Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Metric Distortion of Committee Election on Perturbation Stable Instances

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΔΗΜΗΤΡΙΟΣ ΝΙΚΟΛΑΟΣ ΑΒΡΑΜΙΔΗΣ

Επιβλέπων : Δημήτριος Φωτάκης
Καθηγητής Ε.Μ.Π.

Αθήνα, Σεπτέμβριος 2025

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Metric Distortion of Committee Election on Perturbation Stable Instances

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΔΗΜΗΤΡΙΟΣ ΝΙΚΟΛΑΟΣ ΑΒΡΑΜΙΔΗΣ

Επιβλέπων : Δημήτριος Φωτάκης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 9η Σεπτεμβρίου 2025.

Δημήτριος Φωτάκης Καθηγητής Ε.Μ.Π.	Αριστείδης Παγουρτζής Καθηγητής Ε.Μ.Π.	Ευάγγελος Μαρκάκης Καθηγητής Ο.Π.Α.

Αθήνα, Σεπτέμβριος 2025

Δημήτριος Νικόλαος Αβραμίδης

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Περίληψη

Στην παρούσα διπλωματική εργασία μελετάμε τον σχεδιασμό κανόνων ψηφοφορίας για την εκλογή επιτροπής υπό μετρικές προτιμήσεις, υπολογίζοντας τη μετρική παραμόρφωση αυτών όταν το υποκείμενο πρόβλημα ομαδοποίησης ικανοποιεί την ιδιότητα της *σταθερότητας διαταραχής*.

Θεωρούμε ένα σύνολο n ψηφοφόρων και ένα σύνολο m υποψηφίων, οι οποίοι είναι τοποθετημένοι σε κάποιον μετρικό χώρο. Στόχος μας είναι η εκλογή μιας επιτροπής k μελών που ελαχιστοποιεί το *κοινωνικό κόστος*, δηλαδή το άθροισμα των αποστάσεων των ψηφοφόρων από το πλησιέστερο μέλος της επιτροπής. Ωστόσο, υποθέτουμε ότι έχουμε πρόσβαση μόνο σε κατάταξεις προτιμήσεων των ψηφοφόρων και όχι στις ακριβείς αποστάσεις.

Η *μετρική παραμόρφωση* ενός κανόνα ψηφοφορίας μετρά το λόγο του κόστους της λύσης που επιλέγεται από τον κανόνα, στη χειρότερη περίπτωση, προς το ελάχιστο δυνατό κοινωνικό κόστος. Παρουσιάζουμε γνωστά αποτελέσματα από τη βιβλιογραφία που παρέχουν άνω και κάτω φράγματα για τη μετρική παραμόρφωση και περιγράφουμε κανόνες που πετυχαίνουν σταθερή μετρική παραμόρφωση με περιορισμένο αριθμό ερωτημάτων αποστάσεων.

Στη συνέχεια, μελετούμε πώς δομικές ιδιότητες που προκύπτουν από τη σταθερότητα διαταραχής μπορούν να αξιοποιηθούν στον σχεδιασμό πιο αποδοτικών αλγορίθμων. Εστιάζουμε στην περίπτωση όπου $k \geq 3$ και το κόστος κάθε ψηφοφόρου ορίζεται ως η απόστασή του από το πλησιέστερο μέλος της επιτροπής. Προηγούμενες εργασίες έχουν δείξει ότι η μετρική παραμόρφωση είναι γενικά μη φραγμένη σε αυτό το πλαίσιο, δίχως ερωτήματα για τις απόστασεις, και ότι απαιτούνται $O(\text{poly}(\log n, k))$ ερωτήματα για ακριβείς αποστάσεις ώστε να επιτευχθεί σταθερή παραμόρφωση. Οι αλγόριθμοι που παρουσιάζουμε επιτυγχάνουν σταθερή μετρική παραμόρφωση χρησιμοποιώντας μόλις $O(2^k)$ ερωτήματα απόστασης, δηλαδή αριθμό ανεξάρτητο από τον αριθμό των ψηφοφόρων.

Λέξεις κλειδιά

Υπολογιστική Θεωρία Κοινωνικής Επιλογής, Εκλογή με πολλούς Νικητές, Εκλογή Επιτροπής, Κανόνες ψηφοφορίας, Μετρική Παραμόρφωση, Σταθερότητα Διαταραχής

Abstract

In this thesis, we study the design of voting rules for committee selection under metric preferences, and we analyze their metric distortion when the underlying clustering problem satisfies the property of *perturbation stability*.

We consider a set of n voters and a set of m candidates, both embedded in some metric space. Our goal is to elect a committee of k members that minimizes the *social cost*, defined as the sum of the distances of all voters to their closest committee member. However, we assume that we only have access to the voters' preference rankings and not to the exact distances.

The *metric distortion* of a voting rule measures, in the worst case, the ratio between the cost of the committee selected by the rule and the minimum possible social cost. We present known results from the literature that provide upper and lower bounds on metric distortion, and we describe rules that achieve constant metric distortion using a limited number of distance queries.

Subsequently, we study how structural properties arising from perturbation stability can be exploited in the design of more efficient algorithms. We focus on the case where $k \geq 3$ and each voter's cost is defined as their distance to the nearest committee member. Previous work has shown that metric distortion is generally unbounded in this setting without access to distance queries, and that $O(\text{poly}(\log n, k))$ distance queries are required to guarantee constant distortion. The algorithms we present achieve constant metric distortion using only $O(2^k)$ distance queries, i.e., a number independent of the number of voters.

Key words

Computational Social Choice, Multi-winner Election, Committee Election, Voting Rules, Metric Distortion, Perturbation Stability.

Ευχαριστίες

Πρωτίστως θα ήθελα να ευχαριστήσω τον καθηγητή κ. Δημήτρη Φωτάκη, για την ευκαιρία που μου παρείχε να ασχοληθώ με το συγκεκριμένο αντικείμενο, και τη μετάδοση του πάθους του για την αλγοριθμική θεωρία παιγνίων και τη θεωρητική πληροφορική μέσα από τα μαθήματά του, για την εμπιστοσύνη που μου έδειξε αναλαμβάνοντας την επίβλεψη της παρούσας διπλωματικής εργασίας και συνολικά για τη βοήθεια και την υποστήριξη που μου παρείχε. Ευχαριστώ επίσης τους καθηγητές κ. Αριστείδη Παγουρτζή και κ. Ευάγγελο Μαρκάκη που συμμετείχαν στην τριμελή επιτροπή. Επιπλέον θα ήθελα να ευχαριστήσω τα μέλη του εργαστηρίου Corelab, για τη βοήθεια τους. Τέλος, ευχαριστώ ιδιαίτερα τους γονείς μου, τον αδερφό μου και τους κοντινούς φίλους μου για τη στήριξη τους την περίοδο των σπουδών μου και συνολικά στη ζωή μου.

Δημήτριος Νικόλαος Αβραμίδης,

Αθήνα, 9η Σεπτεμβρίου 2025

Contents

CHAPTER 1

Εκτενής Ελληνική Περίληψη

Η θεωρία της κοινωνικής επιλογής μελετά πώς οι ατομικές προτιμήσεις μπορούν να συγκεντρωθούν σε μία συλλογική απόφαση [28]. Αν και δεν αποτελεί το μοναδικό πλαίσιο, ένα ιδιαίτερα διαδεδομένο μοντέλο για αυτή την ανάλυση είναι οι εκλογές, όπου οι συμμετέχοντες, γνωστοί ως *ψηφοφόροι*, εκφράζουν προτιμήσεις πάνω σε ένα σύνολο εναλλακτικών, που αποκαλούνται *υποψήφιοι*. Ένας *κανόνας ψηφοφορίας* λαμβάνει ως είσοδο τις προτιμήσεις των ψηφοφόρων και επιλέγει έναν υποψήφιο (ή μια k-μελή επιτροπή) ως νικητή.

