Deliberation via Matching

Anonymous Authors

Contents

Abstract

We study deliberative social choice, where voters refine their preferences through small-group discussions before collective aggregation. We introduce a simple and transparent *deliberation via matching* protocol: for each pair of candidates, we form an arbitrary maximum matching among voters who disagree on that pair, and each matched pair deliberates. The resulting preferences are then appropriately weighted and aggregated using the weighted uncovered set tournament rule.

Within the metric distortion framework, our protocol achieves distortion 3, hence breaking the lower bound of 3.11 for tournament rules without deliberation and matching the lower bound for deterministic rules without deliberation. Beyond this quantitative improvement, our main contribution is a novel *geometric analysis* of deliberation: Evaluating distortion naturally reduces to a non-linear program, and we give an explicit characterization of its worst-case instances by showing they remain vertices of the feasible polytope of instances. This characterization yields a closed-form proof of the distortion bound, avoiding the numerical optimization required in prior work, and provides a general analytical framework for studying the distortion of other deliberative protocols.

1 Introduction

Collective decision-making lies at the core of both democratic governance and algorithmic social choice. Classical models assume that voters possess fixed, exogenous preferences over a set of alternatives, which are then aggregated through a social choice rule. Yet in practice, preferences are rarely static: individuals deliberate, exchange arguments, and frequently revise their views in response to others' reasoning. A large body of research in deliberative democracy—most notably *deliberative polling* and *citizens' assemblies* pioneered by Fishkin and colleagues—shows that when individuals are given balanced information and structured opportunities for discussion, their policy preferences can shift substantially and become more informed [15, 19]. These findings underscore a broader normative intuition: collective decisions should emerge from public reasoning rather than isolated votes.

At the same time, empirical work indicates that *deliberation is most effective in small groups*. Large assemblies or unstructured online forums often suffer from coordination challenges, conformity pressures, and polarization effects, where participants reinforce existing biases [6, 13, 25]. In contrast, small, balanced groups promote reasoned exchange and opinion updating [19, 16], while maintaining manageability and diversity of perspectives. Beyond these empirical considerations, small-group deliberation is also more *practical* in large-scale settings: it can be implemented in parallel, either through many simultaneous discussions among pairs or triads of participants, or via automated or AI-assisted mediators [5, 21, 1, 8, 7]. These advantages motivate theoretical models that capture the benefits of *structured, small-group deliberation* rather than full-group discussion.

Recent theoretical work has begun to formalize this intuition [14, 18]. In these models, voters engage in local discussions that modify their ordinal preferences, and the resulting rankings are then aggregated using a social choice rule. Such frameworks allow us to ask a fundamental algorithmic question:

Can structured, small-group deliberation provably improve the efficiency of collective decisions?

We study this question through the lens of the *metric distortion* framework [2], a quantitative model for evaluating the efficiency of social choice rules. In this framework, both voters and alternatives are embedded in an unknown metric space that captures their underlying preferences: Voters prefer alternatives that are closer to them in this latent geometry. A social choice rule, which only observes voters' ordinal rankings over alternatives, selects a single winner. The *distortion* of a rule measures how far the chosen winner can be, in the worst case, from the welfare-optimal alternative that minimizes the total distance to all voters. Thus, a smaller distortion indicates a decision rule that better preserves social welfare despite having only ordinal information.

Within this setting, it is known that any deterministic rule must incur a distortion of at least 3 [2, 24, 17, 22]. A prominent and well-studied subclass of such rules are *tournament rules*, which base their decision on the outcomes of pairwise contests between alternatives. Tournament rules are appealing not only for their analytical simplicity but also for their low *cognitive complexity*: Voters need only compare two alternatives at a time rather than rank all options simultaneously. However, any tournament rule (that only uses pairwise information about canddidates) has a lower bound of 3.11 on distortion [11], slightly worse than the deterministic optimum. This motivates the following question:

Can small-group deliberation, where voters refine their pairwise comparisons through discussion, improve the distortion of tournament rules while preserving their simplicity?

The recent work of [18] provided the first affirmative answer for *three-person deliberation*. In their model, when a small group of voters deliberates, they collectively choose between any two alternatives by favoring the one with the smaller *average distance* in the latent metric, that is, the alternative closer to the group's barycenter. Aggregating the resulting pairwise outcomes through the well-known Copeland

tournament rule [27, 23], they showed that such *averaging-based deliberation* with groups of size at least 3 can achieve distortion strictly better than 3, thereby surpassing the lower bounds for both tournament and general social choice rules without deliberation. This result established that structured local deliberation can provably improve social welfare. However, their analysis relied on solving a high-dimensional non-convex program numerically, and importantly, left open both the analytical tractability and the effectiveness of *two-person deliberation*, the smallest and most practical form of discussion.

Deliberation via Matching. In this paper, we propose a simpler and more natural model of deliberation based on *pairwise discussions* (groups of size 2). Our protocol, called *deliberation via matching*, proceeds as follows. For every pair of candidates (X, Y), we form a maximum matching among voters who disagree on their relative ranking, and each matched pair deliberates. The result of each deliberation updates their pairwise preference according to the sum of the latent d in the underlying metric. These refined pairwise preferences are then aggregated using the *weighted uncovered set* tournament rule [24, 20]. A scalar parameter $w > 0$ controls the influence of deliberation: each matched pair contributes weight w to its joint outcome, while unmatched individual votes retain unit weight. This protocol differs from prior models [18] that required all groups of a fixed size to deliberate, in that (i) the protocol is simpler to state and analyze, and (ii) it allows precise control over how individual votes and pairwise deliberations are weighted when constructing the tournament graph.

Within the metric distortion framework, we prove the following main theorem:

Theorem 1.1 (Informal Main Theorem)**.** *The deliberation-via-matching protocol with pairwise (two-person) deliberation achieves a metric distortion of* 3.

This bound breaks the 3.11 lower bound for tournament rules without deliberation and matching the deterministic optimum of 3 for any social choice rule without deliberation, showing that even the minimum group size in deliberation provably improves the distortion of tournament rules.

Technical Contribution. Beyond the quantitative bound, our main technical contribution is to develop a novel analytical framework for reasoning about deliberation. As pointed out in [18], the key difficulty with analyzing deliberative protocols is that the distortion is the solution to a non-linear and non-convex program over the distribution of voter–candidate distances. This is in contrast to classical social choice, where distortion is often the solution to a linear program. For the deliberation-via-matching protocol, we prove that all extremal instances (that maximize distortion) are vertices of the convex hull of the non-convex set of feasible instances. This step is non-trivial, since not only are the constraints non-convex, but the distortion objective is as well. Each extremal vertex now admits a clean geometric interpretation: Voters occupy only three distinct locations in the latent metric, and the worst-case matching pairs voters in a simple, deterministic pattern. This reduction yields an exact, closed-form proof of the distortion bound and replaces the numerical optimization used in prior work [18] with a scalable and interpretable geometric analysis.

As a warm-up, we analyze the special case with only two alternatives. In this setting, the deliberation-via-matchings rule admits a direct and elegant analysis: by pairing voters who disagree and letting each pair support the alternative with the smaller total distance, we show that any winner must be backed, in effect, by at least two-thirds of the electorate. This immediately yields a distortion bound of 2, improving upon the classic bound of 3 for deterministic rules without deliberation in this case. Our analysis for multiple alternatives builds on this argument. Indeed, the tournament rules we build on [24, 20] needs bounding the distortion for 3 candidates, where each pairwise comparison between candidates behaves like the two-alternative case. The resulting optimization problem can thus be viewed as a higher-dimensional analogue of the two-candidate distortion bound. However, this optimization becomes significantly more non-trivial and needs a careful construction and proof of the extremal solutions, as discussed above.

Taken together, our results suggest that small-group deliberation can be both *powerful* and *tractable*: Even minimal pairwise interactions suffice to improve the performance of the well-studied tournament rules. More broadly, our geometric characterization provides a new methodological foundation for analyzing deliberative extensions of social choice mechanisms.

1.1 Related Work

Our work lies at the intersection of metric distortion, deliberative social choice, and sampling-based decision mechanisms. We briefly touch on the most relevant lines of research.

Metric Distortion and Tournament Rules. The metric distortion framework was introduced by Anshelevich et al. [2], building on earlier work by Procaccia and Rosenschein [26], to study how well deterministic voting rules can approximate the social optimum when only ordinal information is available. They showed that the Copeland rule has distortion at most 5, and that no deterministic rule can achieve distortion below 3. Later work tightened the upper bound to 3 via novel social choice rules such as the *matching uncovered set* [17] and *plurality veto* [22]. For randomized voting rules, the work of [4] showed a lower bound of 2. This lower bound was subsequently improved to 2.11 by [10]. An upper bound of 3 follows from random dictatorship [4], and this was improved to 2.74 in [12]. We refer the reader to [3] for a survey.

A particularly important subclass of deterministic voting rules are *tournament rules*, which make decisions based solely on the outcomes of pairwise majority contests between candidates. Tournament rules are attractive because they rely only on pairwise comparisons, requiring voters to reason about two alternatives at a time. However, despite their simplicity, tournament rules are fundamentally limited. On the upper side, Munagala and Wang [24] and Kempe [20] defined a class of *weighted tournament rules*, in which pairwise majority margins are aggregated with varying strengths. Their specific rule, the weighted uncovered set, achieves distortion at most $2 + \sqrt{5} \approx 4.236$. Subsequently, Charikar et al. [11] gave a construction with an improved upper bound of 3.94, and proved a lower bound of approximately 3.11 on the distortion of any deterministic tournament rule. This results shows that, in the absence of deliberation, tournament rules cannot match the optimal distortion of 3 achievable by general deterministic mechanisms.

Our work revisits this barrier through the lens of deliberation. We show that by allowing pairs of voters to refine their comparisons before aggregation, a tournament-based rule can in fact achieve the optimal distortion bound of 3, thereby escaping the classical 3.11 limit for non-deliberative tournaments.

Deliberative Social Choice and Sortition. The idea that discussion can improve collective decisions has a long pedigree in political philosophy and deliberative democracy, for example through Fishkin's deliberative polling and citizens' assemblies [15, 19]. Many deliberative systems in practice use *sortition*—random sampling of participants into discussion bodies—to reduce biases and improve legitimacy.

Several theoretical models of deliberation within the metric distortion framework have been recently proposed. Caragiannis et al. [9] examine models of sortition where a large random sample of voters deliberates to compute a consensus or median point, achieving logarithmic (in the number of alternatives) bound on the sample size required to attain distortion arbitrarily close to one. However, these approaches typically assume a single large deliberative body, which raises issues of coordination and bias in practice. In contrast, our focus is on *small-group deliberation* rather than sortition. Here, Fain et al. [14] studied a two-person bargaining model under metric preferences, while Goel, Goyal, and Munagala [18] proposed a general model in which voters deliberate in groups of size up to k, using an averaging rule over the latent metric before aggregating via a tournament rule. They showed that groups of size 3 suffice to beat the classical deterministic distortion bound, with an analysis that relied on numerical optimization.

Our approach differs in three key ways: (i) we use only two-person deliberation, the smallest possible interaction, (ii) the protocol precisely controls how to weigh individual votes versus the outcome of delib-

erations, and (iii) our analysis is fully geometric and explicit, not reliant on black-box numerical solvers. Because two-person deliberation is more feasible, both cognitively and in distributed implementations, our protocol provides a more natural and scalable path toward integrating deliberation into social choice.

