cdot \sqrt{\frac{\tilde{m}}{\pi \cdot n_1}} \end{aligned} \tag{2}$$

Setting $t = \frac{\sqrt{n_1}}{\tilde{m}^{3.5}}$ and taking the union bound over all possible pairs (i,j) gives $D^1_{1,2} = \Omega(\frac{n_1}{m^{3.5}})$ with probability $1 - O(1/\tilde{m}) = 1 - o(1)$, where the last equality follows from the concentration bound on $\tilde{m}$. □

A concentration bound on $D^2_{1,j}$ (for all $c_j \in A \setminus \{c_1\}$) follows from a Chernoff bound and a union bound over all j:

$$\Pr[D^2_{1,j} \leq 2\sqrt{\frac{n_2}{\tilde{m}}} \cdot \log m \text{ for all } j \geq 2] = 1 - o(1) \tag{3}$$

We now summarize by showing that, w.h.p., $D^1_{1,2} > D^2_{1,j}$:

$$\frac{\sqrt{n_1}}{\tilde{m}^{3.5}} > \frac{\sqrt{n_2} \cdot \log m}{\sqrt{\tilde{m}}} \tag{4}$$

As $m > \tilde{m}$ and $\sqrt{m} > 1$, it suffices to show:

$$\frac{\sqrt{n - 2n \cdot (1-p)^k}}{m^{3.5}} > \sqrt{2n \cdot (1-p)^k} \cdot \log m \tag{5}$$

The above holds (for n, m sufficiently large) if we set $(1-p)^k = m^{-8}$, which gives $k = O(\log m)$, as required. □

A matching lower-bound shows this upper-bound is tight:

Theorem 6 (Avg. case lower bound)**.** *If $n = \exp(\Omega(m))$, $k = \Omega(\log m)$ is necessary for top-k voting to produce the true plurality winner w.h.p. under impartial culture.*

Proof. The proof is largely symmetric to the proof of the upper-bound. We use the same notation as in the previous proof. We first provide an upper-bound on the difference between the score of the highest-ranking candidate and the second-highest. As before, order C based on their scores in V_1: $sc_1^P(c_1) > sc_1^P(c_2) \dots$ (for completeness, let unavailable candidates have score 0). Also, recall that $D^1_{i,j} = sc_1^P(c_i) - sc_1^P(c_j)$. The following lemma asserts that the top two scores are likely to be close to one another.[2]

Lemma 7. *$D^1_{1,2} = O(\sqrt{\frac{n \log^2 \log m}{m \log m}}) = o(\sqrt{\frac{n}{m}})$ w.h.p.*

Proof. Let $A \subseteq C$ be the available set $(|A| = \tilde{m})$, and partition A into two (roughly) equal size sets: $A_1, A_2 \subset A$, such that $|A_1| = \lfloor \tilde{m}/2 \rfloor, |A_2| = \lceil \tilde{m}/2 \rceil$. Define two random variables: $t_1 = \max_{c \in A_1} sc_1^P(c), t_2 = \max_{c \in A_2} sc_1^P(c)$. It it easy to see that $D^1_{1,2} \leq |t_1 - t_2|$, so we prove the claim by upper-bounding the r.h.s. of the inequality. The number of the votes in V_1 that rank candidates in A_1 (A_2) first is bounded away from $n_1/2$ by $O(\sqrt{n_1})$ w.h.p. So the score of each candidate in A_1 and A_2 is distributed according to a typical balls-and-bins process, in which $n_1/2 \pm o(n_1)$ balls are thrown into $\tilde{m}/2$ bins, at random. Using Thm. 1 of [Raab and Steger, 1998], we have $|t_i - \mathbb{E}[t_i]| = \Theta\left(\sqrt{\frac{n_1 \log \tilde{m}}{\tilde{m}}} \sqrt{(1 - (1+\epsilon)\frac{\log\log\tilde{m}}{2\log\tilde{m}})}\right)$, for $\epsilon > 0$ w.h.p., for $i = 1, 2$. Using our bounds on $\tilde{m}, n_1$, and the approximation $\sqrt{1-x} = 1 - \Theta(x)$, we derive $|t_1 - t_2| = O(\sqrt{\frac{n \cdot \log m}{m} \frac{\log logm}{\log m}}) = O(\frac{n \log^2 \log m}{\log m})$, w.h.p. □

Lemma 8. *Let $k = o(\log m)$. Then $D^2_{2,1} = \Omega(\sqrt{n/m})$ with constant probability.*

[2] We thank Neal Young for the idea of the proof.

The proof is similar to that of Lemma 4. □

To summarize, we see that top-k voting can be very effective for plurality voting with the possibility of unavailable candidates under the impartial culture model, requiring elicitation of only the $O(\log m)$ most-preferred candidates from each voter to ensure the correct winner w.h.p. (this upper bound is tight). If one wants worst-case assurances, this same bound suffices for "small" elections (with a number of voters polynomial in n); but for "large" elections (with an exponential number of voters), top-k voting offers no savings.

4 Top-k Voting and Borda Scoring

We now turn our attention to Borda scoring, and provide similar results. As with plurality, we begin with a worst-case lower bound on k. We note that the following result follows quite directly from a general result on the (deterministic) communication complexity of any rank-based voting rule: Conitzer and Sandholm [2005] show that such rules require $O(nm \log m)$ bits of communication in the worst case (i.e., essentially elicitation of full rankings). However, we provide a direct construction for Borda.