Στην ιδανική περίπτωση όπου κάθε ψηφοφόρος αναθέτει σε κάθε υποψήφιο μια αριθμητική (καρδινάλια) χρησιμότητα, ένας φυσικός στόχος είναι να επιλεγεί ο υποψήφιος (ή η επιτροπή) που μεγιστοποιεί τη *κοινωνική ευημερία*—δηλαδή το άθροισμα των χρησιμοτήτων όλων των ψηφοφόρων. Δεδομένης πλήρους πρόσβασης στις τιμές αυτές, η επίλυση του προβλήματος είναι υπολογιστικά τετριμμένη: αρκεί να υπολογίσουμε το συνολικό άθροισμα για κάθε υποψήφιο (ή επιτροπή) και να επιλέξουμε αυτόν (ή αυτήν) με τη μέγιστη τιμή.

Ωστόσο, στις περισσότερες ρεαλιστικές εφαρμογές, η ακριβής αποτύπωση αριθμητικών χρησιμοτήτων είναι ανέφικτη λόγω γνωστικών και χρονικών περιορισμών. Ως αποτέλεσμα, οι περισσότεροι μηχανισμοί ψηφοφορίας βασίζονται σε *διατακτικές* εισόδους, όπου οι ψηφοφόροι παρέχουν μόνο κατατάξεις των υποψηφίων, αντί για ρητές αριθμητικές αξιολογήσεις. Αυτή η απώλεια πληροφορίας συνεπάγεται ότι κανένας κανόνας δεν μπορεί, γενικά, να εγγυηθεί την επιλογή του υποψηφίου (ή της επιτροπής) που μεγιστοποιεί την κοινωνική ευημερία. Η πρόκληση αυτή παρουσιάζει αναλογίες με προβλήματα από τη θεωρία των *προσεγγιστικών αλγορίθμων* [74] και των *online αλγορίθμων* [23], όπου ο στόχος είναι η λήψη αποφάσεων με περιορισμένη ή ατελή πληροφορία.

Για την ποσοτικοποίηση της απώλειας απόδοσης λόγω της διατακτικής πληροφορίας, οι Procaccia και Rosenschein [69] εισήγαγαν την έννοια της *παραμόρφωσης*. Η παραμόρφωση ενός κανόνα ψηφοφορίας ορίζεται ως ο μέγιστος λόγος, στη χειρότερη περίπτωση, της κοινωνικής ευημερίας του βέλτιστου υποψηφίου (ή επιτροπή) προς την ευημερία του υποψηφίου (ή επιτροπής) που επιλέγει ο κανόνας. Η μετρική αυτή έχει αναδειχθεί σε θεμελιώδες εργαλείο για την ανάλυση διατακτικών μηχανισμών, καθώς επιτρέπει τη σύγκριση κανόνων ως προς την ικανότητά τους να προσεγγίζουν το βέλτιστο αποτέλεσμα, και έχει οδηγήσει στον σχεδιασμό νέων κανόνων με στόχο τη μείωση της παραμόρφωσης. Παρ' όλα αυτά, ισχυρά αποτελέσματα αδυναμίας δείχνουν ότι η παραμόρφωση μπορεί να είναι σημαντικά υψηλή, ακόμη και για τυχαιοποιημένους κανόνες [26, 30].

Για να αποκτήσουμε πιο ουσιαστικές εγγυήσεις, οι Anshelevich et al. [8] πρότειναν ένα πιο δομημένο μοντέλο, στο οποίο τόσο οι ψηφοφόροι όσο και οι υποψήφιοι τοποθετούνται σε έναν μετρικό χώρο. Σε αυτό το πλαίσιο, η απόσταση μεταξύ ενός ψηφοφόρου και ενός υποψηφίου αναπαριστά το κόστος που υφίσταται ο ψηφοφόρος αν εκλεγεί ο συγκεκριμένος υποψήφιος, και υποθέτουμε ότι οι ψηφοφόροι

προτιμούν υποψηφίους που είναι πιο κοντά τους. Το μοντέλο αυτό είναι ιδιαίτερα εύστοχο σε εφαρμογές όπως οι πολιτικές εκλογές, όπου οι προτιμήσεις διαμορφώνονται βάσει *ιδεολογικής απόστασης*—δηλαδή του βαθμού στον οποίο οι απόψεις του υποψηφίου ταυτίζονται με τις πεποιθήσεις του ψηφοφόρου. Υπό αυτό το πρίσμα, η κοινωνική ευημερία αντικαθίσταται από το *κοινωνικό κόστος*, που ορίζεται ως το άθροισμα των αποστάσεων όλων των ψηφοφόρων από τον εκλεγμένο υποψήφιο (ή επιτροπή), και ο στόχος γίνεται η ελαχιστοποίησή του.

Αν και το μετρικό μοντέλο οδηγεί σε αυστηρότερα άνω και κάτω φράγματα παραμόρφωσης σε σχέση με το γενικό διατακτικό πλαίσιο, τα εμπόδια παραμένουν. Είναι γνωστό ότι κανένας ντετερμινιστικός διατακτικός μηχανισμός δεν μπορεί να επιτύχει παραμόρφωση μικρότερη από 3 στην εκλογή ενός νικητή [8], ενώ για $k \geq 3$ η παραμόρφωση μπορεί να είναι απεριόριστη [33], ακόμη και σε μετρικά περιβάλλοντα. Τα παραπάνω καταδεικνύουν ένα βαθύ χάσμα ανάμεσα σε αυτό που είναι εφικτό με μόνο διατακτική πληροφορία και σε αυτό που μπορεί να επιτευχθεί με καρδινάλια δεδομένα.

Για την υπέρβαση αυτών των περιορισμών, έχει προταθεί η ενίσχυση του μοντέλου με *περιορισμένη πρόσβαση σε αποστάσεις*, υπό τη μορφή στοχευμένων *ερωτημάτων απόστασης*. Η ιδέα είναι να διατηρηθεί η απλότητα των διατακτικών κανόνων, αλλά να επιτραπεί η στρατηγική χρήση ενός μικρού αριθμού ερωτημάτων, ώστε να βελτιωθεί σημαντικά η ακρίβεια του αποτελέσματος. Ο συνδυασμός αυτού του περιορισμένου καρδινάλιου μοντέλου με μετρικές προτιμήσεις συγκροτεί το βασικό πλαίσιο αυτής της διπλωματικής εργασίας, εντός του οποίου μελετούμε πώς η πληροφορία και η δομή μπορούν να αξιοποιηθούν από κοινού για τον σχεδιασμό μηχανισμών με αποδεδειγμένα χαμηλή παραμόρφωση.

Αξίζει να σημειωθεί ότι το μετρικό αυτό πλαίσιο υποδηλώνει μια φυσική γεωμετρική δομή που ευθυγραμμίζεται με ένα υποκείμενο *πρόβλημα ομαδοποίησης*: οι ψηφοφόροι τείνουν να συγκεντρώνονται γύρω από τους υποψηφίους που προτιμούν, και ο στόχος της ελαχιστοποίησης του κοινωνικού κόστους αντιστοιχεί στην επιλογή αντιπροσωπευτικών κέντρων (υποψηφίων) για αυτές τις ομάδες. Υπό αυτήν την οπτική, η επιλογή υποψηφίου (ή επιτροπής) μπορεί να θεωρηθεί ως ένα πρόβλημα *ομαδοποίησης με περιορισμένη πληροφόρηση*, ανάλογο με προβλήματα k-median και k-center.