2 Preliminaries

We begin by reviewing the metric distortion framework and the class of tournament rules used in our analysis, following [2, 24, 18].

Metric Distortion Framework. Let $C = \{c_1, \ldots, c_m\}$ denote a finite set of m candidates (alternatives), and let V denote a finite set of n voters. Each voter $v \in V$ has a ranking over the candidates that is *consistent* with an underlying latent metric space $(\mathcal{M}, d)$ that contains both voters and candidates as points. If v ranks candidate X higher than Y, then $d(v, X) \leqslant d(v, Y)$. The metric d is not known to the social planner, who only observes the ordinal rankings induced by it. For any two candidates $X, Y \in C$, let XY denote the set of voters who prefer X to Y, with cardinality $|XY|$. Should ties exist, i.e., $d(v, X) = d(v, Y)$, we handle them in any consistent way that counts each tied voter toward exactly one of XY, YX.

For any candidate $X \in C$, define its *social cost*

$$SC(X) \;=\; \sum_{v \in V} d(v, X),$$

that is, the total distance from all voters to c. Let $X^* = \arg\min_{X \in C} SC(X)$ denote the socially optimal (1-median) alternative. Given a social choice rule $\mathcal{S}$ that maps the profile of rankings to a winning candidate $\mathcal{S}(V)$, the *distortion* of $\mathcal{S}$ is defined as

$$\text{Distortion}(\mathcal{S}) \;=\; \sup_{d \text{ consistent}} \frac{SC(\mathcal{S}(V))}{SC(X^*)}.$$

A smaller distortion indicates that $\mathcal{S}$ achieves better welfare despite only knowing ordinal information.

Tournament Rules. A *tournament graph* on the candidates is a complete directed graph, with weights $f(XY) \in [0, 1]$ for each directed edge $X \to Y$, so that for every pair of candidates (X, Y), we have $f(XY) + f(YX) = 1$. In the setting without deliberation, $f(XY)$ represents the fraction of voters that prefer X over Y; however, the weights we construct later will also reflect the outcome of deliberation. A tournament rule takes such a weighted graph as input and outputs the winning candidate.

Among many tournament-based social choice rules, we focus on the *λ-weighted uncovered set (WUS)* rule of [24, 20]. Given a tournament with edge weights $f(XY) \in [0, 1]$, define that a candidate X *λ-covers* candidate Y if either

1. $f(XY) \geq 1 - \lambda$, or
2. there exists a third candidate Z such that $f(XZ) \geq 1 - \lambda$ and $f(ZY) \geq \lambda$.

The *λ-weighted uncovered set* WUS_λ consists of all candidates not λ-covered by any other. It is known that for $\lambda \in [1/2, 1)$, WUS_λ is nonempty [24]. Further, if $f(XY)$ is the fraction of voters that prefer X to Y (no deliberation), then the rule selecting any candidate from WUS_λ achieves distortion at most $2 + \sqrt{5} \approx 4.236$ for an appropriate choice of λ [24, 20].

Small-Group Deliberation. We next recall the 2-person deliberation model with *averaging* introduced in [18], which serves as a baseline for our analysis. A deliberation involves two voters u, v and a pair of candidates (X, Y). Under the *averaging model*, the pair collectively supports the alternative with smaller total distance, or equivalently,

$$X \text{ wins against } Y \quad \text{iff} \quad d(u, X) + d(v, X) \leq d(u, Y) + d(v, Y).$$

3 Deliberation via Matching Protocol

We now describe our main protocol, *Deliberation via Matching*, which implements two-person deliberation between voters who disagree on a pair of candidates. The protocol defines a weighted tournament over candidates, parameterized by a deliberation weight $w \geqslant 0$ and the λ-weighted uncovered set parameter $\lambda \in [1/2, 1)$. These parameters will be optimized later.

Matching Step. Fix two distinct candidates $X, Y \in C$. Let XY denote the set of voters who prefer X to Y, and YX denote those who prefer Y to X.

Form an arbitrary maximum matching M_{XY} between voters in XY and voters in YX; that is, select $|M_{XY}| = \min\{|XY|, |YX|\}$ disjoint pairs (u_i, v_i) with $u_i \in XY$ and $v_i \in YX$ for $i = 1, \cdots, |M_{XY}|$. Each pair (u_i, v_i) represents a two-person deliberation between voters with opposing preferences on (X, Y). Any remaining voters (those not matched) are said to be *unmatched*. Note all unmatched voters must have the same preference: either they all prefer X (if $|XY| \geqslant |YX|$) or all prefer Y (if $|XY| < |YX|$).

In the averaging model of deliberation, let W_{XY} denote the number of matched pairs that favor A, and $W_{YX} = |M_{XY}| - W_{XY}$ the number that favor B.

To ease notation, when context is clear, we use M_{XY} to denote both a given maximal matching between (XY, YX) and its matching size $|M_{XY}| = \min\{|XY|, |YX|\}$.

Weighted Aggregation and Tournament. We define the *weighted pairwise score* of X against Y as

$$\text{score}(XY; w) = \frac{|XY| + w \cdot W_{XY}}{n},$$

and symmetrically $\text{score}(YX; w) = (|YX| + w \cdot W_{YX})/n$. We divide by n so that the score() function is independent of n, the number of voters. The total score for the pair (X, Y) is therefore $\text{score}(XY; w) + \text{score}(YX; w) = 1 + w \cdot M_{XY}/n$. We define the normalized score to be

$$f(XY; w) = \frac{\text{score}(XY; w)}{\text{score}(XY; w) + \text{score}(YX; w)} \tag{1}$$

and $f(YX; w)$ likewise so that $f(XY; w) + f(YX; w) = 1$. When the context is clear (e.g. w is a prescribed constant), we may simply write $f(XY; w)$ and $\text{score}(XY; w)$ as $f(XY)$ and $\text{score}(XY)$.

Applying the above procedure to every ordered pair of candidates (X, Y) defines a weighted tournament graph on C. The final collective decision is obtained by applying the λ-weighted uncovered set rule WUS_λ (as defined in Section 2) to this tournament.

Parameters. The protocol is governed by two parameters:

- the *deliberation weight* $w \geqslant 0$, controlling the relative influence of two-person deliberation outcomes versus individual preferences, and

- the *uncovering parameter* $\lambda \in [1/2, 1)$, which determines the strength of the dominance condition used in the λ-weighted uncovered set rule.

When $w = 0$, the protocol reduces to a standard tournament rule without deliberation. As w increases, the outcomes of matched deliberations receive greater emphasis, interpolating smoothly between non-deliberative aggregation and fully deliberative pairwise refinement.

4 Warm-up: Distortion Analysis for Two Candidates

The analysis begins with the simplest nontrivial setting of only two candidates. Studying this case serves two purposes. First, it provides a sharp quantitative improvement over the non-deliberative setting: in the absence of deliberation, any deterministic social choice rule has a worst-case distortion of 3 [2], while we show that the deliberation-via-matching protocol achieves a distortion of 2. Second, the two-candidate model isolates the geometric effect of pairwise deliberation without the additional complexity of tournament aggregation. It therefore acts as a warm-up for the more general analysis in the following sections, where we extend the same reasoning to multiple candidates and show that the overall protocol achieves distortion 3.

4.1 Preliminaries

We now recall the notation specialized to this two-candidate setting. Let the candidates be A and B, separated by distance $d(A, B)$ in the latent metric. Let AB (respectively BA) denote the set of voters who prefer A (respectively B), so that $|AB| + |BA| = n$, the total number of voters. Let M denote the arbitrary matching formed between voters in AB and those in BA according to the deliberation-via-matching protocol (Section 3). Each matched pair $(u, v) \in M \in AB \times BA$ deliberates between A and B and supports the alternative with the smaller total distance to the pair. Define

$$M_A = \{(u, v) \in M : A \text{ wins}\} = \{(u, v) \in M : d(u, A) + d(v, A) \leqslant d(u, B) + d(v, B)\}$$

$$M_B = \{(u, v) \in M : B \text{ wins}\} = \{(u, v) \in M : d(u, A) + d(v, A) > d(u, B) + d(v, B)\}.$$

Observe M_A, M_B partition M, and recall that the number of A-wins pairs (resp. B-win pairs) are $W_A = |M_A|$ (resp. $W_B = |M_B|$) by definition. The electorate now splits into three types of voters: (i) Those that contribute to A-wins, grouped as pairs from $AB \times BA$; (ii) Those that contribute to B-wins, also grouped as pairs; and (iii) Unmatched voters, all of whom belong to AB if $|AB| \geqslant |BA|$ and BA otherwise.

In the protocol in Section 3, we will set $\lambda = 1/2$ and $w = 1$. This means we set

$$\text{score}(AB) = \frac{|AB| + W_{AB}}{n},$$

and apply the Copeland rule with $f(AB) = \text{score}(AB)/(\text{score}(AB) + \text{score}(BA))$, so that A is the winner if $\text{score}(AB) \geqslant \text{score}(BA)$, and B is the winner otherwise.

We note that the classic Copeland rule declares A as the winner if and only if $|AB| \geqslant |BA|$; it is well known that this rule, as well as any other deterministic rule relying solely on ordinal information, has distortion $\geqslant 3$ even on two candidates [2]. With deliberation, we instead declare A as the winner if and only if $|AB| + W_A \geqslant |BA| + W_B$, and we show this simple change leads to an improved distortion of 2.

4.2 Analysis of the Copeland Rule

Assume A is the winner. To bound the distortion, we aim to maximize $SC(A)/SC(B)$.

Upper-bounding $SC(A)$. For every voter v, by triangle inequality

$$d(v,A) \leqslant d(v,B) + \mathbf{1}[v \in BA] \cdot d(A,B) = \begin{cases} d(v,B) & \text{if } v \in AB \\ d(v,B) + d(B,A) & \text{if } v \in BA. \end{cases} \tag{2}$$

Based on the outcomes of the matching, we split $SC(A)$ into three sums and analyze them separately.

$$SC(A) = \sum_{(u,v)\in M_A} [d(u,A) + d(v,A)] + \sum_{(u,v)\in M_B} [d(u,A) + d(v,A)] + \sum_{v \text{ unmatched}} d(v,A).$$

- For $(u,v) \in M_A$: as A wins the deliberation, $d(u,A) + d(v,A) \leqslant d(u,B) + d(v,B)$.
- For $(u,v) \in M_B$: assume $u \in AB$ and $v \in BA$, so that the corresponding applications of Equation (2) give $d(u,A) + d(v,A) \leqslant d(u,B) + d(v,B) + d(A,B)$.
- Equation (2) is also directly applicable on the sum over unmatched voters.