Theorem 9 (Worst case lower bound)**.** *Top-k voting requires $k = \Omega(m)$ to determine the correct Borda winner w.h.p. in the worst-case, even when $p = 1$.*

Proof. Assume for simplicity that $|C|$ is odd and larger than 5. Let A be an available set and $C = \{c\} \cup A$ for some designated candidate c. Let π be an arbitrary ordering of A, and π^r its reverse. Let (π_1, π_2) be a profile with two votes, where π_1 and π_2 are obtained by placing c between the candidates ranked in positions $\frac{m-1}{2} - 1$ and $\frac{m-1}{2}$ in each of π and π^r. If $k = \frac{m-1}{2} - 1$, c will not be the top-k Borda winner, even though it is the true Borda winner—its average score is $\frac{m+1}{2}$, whereas the average score of all other candidates is $\frac{m-1}{2}$. □

We now provide an average-case lower bound on k under the impartial culture assumption.

Theorem 10 (Avg. case lower bound)**.** *If $n = \Omega(m^3 \cdot \log m)$, then $k = \Omega(m/\log m)$ is necessary for top-k voting to produce the true Borda winner w.h.p. under impartial culture, even when $p = 1$.*

Proof. The proof idea will be similar to the proof we give for the plurality voting rule: we will upper bound the *observed* difference in score between the winner according to the top-k voting, and any other candidate. Then, we will show that with constant probability this difference between the score of the winner and that of the candidate with the second highest score is eliminated as a result of discounted votes. Given the *real* Borda scores $sc_i^B(\cdot)$ of the candidates in vote π_i, let $\alpha_i(c) = sc_i^B(c)$ if $sc_i^B(c) \geq m - k$, and $\alpha_i = 0$, otherwise. That is, $\alpha_i(c)$ is the Borda score of c according to top-k voting. Similarly, $\beta_i(c) = sc_i^B(c)$ if $sc_i^B(c) < m - k$ and $\beta_i(c) = 0$ otherwise; i.e., the extra score "missed" due to top-k voting. We let $\alpha(c) = \sum_{i \in N} \alpha_i(c)$ and $\beta(c) = \sum_{i \in N} \beta_i(c)$. Finally, for two distinct candidates $c, c' \in C$, $D^T(c, c') = \alpha(c) - \alpha(c)$, and $D^B(c, c') = \beta(c) - \beta(c')$. We will show that if c and c' are the highest and second highest ranking candidates according to top-k voting, with constant probability, for $o(m/\log m)$, $D^T(c, c') < D^B(c', c)$.

Lemma 11. *If $k = o(m/\log m)$, then for all $c, c' \in C$, $D^T(c, c') = o(\sqrt{n}\frac{m}{\log m})$ with high probability.*

Proof. We prove the lemma via a simple concentration bound. Clearly $\mathbb{E}[D^T(c, c')] = 0$. Bounding the variance:

$$
\begin{aligned}
Var[D^T(c, c')] &= \frac{n}{m(m-1)} \sum_{t_c=0}^{k} \sum_{t_{c'}=0}^{k} (t_{c'} - t_c)^2 \\
&= \frac{n}{m-1} \sum_{t=0}^{k} t^2 - \frac{2}{m(m-1)} \left(\textstyle\sum_{t=0}^{k} t\right)^2 \\
&= n \cdot \frac{(k+1)(4mk^2 + 2mk - 3k^3 - 3k^2)}{6m(m-1)} \\
&= O(\frac{nk^3}{m}) = \frac{\alpha n k^3}{m} \qquad (6)
\end{aligned}
$$

where the last inequality was obtained by assuming that $k = o(m)$.

Now, using the Bernstein inequality, we get that for any $c, c' \in C$:

$$
\begin{aligned}
Pr[|D^T(c, c')| &\leq \sqrt{5 \log m}\sqrt{\frac{\alpha k^3}{m}}] \\
&\leq 2 \exp\left(-\frac{5 \log m \cdot \alpha n k^3/m}{\alpha n k^3 + \frac{m}{3} \cdot \sqrt{\frac{\alpha \log m \cdot n k^3}{m}}}\right) \\
&\leq 2 \exp\left(\frac{5 \log m \cdot \alpha n k^3/m}{2\alpha n k^3/m}\right) = \frac{2}{m^{2.5}} \qquad (7)
\end{aligned}
$$

where the second inequality follows by assuming: $n \geq \frac{m^3 \cdot \log m}{9 \cdot \alpha \cdot k^3}$. The lemma follows by taking the union bound over all $O(n^2)$ pairs (c, c'), and plugging $k = o(k/\log m)$. □

Next, we show that this gap in observed scores of, among other pairs, the highest and second-highest scores , can be closed due to uncounted scores.

Lemma 12. *If $k = O(m/\log m)$ then $D^B(c', c) = \Omega(m\sqrt{n})$ with constant probability, where c and c' are the candidates with the highest and second-highest scores.*

Proof. We prove the lemma as before by showing that the difference $D^B(c', c)$ can be well approximated using the normal distribution.

Claim 13. *If $k = O(m/\log m)$, $Var[D^B(c, c')] = \Omega(m^2 \cdot n)$, with constant probability.*

First, conditioning on the first and second ranking candidates c_1, c_2 (WLOG), we divide V into four sets: $V_1 = \{\pi_i \in V : \pi_i(c), \pi_i(c') \leq k\}$, $V_2 = \{\pi_i \in V : \pi_i(c) \leq k, \pi_i(c') > k\}$, $V_3 = \{\pi_i \in V : \pi_i(c) > k, \pi_i(c') \leq k\}$, $V_4 = \{\pi_i \in V : \pi_i(c), \pi_i(c') > k\}$. Furthermore, we let n_1, n_2, n_3, and n_4 denote the sizes of V_1, V_2, V_3 and V_4, respectively.