Σε αυτό το πλαίσιο, δεδομένου ότι τα προβλήματα ομαδοποίησης ανήκουν στη κλάση NP, είναι εύλογο να εξετάσουμε παραδοχές που εκφράζουν ρεαλιστικές δομικές ιδιότητες των προτιμήσεων, οι οποίες διευκόλυνουν στην επίλυση αυτών. Μία τέτοια παραδοχή είναι η *σταθερότητα διαταραχής* [21, 20], η οποία υποδηλώνει ότι η βέλτιστη λύση παραμένει αμετάβλητη υπό μικρές παραμορφώσεις στις αποστάσεις. Στο περιβάλλον της κοινωνικής επιλογής, αυτό αντανακλά την ιδέα ότι οι εκλογικές βάσεις των υποψηφίων είναι σαφώς διαχωρισμένες και ανθεκτικές σε θόρυβο ή μικρές μεταβολές αντίληψης. Οι ψηφοφόροι παραμένουν κοντά στους προτιμώμενους υποψηφίους τους ακόμη και υπό μικρές μεταβολές, γεγονός που επιτρέπει την ανάπτυξη πιο σταθερών και αποδοτικών αλγορίθμων.

Ο στόχος αυτής της διπλωματικής είναι να εξερευνήσει πώς τέτοιες υποθέσεις σταθερότητας μπορούν να αξιοποιηθούν για την κατασκευή κανόνων με μικρή παραμόρφωση και ελάχιστες απαιτήσεις πληροφορίας, ιδίως στο πολυ-νικητήριο πλαίσιο όπου οι κλασικές προσεγγίσεις αποτυγχάνουν.

1.1 Κανόνες Ψηφοφορίας και Παραμόρφωση

Τα κλασικά μοντέλα στην θεωρία κοινωνικής επιλογής αναπαριστούν την είσοδο κάθε ψηφοφόρου ως μία αυστηρή κατάταξη πάνω στο σύνολο των υποψηφίων, και ο ρόλος του κανόνα ψηφοφορίας είναι να επεξεργαστεί αυτές τις κατατάξεις και να επιστρέψει έναν νικητή. Η διατύπωση αυτή αντανακλά τον φυσικό τρόπο με τον οποίο οι άνθρωποι εκφράζουν προτιμήσεις—με την διάταξη εναλλακτικών—αντί να αποδίδουν ακριβείς αριθμητικές τιμές. Ελλείψει καρδινάλιων χρησιμοτήτων, μια κλασική μεθοδολογία για την αξιολόγηση κανόνων είναι η *αξιωματική προσέγγιση*, όπου επιθυμητές αρχές τυποποιούνται ως αξιώματα και οι κανόνες αξιολογούνται ανάλογα με το ποια από αυτά ικανοποιούν. Εμβληματικές συνεισφορές σε αυτό το πλαίσιο περιλαμβάνουν τα θεωρήματα αδυνατότητας των Ar-

row [14] και Gibbard [49], τον χαρακτηρισμό του May [64], το αποτέλεσμα του Satterthwaite [72], καθώς και την εργασία του Young [78]. Ένα ευρύτερο περίγραμμα της αξιωματικής παράδοσης παρέχεται στην επισκόπηση του Zwicker [79].

Σε αντίθεση με την αξιωματική σκοπιά, η παρούσα διπλωματική υιοθετεί την *ωφελιμιστική* οπτική, με ρίζες στη θεωρία παιγνίων [76] και τον αλγοριθμικό σχεδιασμό μηχανισμών [68]. Σύμφωνα με αυτήν, η προτίμηση κάθε ψηφοφόρου μοντελοποιείται ως μία *συνάρτηση χρησιμότητας* πάνω στο σύνολο των υποψηφίων η οποία λαμβάνει πραγματικές τιμές, και ο συλλογικός στόχος είναι η επιλογή του αποτελέσματος που μεγιστοποιεί τη συνολική χρησιμότητα—την *κοινωνική ευημερία*. Αν και το μοντέλο αυτό δεν είναι κατάλληλο για κάθε εκλογικό σενάριο—ιδίως όταν οι χρησιμότητες δεν είναι συγκρίσιμες μεταξύ ατόμων—παραμένει ιδιαίτερα σχετικό σε πολλές πρακτικές εφαρμογές. Παραδείγματα αποτελούν συστήματα συστάσεων και πλατφόρμες ηλεκτρονικού εμπορίου, όπου οι χρήστες αξιολογούν εσωτερικά τις επιλογές με καρδινάλια κριτήρια, ακόμη κι αν δεν τα αναφέρουν ρητά. Όπως επισημαίνουν οι Boutilier et al. [27], παρότι αυτές οι χρησιμότητες συνήθως παραμένουν «κρυφές», οι πράκτορες μπορούν να παρέχουν διατακτικές κατατάξεις που ευθυγραμμίζονται με τις υποκείμενες προτιμήσεις τους. Επιπλέον, ευρήματα από τη συμπεριφορική επιστήμη υποστηρίζουν ότι οι άνθρωποι δυσκολεύονται να αποδώσουν ακριβείς αριθμούς στις επιλογές τους, γεγονός που ενισχύει την πρακτική αναγκαιότητα εργασίας με διατακτικά δεδομένα.

Όταν διαθέτουμε μόνο διατακτικές κατατάξεις, είναι γενικά αδύνατο ένας κανόνας να αναγνωρίζει πάντοτε τον υποψήφιο που μεγιστοποιεί την κοινωνική ευημερία λόγω έλλειψης πληροφορίας. Το γεγονός αυτό υποδεικνύει μια αλγοριθμική ερμηνεία των κανόνων ψηφοφορίας: μπορούν να θεωρηθούν ως *αλγόριθμοι προσέγγισης* που επιδιώκουν αποτελέσματα σχεδόν βέλτιστα υπό περιορισμένη πληροφορία. Τη σκοπιά αυτή εισήγαγαν οι Procaccia και Rosenschein [69], προτείνοντας την έννοια της *παραμόρφωσης* για την ποσοτική αξιολόγηση της αποτελεσματικότητας ενός κανόνα. Η παραμόρφωση ορίζεται ως ο λόγος, στη χειρότερη περίπτωση, μεταξύ της κοινωνικής ευημερίας του βέλτιστου υποψηφίου (ή επιτροπής) και εκείνης του υποψηφίου (ή επιτροπής) που επιλέγει ο κανόνας. Το πλαίσιο αυτό επιτρέπει αυστηρή, αριθμητική σύγκριση κανόνων—όπου μικρότερη παραμόρφωση σημαίνει καλύτερη απόδοση. Αναλυτική επισκόπηση των κύριων εξελίξεων παρέχεται στο survey των Anshelevich et al. [9].

Τυπικός ορισμός και ορολογία. Έστω V το σύνολο των n ψηφοφόρων και C το σύνολο των m υποψηφίων. Κάθε ψηφοφόρος $v \in V$ διαθέτει συνάρτηση χρησιμότητας $u_v : C \to \mathbb{R}_{\geq 0}$, και η κοινωνική ευημερία ενός υποψηφίου $c \in C$ ορίζεται ως

$$W(c) \;=\; \sum_{v \in V} u_v(c).$$

Ένα προφίλ κατατάξεων $\succ := (\succ_v)_{v \in V}$ είναι *συμβατό* με τις χρησιμότητες $(u_v)_{v \in V}$ αν για κάθε v και $c, c' \in C$ ισχύει $c \succeq_v c' \Rightarrow u_v(c) \geq u_v(c')$. Δεδομένου ότι οι χρησιμότητες είναι ορισμένες μέχρι θετικό γραμμικό μετασχηματισμό, απαιτείται μια *κανονικοποίηση* (π.χ. *unit-sum*: $\sum_c u_v(c) = 1$ για κάθε v, ή *unit-range*: $\max_c u_v(c) - \min_c u_v(c) = 1$) ώστε ο λόγος να είναι νοηματοδοτημένος.