Observe that the total additional copies of $d(A,B)$ that appear in $SC(A)$ equals W_B plus number of unmatched BA voters; this is equivalent to $|BA| - W_A$. Hence,

$$SC(A) \leqslant SC(B) + (|BA| - W_A) \cdot d(A,B). \tag{3}$$

Lower-bounding $SC(B)$. For any pair $(u,v) \in M_A$, deliberation constraint plus triangle inequality imply

$$\begin{cases} d(u,B) + d(v,B) \geqslant d(u,A) + d(v,A) \\ d(u,A) + d(u,B) \geqslant d(A,B) \\ d(v,A) + d(v,B) \geqslant d(A,B) \end{cases} \qquad \Longrightarrow \qquad d(u,B) + d(v,B) \geqslant d(A,B)$$

We now lower bound $SC(B)$ as follows:

- Each $(u,v) \in M_A$ contributes $d(A,B)$ to $SC(B)$, and there are W_A such pairs.
- The remaining $|AB| - W_A$ voters in AB each contribute at least $d(A,B)/2$ to $SC(B)$, since $d(v,A) \leq d(v,B)$ and $d(v,B) - d(v,A) \geq d(A,B)$, which imply $d(v,B) \geq d(A,B)/2$.

Therefore,

$$SC(B) \geqslant W_A \cdot d(A,B) + (|AB| - W_A) \cdot d(A,B)/2 = (|AB| + W_A)/2 \cdot d(A,B). \tag{4}$$

Combining Equation (3) and Equation (4), we see that

$$\begin{aligned} \frac{SC(A)}{SC(B)} &\leqslant \frac{SC(B) + (|BA| - W_A) \cdot d(A,B)}{SC(B)} \leqslant 1 + \frac{(|BA| - W_A) \cdot d(A,B)}{(|AB| + W_A)/2 \cdot d(A,B)} \\ &= 1 + \frac{2(|BA| - W_A)}{|AB| + W_A} = \frac{2n}{|AB| + W_A} - 1 = \frac{2}{\text{score}(AB)} - 1. \end{aligned} \tag{5}$$

The above inequality holds for two candidates A, B regardless of the choice of (λ, w) used in the protocol in Section 3. We now analyze the protocol when $\lambda = 0.5$ and $w = 1\text{m}$ that is the Copeland Rule.

Theorem 4.1. *The metric distortion of the deliberation via matching protocol with the Copeland Rule for any 2-candidate instance is bounded by* 2.

Proof. By Equation (5), it suffices to show that if A wins, then $\text{score}(AB) \geqslant 2/3$. To prove this claim, we first assume $|AB| \leqslant |BA|$, so that $|AB| = W_A + W_B$. Since A is the winner,

$$|AB| + W_A \geqslant |BA| + W_B = |BA| + (|AB| - W_A) = n - W_A \qquad \Rightarrow \qquad 2W_A \geqslant |BA| = n - |AB|.$$

But we also have $W_A \leqslant |AB|$, so that $|AB| \geqslant n/3$. Hence

$$n \cdot \text{score}(AB) = |AB| + W_A \geqslant |AB| + \frac{n - |AB|}{2} = \frac{n + |AB|}{2} \geqslant \frac{n + (n/3)}{2} = \frac{2n}{3}.$$

If instead $|AB| \geqslant |BA|$ so that $|BA| = W_A + W_B$ and $|AB| \geqslant n/2$, then since A is the winner,

$$|AB| + W_A \geqslant |BA| + W_B = |BA| + (|BA| - W_A) = 2|BA| - W_A \qquad \Rightarrow \qquad 2W_A \geqslant 2|BA| - |AB|.$$

If $|AB| \geqslant 2n/3$ there is nothing to show, so we assume $n/2 \leqslant |AB| \leqslant 2n/3$. In this case, the above inequality becomes $2W_A \geqslant 2(n - |AB|) - |AB| = 2n - 3|AB|$. Then,

$$n \cdot \text{score}(A) = |AB| + W_A \geqslant |AB| + \frac{2n - 3|AB|}{2} = \frac{2n - |AB|}{2} \geqslant \frac{2n - 2n/3}{2} = \frac{2n}{3}. \qquad \square$$

5 Distortion Analysis for General Number of Candidates

Recall for any ordered pair of candidates (X, Y), we defined

$$\text{score}(XY; w) = |XY| + w \cdot W_{XY}, \qquad f(XY; w) = \frac{\text{score}(XY; w)}{\text{score}(XY; w) + \text{score}(YX; w)} \tag{6}$$

as in Equation (1), where $|XY|$ is number of voters preferring X to Y, W_{XY} is the number of deliberation pairs that favor X, and $w \geqslant 0$ controls the weight placed on the deliberative outcomes. We then select a winner using the λ-weighted uncovered set rule on this tournament by selecting any voter in the *λ-weighted uncovered set* WUS_λ as the winner (cf. Section 2). Throughout this section, we write $f(XY)$ and $\text{score}(XY)$, with the w-dependence implicit whenever the context is clear.

[Qilin: Fix parameters early on? This won't be used until Theorem 5.14 but makes the analysis of **??** a lot easier (avoids edge cases).] While the proof uses (λ, w) symbolically, throughout this section, we let (λ, w) take the following values for reasons that will become clear later (Theorem 5.15):

$$\boxed{\lambda^* = \frac{3 - \sqrt{3}}{2} \approx 0.6339, \qquad w^* = \sqrt{3} - 1 \approx 0.7321.} \tag{7}$$

Using the analysis technique for uncovered set tournament rules in [2, 24], suppose B is the optimal candidate, and let A be the outcome of our protocol. Then there exists another candidate C so that the uncovered set property holds: Either $f(AB) \geq 1 - \lambda$ directly, or $f(AC) \geq \lambda$ and $f(CB) \geq \lambda$. It therefore suffices to consider three such candidates A, B, C and the worst-case instance over these as:

$$\begin{aligned} \text{Distortion} = \quad & \sup \ \frac{SC(A)}{SC(B)} \\ \text{Subject to} \quad & \text{either} \ \ (f(AB) \geqslant 1 - \lambda), \\ & \text{or} \quad \big(f(AC) \geqslant 1 - \lambda \text{ and } f(CB) \geqslant \lambda\big). \end{aligned} \tag{8}$$

We now consider the two cases.

[Qilin: If we decide to introduce parameter choices early in Equation (7), then maybe we should explicitly solve case 1.]

Case 1: $f(AB) \geq 1 - \lambda$. This means we need to upper bound $SC(A)/SC(B)$ when

$$\text{score}(AB; w) \geq (1 - \lambda)\,(\text{score}(AB; w) + \text{score}(BA; w))\,.$$

We use Equation (5) to express $SC(A)/SC(B)$ as a function of $\text{score}(AB; 1) = |AB| + W_A$. This expression is independent of the function $f(AB)$ used to determine the winner. Therefore, to upper bound the distortion for our parameters (λ, w), it suffices to deduce a lower bound on $\text{score}(AB; 1)$. This can be obtained via the following non-linear program, where the variables are a, w_a:

$$\begin{aligned} \text{Minimize} \quad & \text{score}(AB; 1) = a + w_a \\ \text{Subject to} \quad & 0 \leqslant a \leqslant 1 \\ & 0 \leqslant w_a \leqslant \min\{a, 1 - a\} \\ & a + w \cdot w_a \geqslant (1 - \lambda) \cdot (1 + w \cdot \min\{a, 1 - a\}) \end{aligned} \tag{9}$$

where $a = |AB|/n$ is the fraction of AB voters, and $w_a = W_A/n$ is the fraction of deliberations favoring A (as a fraction of n). The last constraint ensures $f(AB) \geqslant 1 - \lambda$. Note that $W_{AB} + W_{BA} = \min(a, 1 - a)$, the matching size. If the above program is optimum θ, then the distortion is upper bounded by $2/\theta - 1$. We will choose (λ, w) after solving the second case below, and for that setting of parameter, solve the above program, which can be reduced to solving two LPs.

Case 2. In the rest of this section, we focus on the other case, where $f(AC) \geqslant 1 - \lambda$ and $f(CB) \geqslant \lambda$. Bounding the distortion in this case forms the remainder of the proof.

5.1 Measure Space Relaxation

The remaining proof focuses on three candidates A, B, C, and upper bounds distortion $SC(A)/SC(B)$, subject to the second constraint in Equation (8). Our distortion proof proceeds by a sequence of relaxations that do not decrease distortion, but enable us to find a supremum with nice structure.

We first relax the finite set of voters to a set of voters with measure 1; the finite voter case remains a special case, so that showing a distortion bound in the latter model suffices. In this model, the *electorate* is a probability space (V, μ) with $\mu(V) = 1$. For a measurable set $S \subset V$, write $|S| = \mu(S)$. An *instance* I consists of an electorate V, three *candidates* A, B, C, and a (possibly infinite) metric defined on $V \cup \{A, B, C\}$. [Kamesh: Formally, for given $f = d(A, B)$, $g = d(B, C)$, and $h = d(A, C)$ that satisfy the triangle inequality, consider the set of constraints $\mathcal{P}(f, g, h)$:

$$x + y \geq f \geq |x - y|, \qquad y + z \geq g \geq |y - z|, \qquad x + z \geq h \geq |x - z|, \qquad x, y, z \geq 0.$$

Any voter v can be parameterized by a triple $(x, y, z) \in \mathcal{P}(f, g, h)$, where $x = d(v, A)$, $y = d(v, B)$, and $z = d(v, C)$. Once this is specified, the distances between voters can be made equal to the smallest distance through the candidates as:

$$d(u, v) = \min_{X, Y \in \{A, B, C\}} (d(u, X) + d(X, Y) + d(Y, v))$$

and this will satisfy triangle inequality. The space over which we optimize distortion is the space of all $f, g, h \geq 0$ with $f + g \geq h \geq |f - g|$, all measures μ over $\mathcal{P}(f, g, h)$, all matchings of relevant voters for that measure for each pair of candidates. Each point in this space defines a preference profile and an outcome of deliberation, and we take the supremum of the distortion over all these points. We will assume that there are no ties, so that the set of points where any two of x, y, z are equal is zero.]

We define the *social cost* of a candidate X as $SC(X) = \int_V d(v, X)\ \mathrm{d}\mu(v)$. For candidates $X, Y \in \{A, B, C\}$ and $v \in V$, define the *intensity* $\Delta_{XY}(v) = d(v, Y) - d(v, X)$ which quantifies how much v prefers X over Y. The preference set partitions remain unchanged: $XY = \{v : \Delta_{XY}(v) \geqslant 0\}$ and $YX = \{v : \Delta_{XY}(v) < 0\}$, and let $m_{XY} = \min\{|XY|, |YX|\}$. A *maximal matching* between XY and YX is a finite product measure γ on $XY \times YX$ such that there exist measurable $S \subset XY, T \subset YX$ with $|S| = |T| = m_{XY}$, and

$$\gamma(F \times YX) = |F \cap S|, \qquad \gamma(XY \times G) = |G \cap T|, \qquad \text{for all measurable } F, G.$$

Intuitively, this matching matches S against T, which we call *marginals*. Given a maximal matching γ, the *win mass* for X is a direct analog of the number of X-win pairs in the discrete case and is defined as

$$W_{XY}(\gamma) = \gamma(\{(u, v) : \Delta_{XY}(u) + \Delta_{XY}(v) \geqslant 0\}).$$

An *X-optimal* matching is a maximizer $\gamma^* = \arg\max_\gamma W_{XY}(\gamma)$. We provide a high-level structural overview of X-optimal matchings below; a formal proof can be found in Theorem A.1.