By assuming that $k = O(m)$ and using a concentration argument, we get that $n_1 = n_2 = n_3 = n_4 = \Theta(n)$ with high probability. As before, we lower-bound the variance of $D^B(c, c')$.

$$\begin{aligned}\mathbb{E}[(D^B(c,c'))^2] &= \frac{n_2+n_3}{m-k}\sum_{t=0}^{m-k-2} t^2 \\ &+ \frac{n_4}{(m-k)(m-k-1)}\sum_{t_c=0}^{m-k}\sum_{t_{c'}=0}^{m-k}(t_c - t_{c'})^2 \\ &= n_4 \cdot \frac{(m-k)^3-(m-k)}{6\cdot(m-k-1)} \\ &+ \frac{n_2+n_3}{6}\cdot(m-k+1)(2m-2k+1)\end{aligned}$$

Assuming only $k = \beta \cdot m$ and using the concentration bounds for n_2, n_3, and n_4 suffices to get $\mathbb{E}[(D^B(c,c'))^2] = \Theta(m^2 \cdot n)$.

On the other hand, $\mathbb{E}[D^B(c,c')] = (n_2 - n_3) \cdot \frac{m-k}{2}$. As $Var[D^B(c,c')] = \mathbb{E}[D^B(c,c')^2] - \mathbb{E}[D^B(c,c')]^2$, upper bounding $\mathbb{E}[D^B(c,c')]$ will prove the claim. First, notice that $\mathbb{E}[n_2 - n_3] = 0$. Furthermore, notice that $Var(n_2 - n_3) \leq n\frac{k\cdot(m-k)}{m(m-1)}$, as for every vote, the probability that only one of $\{c, c'\}$ is in the top-k is $\frac{k\cdot(m-k)}{m(m-1)}$. Plugging $k = O(m/\log m)$ implies $Var[n_2 - n_3] \leq \beta\frac{n}{\log m}$, w.h.p. By a straightforward tail-bound we get that $|n_2 - n_3| = O(\sqrt{\frac{n}{\log m}})$ with constant probability. Plugging this bound into the term for $\mathbb{E}[D^B(c,c')]$ gives that $Var[D^B(c,c')] = \Omega(m^2 \cdot n)$ with constant probability.

As done for the Plurality rule, we turn to the Berry-Esseen to lower bound $D^B(c,c')$. First, we lower bound the third moment of $D^B(c,c')$.

Claim 14. *The third moment* $\rho = \mathbb{E}[|(D^B(c,c') - \mathbb{E}[D^B(c,c')])^3|] = O(m^3 \cdot n^2)$.

To prove the claim, we note that $|D^B(c,c')| \leq m \cdot n$, as the difference between the Borda score of any two candidates is at most $m \cdot n$. Also, by a slightly stronger tail-bound, we get that $|n_2 - n_3| \leq n\sqrt{m}$, w.h.p., which implies that $\mathbb{E}[D^B(c,c')] = O(n\sqrt{m})$. Expanding the term for ρ, we obtain:

$$\begin{aligned}\rho &= \mathbb{E}[|(D^B(c,c') - \mathbb{E}[D^B(c,c')])^3|] \\ &\leq \mathbb{E}[|D^B(c,c')|^3] + |\mathbb{E}[D^B(c,c')]^3| \\ &+ 3\mathbb{E}[D^B(c,c')^2]\cdot\mathbb{E}[|D^B(c,c')|] \\ &+ 3\mathbb{E}[|D^B(c,c')|]\cdot\mathbb{E}[D^B(c,c')]^2 \qquad (8)\end{aligned}$$

By a simple tail bound, with probability $1 - O(1/m)$, $|D^B(c,c') - \mathbb{E}[D^B(c,c')]| \leq O(\sqrt{\log m} \cdot \sqrt{Var(D^B(c,c'))}) = O(\frac{m\cdot\sqrt{n}}{\log m}\sqrt{\log m}) = O(m\sqrt{\frac{n}{\log m}})$.

So: $\mathbb{E}[D^B(c,c')^3] \leq O(m\sqrt{n/\log m}) + O(\frac{1}{m}\cdot m \cdot n) = O(m\sqrt{n/\log m})$ (we have used the upper bound on the total difference).

Thus:

$$\begin{aligned}\rho &\leq O(m\sqrt{n/\log m}) + O(m\sqrt{n/\log m}) \\ &+ m^3 \cdot n^{1.5} + O(m^3 \cdot n^{1.5}) + O(m^3 \cdot n^2)\end{aligned}$$

Thus, for a normal distribution Φ centered around $\mathbb{E}[D^B(c,c')]$, we have:

$$\begin{aligned}&\sup_x |Pr[D^B(c,c') = x] - \Phi(x)| < \frac{C\rho}{\sigma^3\sqrt{n}} \\ &= \frac{O(m^3\cdot n^2)}{\Theta(m^3\cdot n^{1.5})\cdot\sqrt{n}} = O(\sqrt{1/m}), \text{with constant probability}\end{aligned}$$

Using the properties of the normal distribution, we get that with constant probability $D^B(c,c') = \Omega(m\sqrt{n})$ (as $|\mathbb{E}[D^B(c',c)]| = O(m\sqrt{n})$ and $Var = \Omega(m\sqrt{n})$ with constant probability). □

Combining Lemma 11 and Lemma 12 we get that $D(c,c') = D^B(c,c') + D^T(c,c') < 0$ with constant probability, which proves the theorem. □

To summarize, top-k voting cannot ease the elicitation burden in Borda elections in the worst case. Under impartial culture, there is hope for *some* elicitation savings for elections of reasonable size, as indicated by our lower bound of $k = \Omega(m/\log m)$, which suggests that $O(m/\log m)$ might suffice. But these savings are not nearly as substantial as in the case of plurality, nor are they guaranteed without a matching upper bound. A matching upper bound, or a stronger lower bound—for instance, perhaps our proof could be strengthened to give a lower bound of $\Omega(m)$—is an important result needed to complete the picture regarding Borda under impartial culture. Despite this, we will see below that top-k voting can, in fact, help substantially in Borda voting under other, more realistic preference distributions.