Παραμόρφωση Ένας ντετερμινιστικός κανόνας f που δέχεται προφίλ κατατάξεων $\succ$ και επιστρέφει υποψήφιο $f(\succ) \in C$ έχει *παραμόρφωση*

$$\mathrm{dist}(f) \;=\; \sup_{\succ} \; \sup_{\substack{(u_v) \text{ συμβ. με } \succ \\ \text{και κανονικ.}}} \frac{\max_{c \in C} W(c)}{W\big(f(\succ)\big)}.$$

Για τυχαιοποιημένο κανόνα $\mathcal{R}$ που παράγει κατανομή $\mathcal{R}(\succ)$ πάνω στο C, η παραμόρφωση ορίζεται ως

$$\operatorname{dist}(\mathcal{R}) \;=\; \sup_{\succ} \sup_{\substack{(u_v)\ \text{συμβ. με}\ \succ \\ \text{και κανονικ.}}} \frac{\max_{c\in C} W(c)}{\mathbb{E}_{X\sim\mathcal{R}(\succ)}[\,W(X)\,]}.$$

Η κανονικοποίηση είναι κρίσιμη: χωρίς αυτήν, μια απλή ομοιόμορφη κλιμάκωση των χρησιμοτήτων του ενός ψηφοφόρου θα μπορούσε να εκτινάξει αυθαίρετα τον λόγο, καθιστώντας την παραμόρφωση απροσδιόριστη. Τα συνήθη πρωτόκολλα (*unit-sum*, *unit-range*) θέτουν όλους τους ψηφοφόρους στην ίδια «κλίμακα», επιτρέποντας δίκαιη σύγκριση κανόνων.

Στο γενικότερο περιβάλλον—όπου οι χρησιμότητες είναι αυθαίρετες—οι Procaccia και Rosenschein [69] έδειξαν ότι η παραμόρφωση μπορεί να είναι απεριόριστη, ακόμη και για απλούς και ευρέως χρησιμοποιούμενους κανόνες. Για την αποφυγή αυτής της παθογένειας, εισήγαγαν υποθέσεις κανονικοποίησης (π.χ. *unit-sum*). Ακόμη και τότε, η παραμόρφωση παραμένει υψηλή: για ντετερμινιστικούς κανόνες, το βέλτιστο εφικτό είναι τάξης $\Theta(m^2)$ [32, 31], ενώ οι τυχαιοποιημένοι κανόνες μπορούν να τη μειώσουν σε $\tilde{O}(\sqrt{m})$ [27]. Τα αποτελέσματα αυτά αναδεικνύουν τα όρια των διατακτικών κανόνων σε απεριόριστα περιβάλλοντα και κινητροδοτούν τη στροφή προς πιο δομημένα μοντέλα—όπως το μετρικό πλαίσιο που εξετάζεται στη συνέχεια σε αυτήν τη διπλωματική.

Παραπέρα πλαίσια και επεκτάσεις. Για πληρότητα, αναφέρουμε ότι η παραμόρφωση έχει μελετηθεί και σε πληθώρα συναφών ρυθμίσεων: πρωτόκολλα με *περιορισμένη επικοινωνία* [62, 63], *κατανεμημένα* περιβάλλοντα [46], *μερικές προτιμήσεις* (incomplete/partial preferences) [24], καθώς και *στόχους δικαιοσύνης* και ποικιλίας [39]. Οι γραμμές αυτές ενισχύουν την εικόνα ότι η παραμόρφωση λειτουργεί ως ενιαίο εργαλείο αξιολόγησης κάτω από πολλαπλούς επιχειρησιακούς περιορισμούς—πληροφοριακούς, επικοινωνιακούς ή υπολογιστικούς.

Σύνδεση με μετρικές προτιμήσεις. Η μετάβαση σε *μετρικά* μοντέλα προτιμήσεων επιτρέπει την επιβολή γεωμετρικής δομής (αποστάσεις) που συχνά αντανακλά ρεαλιστικά σενάρια (π.χ. ιδεολογικές αποστάσεις).Η ωφελιμιστική αποτίμηση (μέγιστη ευημερία) έχει ένα φυσικό ανάλογο *κόστους* (ελάχιστο κοινωνικό κόστος) και αντίστοιχη έννοια παραμόρφωσης βασισμένης σε κόστη. Στα επόμενα κεφάλαια, αξιοποιούμε αυτή τη δομή—και, όπου χρειάζεται, περιορισμένη πρόσβαση σε καρδινάλια στοιχεία μέσω ερωτημάτων—για να επιτύχουμε ουσιαστικά καλύτερες εγγυήσεις.

1.2 Μετρική Παραμόρφωση Καθαρά Διατακτικών Κανόνων Ψηφοφορίας

Η παρούσα διπλωματική εργασία επικεντρώνεται στο πλαίσιο της *μετρικής παραμόρφωσης*, όπως αυτό εισήχθη από τους Anshelevich et al. [8], το οποίο μοντελοποιεί τους ψηφοφόρους και τους υποψηφίους ως σημεία σε έναν αφηρημένο μετρικό χώρο. Η βασική υπόθεση είναι ότι οι ψηφοφόροι προτιμούν υποψηφίους που βρίσκονται πλησιέστερα σε αυτούς στον μετρικό χώρο, αντικατοπτρίζοντας την διαισθητική ιδέα ότι η εγγύτητα αντιστοιχεί σε μεγαλύτερη ευθυγράμμιση απόψεων ή προτιμήσεων. Υπό αυτό το πρίσμα, ο στόχος του κάθε ψηφοφόρου—η ελαχιστοποίηση της απόστασης από τον εκλεγμένο υποψήφιο—συμβαδίζει με την ωφελιμιστική επιδίωξη της μέγιστης χρησιμότητας. Η οπτική αυτή ευθυγραμμίζεται με την παράδοση των *χωρικών μοντέλων ψηφοφορίας*, τα οποία έχουν μελετηθεί εκτενώς στην πολιτική επιστήμη [42, 13, 65, 38, 73], όπου η στάση ενός ψηφοφόρου σε ένα πολιτικό ή κοινωνικό ζήτημα αντιστοιχεί σε μία θέση πάνω σε έναν άξονα, π.χ. έναν μονοδιάστατο ιδεολογικό άξονα αριστερά–δεξιά. Ενώ τα κλασικά μοντέλα αυτής της κατηγορίας υιοθετούν ευκλείδεια

γεωμετρία χαμηλής διάστασης, το πλαίσιο που εξετάζουμε επιτρέπει πιο γενικούς μετρικούς χώρους, ώστε να αναπαρίσταται πιο ευέλικτα και ρεαλιστικά η ποικιλομορφία των προτιμήσεων.

Ας ορίσουμε τυπικά τα βασικά συστατικά που θα χρησιμοποιηθούν σε όλη τη διάρκεια της εργασίας. Έστω V και C τα πεπερασμένα σύνολα των *ψηφοφόρων* και των *υποψηφίων*, αντιστοίχως, με $n = |V|$ και $m = |C|$. Οι ψηφοφόροι συμβολίζονται συνήθως με $u, v \in V$ και οι υποψήφιοι με $c, x, y \in C$.

Υποθέτουμε ότι όλοι οι ψηφοφόροι και υποψήφιοι τοποθετούνται σε έναν μετρικό χώρο (X, d), όπου $d : (V \cup C) \times (V \cup C) \to \mathbb{R}_{\geq 0}$ είναι συνάρτηση απόστασης που ικανοποιεί τις ιδιότητες της μετρικής (μη-αρνητικότητα, συμμετρία, τριγωνική ανισότητα). Η απόσταση $d(v, c)$ ερμηνεύεται ως το *κόστος* ή *δυσχέρεια* που βιώνει ο ψηφοφόρος v αν εκλεγεί ο υποψήφιος c.

Για ένα υποσύνολο $W \subseteq C$ μεγέθους k, το οποίο αποκαλούμε *επιτροπή*, το *κοινωνικό κόστος* της επιτροπής W ορίζεται ως:

$$\mathrm{SC}(W, d) = \sum_{v \in V} \min_{c \in W} d(v, c),$$

δηλαδή κάθε ψηφοφόρος αντιστοιχίζεται στον πλησιέστερο υποψήφιο της επιτροπής. Όταν η μετρική d είναι σαφής από τα συμφραζόμενα, γράφουμε απλώς $\mathrm{SC}(W)$.

Μία τριπλέτα (V, C, d) ονομάζεται *στιγμιότυπο* (*instance*). Για κάθε δύο υποψηφίους $c, c' \in C$, λέμε ότι ο ψηφοφόρος $v \in V$ *προτιμά* τον c από τον c', και γράφουμε $c \succ_v c'$, αν ισχύει $d(v, c) < d(v, c')$.