Given X, Y and a maximal matching γ on X, Y and weight w, we define $f(XY) = f(XY; w)$ as in Equation (6), but overloaded with $|\cdot|$ denoting the measure and W_{XY} the win-mass defined via the product measure. The objective is to compute $\sup SC(A)/SC(B)$ subject to $f(AC) \geqslant 1 - \lambda, f(CB) \geqslant \lambda$.

5.2 Structures of X-optimal Matchings

A fixed instance I may admit many (AC, CA) (resp. (CB, BC)) matchings and thus potentially different values for W_{AC} (resp. W_{CB}). Then the values of $f(AC)$ (resp. $f(CB)$) need not be unique for I. It follows that the set $\{I : f(AC) \geqslant 1 - \lambda, f(CB) \geqslant \lambda\}$ is the largest when for any instance I, we always choose $f(AC)$ and $f(CB)$ to the largest admissible values. It is easy to see that these are exactly attained by the A-optimal (AC, CA) and the C-optimal (CB, BC) matchings, respectively. Consequently, to find $\sup SC(A)/SC(B)$, we will from now on define

$$f(XY) = \frac{|XY| + w \cdot W_{XY}(\gamma^*)}{1 + w \cdot m_{XY}} \qquad \text{where } \gamma^* \text{ is the } X\text{-optimal } (XY, YX) \text{ maximal matching,} \tag{10}$$

and we seek to upper bound, with the updated $f(\cdot)$-constraints, the same objective:

$$\sup \frac{SC(A)}{SC(B)} \qquad \text{subject to} \qquad f(AC) \geqslant 1 - \lambda, f(CB) \geqslant \lambda.$$

[Qilin: To Kamesh: on second thought, I think we can move all of these cumbersome definitions to the appendix. In the proof of Theorem 5.3, though, we need to cite definitions of $Q^{\pm}$ from the appendix, as well as the closed form formula for α as in Theorem A.1. All I'll be using are continuity of things that define α to justify continuity of α. No crazy math. Is this (citing things defined only in appendix) okay practice in TCS?]

Existence and Characterization. Given a pair of candidates X, Y, we now prove the existence and characterize the structure of an X-optimal matching γ^* of (XY, YX). Intuitively, this matching is achieved by greedily pairing each pro-X voter with the most pro-Y voter such that the pair still prefers X. [Kamesh: We should have all definitions that we subsequently need here. I think the current writing is fine. It just needs more clarity on what the matching actually is.]

Formally, let Δ_{XY} be defined as above. Let $m = m_{XY} = \min\{|XY|, |YX|\}$. We work w.l.o.g. under $|XY| \leqslant |YX|$ (the other case is symmetric). For $t \geqslant 0$, we define the *tail sets* as

$$\varphi^+(t) = \{u \in XY : \Delta_{XY}(u) \geqslant t\}, \qquad \varphi^-(t) = \{v \in YX : -\Delta_{XY}(v) \geqslant t\}$$

to be the subsets of XY, YX whose Δ intensity is at least as strong as t in magnitude.

Theorem 5.1 (Characterization of X-Optimal Matchings)**.** *With S, T, Q^+, Q^- as above, define the* ***minimal shift***

$$\alpha = \inf\{s \in [0, m] : Q^+(q) \geqslant Q^-(q - s) \text{ for a.e. } q \in [s, m]\}. \tag{11}$$

Let γ be any matching on (XY, YX). Then γ is guaranteed to have X-loss mass of at least α, i.e., the X-win mass $W_{XY}(\gamma) \leqslant m - \alpha$. Moreover, there exists a γ^* attaining this equality, so in particular an X-optimal matching γ^* exists.

5.3 Tightness of the Constraints

[Qilin: I would like to make this section more compact and make the following claims:

Claim 1. Given an instance I, there exists an instance I' with measure μ' where (no-deliberation-ties (at any level, not just 0)), f-values are no lower, and distortion is at least old distortion minus $O(\epsilon)$. The proof of this is essentially Theorem 5.3 step 1 applied twice (to break ties in CB delib, then to AC).

Claim 2. Given an instance I' obtained from Claim 1's transformation, there exists another instance I'' such that $f(AC) = 1 - \lambda$, $f(CB) = \lambda$, and distortion of $I'' \geqslant$ that of I'. The proof of this is essentially the remaining steps of Theorem 5.3.

These two combined justify that it is WLOG to assume both $f()$ constraints are tight *and* the (no-deliberation-ties) assumption.

If this proposed edit looks good, I will rewrite. (Also check the comment on previous page regarding matching; it determines how I phrase this sesction.)] [Kamesh: I don't think this needs a rewrite. It is clear enough. The previous section needs more clarity.]

In this section, we argue that in order to find $\sup SC(A)/SC(B)$ subject to $f(AC) \geqslant 1-\lambda$ and $f(CB) \geqslant \lambda$, it suffices to restrict our attention to instances where $f(AC) = 1 - \lambda$ and $f(CB) = \lambda$.

The lemma below makes one of the constraints tight.

Lemma 5.2. *It suffices to assume that at least one $f(\cdot)$ constraint is tight, i.e.:*

$$\sup_{\substack{f(AC)\geqslant 1-\lambda \\ f(CB)\geqslant\lambda}} \frac{SC(A)}{SC(B)} = \max\left\{ \sup_{\substack{f(AC)=1-\lambda \\ f(CB)\geqslant\lambda}} \frac{SC(A)}{SC(B)}, \sup_{\substack{f(AC)\geqslant 1-\lambda \\ f(CB)=\lambda}} \frac{SC(A)}{SC(B)} \right\}$$

Proof. Pick any instance with $f(AC) > 1 - \lambda$ and $f(CB) > \lambda$. Continuously pad voters at B (and re-normalize) so that $SC(A)/SC(B)$ strictly increases. [Kamesh: Is the next sentence correct?] This strictly decreases $|CB|$ and weakly decreases $W_{CB}(\gamma^*)$, which strictly decreases $f(CB)$. We stop when either $f(AC)$ hits $1 - \lambda$ or $f(CB)$ hits λ. □

The next two lemmas together will make both constraints tight.

Lemma 5.3. $$\sup_{\substack{f(AC)=1-\lambda \\ f(CB)\geqslant\lambda}} \frac{SC(A)}{SC(B)} = \sup_{\substack{f(AC)=1-\lambda \\ f(CB)=\lambda}} \frac{SC(A)}{SC(B)}.$$

Proof. Throughout, write $\Delta(v) = d(v, B) - d(v, C)$. The key idea of the proof is to change the metric space continuously, so that $f(CB)$ continuously as well. This will first require removing atomic masses from the voters by spreading out their location in the metric space (that is, replace each voter with a small ball). Subsequently, the change in distances will be linear.

Step 1. Breaking ties via smoothing. We first remove point masses from the measure space over voters, while keeping $f(AC)$ fixed and weakly increasing $f(CB)$. Fix $\epsilon > 0$ and define

$$V' = V \times (0, \epsilon), \qquad \mathrm{d}\mu'(v, r) = \mathrm{d}\mu(v)\frac{\mathrm{d}r}{\epsilon}.$$

Define a metric d' on V' by

$$d'((v,r),A) = d(v,A) + r, \qquad d'((v,r),C) = d(v,C) + r, \qquad d'((v,r),B) = d(v,B) + \epsilon$$
$$d'(A,B) = d(A,B) + \epsilon, \qquad d'(C,B) = d(C,B) + \epsilon, \qquad d'(A,C) = d(A,C).$$

Intuitively, for each voter v, we "spread" it uniformly onto a continuum of $(0, \epsilon)$ and preserve the total mass. Observe that the changes made to $d'(\cdot, A)$ and $d'(\cdot, C)$ are identical, so the sets AC, CA, the $A - C$ maximal matching outcomes, as well as $f(AC)$, all remain unchanged. On the other hand,

$$\Delta'(v,r) = d'((v,r),B) - d'((v,r),C) = \Delta(v) + (\epsilon - r) > \Delta(v),$$

so every voter's CB intensity strictly increases. Under the C-optimal matching, $f(CB)$ cannot decrease. Further, we observe that (CB, BC)-related level sets have measure zero. In particular, the distribution of Δ' is non-atomic and $\mu'(\{\Delta' = 0\}) = 0$. [Kamesh: The next statement is trivially true if we start with a finite set of voters and do the smoothing. I am not seeing why it is true in general for densities – that's more non-trivial but we don't need it here. One option is to do the continuous business here, and keep it discrete till here.] This also means that for two voters $u \in CB$ and $v \in BC$, the measure on the their total intensity $\Delta'(u, r_1) + \Delta'(v, r_2)$ is non-atomic as well. Furthermore, because $\mathbb{E}[r] = \epsilon/2$, we have $SC'(A) = SC(A) + \mathbb{E}[r] = SC(A) + \epsilon/2$ and $SC'(B) = SC(B) + \epsilon$. Hence in this step the distortion changes by $SC'(A)/SC'(B) - SC(A)/SC(B) = O(\epsilon)$.

Step 2. Monotonically decreasing $f(CB)$. We now describe an operation that preserves $f(AC) = 1 - \lambda$ but continuously decreases $f(CB)$. For $t \geqslant 0$, we define

$$d_t((v,r),A) = d'((v,r),A) + t \qquad d_t((v,r),C) = d'((v,r),C) + t \qquad d_t((v,r),B) = d'((v,r),B)$$
$$d_t(A,B) = d'(A,B) + t \qquad d_t(C,B) = d'(C,B) + t \qquad d_t(A,C) = d'(A,C) + 2t \tag{12}$$

Once again, the changes to $d_t(\cdot, A)$ and $d_t(\cdot, C)$ are identical, so that $f(AC; t)$ remains unchanged. We focus on $f(CB; t)$ now. For every (v, r) we have

$$\Delta_T(v,r) = d_t((v,r),B) - d_t((v,r),C) = \Delta'(v,r) - t.$$

Consider the partition of V by $CB_t = \{(v,r) : \Delta'(v,r) \geqslant t\}$ and $BC_t = \{(v,r) : \Delta'(v,r) < t\}$. These represent the new mass of voters who prefer C to B and vice versa, respectibely. Let $m_t = \min\{|CB_t|, |BC_t|\}$ denote the new matching size. Consider constructing the C-optimal (CB, BC) matching using the quantile argument in Theorem A.1. Let Q_t^+ be the ascending quantile of Δ' on CB_t and Q_t^- the ascending quantile of $-\Delta'$ on BC_t, both parameterized on $[0, m_t]$. By definition, a matched pair $((u^+, r^+), (u^-, r^-)) \in CB \times BC$ at rank q is a C-win if and only if

$$(\Delta'(u^+, r^+) - t) + (\Delta'(u^-, r^-) - t) \geqslant 0 \iff Q_t^+(q) - Q_t^-(q) \geqslant 2t.$$

The C-win mass $W_C(t)$ in the matching and the score $f(CB; t)$ can be respectively written as

$$W_C(t) = \int_0^{m_t} \mathbf{1}\{Q_t^+(q) - Q_t^-(q) \geqslant 2t\}\, \mathrm{d}q, \qquad f(CB;t) = \frac{|CB_t| + w \cdot W_C(t)}{1 + w \cdot m_t}. \tag{13}$$

Step 3. Continuity. We now argue that $t \mapsto W_C(t)$ is continuous. Since $t \mapsto m_t$ is also continuous, this establishes that $t \mapsto f(CB;t)$ is continuous. By assumption, when $t = 0$, we have $f(CB;t=0) > \lambda$. As $t \to \infty$, almost every voter will rank B over C according to Equation (12), at which point $f(CB;t) \to 0$. The intermediate value theorem implies there exists t^* such that $f(CB;t^*) = \lambda$. Furthermore,

$$\frac{SC_t(A)}{SC_t(B)} \geqslant \frac{SC'(A)}{SC'(B)} \geqslant \frac{SC(A)}{SC(B)} - O(\epsilon),$$

where the first $\geqslant$ is by Equation (12) and the second is by the last line of Step 1. This completes the proof since ϵ is arbitrary.