5 Zero-elicitation Protocols

It is widely recognized that the impartial culture assumption does not provide a realistic model of real-world preferences or voting situations [Regenwetter *et al.*, 2006]. For this reason, exploring the ability to limit elicitation under other, more realistic probabilistic models of voter preference is of great import. We consider one such model in this work, namely the Mallows model, since it allows us to generalize the impartial culture model (which is a special case) by simply varying the dispersion or degree of concentration of voter preferences in a natural way. While we do not claim that the Mallows model is an ideal model for all social choice situations (though it serves as an important backbone for mixture models of preferences [Murphy and Martin, 2003; Busse *et al.*, 2007; Lu and Boutilier, 2011a]), it represents an important starting point for the broader investigation of top-k voting.

In this section, we theoretically analyze the special case of *zero elicitation protocols*—that is, top-k voting when we set $k = 0$—under Mallows model distributions. Specifically, we ask how concentrated voter preferences need to be—what dispersion values φ suffice—to ensure that correct plurality and Borda winners can be selected w.h.p. *without eliciting any information from voters*. For ease of presentation, we assume $p = 1$ (i.e., all candidates are available); however, our proofs can be modified to accommodate $p < 1$, using simple applications of Chernoff and union bounds to account for missing candidates. In the next section, we empirically analyze top-k

voting for both zero elicitation and more general values of k under Mallows models.

It is important to recognize that voting is often used for two distinct purposes, aggregation of *preferences* as discussed above, and aggregation of *information* [de Caritat marquis de Condorcet, 1785; Young, 1995]. In the latter case, it is often assumed that some true (objective) *latent* ranking of alternatives gives rise to the reported rankings of voters, with the aim of recovering this latent ranking from the votes (e.g., using some form of maximum likelihood estimation). In such a case, having the mean ranking $\hat{\pi}$ given *a priori* via a Mallows model leaves no reason to actually elicit votes (since the ranking to be estimated is given as input). However, voting does offer value when aggregating preferences: The ranking $\hat{\pi}$ may represent, for example, an ordering of candidates based on some observable characteristic that correlates voter preferences, but does not actually determine them. If our aim is to maximize societal satisfaction using a specific voting rule (as opposed to estimating the objective ranking itself), then preference elicitation is generally needed. Our aim in this section is analyze how concentrated preferences need to be to support preference aggregation with *no elicitation*.

Assume a Mallows model $(\hat{\pi}, \varphi)$ over m candidates C. With no elicitation, the candidate with the expected highest (plurality or Borda) score is obviously the highest ranked candidate $\hat{\pi}(1)$, and it has the highest probability of winning if $\varphi < 1$ (if $\varphi = 1$, all candidates are equally likely to be winners). Under plurality voting, we can show that with a sufficiently large voter population, this approach performs well.

Theorem 15. *If* $n = \Omega\left(\frac{\log m(1-\varphi^m)}{(1-\varphi)^3}\right)$*, then the highest-ranked candidate* $\hat{\pi}(1)$ *is the plurality winner w.h.p.*

Thm. 15 can be proven using the Bernstein inequality and union bound to bound the probability that the highest-ranked candidate in $\hat{\pi}$ is dominated by another.

Proof. Relating to our previous notation, we let $D_i = sc^P(c_1) - sc^P(c_i)$. Also, we assume without loss of generality that the candidates in C are numbered based on their position in the reduced reference ranking $\hat{\pi}$.

It follows from the definition of the Mallows distribution that $Pr[\pi^{-1}(1)| = c_i] = \varphi^{i-1}/Z_m$. Thus, $\mathbb{E}[D_i] = \frac{n(1-\varphi^{i-1})}{Z_m}$, and $\mathbb{E}[D_i^2] = \frac{n(1+\varphi^{i-1})}{Z_m}$.

Now, by the union bound, we have

$$Pr[D_i > 0, \forall i \neq 1] \geq 1 - \sum_{i \neq 1} Pr[D_i < 0]$$

As before, we can treat D_i as the sum of n random variables $X_1, \ldots, X_n$, where for each $i = 1, \ldots, n$, $X_i = 1$ with probability φ^{i-1}/Z_m, $X_i = -1$ with probability $1/Z_m$, and $X_i = 0$ otherwise. By the Bernstein inequality, we have

$$\begin{aligned} Pr[D_i > 0] &\leq \exp\left(-\tfrac{\mathbb{E}[D_i]^2}{2(\mathbb{E}[D_i^2] + \frac{1}{3}\mathbb{E}[D_i])}\right) \\ &\leq \exp\left(-\tfrac{3\mathbb{E}[D_i]^2}{8\mathbb{E}[D_i^2]}\right) \\ &\leq \exp\left(-\tfrac{3n(1-\varphi)(1-\varphi^{i-1})^2}{8(1-\varphi^m)}\right) \\ &\leq \exp\left(-\tfrac{3n(1-\varphi)^3}{8(1-\varphi^m)}\right) \end{aligned}$$

Applying the union-bound gives:

$$Pr[D_i < 0, \forall i > 1] \geq 1 - \exp\left(\log m - \tfrac{3n(1-\varphi)^3}{8(1-\varphi^m)}\right)$$

Solving for n and using the bound on m concludes the proof. □

We can derive a similar bound for Borda voting.