Προφίλ Προτιμήσεων Ένα *προφίλ προτιμήσεων* $\succ := (\succ_v)_{v \in V}$ είναι μία n-άδα αυστηρών ολικών διατάξεων πάνω στο σύνολο των υποψηφίων. Δηλαδή, για κάθε $v \in V$, η σχέση $\succ_v$ είναι μια πλήρης και ασύμμετρη κατάταξη των υποψηφίων C, όπου $c \succ_v c'$ σημαίνει ότι ο ψηφοφόρος v προτιμά αυστηρά τον c από τον c'.

Λέμε ότι μία μετρική d είναι *συμβατή* με το προφίλ $\succ$, και γράφουμε $d \rhd \succ$, αν για κάθε $v \in V$, $c \succ_v c' \Rightarrow d(v, c) < d(v, c')$. Δηλαδή, η μετρική αποτυπώνει επακριβώς την κατάταξη των προτιμήσεων.

Το βασικό πρόβλημα που μας ενδιαφέρει είναι το εξής: ένας αλγόριθμος ALG, γνωστός και ως *κανόνας ψηφοφορίας*, λαμβάνει ως είσοδο ένα προφίλ $\succ$ που είναι συμβατό με κάποια άγνωστη μετρική d, και καλείται να επιλέξει επιτροπή $W \subseteq C$ μεγέθους k, με στόχο την ελαχιστοποίηση του κοινωνικού κόστους $\mathrm{SC}(W, d)$, παρότι δεν έχει πρόσβαση στις πραγματικές αποστάσεις.

Η *παραμόρφωση* ενός αλγορίθμου/κανόνα ALG ορίζεται ως ο λόγος της χειρότερης δυνατής απόδοσής του προς το βέλτιστο. Συγκεκριμένα,

$$\mathrm{distortion}(\mathrm{ALG}) = \sup_{\succ} \sup_{d \rhd \succ} \frac{\mathrm{SC}(\mathrm{ALG}(\succ), d)}{\mathrm{SC}(W^*(d), d)},$$

όπου το $\sup_{\succ}$ τρέχει πάνω σε όλα τα προφίλ προτιμήσεων και το $\sup_{d \rhd \succ}$ πάνω σε όλες τις μετρικές d που είναι *συμβατές* με το προφίλ $\succ$. Εδώ $W^*(d)$ δηλώνει την επιτροπή μεγέθους k που ελαχιστοποιεί το κοινωνικό κόστος υπό τη μετρική d.

Στο Κεφάλαιο 3, αναλύουμε τις δυνατότητες και τους περιορισμούς καθαρά διατακτικών κανόνων ψηφοφορίας υπό μετρικές προτιμήσεις, εστιάζοντας στο πλαίσιο παραμόρφωσης που πρότειναν οι Anshelevich et al. [8]. Ο αλγόριθμος δεν έχει πρόσβαση στις αποστάσεις αλλά οφείλει να επιλέξει μια επιτροπή με μικρό κοινωνικό κόστος, και το χάσμα από το βέλτιστο μετράται μέσω της μετρικής παραμόρφωσης—της βασικής μετρικής μελέτης της παρούσας εργασίας.

Η περίπτωση ενός νικητή. Ξεκινάμε από την απλή περίπτωση $k = 1$. Οι Anshelevich et al. [8] έδειξαν ότι κανένας ντετερμινιστικός κανόνας δεν μπορεί να επιτύχει μετρική παραμόρφωση μικρότερη

από 3, ακόμη και στην περίπτωση με μόλις δύο υποψηφίους. Συγκεκριμένα, τοποθετούν τους υποψηφίους σε μια ευθεία, έτσι ώστε οι μισοί ψηφοφόροι να ταυτίζονται με τη θέση του πρώτου υποψηφίου, ενώ οι υπόλοιποι μισοί να βρίσκονται σε θέση τέτοια ώστε να απέχουν κατά $\epsilon > 0$ λιγότερο από τον δεύτερο υποψήφιο σε σχέση με τον πρώτο. Καθώς $\epsilon \to 0$, ο λόγος του κοινωνικού κόστους στη χειρότερη περίπτωση οποιασδήποτε ντετερμινιστικής επιλογής προς το βέλτιστο τείνει στο 3, καθώς το προφίλ προτιμήσεων $\succ$ δεν επιτρέπει διάκριση μεταξύ των δύο υποψηφίων. Το κάτω αυτό φράγμα είναι βέλτιστο: οι Gkatzelis, Halpern και Shah [50] σχεδίασαν κανόνα που επιτυγχάνει ακριβώς παραμόρφωση 3, χρησιμοποιώντας συνδυαστικά επιχειρήματα. Εντυπωσιακά, ο απλός και πρακτικός κανόνας *Πλειοψηφία-Απαγόρευση* (Plurality-Veto), που πρότειναν οι Kızılkaya και Kempe [59], επιτυγχάνει επίσης παραμόρφωση 3 και παρουσιάζεται αναλυτικά μαζί με απλή απόδειξη.

Η περίπτωση πολλών νικητών. Η πολυπλοκότητα αυξάνεται δραστικά στην περίπτωση $k \geq 2$. Πλέον υπάρχουν διαφορετικοί τρόποι να ορίσουμε το κόστος ενός ψηφοφόρου απέναντι σε επιτροπή: άθροισμα αποστάσεων, ελάχιστη απόσταση, κ.λπ., με διαφορετικές ερμηνείες και ιδιότητες [41, 44]. Στην παρούσα εργασία, εστιάζουμε στο μοντέλο όπου το κόστος κάθε ψηφοφόρου είναι η απόσταση από το πλησιέστερο μέλος της επιτροπής, σε συμφωνία με τους κανόνες των Chamberlin–Courant [34] και Monroe [66].

Η δυσκολία του προβλήματος είναι εμφανής. Οι Caragiannis, Shah και Voudouris [33] ανέδειξαν μια *τριχοτομία* στην παραμόρφωση, ανάλογα με τον τρόπο υπολογισμού του κόστους ψηφοφόρου. Προσαρμόζοντας τα αποτελέσματά τους στην περίπτωση που το κόστος καθε ψηφοφόρου ορίζεται ως η απόσταση του απο το κοντινότερο για αυτόν μέλος της επιτροπής, αποδεικνύεται ότι:

- Για $k \geq 3$, η παραμόρφωση κάθε ντετερμινιστικού κανόνα είναι *μη φραγμένη*.
- Για $k = 2$, η παραμόρφωση είναι φραγμένη αλλά αυξάνεται γραμμικά με τον αριθμό των ψηφοφόρων· παρουσιάζουμε τον αλγόριθμο PolarOpposites, που επιτυγχάνει παραμόρφωση $O(n)$.
- Επιπλέον, αποδεικνύεται ότι ακόμα και τυχαιοποιημένοι κανόνες έχουν παραμόρφωση $\Omega(n)$ για $k = 2$.

Τα ευρήματα αυτά αναδεικνύουν το χάσμα ανάμεσα στην εκλογή ενός και περισσοτέρων υποψηφίων και καταδεικνύουν τα θεμελιώδη όρια των καθαρά διατακτικών μεθόδων. Στην απόδειξη των κάτω φραγμάτων, τόσο για την περίπτωση δύο νικητών όσο και για τρεις ή περισσότερους, η βασική δυσκολία έγκειται στο ότι οποιοσδήποτε ντετερμινιστικός αλγόριθμος που βασίζεται αποκλειστικά στη διατακτική πληροφορία $\succ$ δεν μπορεί να διαπιστώσει αν —και ποιοι— υποψήφιοι βρίσκονται στην ίδια θέση. Το ίδιο προφίλ προτιμήσεων μπορεί να προκύπτει από διαφορετικές μετρικές υλοποιήσεις, στις οποίες οι θέσεις διαφορετικών ζευγών ή υποσυνόλων υποψηφίων ταυτίζονται. Αυτή η αδυναμία διάκρισης υπογραμμίζει την ανάγκη για εμπλουτισμένα μοντέλα με πρόσβαση σε μερική καρδινάλια πληροφορία ή/και τη χρήση τυχαιοποίησης.