We now prove the continuity of $t \mapsto W_C(t)$. Let $t_n \to t$ be any sequence of values; we show $W_C(t_n) \to W_C(t)$. Because Δ' is made non-atomic by Step 1, its CDF $F(x) = |\{\Delta' \leqslant x\}|$ is continuous. Thus $|CB_t| = 1 - F(t)$, $|BC_t| = F(t)$, and m_t vary continuously in t. The quantiles $Q_t^{\pm}(\cdot)$ are also continuous a.e. Crucially, from Step 1, the pair-sum tie set $\{q \in [0, m_t] : Q_t^+(q) - Q_t^-(q) = 2t\}$ has Lebesgue measure 0, so the indicators converge pointwise a.e.:

$$\mathbf{1}\{Q_{t_n}^+(q) - Q_{t_n}^-(q) \geqslant 2t_n\} \to \mathbf{1}\{Q_t^+(q) - Q_t^-(q) \geqslant 2t\}.$$

The integrands of W_C are bounded by 1, so by dominated convergence, $W_C(t_n) \to W_C(t)$, as desired. □

The proof of the the theorem below is nearly identical to the previous theorem.

Lemma 5.4. $$\sup_{\substack{f(AC)\geqslant 1-\lambda \\ f(CB)=\lambda}} \frac{SC(A)}{SC(B)} = \sup_{\substack{f(AC)=1-\lambda \\ f(CB)=\lambda}} \frac{SC(A)}{SC(B)}.$$

Proof. We only outline the changes made to d' and d_t. To define d' we let

$$d'((v,r),A) = d(v,A), \qquad d'((v,r),C) = d(v,C) + r, \qquad d'((v,r),B) = d(v,B) + r.$$

We realize this by defining $d'(A,B) = d(A,B) + \epsilon$, $d'(C,B) = d(C,B)$, and $d'(A,C) = d(A,C) + \epsilon$. For the second step, for $t \geqslant 0$ we let

$$d_t((v,r),A) = d'((v,r),A) + t, \qquad d_t((v,r),C) = d'((v,r),C), \qquad d_t((v,r),B) = d'((v,r),B),$$

with $d_t(A,B) = d'(A,B) + t$, $d_t(C,B) = d'(C,B)$, and $d_t(A,C) = d'(A,C) + t$. The rest of the proof remains the same, and we omit the details. □

Combining the previous three claims, we achieve the following reduction, which we will assume to hold for the rest of this section.

Theorem 5.5.

$$\sup_{\substack{f(AC)\geqslant 1-\lambda \\ f(CB)\geqslant \lambda}} \frac{SC(A)}{SC(B)} = \sup_{\substack{f(AC)=1-\lambda \\ f(CB)=\lambda}} \frac{SC(A)}{SC(B)}.$$

5.4 Optimization Reductions: Convexification, Coupling, & Closure Space

[Qilin: Rework starts from this section. In addition to convexification, I introduced two new definitions: (i) given (X,Y), the best coupling of X,Y that gives the smallest $\mathbb{E}Z$ is given in Theorem 5.9; (ii) the notion of "closure" is introduced in Theorem 5.11, where we just allow arbitrary apportioning of preference and deliberation ties (and no longer assume they have measure 0, since clearly an optimal solution should sit right on these ties).]

Despite the highly nonlinear and nonconvex nature of the deliberation-via-matching protocol, for instance the $\min(\cdot,\cdot)$ function involved in the matching size, in this section, we pave a path towards a systematic, tractable analysis.

Lemma 5.6 (1-dimensional marginal sufficiency of $f(\cdot)$ constraints)**.** *Define three variables X, Y, Z by*

$$X(v) = d(v, C) - d(v, A), \qquad Y(v) = d(v, B) - d(v, C), \qquad Z(v) = d(v, C). \tag{14}$$

Then X encodes all information needed to determine the (AC, CA) deliberation: $|AC|, |CA|$, as well as the distribution of intensities on V. Consequently, we may compute $f(AC)$ give X. Likewise, Y encodes $f(CB)$. It also follows from the definitions that $SC(A)/SC(B) = [\mathbb{E}Z - \mathbb{E}X]/[\mathbb{E}Z + \mathbb{E}Y]$.

Proof. Immediate from Theorem A.1, which requires nothing more than the distribution of voters' preference intensities. □

We note that Theorem 5.6 does *not* assert independence of X, Y, and Z. Instead, it only implies that the variables X, Y are useful tools that we can use to characterize the outcomes of deliberations, thus providing an alternate perspective to the problem of interest. There need not be a one-to-one correpsondence between instances I and triples (X, Y, Z).

Notation clarification. Up until Theorem 5.6, uppercase X, Y have mostly been used to denote arbitrary candidates $\in \{A, B, C\}$. From now on, we will repeatedly treat X, Y, Z as random variable defined according to Equation (14) unless indicated otherwise.

Remark 5.7. From now on, for any instance that we work with, we may assume $\mathbb{E}X + \mathbb{E}Y < 0$; otherwise, $SC(A) \leqslant SC(B)$ and the instance is irrelevant for finding supremum distortion. Since the function $z \mapsto [z - \mathbb{E}X]/[z + \mathbb{E}Y]$ is decreasing in z, given fixed (X, Y), to maximize distortion, it suffices find the pointwise minimum Z that is metric feasible in the sense that Equation (14) can be realized in some latent metric space. In the next lemma, we establish a necessary and sufficient condition.

[Qilin: To Kamesh: update 10/20 — I have rephrased Theorem 5.9. Hopefully it's easier to follow now. I think Theorem 5.8 is now pretty straightforward (just verify triangle-ineqs), and hopefully so is the new Theorem 5.9. As a sanity check, I have run many simulations, all of which suggest that counter monotone coupling helps minimize $\mathbb{E}Z$.]

Lemma 5.8. *Fix real-valued functions X, Y on the electorate V. For any real-valued function Z on V, in order for (X, Y, Z) to be be realized by some metric d via Equation (14), it is necessary and sufficient that*

$$Z(v) \geqslant Z_{\min}(v) = \max\left\{\frac{\|X\|_\infty + X(v)}{2}, \frac{\|Y\|_\infty - Y(v)}{2}, \frac{\|X + Y\|_\infty - (Y(v) - X(v))}{2}\right\} \qquad \textit{for all } v. \tag{15}$$

Proof. We first prove necessity. Because d is nonnegative, we must have $d(v, C) = Z(v) \geqslant 0, d(v, A) = Z(v) - X(v) \geqslant 0$, and $d(v, B) = Z(v) + Y(v) \geqslant 0$ from Equation (14). Triangle inequalities for (v, A, C) imply

$$|d(v, A) - d(v, C)| = |X(v)| \leqslant d(A, C) \leqslant d(v, A) + d(v, C) = 2Z(v) - X(v).$$

Taking supremum over the first $\leqslant$ gives $d(A, C) \geqslant \|X\|_\infty$; combining with the second $\leqslant$,

$$2Z(v) - X(v) \geqslant \|X\|_\infty \qquad \text{so} \qquad Z(v) \geqslant \frac{\|X\|_\infty + X(v)}{2}. \tag{16}$$

The remaining two terms can be obtained analogously by enforcing triangle inequalities on (v, B, C) and (v, A, B), respectively.

For sufficiency, assume Equation (15) and define $d(A, C) = \|X\|_\infty, d(B, C) = \|Y\|_\infty$, and $d(A, B) = \|X + Y\|_\infty$. Then (A, B, C) satisfy triangle inequalities, and for each voter, the inequalities established in the necessity part flip to become upper bounds: for (v, A, C), we have

$$|X(v)| \leqslant d(A, C) \leqslant 2Z(v) - X(v)$$

and likewise for (v, B, C) and (v, A, B), so triangle inequalities also hold among these pairs. We remarked in Section 5.1 that this suffices for demonstrating realizability. □

Unless otherwise indicated, given a pair (X, Y) defined on V, we will from now on default to defining Z as in Equation (15).

[Kamesh: The next statements should go into the statement of the lemma below.]

Another perspective of viewing X, Y is to view them as functions induced by *distributions* $\mathcal{D}_X, \mathcal{D}_Y$ on V. By Theorem 5.6, as long as $X_1, X_2 \sim \mathcal{D}_X$ are identically distributed, the value $f(AC)$ is determined; likewise for $Y_1, Y_2 \sim \mathcal{D}_Y$. Therefore, to upper bound $SC(A)/SC(B)$ while preserving $f(AC), f(CB)$, it is safe to analyze different *couplings* of $\{(X(v), Y(v))\}_{v \in V}$, with $X \sim \mathcal{D}_X, Y \sim \mathcal{D}_Y$, and focus on the ones that give the optimal Z (smallest $\mathbb{E}Z$) according to Equation (15). (Recall from Theorem 5.7 that with $\mathbb{E}X, \mathbb{E}Y$ fixed, and $\mathbb{E}X + \mathbb{E}Y < 0$ assumed, the smaller $\mathbb{E}Z$ is, the larger the distortion.) As a toy example, consider a space with two discrete voters v_1, v_2. Let $\mathcal{D}_X$ takes values ± 1 each with probability $1/2$, and $\mathcal{D}_X$ takes value in $\{0, 1\}$ each w.p. $1/2$. The coupling $\{(X(v_1), Y(v_1)) = (1, 0), (X(v_2), Y(v_2)) = (-1, 1)\}$ gives rise to a set of constraints on Z that is most likely different from that of the coupling $\{(X(v_1), Y(v_1)) = (1, 1), (X(v_2), Y(v_2)) = (-1, 0)\}$, even though the set sizes like $|AC|$ and deliberation outcomes are identical.

[Qilin: Saying X, Y as distirbutions on V seems necessary here, but I dislike this formality... Almost feels like a probability theory statement. Suggestions? Anyways, the key of the following lemma is to minimize $\|X + Y\|_\infty$ which intuitively suggests we should pair positive X with negative Y's and vice versa.]

[Kamesh: This is very nice. It seems correct.]

Lemma 5.9 (Counter-monotone coupling of X, Y)**.** *Fix distributions $\mathcal{D}_X, \mathcal{D}_Y$ on V and consider any coupling (X, Y) with $X \sim \mathcal{D}_X, Y \sim \mathcal{D}_Y$. Among all instances whose X, Y follow distributions $\mathcal{D}_X, \mathcal{D}_Y$, there exists one with maximal distortion that couples (X, Y) counter-monotonically: Informally, it couples the largest $X(v)$ with the smallest / most negative $Y(v)$ and so on.*

Proof. We prove the discrete case only; a continuous argument can be proven analogously with discrete rankings replaced by quantiless. The proof proceeds in two steps. Given $X \in \mathcal{D}_X, Y \in \mathcal{D}_Y$, their expectations are fixed. Therefore, by Theorem 5.7 it suffices to minimize $\mathbb{E}Z$ subject to Equation (15).