Theorem 16. *If* $n \geq \Gamma(\varphi)\ln m$*, where* $\Gamma(\varphi) = (8(1+\varphi)^2(1-\varphi)^3 + (1+\varphi))/(1-\varphi)^7$*, then the highest-ranked candidate* $\hat{\pi}(1)$ *is the Borda winner w.h.p.*

Sketch of Proof As before, assume without loss of generality that the candidates are numbered according to their rank in $\hat{\pi}$. We make use of the following straightforward lemma:

Lemma 17. *For every* $i \leq m-1$*, and* $1 \leq t_1 < t_2 \leq m$*,* $\varphi \cdot Pr[\pi(t_1) = c_i, \pi(t_2) = c_{i+1}] = Pr_\pi[\pi(t_2) = c_i, \pi(t_1) = c_{i+1}]$

The lemma follows from a simple coupling argument and the definition of the Mallows distribution. As before, we let $D_i = sc^B(c_1) - sc^B(c_i)$. Lemma 20 implies that the expected Borda score $\mathbb{E}[c_i]$ are a non-increasing with i:

Corollary 18. *For every* $2 < i \leq m$*,* $\mathbb{E}[D_i] \geq \mathbb{E}[D_2]$*.*

We proceed as before by bounding the probability that the score of c_1 is lower than the Borda score of some other candidate; i.e., that $D_i < 0$ for some $i > 2$.

We can now bound $\mathbb{E}[D_2]$ and the k'th moment of D_2.

Lemma 19. *The following bounds hold for the expectation* $\mathbb{E}[D_2]$*, the second, and the* k*'th moments of* D_2*:*

1. $\frac{n(1-\varphi)}{1+\varphi} \leq \mathbb{E}[D_2] \leq \frac{n}{1+\varphi}$.
2. $n \leq \mathbb{E}[D_2^2] \leq 2n/(1-\varphi)^2$*; and*
3. $\mathbb{E}[|D_2|^k] \leq k!n/((1-\varphi)^k(1-\varphi^{m-1})(1-\varphi^m))$.

Proof. Using the definition of the Mallows distribution and Lemma 20 we obtain $\mathbb{E}[D_2] = n\frac{\varphi - m\cdot\varphi^m + m\cdot\varphi^{m+2} - \varphi^{2m+1}}{(1+\varphi)(1-\varphi^m)(\varphi-\varphi^m)}$. By observing that the above expectation is nondecreasing in m, setting $m = 2$ and taking the limit $m \to \infty$, we obtain the first part of the lemma. Also, using the definition of the Mallows distribution, we have:

$$\mathbb{E}[|D_2|^k] = \frac{n}{Z_m \cdot Z_{m-1}} \sum_{1 \leq t_1 < t_2 \leq m} (t_2 - t_1)^k (1+\varphi)\varphi^{t_1+t_2-3}$$

$$= \frac{n(1+\varphi)(1-\varphi)^2}{\varphi^3(1-\varphi^{m-1})(1-\varphi^m)} \sum_{d=1}^{m-1} d^k \varphi^d \sum_{t_1=1}^{m-d} \varphi^{2t_1} \tag{9}$$

Applying similar methods to Eq.12 gives parts (2) and (3). □

In order to use Bernstein's inequality, we need to find a constant c such that $\mathbb{E}[|D_2|^k] \leq 0.5 \cdot k!\mathbb{E}[D_2^2] \cdot c^{k-2}$. It can be verified that $c = 2/(1-\varphi)^5$ satisfies this inequality. Now, applying Bernstein's inequality we get:

$$Pr[D_2 < 0] \leq \exp\left(-\tfrac{n(1-\varphi)^7}{4((1-\varphi)^3(1+\varphi)^2 + (1+\varphi))}\right) \tag{10}$$

Applying Corollary 21 and taking the union bound over the m candidates gives the bound on n. □

Notice that our proof implies an even stronger result, that for a sufficiently large population of voters, the entire *ranking* induced by the Borda scores of the candidates corresponds to the reference ranking.

Proof. Our proof implies an even stronger result, that for a sufficiently large population of voters, the ranking induced by the Borda scores of the candidates corresponds to the reference ranking. As before, assume without loss of generality that the candidates are numbered according to their rank in $\hat{\pi}$. We will make use of the following straightforward lemma:

Lemma 20. *For every* $i = 1, \ldots, m-1$, *and* $1 \leq t_1 < t_2 \leq m$, $\varphi \cdot Pr[\pi(t_1) = c_i, \pi(t_2) = c_{i+1}] = Pr_\pi[\pi(t_2) = c_i, \pi(t_1) = c_{i+1}]$

The lemma follows from a simple coupling argument and the definition of the Mallows distribution. As before, we let $D_i = sc^B(c_1) - sc^B(c_i)$. As a corollary, Lemma 20 implies that the expected Borda score $\mathbb{E}[c_i]$ are a non-increasing with i. This gives the following corollary

Corollary 21. *For every* $1 \leq i < j \leq m$, $\mathbb{E}[D_j] \geq \mathbb{E}[D_i]$.

We proceed as before by bounding the probability that the score of c_1 is lower than the Borda score of some other candidate; i.e., that $D_i < 0$ for some $i > 2$. The implication of Corollary 21 is that $Pr[D_i < 0] \leq Pr[D_2 < 0]$ for all $i = 2, \ldots, m$.

We now provide bounds for $\mathbb{E}[D_2]$ and the k'th moment of D_2.

Lemma 22. *The following bounds hold for the expectation* $\mathbb{E}[D_2]$*, the second, and the k'th moments of* D_2*:*

1. $\frac{n(1-\varphi)}{1+\varphi} \leq \mathbb{E}[D_2] \leq \frac{n}{1+\varphi}$

2. $n \leq \mathbb{E}[D_2^2] \leq 2n/(1-\varphi)^2$*, and more generally:*

3. $\mathbb{E}[|D_2|^k] \leq k!n/((1-\varphi)^k(1-\varphi^{m-1})(1-\varphi^m))$.