Τυχαιοποίηση στον μονο-νικητήριο κανόνα. Παρότι η παραμόρφωση 3 είναι αυστηρή για ντετερμινιστικούς κανόνες, ένα σημαντικό ανοιχτό πρόβλημα είναι το βέλτιστο που μπορεί να επιτευχθεί από τυχαιοποιημένους κανόνες. Οι Anshelevich και Postl [11] έθεσαν κάτω φράγμα 2, και έδειξαν ότι ο κανόνας Random Dictatorship έχει παραμόρφωση < 3, αν και πλησιάζει το 3 όταν το n αυξάνεται. Μεταγενέστερα έργα [43, 58] πρότειναν εναλλακτικούς τυχαιοποιημένους μηχανισμούς με παραμόρφωση που συγκλίνει στο 3 καθώς το πλήθος υποψηφίων αυξάνεται.

Πιο πρόσφατα, οι Charikar και Ramakrishnan [35] ανέβασαν το κάτω φράγμα στο 2.1126, ενώ οι Charikar et al. [36] σχεδίασαν μηχανισμό με παραμόρφωση το πολύ 2.753, σπάζοντας το ιστορικό φράγμα του 3.

Παρακάτω συνοψίζουμε τα γνωστά αποτελέσματα:

Μέγεθος Επιτροπής k	**Πλαίσιο**	**Καλύτερη Παραμόρφωση**
$k = 1$	Ντετερμινιστικό	3 (αυστηρό)
$k = 1$	Τυχαιοποιημένο	$[2,\ 3 - \frac{2}{n}]$
$k = 2$	Ντετερμινιστικό	$\Theta(n)$
$k = 2$	Τυχαιοποιημένο	$\Omega(n)$
$k \geq 3$	Ντετερμινιστικό	Μη φραγμένο
$k \geq 3$	Τυχαιοποιημένο	Μη φραγμένο

Table 1.1: Γνωστά φράγματα παραμόρφωσης για καθαρά διατακτικούς κανόνες υπό μετρικές προτιμήσεις.

Περαιτέρω επεκτάσεις. Πέρα από τις βασικές ρυθμίσεις, έχουν προταθεί ποικίλες επεκτάσεις του πλαισίου παραμόρφωσης. Οι Goel, Lee και Shah [51] εξετάζουν υβριδικά μοντέλα που συνδυάζουν μετρική παραμόρφωση με καρδινάλιους στόχους. Άλλες εργασίες, όπως αυτές των [1, 58], διερευνούν το εμπόριο ανάμεσα σε απόδοση και πληροφοριακή πολυπλοκότητα.

Επιπλέον, η παραμόρφωση έχει μελετηθεί σε κατανεμημένα περιβάλλοντα [10], υπό περιορισμούς ειλικρίνειας [45], ή σε πλαίσια με ελλιπή ή μερική πληροφορία προτιμήσεων [11, 55, 43, 25, 6]. Όλες αυτές οι επεκτάσεις υπογραμμίζουν τη βαθιά αλληλεπίδραση ανάμεσα σε υπόθεση πληροφορίας, υπολογιστική πολυπλοκότητα και εγγυήσεις απόδοσης στην κοινωνική επιλογή.

1.3 Μετρική παραμόρφωση αλγορίθμων με πρόσβαση σε καρδινάλια ερωτήματα

Τα ισχυρά αποτελέσματα αδυναμίας που έχουν εδραιωθεί για τους καθαρά διατακτικούς αλγορίθμους—ιδιαίτερα η μη φραγμένη παραμόρφωση για $k \geq 3$—θέτουν το φυσικό ερώτημα: μπορεί η περιορισμένη πρόσβαση στη μετρική πληροφορία να αποκαταστήσει ουσιαστικές εγγυήσεις σε εκλογές πολλαπλών νικητών; Παρακινούμενοι από πρόσφατη πρόοδο στα μοντέλα με ερωτήματα [3, 5, 4], μελετούμε την επίδραση του εμπλουτισμού των εκλογικών αλγορίθμων με έναν μικρό αριθμό *ερωτημάτων απόστασης* προς τους ψηφοφόρους. Ο εμπλουτισμός αυτός επιτρέπει στον αλγόριθμο να έχει περιορισμένη πρόσβαση σε ακριβείς αποστάσεις, ενώ συνεχίζει να λειτουργεί κυρίως πάνω σε διατακτικές προτιμήσεις.

Στο Κεφάλαιο 4 εξετάζουμε πρόσφατη εργασία των Fotakis et al. [47], η οποία μελετά την αντιστάθμιση μεταξύ παραμόρφωσης και πολυπλοκότητας ερωτημάτων στη μονοδιάστατη Ευκλείδεια περίπτωση, όπου το σύνολο των ψηφοφόρων δεν χρειάζεται να συμπίπτει με το σύνολο των υποψηφίων. Το *εξισωτικό κόστος* (egalitarian cost) μιας επιτροπής ορίζεται ως το μέγιστο κόστος που υφίσταται οποιοσδήποτε ψηφοφόρος, δηλαδή η απόσταση της χειρότερα εξυπηρετούμενης ψηφοφόρου από το πλησιέστερο μέλος της επιτροπής.

Το πρώτο τους αποτέλεσμα εδραιώνει ένα ισχυρό κάτω φράγμα: κάθε ντετερμινιστικός αλγόριθμος που χρησιμοποιεί λιγότερα από $k-2$ ερωτήματα απόστασης μπορεί να έχει *μη φραγμένη* παραμόρφωση, τόσο για το κοινωνικό κόστος όσο και για το εξισωτικό κόστος. Το αποτέλεσμα αυτό υπογραμμίζει

τους έμφυτους περιορισμούς των καθαρά διατακτικών μηχανισμών, ακόμη και σε εξαιρετικά δομημένους μετρικούς χώρους.

Οι συγγραφείς δείχνουν έπειτα ότι το εμπόδιο αυτό μπορεί να ξεπεραστεί με έναν περιορισμένο αριθμό ερωτημάτων. Προσαρμόζουν έναν άπληστο αλγόριθμο εμπνευσμένο από την 2-προσέγγιση του Gonzalez για το k-center, αποδεικνύοντας ότι επιτυγχάνει παραμόρφωση το πολύ $5n$ για το κοινωνικό κόστος και το πολύ 5 για το εξισωτικό κόστος, χρησιμοποιώντας μόνο $O(k)$ ερωτήματα απόστασης. Τα όρια αυτά περιλαμβάνουν προσθετικούς όρους $3n$ και 3, αντίστοιχα, που προκύπτουν από το γεγονός ότι τα σύνολα ψηφοφόρων και υποψηφίων δεν συμπίπτουν.

Βασιζόμενοι σε αυτό, εισάγουν μία πιο εξελιγμένη μέθοδο βασισμένη στην έννοια των (ℓ, β)-δικριτηριακών λύσεων, δηλαδή μικρών υποσυνόλων υποψηφίων των οποίων η δομή προσεγγίζει εκείνη μιας βέλτιστης επιτροπής. Μέσω μίας ιεραρχικής διαδικασίας διαμέρισης, κατασκευάζουν μία $(O(k \log n), 2)$-δικριτηριακή λύση χρησιμοποιώντας μόνο $O(k \log n)$ ερωτήματα. Η μείωση αυτή επιτρέπει την εφαρμογή δυναμικού προγραμματισμού σε ένα περιορισμένο πεδίο, οδηγώντας σε μία επιτροπή με σταθερή παραμόρφωση (το πολύ 5) διατηρώντας πολυωνυμική χρονική πολυπλοκότητα και υπογραμμική χρήση ερωτημάτων.