We prove the claim via a classical application of the exchange argument: as long as the coupling involves pairs $(x_1, y_1) = (X(u_1), Y(v_1))$ and $(x_2, y_2) = (X(u_2), Y(v_2))$ with $x_1 < x_2, y_1 < y_2$, swapping them (pairing x_1 with y_2 and x_2 with y_1) does not harm the goal of minimizing $\mathbb{E}Z$. Iterating this local improvement rule until no "out-of-oder" pairs remain forces us to land at a said counter-monotone coupling. Note that throughout, $\|X\|_\infty, \|Y\|_\infty$ remain unchanged.

Formally, the proof consists of two parts. First, a conditional result: Given a *frozen* baseline $c = \|X+Y\|_\infty$, a local counter-monotone swap never increases $\mathbb{E}Z$. Second, we observe that local swaps indeed do not worsen (increase) $\|X + Y\|_\infty$, so it in fact unconditionally weakly decreases $\mathbb{E}Z$.

Step 1. With $\|X + Y\|_\infty$ fixed, local swap helps. Start with an arbitrary coupling π on (X, Y) and let $c(\pi) = \|X + Y\|_\infty$. Define, treating $c(\pi)$ as a constant, a bivariate function

$$h_c(x, y) = \max\{\|X\|_\infty + x, \|Y\|_\infty - y, c - (y - x)\},$$

so $Z = h_c(X, Y)/2$ pointwise by Equation (15). For $x_1 < x_2, y_1 < y_2$, we claim that

$$h_c(x_1, y_1) + h_c(x_2, y_2) \geqslant h_c(x_1, y_2) + h_c(x_2, y_1), \tag{17}$$

which would immediately imply that the local swap, *assuming fixed* $\|X+Y\|_\infty$, weakly decreases $\mathbb{E}[h_c(X, Y)]$. Fix y and consider the mapping $x \mapsto h_c(x, y)$. Let

$$T(y) = \min\{\|Y\|_\infty - \|X\|_\infty - y, \|Y\|_\infty - c\}.$$

Observe that if $x \leqslant T(y)$, then $h_c(x,y) = \|Y\|_\infty - y$; if $x \geqslant T(y)$, then $\|Y\|_\infty - y$ no longer dominates, and

$$h_c(x,y) = \max\{\|X\|_\infty + x, c + x - y\} = x + \max\{\|X\|_\infty, c - y\}.$$

Consequently, for every fixed y, $x \mapsto h_c(x,y)$ is piecewise linear: constant up to $T(y)$, then slope 1 hereafter. From this, we observe that

$$\Delta(y) = h_c(x_2,y) - h_c(x_1,y) = [x_2 - \max\{x_1, T(y)\}]^+$$

where $[t]^+ = \max\{t,0\}$. Because $T(y)$ is nonincreasing in y, the map $y \mapsto \max\{x_1, T(y)\}$ is also nonincreasing. Therefore, $\Delta(y)$ is nondecreasing in y. For $y_1 < y_2$, we therefore have $\Delta(y_2) \geqslant \Delta(y_1)$. Unwinding this inequality gives Equation (17).

Step 2. Local swap does not raise $\|X+Y\|_\infty$. For the same $x_1 < x_2$, $y_1 < y_2$, because

$$\max\{|x_1 + y_2|, |x_2 + y_1|\} \leqslant \max\{|x_1 + y_1|, |x_2 + y_2|\},$$

the swap does not increase $\|X+Y\|_\infty$. As $h_c(x,y)$ is nondecreasing in c, this proves that local swaps indeed always help, *unconditionally*. We therefore iteratively perform local swaps until no swap is available, which happens precisely when the resulting coupling is counter-monotone. It remains to notice that this gives the global minimizer of $\mathbb{E}Z$. For a coupling π we define $c(\pi) = \|X+Y\|_\infty$ under it. Let $c^* = \min_\pi c(\pi)$ and let π^* be a counter-monotone coupling with $c(\pi^*) = c^*$. Then, for an arbitrary π,

$$\underbrace{\mathbb{E}_\pi[h_{c(\pi)}(X,Y)]}_{\mathbb{E}Z \text{ under } \pi} \geqslant \underbrace{\mathbb{E}_\pi[h_{c^*}(X,Y)]}_{\text{since } c^* \leqslant c(\pi)} \geqslant \underbrace{\mathbb{E}_{\pi^*}[h_{c^*}(X,Y)]}_{\text{Step 2}}.$$

Therefore π^* is an optimal coupling we seek, and the proof is complete. □

[Qilin: Some transition sentences here TBD. Also, we essentially only used convexity to (formally) rule out continuous, tie-free instances as maximizers. I don't think our new **????** still uses convexification. It is however this idea of convexification that brings out closure (Theorem 5.11) which is crucial for subsequent analyses. I'm sure there's a better story to make up here.]

[Kamesh: I don't think all this is needed. For any distortion θ, write it as maximizing numerator minus θ times denominator, and checking if the max is < 0. Clearly the max is decreasing in Z. Also, we should normalize by setting the norm of X and Y to be fixed numbers, like 1 and B. That avoids the unboundedness problem.]

[Qilin: Yes, something along this normalization argument should suffice. We'll revisit this later. The only place we use Theorem 5.10 is Theorem 5.13 where we argue the need to bring out closure (allow again atomic ties and break them arbitrarily reasonable way). Then the rest of the proof focus entirely on directly improving $[\mathbb{E}Z - \mathbb{E}X]/[\mathbb{E}X + \mathbb{E}Y]$ by, say, decreasing any of $\mathbb{E}X, \mathbb{E}Y, \mathbb{E}Z$, without relying on the monotonicity of $R(x,y,z)$ along any line segments.]

Lemma 5.10 (Convexification of objective)**.** *We associate every feasible instance I with a triple*

$$t(I) = (x(I), y(I), z(I)) = (\mathbb{E}X, \mathbb{E}Y, \mathbb{E}Z) \in \mathbb{R}^3$$

according to Equation (15), and write distortion as $R(x,y,z) = (z-x)/(z+y)$ where $z - x \geqslant 0$ and $z + y = 1$ (we normalize the social cost of B to 1). Then, given a collection $S = \{t(I)\} = \{(x(I), y(I), z(I))\}$ of instance-induced triples,

$$\sup_S R(x,y,z) = \sup_{\text{conv}(S)} R(x,y,z). \tag{18}$$

In particular, on $\mathcal{I} = \{t(I) : I \text{ satisfies } f(AC) = 1 - \lambda, f(CB) = \lambda\}$, the collection of feasible instances of interest, to find the supremum distortion over $\mathcal{T} = \{t(I) : I \in \mathcal{I}\}$, it suffices to evaluate R over the boundary of the closure of $\text{conv}(\mathcal{T})$.

Proof. Consider any line segment in the domain, parameterized as $(x(t), y(t), z(t)) = (x_0, y_0, z_0) + \lambda(a, b, c)$ for $0 \leqslant \lambda \leqslant 1$. Along this segment,

$$R(x(t), y(t), z(t)) = \frac{(z_0 - x_0) + (c - a)t}{(z_0 + y_0) + (c + b)t}$$

whose derivative over t equals

$$R'(x(t), y(t), z(t)) = \frac{(c - a)(z_0 + y_0) - (c + b)(z_0 - x_0)}{(z_0 + y_0 + (c + b)t)^2}$$

whose sign is independent of t. Therefore, R is monotone along any line segment, and it follows that optimizing R over S may be convexified into optimizing over $\text{conv}(S)$; if $\text{conv}(S)$ is closed, then R attains a maximizer on its boundary, and if S is open, then the supremum is attained somewhere on the boundary of the closure of $\text{conv}(\mathcal{T})$. □

An important caveat of the convexification argument concerns the closure of all (x, y, z) subject the constraints $f(AC) = 1 - \lambda, f(CB) = \lambda$: the feasible image $\mathcal{T} \in \mathbb{R}^3$ need not be closed. Indeed, sequences of tie-free, equally-feasible instances as in (earlier) can approach, in terms of $t(I)$, may very well converge to a profile where the preference and/or deliberation tie sets accumulate nonzero mass. Therefore, for the following geometric arguments, it is natural to pass to the Euclidean closure of $\mathcal{T}$, denoted $\text{cl}(\mathcal{T})$. We now formalize the closure of $\mathcal{T}$. Recall $\mathcal{I}$ be the set of instances we currently consider ($f(AC) = 1-\lambda, f(CB) = \lambda$, in the measure space and tie-free perspective.)

[Qilin: This just says we allow preference and deliberation ties to now have nonzero measure, and that any way of apportioning them are acceptable.]

Definition 5.11 (Closure of $\mathcal{I}$)**.** An instance $\hat{I}$ (allowed to have nonzero measure on preference ties and deliberation sum-ties) belongs to $\text{cl}(\mathcal{I})$ if the following conditions hold:

- There exist measurable *preference tie-splitting rules*

$$\theta_{XY} : \{v : \Delta_{XY} = 0\} \to [0, 1], \qquad XY \in \{AC, CA, CB, BC\} \tag{19}$$

  with $\theta_{AC} + \theta_{CA} = \theta_{CB} + \theta_{BC} = 1$.

- There exist measurable *deliberation tie-splitting rules*

$$\beta_{XY} : \{(u, v) \in XY \times YX : \Delta_{XY}(u) + \Delta_{XY}(v) = 0\} \to [0, 1], \tag{20}$$

  with $\beta_{AC} + \beta_{CA} = \beta_{CB} + \beta_{BC} = 1$.

- The tie-splitting rules are such that if we attribute a θ-fraction of XY-preference ties to XY and a β-fraction of XY-derliberation preference ties as X-wins, then $f(AC) = 1 - \lambda, f(CB) = \lambda$.

Theorem 5.12. $\{t(I) : I \in \text{cl}(\mathcal{I})\}$ *is the Euclidean closure of* $\mathcal{T} = \{t(I) : I \in \mathcal{I}\}$. *Hence we will denote the first set as* cl(T) *from now on.*

Proof. Currently omitted. [Qilin: This is very easy to fill in. Low priority.] □

[Qilin: The following theorem may or may not be needed anymore. In fact we should discuss about the structure of the paper — tie-free relaxation is still crucially needed to argue for Theorem 5.5, but it's less of importance after that. I believe **????** do not strictly require Theorem 5.13.]

Theorem 5.13 (No tie-free instance is a maximizer)**.** *For every $I \in \mathcal{I}$ (where the no-tie assumption holds), there exist distinct $I^+, I^- \in \mathcal{I}$ such that $t(I) = (t(I^+) + t(I^-))/2$. Consequently, $t(I)$ is not a maximizer of R over* $\mathrm{cl}(\mathrm{conv}(\mathcal{T}))$.