Proof. By definition,

$$\mathbb{E}[D_2] = n \sum_{t_1, t_2 \in [m]} (t_1 - t_2) Pr[\pi(t_1) = c_1, \pi(t_2) = c_2]$$

$$= n \sum_{1 \leq t_1 < t_2 \leq t_2} (t_2 - t1)(1-\varphi) Pr[\pi(t_1) = c_1, \pi(t_2) = c_2]$$

$$= n(1-\varphi) \sum_{1 \leq t_1 < t_2 \leq m} (t_2 - t_1) \frac{\varphi^{t_1+t_2-3}}{(\sum_{i=0}^{m-2} \varphi^i)(\sum_{i=0}^{m-1} \varphi^i)}$$

$$= n \frac{\varphi - m \cdot \varphi^m + m \cdot \varphi^{m+2} - \varphi^{2m+1}}{(1+\varphi)(1-\varphi^m)(\varphi - \varphi^m)} \tag{11}$$

where the first equality follows from Lemma 20, and the second one follows from the definition of the Mallows distribution. By observing that the above expectation is non-decreasing with the value of m, [3] and plugging $m = 2$, we get the lower-bound on $\mathbb{E}[D_2]$. The upper-bound on $\mathbb{E}[D_2]$ is obtained by taking the limit of $m \to \infty$ in Eq.11.

Now, by definition:

$$\mathbb{E}[|D_2|^k] = \frac{n}{Z_m \cdot Z_{m-1}} \sum_{1 \leq t_1 < t_2 \leq m} (t_2 - t_1)^k (1+\varphi) \varphi^{t_1+t_2-3}$$

$$= \frac{n(1+\varphi)(1-\varphi)^2}{\varphi^3(1-\varphi^{m-1})(1-\varphi^m)} \sum_{d=1}^{m-1} d^k \varphi^d \sum_{t_1=1}^{m-d} \varphi^{2t_1} \tag{12}$$

The lower-bound on the second moment of D_2 is obtained by plugging $k = 2$, and applying the assumption of $m \geq 2$ in Eq.12 (using the monotonicity of the moment of D_2 as done before for the expectation) :

$$\mathbb{E}[D_2^2] \geq \frac{n(1+\varphi)(1-\varphi)^2}{\varphi^3(1-\varphi)(1-\varphi^2)} \varphi^3 = n$$

In order to upper-bound the k'th moment of D_2, we extend the inner and outer sums:

$$\mathbb{E}[|D_2|^k] \leq \frac{n(1+\varphi)(1-\varphi)^2}{\varphi^3(1-\varphi^{m-1})(1-\varphi^m)} \sum_{d=1}^{\infty} d^k \varphi^d \sum_{t_1=1}^{\infty} \varphi^{2t_1} =$$

$$= \frac{n(1+\varphi)(1-\varphi)^2}{\varphi^3(1-\varphi^{m-1})(1-\varphi^m)} \frac{\varphi^2}{1-\varphi^2} \sum_{d=1}^{\infty} d^k \varphi^d$$

$$= \frac{n(1-\varphi)}{\varphi(1-\varphi^{m-1})(1-\varphi^m)} \sum_{d=1}^{\infty} d^k \varphi^d$$

$$= \frac{n(1-\varphi)}{\varphi(1-\varphi^{m-1})(1-\varphi^m)} \sum_{d=1}^{\infty} k! \binom{d+k-1}{k} \varphi^d$$

$$= \frac{n(1-\varphi)}{\varphi(1-\varphi^{m-1})(1-\varphi^m)} \cdot \frac{k!\varphi}{(1-\varphi)^{k+1}}$$

$$= \frac{k!n}{(1-\varphi)^k(1-\varphi^{m-1})(1-\varphi^m)} \tag{13}$$

Plugging $k = 2$ and taking the limit $m \to \infty$ gives the upper-bound on the second moment. □

In order to use Bernstein's inequality, we need to find a constant c such that

$$\mathbb{E}[|D_2|^k] \leq 0.5 \cdot k!\mathbb{E}[D_2^2] \cdot c^{k-2}$$

It can be verified that $c = 2/(1-\varphi)^5$ satisfies this inequality. Now, applying Bernstein's inequality we get:

$$Pr[D_2 < 0] \leq \exp\left(-\frac{\mathbb{E}[D_2]^2}{2(\mathbb{E}[D_2^2] + c\mathbb{E}[D_2])}\right)$$

$$\leq \exp\left(-\frac{n^2(1-\varphi)^2/(1+\varphi)^2}{2(\frac{2n}{(1-\varphi)^2} + \frac{2}{(1-\varphi)^5} \cdot \frac{n}{1+\varphi})}\right)$$

$$= \exp\left(-\frac{n(1-\varphi)^7}{4((1-\varphi)^3(1+\varphi)^2 + (1+\varphi))}\right) \tag{14}$$

Applying Corollary 21 and taking the union bound over the m candidates gives the bound on n. □

6 Empirical Results

The bounds above provide some theoretical justification for the use of top-k voting; however, they do not prescribe precise values for the choice of k with respect to specific priors

[3] This observation can be easily proved using the Repeated Insertion model **CITE DOIGNON here**, which is equivalent to the Mallows distribution.

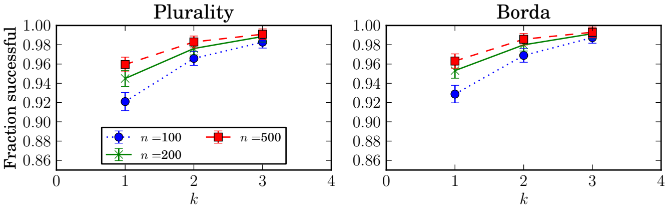

Figure 1: Correctness of top-k voting: $m = 10$, varying k and n.

and election sizes (m, n). In this section we present simulation results for small elections with $m = 10$ candidates and $n = 100$ to 5000 voters to illustrate the probability of correct winner selection in both plurality and Borda elections using top-k voting for several values of k (including zero elicitation), under Mallows models with a range of dispersion value φ. In our experiments, we generate 10,000 random preference profiles for each parameter setting by drawing voter rankings i.i.d. from the appropriate Mallows model, and measure the fraction of such profiles in which top-k delivers the true winning candidate. We assume a candidate availability probability of $p = 0.5$ throughout, except for results concerning zero-elicitation (in which all candidates are available).