Στο Κεφάλαιο 5 εξετάζουμε πρόσφατη εργασία των Burkhardt et al. [29], η οποία μελετά την αντιστάθμιση μεταξύ παραμόρφωσης και πολυπλοκότητας ερωτημάτων σε γενικούς μετρικούς χώρους, όπου δεν μπορούν να γίνουν δομικές παραδοχές (όπως η Ευκλείδεια γεωμετρία). Σε αυτό το πλαίσιο, η έλλειψη ολικής διάταξης επιβάλλει το σύνολο των υποψηφίων να συμπίπτει με το σύνολο των ψηφοφόρων προκειμένου να προκύψουν ουσιώδη αποτελέσματα, και όλες οι αποστάσεις μεταξύ πρακτόρων πρέπει να αποκτηθούν ρητά μέσω ερωτημάτων.

Παρότι δεν αναλύουμε λεπτομερώς τα κάτω φράγματά τους σε αυτή τη διατριβή, αξίζει να σημειωθεί ότι εδραιώνουν ισχυρά αποτελέσματα αδυναμίας: κανένας αλγόριθμος που χρησιμοποιεί λιγότερα από $O(k)$ ερωτήματα απόστασης δεν μπορεί να εγγυηθεί φραγμένη παραμόρφωση για οποιονδήποτε (k, z)-στόχο ομαδοποίησης. Επιπλέον, η επίτευξη σταθερής παραμόρφωσης σε σχέση με το κοινωνικό κόστος απαιτεί τουλάχιστον $\Omega(k + \log \log n)$ ερωτήματα όταν το k είναι μεταβλητό, και τουλάχιστον $\Omega(k \cdot 2^{\log^* n})$ όταν το k είναι σταθερό. Τα αποτελέσματα αυτά ενισχύουν την αναγκαιότητα περιορισμένης πρόσβασης στη μετρική πληροφορία, ακόμη και όταν συνδυάζεται με πλήρη διατακτική πληροφόρηση.

Στην θετική πλευρά, και με ιδιαίτερη σημασία για τη διατριβή αυτή, δείχνουν ότι $O(k)$ ερωτήματα απόστασης επαρκούν για την επίτευξη φραγμένης παραμόρφωσης. Συγκεκριμένα, μία προσεκτικά σχεδιασμένη άπληστη διαδικασία, εμπνευσμένη από τον αλγόριθμο του Gonzalez για το k-center, εξασφαλίζει παραμόρφωση το πολύ 4 για το εξισωτικό κόστος και $4n$ για το κοινωνικό κόστος, εκτελώντας μόνο $2k$ ερωτήματα.

Βασιζόμενοι σε αυτό το πλαίσιο, οι συγγραφείς προτείνουν έναν πιο εξελιγμένο αλγόριθμο για τον στόχο του k-median, αξιοποιώντας την έννοια των (ℓ, β)-δικριτηριακών λύσεων και ένα δακτυλιοειδές σχήμα ιεραρχικής διαμέρισης. Η κεντρική συμβολή είναι ένας αλγόριθμος που επιτυγχάνει σταθερή αναμενόμενη παραμόρφωση χρησιμοποιώντας μόνο $O(k^4 \log^5 n)$ ερωτήματα απόστασης. Η βασική ιδέα είναι να προσομοιώσει τη διαδικασία δειγματοληψίας του k-median++ προσεγγίζοντας την κατανομή που βασίζεται στις αποστάσεις μέσω δακτυλιοειδών αποσυνθέσεων, όπου για τα σημεία κάθε δακτυλίου απαιτείται μόνο ένα ερώτημα απόστασης. Ένα ενισχυμένο σχήμα δειγματοληψίας διασφαλίζει ότι περιοχές με υψηλό κόστος είναι πιθανό να επιλεγούν, και μία γεωμετρική ανάλυση μείωσης του ακάλυπτου κόστους εγγυάται ταχεία σύγκλιση. Τέλος, ένα βήμα μείωσης από δικριτηριακή σε κανονική λύση μετατρέπει το προκύπτον σύνολο σε έγκυρη επιτροπή μεγέθους k, διατηρώντας τη σταθερή προσέγγιση. Ο τελικός αλγόριθμος είναι αποδοτικός ως προς τα ερωτήματα και λειτουργεί αποκλειστικά υπό διατακτική πρόσβαση με περιορισμένη μετρική πληροφορία.

Συνολικά, τα αποτελέσματα αυτά δείχνουν ότι η στρατηγική χρήση ενός μικρού αριθμού ερωτημάτων απόστασης επαρκεί για να ξεπεραστεί το εμπόδιο της μη φραγμένης παραμόρφωσης σε εκλογές

πολλαπλών νικητών. Αναδεικνύουν τη δύναμη των υβριδικών μοντέλων που συνδυάζουν διατακτικές προτιμήσεις με περιορισμένη καρδινάλια πρόσβαση, και παρέχουν έναν συγκεκριμένο δρόμο προς τον σχεδιασμό εκλογικών κανόνων που είναι ταυτόχρονα πληροφοριακοί και αποδοτικοί.

1.4 Ομαδοποίηση και Σταθερότητα

Λαμβάνοντας υπόψη τη γεωμετρική ερμηνεία των μετρικών προτιμήσεων, το πρόβλημα της εκλογής μιας κοινωνικά βέλτιστης επιτροπής μπορεί να ιδωθεί φυσικά ως ένα πρόβλημα ομαδοποίησης: κάθε ψηφοφόρος αντιστοιχίζεται στο πλησιέστερο μέλος της επιτροπής, και στόχος είναι η ελαχιστοποίηση του συνολικού κόστους αυτών των αντιστοιχίσεων. Αυτή η προοπτική αποκαλύπτει έναν ισχυρό δεσμό μεταξύ της πολυμελούς εκλογής και των κλασικών στόχων της ομαδοποίησης, όπως το k-median και το k-center, και παρέχει κίνητρο για τη χρήση εργαλείων από τη θεωρία της ομαδοποίησης στον σχεδιασμό και την ανάλυση κανόνων ψηφοφορίας.

Δυστυχώς, αυτοί οι στόχοι είναι υπολογιστικά δύσκολο να βελτιστοποιηθούν (NP-hard) σε γενικούς μετρικούς χώρους, ενώ η παραμόρφωση μπορεί να είναι απεριόριστη στη χειρότερη περίπτωση — ακόμα και με πλήρη πρόσβαση στις αποστάσεις. Για να ξεπεράσουμε αυτούς τους περιορισμούς, υιοθετούμε μια προσέγγιση *πέρα από τη χειρότερη περίπτωση* βασισμένη σε δομικές υποθέσεις. Συγκεκριμένα, εστιάζουμε στη *σταθερότητα διαταραχής* (perturbation stability), μια ευρέως μελετημένη υπόθεση στην ομαδοποίηση, η οποία απαιτεί η βέλτιστη λύση να παραμένει αμετάβλητη υπό περιορισμένες πολλαπλασιαστικές διαταραχές της μετρικής. Η υπόθεση αυτή αντικατοπτρίζει τη διαίσθηση ότι πολλές πραγματικές περιπτώσεις έχουν ισχυρή υποκείμενη δομή, την οποία μπορούμε να εκμεταλλευτούμε αλγοριθμικά.

Στο Κεφάλαιο 6, εξετάζουμε τις συνέπειες της σταθερότητας διαταραχής στην επίλυση του $k-median$ προβλήματος ομαδοποίησης. Ξεκινάμε ορίζοντας επίσημα την έννοια της γ-σταθερότητας διαταραχών και μελετάμε τις δομικές ιδιότητες που αυτή συνεπάγεται, όπως:

- την *γ-εγγύτητα στο κέντρο* (γ-center proximity), η οποία εξασφαλίζει ότι κάθε σημείο είναι σημαντικά πιο κοντά στο δικό του κέντρο από οποιοδήποτε άλλο·
- την *ασθενή εγγύτητα στο κέντρο* (weak center proximity), μια χαλάρωση που ισχύει για τα περισσότερα σημεία, αλλά όχι απαραίτητα για όλα·
- τον *διαχωρισμό των ομάδων* (cluster separation), ο οποίος διασφαλίζει ότι οι ομάδες είναι καλά απομονωμένες·
- και την ιδιότητα της *ελάχιστης σταθερότητας* (min-stability), η οποία σημαίνει ότι η βέλτιστη ομαδοποίηση αντιστοιχεί σε ένα «κλάδεμα» του δέντρου single-linkage.

