

**A Maximum Likelihood Approach For Selecting Sets of Alternatives**

Ariel D. Procaccia and Sashank J. Reddi and Nisarg Shah

School of Computer Science, Carnegie Mellon University.

---

**RESULTS I: INCLUDE-TOP**

**Objective:**

- Select a subset of size \( k \) that maximizes the chances of including the best alternative (which is at the top in the true order \( \sigma^* \)).

**Applications:**

- New Product Development (NPD) - shortlisting a few designs through a market survey.
- Cloud Computing - choosing nodes for speculative execution.
- Crowdsourcing.

**Why simple majority does not work?**

- \( k = 2 - \tau \) and \( \tau \) are selected.
- \( \tau \) is more likely to be the best!

**How to select the optimal \( k \)-subset of candidates?**

- **Theorem:** INCLUDE-TOP is \( NP \)-hard for both noisy orders and noisy comparisons (and hence for noisy choice model) for any \( 1 \leq k \leq m - 1 \).
- The hardness result uses the case when \( y = 0 \), i.e., when there is little noise. But in that case, we don’t need the best because any reasonable rule would work well, if not optimally.
- For the other extreme when there is high noise, it turns out that we can indeed perform optimally in polynomial time.

- **Extended Scoring Method:**
  \[ SC(\sigma) = \sum_{a \neq b} \text{number of pairwise wins} \]

- **Theorem:** When there is high noise (\( y \) is sufficiently large), top \( k \) candidates according to ESM maximize the chances of including the best alternative.
- For noisy orders, reduces to a famous rule known as the Borda count.
- **Q:** Okay, we included the best. What about the other alternatives chosen? Are they very bad? Intuition says NO. If they were likely to be the best, they must be good themselves. Turns out, that is exactly right!

**RESULTS II: TOP-SUBSET**

**Objective:**

- Select a subset of size \( k \) that maximizes the chances that it is the subset of top \( k \) candidates in the true order \( \sigma^* \).
- Appears in practice, for example, in team building.
- Also \( NP \)-hard for both noisy orders and noisy comparisons for any \( 1 \leq k \leq m - 1 \).
- When there is high noise (\( y \) is sufficiently large), top \( k \) candidates according to ESM also form the optimal \( k \)-subset for the TOP-SUBSET objective.

**Implications:**

- Top \( k \) candidates in the ESM do not only ensure that the top candidate is included with high probability, but also ensure that the other candidates are very likely to be the next \( k \) best candidates.

**RESULTS III: TOP-TUPLE**

**Objective:**

- Select an ordered tuple of size \( k \) that maximizes the chances that it is the \( k \)-prefix of the true order \( \sigma^* \).
- **Q:** Why is it not the same as TOP-SUBSET? 5th of (likelihoods of various orders).
- Appears in practice, for example, in selecting a committee (e.g., president, vice-president etc.) and in combining search results from different search engines.
- As before, the objective is \( NP \)-hard for \( 1 \leq k \leq m \) and is tractable when the noise is sufficiently high.
- For \( k=1 \), the objective reduces to finding the candidate which is most likely to be the best candidate and for \( k=m \), it reduces to finding an HLE for the true order.
- Thus for noisy orders with high noise, it reduces to Borda count for \( k=1 \) and Kemeny rule for \( k=m \). In fact, the optimal rules for various values of \( k \) in this case form a continuum between the two voting rules.

**Scored Tuples Method:**

- \( SC(n_1, n_2, \ldots, n_k, D) = \sum_{i=1}^{k} SC(n_i) - \sum_{i<j} \epsilon(n_i, n_j, D) \)
- Choose the \( k \)-tuple that maximizes the score.

---

**DISTRIBUTIONS OF VOTES**

**MODEL 1:** Noisy Comparisons

- \( n \) preferences between each pair of alternatives (a,b).
- \( a \preceq b \) with probability \( p \)
- \( a \succ b \) with probability \( 1-p \)

**MODEL 2:** Noisy Orders (also, Mallows’s Model / Condorcet Noise Model)

- No total orders among the alternatives.
- Probability of an order decreases exponentially as the distance from \( \sigma^* \) increases.
- Pairwise Distance (Alex, Kendall’s Tau Distance)

**Our Generalization: Noisy Choice Model**

- Any preference dataset \( O \) and its distribution \( f \) that satisfy the following properties.
- Need not be total orders or pairwise preferences.
- \( n_m := n \) in \( O \).

**Properties:**

1. \( n_a + n_b = n \), for every \( a \neq b \).
2. \( P(\sigma|O) = \frac{1}{\sum_{\sigma} f(\sigma)} \cdot \prod_{a \neq b} \frac{n_a + n_b}{n} \).
- In the noise level, \( y \rightarrow 0 \) represents completely noise free distribution and \( y = 1 \) represents the uniform distribution.

---

**EXTENSIONS**

- **High Likelihood** (instead of “Maximum Likelihood”)
  - Instead of maximizing the probability of including the top candidate, just ensure that the probability is sufficiently high.
  - More candidates need to be included in the subset to increase the probability.
- **Question:** How many candidates need to be chosen?
  - Preliminary results available for Extended Scoring Method.

**Solving the objectives with high probability**

- Selecting the optimal \( k \)-subset for INCLUDE-TOP is \( NP \)-hard but it can be computed with high probability.
- This approach relies on sampling from an extension of Mallows’s model known as the Generalized Mallows’s Model.
- **Open Question:** Can sampling from the Generalized Mallows’s Model be done in polynomial time?

**Data Allocation**

- Current method: input dataset is assumed to be given and a subset of alternatives is selected optimizing certain objective functions.
- Consider an extension where there are multiple users available with different precision level and there is a limit on the number of evaluations that each user can perform.
- Two processes: allocation (static or dynamic) of alternatives to the users to get the dataset and combining these evaluations to compute the answer.
- **Open Question:** How to design a mechanism that performs both functions in a way that jointly optimizes the objective function?