In all tests of top-k voting with dispersion $\varphi < 0.7$, winner prediction was essentially perfect, even with $k = 1$, regardless of the other parameters. As a consequence, we focus our discussion on values of $\varphi \geq 0.7$. Fig. 1 shows the *success rate* (i.e., rate of correct winner selection) of top-k voting for both plurality and Borda voting, with $\varphi = 0.7$ and $m = 10$, as we vary k and the number of voters. In all cases top-k converges to the correct prediction, and is near-perfect when $k = 3$. With a greater number of voters, performance is better, but the dependence is slight and almost negligible for $k = 3$.

To analyze zero-elicitation, we measured how often the first-ranked candidate in the mallows reference ranking (i.e., the winner under zero-elicitation) is the true election winner under both plurality and Borda voting. We set $m = 10$, and assume $p = 1$ for simplicity. We vary φ and n, and show results averaged over 10,000 elections for each setting in Fig. 2. For $\varphi \leq 0.8$, predictions are near-perfect for $n \geq 700$; and with $\varphi \leq 0.7$, $n \geq 400$ suffices for near 100% accuracy. We note that results are better for Borda than for plurality. For populations with an extremely high degree of dispersion ($\varphi = 0.9$), the success rate for plurality is only 0.8 at $n = 1000$, and the Borda success rate is only 0.92. This is consistent with the trends suggested by our theoretical bounds in the sense that the success probability depends exponentially on φ, which means that it decreases dramatically for larger values of φ.

We also measured how frequently the *entire societal ranking* induced by plurality or Borda voting corresponds to the Mallows reference ranking $\hat{\pi}$. This measures the extent to which $\hat{\pi}$, hence zero-elicitation, accurately reflects the entire societal preference ranking (not just the winner at the top of the ranking). Results are depicted in Fig. 3. Unsurprisingly, the probability of complete ranking accuracy is significantly lower than the probability with which zero-elicitation correctly forecasts just the winner. However, with $n = 1000$,

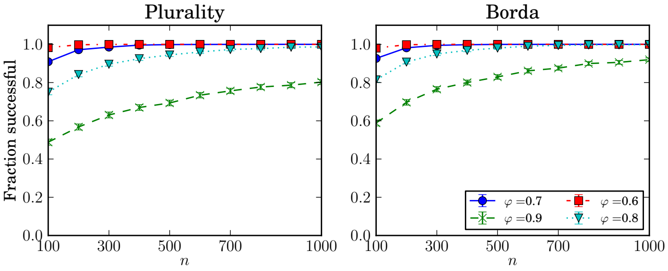

Figure 2: Correctness of zero-elicitation: $m = 10$, varying n, φ.

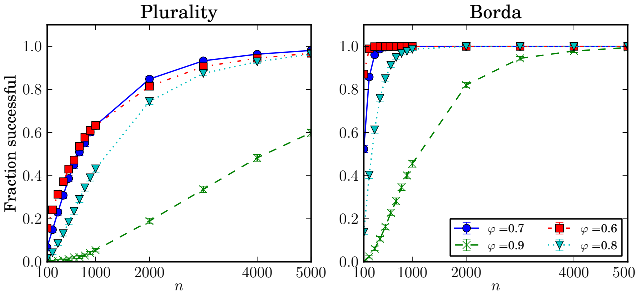

Figure 3: Reconstruction rates of zero-elicitation: $m = 10$, varying n, φ.

almost perfect reconstruction is achieved for Borda scoring when $\varphi \leq 0.8$. Notice that the difference between plurality and Borda is even more pronounced than in winner prediction. Under Borda, $n = 5000$ suffices for accurate assessment of the entire ranking with zero-elicitation, even for $\varphi = 0.9$, while for plurality, results for $\varphi = 0.9$ are much worse (about 0.6), and even for $\varphi = 0.7$ do not reach 100%.

7 Conclusions

We have provided a detailed analysis of top-k voting, allowing for the possibility of unavailable candidates, for both plurality and Borda voting. Our theoretical results place bounds (in some cases tight) on the required values of k needed to determine the correct winner w.h.p., in both a worst-case sense and an average-case sense under impartial culture. We also derived conditions under which zero-elicitation admits correct winner prediction using Mallows models. Our empirical results further demonstrate that relatively small values of k work very well in practice. Even zero-elicitation shows strong promise when preferences exhibit only mild degrees of correlation in elections with a sufficient number of voters.

There are a number of interesting directions for future research. Extending our analysis to other voting rules is of great interest. For example, preliminary results suggest that Copeland exhibits behavior similar to Borda, requiring large k for impartial culture; do certain voting rules exhibit behavior that is intermediate between plurality and Borda? Extending our analysis to a richer class of realistic preference distributions, such as the Plackett-Luce model, or Mallows mixtures, is an important next step, as is testing our approach on real data sets.

A third direction is the investigation of multi-round elicitation protocols [Lu and Boutilier, 2011c], where voting data is elicited in stages, and the protocol terminates when the winner can be determined with high probability. Such protocols are adaptive and dynamic, eliciting information in a given stage conditioned on information gleaned in earlier stages. An important question is whether it is possible to elicit less information on average with such a protocol.

Acknowledgements

Thanks to Neal Young and the anonymous reviewers for helpful suggestions. Boutilier acknowledges the support of NSERC.