Proof. Recall Z is constrained by affine functions in (X, Y). We start by picking $U \subset AC$ with positive measure and then shrink U so that there exist sufficiently small $\epsilon > 0$ where:

(i) One affine constraint of $Z_{\min}$ is tight U with margin $\epsilon > 0$ on all other constraints (i.e., if we perturb U by ϵ, this constraint is still the only one).

(ii) On U, we have $|X| \leqslant \|X\|_\infty - \epsilon$ and $|X + Y| \leqslant \|X + Y\|_\infty - \epsilon$. This ensures our subsequent operations preserve the L^∞ norms of X and Z (Y is unchanged throughout).

(iii) U is either entirely used for the A-optimal matching or entirely unmatched. If matched, assume pair-sum margin is either $\geqslant 3\epsilon$ (A-win) or $\leqslant -3\epsilon$ (A-loss). If unmatched, assume that even if we improve (increase for AC, decrease for CA) the intensity by ϵ, the block remains unmatched.

We argue that the product extension assumption ensures U still has positive measure after shrinking. For (i), the joint law of (X, Y) is absolutely continuous, and so the intersections of the null faces have measure zero. This ensures that we can shrink U accordingly, should it lie on intersections of affine faces. For (ii), the level sets of $|X|, |Y|, |X + Y|$ at their respective L^∞-norms are null in measure, so we can shrink U to stay at least ϵ below these levels. For (iii), $U = (U \cap \{\text{matched}\}) \cup (U \cap \{\text{unmatched}\})$, so one of them must have positive measure; then there must be a positive measure further subset on which the margin is bounded away from 0, regardless of outcome.

Now we define instances $I^\pm$ associated with triples $(X^\pm, Y^\pm, Z^\pm)$, where

$$X^\pm = X \pm \epsilon \mathbf{1}_U, \qquad Y^\pm = Y,$$

and $Z^\pm$ is defined based on $(X^\pm, Y)$ and Equation (15). On U, recall exactly one affine constraint is tight; with respect to this constraint, $\partial Z/\partial X \in \{0, 1/2, 1\}$. With ϵ small, the same affine face remains as the only tight constraint, so there exists $c_U \in \{0, 1/2, 1\}$ with

$$Z^\pm = \begin{cases} Z \pm c_U \cdot \epsilon & \text{on } U \\ Z & \text{everywhere else.} \end{cases}$$

This proves that $t(I) = (t(I^+) + t(I^-))/2$, and because U has nonzero measure, we must have $I^- \neq I^+$. As ϵ is sufficiently small, it preserves all three assumptions above. Hence voter preferences, deliberation outcomes, and $f(\cdot)$ constraints remain unchanged throughout. In particular, $I^\pm \in \mathcal{T}$, and $t(I)$ is not a maximizer of R over $\mathrm{cl}(\mathrm{conv}(\mathcal{T}))$. □

Therefore, R does not attain maximum on $\mathcal{I}$ itself, which justifies why we need the closure space in the first place. (And intuitively, if we disallow ties, there is always room to jitter distances to push for infinitesimal improvements.)

5.5 Structural Characterizations of Optimal Instances

[Qilin: DO NOT READ THIS PART YET. I will instead assume values of (λ, w) and greatly simplify the entire section into one or two short Theorems.]

[Qilin: TODO: some transition here?]

Definition 5.14 (Permissible ranges for $|AC|, |CB|$)**.** Fix a $\lambda \in (1/2, 1)$ and a deliberation weight $w \geqslant 0$. With $f(AC) = 1 - \lambda$, one must have $AC_{\min} \leqslant |AC| \leqslant AC_{\max}$, and likewise, with $f(AC) = \lambda$, $CB_{\min} \leqslant |CB| \leqslant CB_{\max}$

$$AC_{\min} = \frac{1-\lambda}{1+\lambda w} \qquad CB_{\max} = \frac{\lambda(1+w)}{1+\lambda w}$$

$$CB_{\min} = \begin{cases} \dfrac{\lambda - (1-\lambda)w}{1-(1-\lambda)w} & \text{if } w \leqslant \dfrac{2\lambda-1}{1-\lambda} \\ \dfrac{\lambda}{1+(1-\lambda)w} & \text{if } w > \dfrac{2\lambda-1}{1-\lambda} \end{cases} \qquad AC_{\max} = \begin{cases} \dfrac{1-\lambda}{1-(1-\lambda)w} & \text{if } w \leqslant \dfrac{2\lambda-1}{1-\lambda} \\ \dfrac{(1-\lambda)(1+w)}{1+(1-\lambda)w} & \text{if } w > \dfrac{2\lambda-1}{1-\lambda}. \end{cases}$$

These are the quantities that allow the $f(\cdot)$ constraints to be satisfied by winning all deliberations (min) or winning zero deliberation (max). [Qilin: Easily found by solving various equations.]

[Qilin: Should we rewrite our choice of λ, w as λ^*, w^*?]

Remark 5.15. Recall that we assume $(\lambda^*, w^*) = ((3-\sqrt{3})/2, \sqrt{3}-1) \approx (0.6339, 0.7321)$ as in Equation (7). Two reasons justify this choice. For one, by choosing (λ^*, w^*), the following proofs are algebraically simple to follow and a distortion *upper bound* of 3 easily follows. More importantly, for any (λ, w), whether proof is applicable or not, we construct families of instances whose distortion is *at least* 3 under the *deliberation-via-matching* protocol. These combined, we see that choosing (λ^*, w^*) via Equation (7) delivers the *deliberation-via-matching* protocol in its strongest form, all while ensuring elegant proof free of algebraic nightmare.

We now list several algebraic properties of (λ^*, w^*) that will be crucially used later:

(i) $AC^*_{\min} = 0.25, AC^*_{\max} = 0.50, CB^*_{\min} = 0.50$, and $CB^*_{\max} = 0.75$. In particular, $|AC| \leqslant |CA|$ and $|CB| \geqslant |BC|$ always hold for instances meeting the two $f(\cdot)$-constraints.

(ii) Viewing W_A as a function of $|AC|$ subject to $f(AC) = 1 - \lambda^*$, and W_C as a function of $|CB|$ subject to $f(CB) = \lambda^*$, we have the following identities defined on the entirety of feasible ranges:

$$\frac{\mathrm{d}W_A}{\mathrm{d}[|AC|]} = -1, \qquad \frac{\mathrm{d}W_C}{\mathrm{d}[|CB|]} = -2.$$

(iii) For any $|AC| \in [AC^*_{\min}, AC^*_{\max}]$, the A-loss mass (i.e. $|AC| - W_A$) is equal to $2(|AC| - AC^*_{\min})$. Symmetrically, for any $|CB| \in [CB^*_{\min}, CB^*_{\max}]$, the C-win mass W_C equals $2(CB^*_{\max} - |CB|)$.

Proof. (i) follows from definitions. For the AC claim in (ii), $(W_A, |AC|)$ with $|AC| \leqslant 1/2$ must satisfy

$$\frac{|AC| + w^* \cdot W_A}{1 + \underbrace{\min\{|AC|, |CA|\}}_{=|AC|}} = 1 - \lambda^* \qquad \Longrightarrow \qquad W_A = \frac{(1-\lambda^*)(1+w^*\cdot|AC|) - |AC|}{w^*}$$

so the derivative is $(1-\lambda^*) - 1/w^* = -1$. Finally, for the AC-claim in (iii), observe that when $|AC| = AC^*_{\min} = 0.25$ we must have $W_A = 0.25$ (so A-loss is 0), and when $|AC| = AC^*_{\max} = 0.5$, $W_A = 0$ (so A-loss is 0.5). Because $\mathrm{d}W_A/\mathrm{d}[|AC|]$ is constantly -1, along the entire feasible range of $|AC|$, we have $|AC| + W_A = 2AC^*_{\min}$; rearranging gives the claim. The CB-counterparts in (ii) and (iii) are proven analogously. □

Remark 5.16. Suppose X, Y are feasible under (λ^*, w^*) and are coupled counter-monotonically on $[0, 1]$ with X decreasing and Y increasing. Then, $|AC|, |BC| \leqslant 1/2$ by Theorem 5.15(i), and we have the following results:

(i) Let W_A, L_A be the total mass of A-wins and A-losses in the (AC, CA) matching, respectively, so $W_A + L_A = |AC|$. From optimal matching, $[0, 1]$ can be partitioned into five consecutive intervals via $\{0, W_A, |AC|, |AC| + W_A, 2|AC|, 1\}$, with lengths $W_A, L_A, W_A, L_A, 1 - 2|AC|$, respectively. Specifically, the intervals are: $[0, W_A]$ the AC side of A-wins, $[W_A, |AC|]$ the AC side of A-losses, $[|AC|, |AC| + W_A]$ the CA side of A-wins, $[|AC| + W_A, 2|AC|]$ the CA side of A-losses, and $[2|AC|, 1]$ the unmatched (on CA since $|AC| \leqslant 1/2$).

(ii) A direct application of Theorem 5.15(iii) implies that $AC^*_{\min} = 0.25$ is the midpoint of the AC-side A-losses segment $[W_A, |AC|]$.

(iii) Further, the first three intervals add up to $1/2$: with (λ^*, w^*), one always has $|AC| + W_A = 1/2$.

(iv) Likewise, let W_C, L_C be the total mass of C-wins and C-losses in the (CB, BC) matching, so $W_C + L_C = |BC|$. Then, from optimal matching, $[0, 1]$ can be partitioned into five consecutive intervals via $\{0, L_C, |BC|, 1 - |BC|, 1 - |BC| + L_C, 1\}$, with lengths $L_C, W_C, 1 - 2|BC|, L_C, W_C$, respectively. Specifically, the intervals are: $[0, L_C]$ is the BC side of C-losses, $[L_C, |BC|]$ is the BC side of C-wins, $[|BC|, 1 - |BC|]$ is the CB side unmatched, $[1 - |BC|, 1 - |BC| + L_C]$ is the CB side of C-losses, and finally, $[1 - |BC| + L_C, 1]$ is the CB side of C-wins.

(v) Another direct application of Theorem 5.15(iii) implies that $AC^*_{\min} = 0.25$ is also the midpoint of the second segment $[L_C, |BC|]$, which represents the BC-side of C-wins.

(vi) Also, $W_C + 0.5 = 2 \cdot |BC|$, or equivalently $|BC| + L_C = 1/2$.

References

[1] Online deliberation platform. `https://stanforddeliberate.org/`.

[2] Elliot Anshelevich, Onkar Bhardwaj, Edith Elkind, John Postl, and Piotr Skowron. Approximating optimal social choice under metric preferences. *Artif. Intell.*, 264:27–51, 2018.

[3] Elliot Anshelevich, Aris Filos-Ratsikas, Nisarg Shah, and Alexandros A. Voudouris. Distortion in social choice problems: An annotated reading list. *SIGecom Exch.*, 19(1):12–14, 2021.

[4] Elliot Anshelevich and John Postl. Randomized social choice functions under metric preferences. *J. Artif. Intell. Res.*, 58:797–827, 2017.

[5] Lisa P. Argyle, Christopher A Bail, E. Busby, Joshua R Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences of the United States of America*, 120, 2023.

[6] Solomon E. Asch. Opinions and social pressure. *Scientific American*, 193(5):31–35, 1955.

[7] Joshua Ashkinaze, Emily Fry, Narendra Edara, Eric Gilbert, and Ceren Budak. Plurals: A system for guiding LLMs via simulated social ensembles, 2024.