References

[Baldiga and Green, 2011] Katherine A. Baldiga and Jerry R. Green. Assent-maximizing social choice. *Social Choice and Welfare*, pages 1–22, 2011. Online First.

[Bartholdi III *et al.*, 1992] John Bartholdi III, Craig Tovey, and Michael Trick. How hard is it to control an election? *Social Choice and Welfare*, 16(8-9):27–40, 1992.

[Boutilier *et al.*, 2012] Craig Boutilier, Jérôme Lang, Joel Oren, and Héctor Palacios. Robust winners and winner determination policies under candidate uncertainty. In *Proceedings of the Fourth International Workshop on Computational Social Choice (COMSOC-2012)*, Kraków, Poland, 2012.

[Busse *et al.*, 2007] Ludwig M. Busse, Peter Orbanz, and Joachim M. Buhmann. Cluster analysis of heterogeneous rank data. In *Proceedings of the Twenty-fourth International Conference on Machine Learning (ICML-07)*, pages 113–120, 2007.

[Chevaleyre *et al.*, 2010] Yann Chevaleyre, Jérôme Lang, Nicolas Maudet, and Jérôme Monnot. Possible winners when new candidates are added: The case of scoring rules. In *Proceedings of the Twenty-fourth AAAI Conference on Artificial Intelligence (AAAI-10)*, pages 762–767, Atlanta, GA, 2010.

[Conitzer and Sandholm, 2005] Vincent Conitzer and Tuomas Sandholm. Communication complexity of common voting rules. In *Proceedings of the Sixth ACM Conference on Electronic Commerce (EC'05)*, pages 78–87, Vancouver, 2005.

[Conitzer, 2009] Vincent Conitzer. Eliciting single-peaked preferences using comparison queries. *Journal of Artificial Intelligence Research*, 35:161–191, 2009.

[de Caritat marquis de Condorcet, 1785] Jean Antoine Nicolas de Caritat marquis de Condorcet. *Essai sur l'Application de l'Analyse a la Probabilite des Decisions rendues a la Probabilite des Voix*. Paris: L'Imprimerie Royale, 1785.

[Ding and Lin, 2012] Ning Ding and Fangzhen Lin. Voting with partial information: Minimal sets of questions to decide an outcome. In *Proceedings of the Fourth International Workshop on Computational Social Choice (COMSOC-2012)*, Kraków, Poland, 2012.

[Hemaspaandra *et al.*, 2007] Edith Hemaspaandra, Lane Hemaspaandra, and Jörg Rothe. Anyone but him: The complexity of precluding an alternative. *Artificial Intelligence*, 171(5-6):255–285, 2007.

[Kalech *et al.*, 2011] Meir Kalech, Sarit Kraus, Gal A. Kaminka, and Claudia V. Goldman. Practical voting rules with partial information. *Journal of Autonomous Agents and Multi-Agent Systems*, 22(1):151–182, 2011.

[Konczak and Lang, 2005] Kathrin Konczak and Jérôme Lang. Voting procedures with incomplete preferences. In *IJCAI-05 Workshop on Advances in Preference Handling*, pages 124–129, Edinburgh, 2005.

[Lu and Boutilier, 2010] Tyler Lu and Craig Boutilier. The unavailable candidate model: A decision-theoretic view of social choice. In *Proceedings of the Eleventh ACM Conference on Electronic Commerce (EC'10)*, pages 263–274, Cambridge, MA, 2010.

[Lu and Boutilier, 2011a] Tyler Lu and Craig Boutilier. Learning Mallows models with pairwise preferences. In *Proceedings of the Twenty-eighth International Conference on Machine Learning (ICML-11)*, pages 145–152, Bellevue, WA, 2011.

[Lu and Boutilier, 2011b] Tyler Lu and Craig Boutilier. Robust approximation and incremental elicitation in voting protocols. In *Proceedings of the Twenty-second International Joint Conference on Artificial Intelligence (IJCAI-11)*, pages 287–293, Barcelona, 2011.

[Lu and Boutilier, 2011c] Tyler Lu and Craig Boutilier. Vote elicitation with probabilistic preference models: Empirical estimation and cost tradeoffs. In *Proceedings of the Second International Conference on Algorithmic Decision Theory (ADT-11)*, pages 135–149, Piscataway, NJ, 2011.

[Mallows, 1957] Colin L. Mallows. Non-null ranking models. *Biometrika*, 44:114–130, 1957.

[Marden, 1995] John I. Marden. *Analyzing and modeling rank data*. Chapman and Hall, London, 1995.

[Murphy and Martin, 2003] Thomas Brendan Murphy and Donal Martin. Mixtures of distance-based models for ranking data. *Computational Statistics and Data Analysis*, 41:645–655, January 2003.

[Raab and Steger, 1998] Martin Raab and Angelika Steger. "balls into bins" - a simple and tight analysis. In *Proceedings of the Second International Workshop on Randomization and Approximation Techniques in Computer Science*, RANDOM '98, pages 159–170, London, UK, UK, 1998. Springer-Verlag.

[Regenwetter *et al.*, 2006] Michel Regenwetter, Bernard Grofman, A. A. J. Marley, and Ilia Tsetlin. *Behavioral Social Choice: Probabilistic Models, Statistical Inference, and Applications*. Cambridge University Press, Cambridge, 2006.

[Xia and Conitzer, 2008] Lirong Xia and Vincent Conitzer. Determining possible and necessary winners under common voting rules given partial orders. In *Proceedings of the Twenty-third AAAI Conference on Artificial Intelligence (AAAI-08)*, pages 202–207, Chicago, 2008.

[Young, 1995] Peyton Young. Optimal voting rules. *Journal of Economic Perspectives*, 9:51–64, 1995.