[8] Michiel A. Bakker, Martin J. Chadwick, Hannah R. Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matthew M. Botvinick, and Christopher Summerfield. Fine-tuning language models to find agreement among humans with diverse preferences. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NeurIPS '22, 2024.

[9] Ioannis Caragiannis, Evi Micha, and Jannik Peters. Can a few decide for many? The metric distortion of sortition. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

[10] Moses Charikar and Prasanna Ramakrishnan. Metric distortion bounds for randomized social choice. In Joseph (Seffi) Naor and Niv Buchbinder, editors, *Proceedings of the 2022 ACM-SIAM Symposium on Discrete Algorithms, SODA 2022, Virtual Conference / Alexandria, VA, USA, January 9 - 12, 2022*, pages 2986–3004. SIAM, 2022.

[11] Moses Charikar, Prasanna Ramakrishnan, Zihan Tan, and Kangning Wang. Metric distortion for tournament voting and beyond. In Itai Ashlagi and Aaron Roth, editors, *Proceedings of the 26th ACM Conference on Economics and Computation, EC 2025, Stanford University, Stanford, CA, USA, July 7-10, 2025*, pages 790–818. ACM, 2025.

[12] Moses Charikar, Kangning Wang, Prasanna Ramakrishnan, and Hongxun Wu. Breaking the metric voting distortion barrier. In David P. Woodruff, editor, *Proceedings of the 2024 ACM-SIAM Symposium on Discrete Algorithms, SODA 2024, Alexandria, VA, USA, January 7-10, 2024*, pages 1621–1640. SIAM, 2024.

[13] Morton Deutsch and Harold Benjamin Gerard. A study of normative and informational social influences upon individual judgement. *Journal of abnormal psychology*, 51 3:629–36, 1955.

[14] Brandon Fain, Ashish Goel, Kamesh Munagala, and Sukolsak Sakshuwong. Sequential deliberation for social choice. In Nikhil R. Devanur and Pinyan Lu, editors, *Web and Internet Economics - 13th International Conference, WINE 2017, Bangalore, India, December 17-20, 2017, Proceedings*, volume 10660 of *Lecture Notes in Computer Science*, pages 177–190. Springer, 2017.

[15] J.S. Fishkin. *Democracy and Deliberation: New Directions for Democratic Reform*. Yale University Press, 1991.

[16] Bailey Flanigan, Ariel D. Procaccia, and Sven Wang. Distortion under public-spirited voting. In Kevin Leyton-Brown, Jason D. Hartline, and Larry Samuelson, editors, *Proceedings of the 24th ACM Conference on Economics and Computation, EC 2023, London, United Kingdom, July 9-12, 2023*, page 700. ACM, 2023.

[17] Vasilis Gkatzelis, Daniel Halpern, and Nisarg Shah. Resolving the optimal metric distortion conjecture. In Sandy Irani, editor, *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020, Durham, NC, USA, November 16-19, 2020*, pages 1427–1438. IEEE, 2020.

[18] Ashish Goel, Mohak Goyal, and Kamesh Munagala. Metric distortion of small-group deliberation. In Michal Koucký and Nikhil Bansal, editors, *Proceedings of the 57th Annual ACM Symposium on Theory of Computing, STOC 2025, Prague, Czechia, June 23-27, 2025*, pages 1568–1579. ACM, 2025.

[19] Sean Ingham and Ines Levin. Can deliberative minipublics influence public opinion? Theory and experimental evidence. *Political Research Quarterly*, 71(3):654–667, 2018.

[20] David Kempe. An analysis framework for metric voting based on LP duality. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(02):2079–2086, Apr. 2020.

[21] Soomin Kim, Jinsu Eun, Changhoon Oh, Bongwon Suh, and Joonhwan Lee. Bot in the bunch: Facilitating group chat discussion by improving efficiency and participation with a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery.

[22] Fatih Erdem Kizilkaya and David Kempe. Plurality veto: A simple voting rule achieving optimal metric distortion. In Luc De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 349–355. ijcai.org, 2022.

[23] Nicholas R. Miller. Graph-theoretical approaches to the theory of voting. *American Journal of Political Science*, 21(4):769–803, 1977.

[24] Kamesh Munagala and Kangning Wang. Improved metric distortion for deterministic social choice rules. In Anna R. Karlin, Nicole Immorlica, and Ramesh Johari, editors, *Proceedings of the 2019 ACM Conference on Economics and Computation, EC 2019, Phoenix, AZ, USA, June 24-28, 2019*, pages 245–262. ACM, 2019.

[25] Vincent Price, Lilach Nir, and Joseph N. Cappella. Normative and informational influences in online political discussions. *Communication Theory*, 16(1):47–74, 03 2006.

[26] Ariel D. Procaccia and Jeffrey S. Rosenschein. The distortion of cardinal preferences in voting. In *Proceedings of the 10th International Conference on Cooperative Information Agents*, CIA'06, page 317–331, Berlin, Heidelberg, 2006. Springer-Verlag.

[27] George G. Szpiro. *Numbers Rule: The Vexing Mathematics of Democracy, from Plato to the Present*. Princeton University Press, 2010.

A Characterization of X-Optimal Matchings

[Kamesh: The next theorem seems to state an obvious fact about matchings in measure theory. Maybe the proof should go into an appendix?][Qilin: Yes. The characterization of the matching itself has nothing to do with the rest of the distortion proofs, where we only assume the existence of X-optimal matchings.]

Theorem A.1 (Characterization of X-Optimal Matchings)**.** *Fix X, Y, and let $m = m_{XY} = \min\{|XY|, |YX|\}$. Define*

$$\alpha = \sup_{t \geqslant 0} \max\left\{0, \underbrace{\max\{|\varphi^-(t)| - (|YX| - m), 0\}}_{\text{hard } YX \text{ that must be included}} - \underbrace{\min\{|\varphi^+(t)|, m\}}_{\text{strong enough } XY}\right\} \in [0, m]. \tag{21}$$

Let γ be any matching on (XY, YX). Then the X-win mass $W_{XY}(\gamma) \leqslant m - \alpha$. Moreover, there exists a γ^* attaining this equality, so in particular an X-optimal matching γ^* exists ($\sup_\gamma W_{XY}(\gamma)$ is attained).

Proof. We first prove that $W_{XY}(\gamma) \leqslant m - \alpha$. Let γ be any maximal matching with marginals $S \subset XY, T \subset YX$. Since

$$|\varphi^-(t) \cap T| \geqslant |\varphi^-(t)| - |YX \backslash T| = |\varphi^-(t)| - (|YX| - m),$$

we have in particular $|\varphi^-(t) \cap T| \geqslant \max\{0, |\varphi^-(t)| - (|YX| - m)\}$. Intuitively, voters in $\varphi^-(t)$ are the sufficiently pro-Y (with intensity $\leqslant -t$) so they are "hard" to be convinced into W-wins. The inequality above lower bounds the total number of "hard" YX voters that must be present in a maximal matching. On the other hand, $|\varphi^+(t) \cap S| \leqslant \min\{|\varphi^+(t)|, m\}$ upper bounds the number of XY voters that are sufficiently pro-X (with intensity $\geqslant t$) to possibly beat the "hard" voters. Every voter in $\varphi^-(t) \cap T$ must be matched, and an A-loss is guaranteed if matched to a voter outside $\varphi^+(t)$. Hence, the number of *A-losses* is at least $(|\varphi^-(t) \cap T| - |\varphi^+(t) \cap S|)$ which upper bounds Equation (21) after taking supremum over $t \geqslant 0$. Hence $W_{XY}(\gamma) \leqslant m - \alpha$.

We now exhibit a maximal matching γ^* that incurs precisely α X-losses. First assume $|XY| \leqslant |YX|$ so $m = |XY|$. Let S, T be the truncated sets of XY, YX as defined in **??**.

The core idea is simple and straightforward: since α losses are unavoidable anyways, we "sacrifice" the "weakest" α-mass of voters in S and use them to exhaust the "hardest" YX voters in T. Then, greedily, we will use the remaining $m - \alpha$ strongest voters in S to uniformly beat the remaining $(m - \alpha)$ "easier" YX voters in T. [Kamesh: Replace the next sentence with a citation if any.] (This is known in Chinese history as Tian Ji's horse racing strategy.) The rest of the proof is to formalize this in measure theory.

Observe for all t, $|\varphi^-(t) \cap T| = \max\{\varphi^-(t) - (|YX| - m), 0\}$. This allows us to define CDFs on S and T via

$$F^+(t) = |\{u \in S : \Delta_{XY}(u) \leqslant t\}| = |XY| - S^+(t), \qquad F^-(t) = |\{v \in T : -\Delta_{XY}(v) \leqslant t\}| = m - |\varphi^-(t) \cap T|. \tag{22}$$

By the definition of α, we have $F^-(t) \leqslant F^+(t) + \alpha$ for all t. Now let $Q^+(q), Q^-(q)$ be the ascending quantiles of Δ_{XY} on S and of $-\Delta_{XY}$ on T. Define measure-preserving parameterizations $\pi^+ : [0, m] \to S, \pi^- : [0, m] \to T$ with

$$\Delta_{XY}(\pi^+(q)) = Q^+(q), \qquad -\Delta_{XY}(\pi^-(r)) = Q^-(r) \qquad \text{a.e.}$$

and a shift map $f : [0, m] \to [0, m]$ by $f(q) = q - \alpha \mod m$. Finally, set γ^* to be the pushforward of Lebesgue measure on $[0, m]$ by $q \mapsto (\pi^+(q), \pi^-(f(q)))$. By construction, γ^* has marginals S and T.

We claim that for almost every $q \in [\alpha, m]$, $Q^+(q) \geqslant Q^-(q - \alpha)$. Set $t = Q^-(q - \alpha)$. By definition of Q^-, for every $\epsilon > 0$, we have $F^-(t - \epsilon) < q - \alpha$ and $F^-(t) \geqslant q - \alpha$. Recall that $F^-(t) \leqslant F^+(t) + \alpha$, so

$$F^+(t - \epsilon) \leqslant F^-(t - \epsilon) + \alpha < q.$$

Letting $\epsilon \searrow 0$ and using right-continuity of F^+ yields $F^+(t-) \leqslant q$, hence $Q^+(q) \geqslant t$ by the definition of the quantile as a generalized inverse. Hence, for a.e. $q \in [\alpha, m]$ we have

$$\Delta_{XY}(\pi^+(q)) + \Delta_{XY}(\pi^-(f(q))) = Q^+(q) - Q^-(q-\alpha) \geqslant 0,$$

so those pairs are X-wins. Therefore $W_{XY}(\gamma^*) \geqslant m - \alpha$. Combining this with the upper bound $W_{XY} \leqslant m - \alpha$ completes the proof.

If $|XY| \geqslant |YX|$ then we set $T = YX$ and $S \subset XY$ to be the subset of measure m and largest values of Δ_{XY}. The rest of the proof follows analogously. □

B Issues to Keep Track Off

Notations:

- `geqslant` vs. `geq`
- Usage of italics, boldface, and `\emph{}`.
- "WLOG" vs other variants like w.l.o.g.
- Fix `\Cref` environment names: currently Lemma (and others) are cited as Theorem.