Δείχνουμε ότι η γ-εγγύτητα στο κέντρο με $\gamma \geq 2 + \sqrt{3}$ συνεπάγεται ελάχιστη σταθερότητα, συνδέοντας έτσι τοπικές συνθήκες απόστασης με παγκόσμιες δομικές εγγυήσεις.

Αυτές οι δομικές ιδιότητες μας επιτρέπουν να σχεδιάσουμε αποδοτικούς αλγορίθμους που ανακτούν τη βέλτιστη επιτροπή υπό σταθερότητα διαταραχών. Συγκεκριμένα, παρουσιάζουμε δύο αλγοριθμικά πλαίσια που επιτυγχάνουν σε σταθερά στιγμιότυπα:

- **Single-Link++**: Ο αλγόριθμος αυτός κατασκευάζει ένα ελάχιστο δέντρο καλύψεως (MST) πάνω στον μετρικό χώρο και εξετάζει όλες τις k-ομαδοποιήσεις που προκύπτουν αφαιρώντας $k-1$ ακμές. Ανάμεσά τους, επιλέγει εκείνη που ελαχιστοποιεί το k-median κόστος. Δείχνουμε ότι για στιγμιότυπα που ικανοποιούν 2-σταθερότητα διαταραχών, η βέλτιστη ομαδοποίηση αντιστοιχεί σε τέτοιο διαχωρισμό, επιτρέποντας την ανάκτησή της σε πολυωνυμικό χρόνο.

- **Ιεραρχική Ομαδοποίηση με Δυναμικό Προγραμματισμό**: Για περιπτώσεις με $\gamma \geq 2 + \sqrt{3}$, κατασκευάζουμε ένα ιεραρχικό δέντρο ομαδοποίησης με single-linkage. Η βέλτιστη ομαδοποίηση εγγυάται ότι θα εμφανιστεί ως ένα «κλάδεμα» αυτού του δέντρου. Ένας δυναμικός αλγόριθμος αναζητά αποδοτικά το βέλτιστο κλάδεμα, αποδίδοντας ακριβή λύση με πολυπλοκότητα $O(nK^2 + nT(n))$, όπου $T(n)$ είναι το κόστος αξιολόγησης του στόχου ομαδοποίησης σε υποδέντρα.

Και οι δύο αλγόριθμοι εκμεταλλεύονται τη συνδυαστική δομή που συνεπάγεται η σταθερότητα διαταραχών για να υπερβούν τα φράγματα της υπολογιστικής δυσκολίας στη χειρότερη περίπτωση. Στο πλαίσιο της ψηφοφορίας, αυτό δείχνει ότι σταθερά προφίλ προτιμήσεων επιτρέπουν την αποδοτική και με περιορισμένα ερωτήματα επιλογή υψηλής ποιότητας επιτροπών, ακόμη και όταν η πρόσβαση στις αποστάσεις είναι μερική.

1.5 Συνεισφορά

Στο Κεφάλαιο 7, μελετάμε πώς η σταθερότητα διαταραχών μπορεί να αξιοποιηθεί για τον σχεδιασμό αποδοτικών ως προς τα ερωτήματα αλγορίθμων εκλογής επιτροπών σε μετρικούς χώρους. Βασιζόμενοι σε υπάρχουσες δομικές παρατηρήσεις από τη βιβλιογραφία της ομαδοποίησης, περιγράφουμε ένα απλό αλλά γενικό σχήμα μείωσης που εντοπίζει ένα μικρό υποσύνολο υποψηφίων—το οποίο αποκαλούμε *μέτωπο* (frontier)—το οποίο, υπό κατάλληλες υποθέσεις σταθερότητας, εγγυάται ότι περιέχει τη βέλτιστη λύση.

Η βασική παρατήρηση είναι ότι σε γ-σταθερά στιγμιότυπα—ιδίως όταν $\gamma \geq 2 + \sqrt{3}$—η βέλτιστη ομαδοποίηση εμφανίζει ισχυρές ιδιότητες διαχωρισμού και δενδρικής (laminar) δομής. Αυτές οι ιδιότητες μας επιτρέπουν να εντοπίσουμε το μέτωπο χρησιμοποιώντας μόνο διατακτική πληροφορία (ordinal information), χωρίς να χρειαστούμε καθόλου ερωτήματα αποστάσεων. Ως εκ τούτου, μπορούμε να κατασκευάσουμε μια συμπαγή αναπαράσταση του χώρου λύσεων, το μέγεθος της οποίας εξαρτάται μόνο από το μέγεθος της επιτροπής k, και όχι από το πλήθος των υποψηφίων n.

Απεικονίζουμε αυτό το σχήμα μείωσης σε δύο διαφορετικά περιβάλλοντα:

- Στον μονοδιάστατο Ευκλείδειο χώρο, δείχνουμε ότι η μεθόριος οδηγεί σε μια $(2^k - 1, 1)$-δικριτή λύση (bicriteria solution). Εφαρμόζοντας έναν γνωστό δυναμικό αλγόριθμο προγραμματισμού στο μειωμένο στιγμιότυπο, λαμβάνουμε έναν ντετερμινιστικό αλγόριθμο με πολυπλοκότητα $O(2^k)$ ως προς τα ερωτήματα και σταθερή παραμόρφωση.

- Σε γενικούς μετρικούς χώρους, όταν οι ψηφοφόροι και οι υποψήφιοι ταυτίζονται, χρησιμοποιούμε το μέτωπο για να κατασκευάσουμε μια $(2^k - 1, 3)$-δικριτή λύση. Συνδυάζοντάς τη με έναν τυπικό αλγόριθμο προσέγγισης, προκύπτει μια προσέγγιση σταθερής παραμόρφωσης με $O(4^k)$ ερωτήματα αποστάσεων.

Παρότι οι αλγόριθμοι που παρουσιάζουμε αποτελούν προσαρμογές υπαρχόντων μεθόδων, στόχος μας είναι να αναδείξουμε πώς οι δομικές ιδιότητες των στιγμιοτύπων με σταθερότητα διαταραχής μπορούν να καθοδηγήσουν τον σχεδιασμό αποδοτικών ως προς τα ερωτήματα αλγορίθμων ψηφοφορίας. Αυτά τα αποτελέσματα συνιστούν ένα μετριοπαθές βήμα προς τη γεφύρωση του χάσματος μεταξύ θεωρητικής δυσκολίας στη χειρότερη περίπτωση και της πρακτικής δομής, και υποδεικνύουν ευρύτερες δυνατότητες αξιοποίησης της σταθερότητας σε προβλήματα μετρικής κοινωνικής επιλογής.

CHAPTER 2

Introduction

Social choice theory studies how individual preferences can be aggregated into a single collective decision [28]. Although not unique, a common framework for this analysis is the setting of elections, where participants, referred to as *voters*, express preferences over a set of alternatives, referred to as *candidates*, and a function, referred to as a *voting rule*, selects one candidate or a k-committee as the winner based on the reported preferences. In the idealized case where each voter assigns a cardinal utility to each candidate, a natural objective is to select the candidate that maximizes the *social welfare*—the sum of utilities across all voters. Given access to these cardinal utilities, this problem is computationally trivial: we simply compute the sum for each candidate (committee) and output the one with the maximum value.

However, in most real-world scenarios, eliciting precise numerical utilities is infeasible due to cognitive and practical limitations. As a result, voting mechanisms typically operate on *ordinal* inputs, where voters provide rankings rather than explicit utilities. This informational constraint implies that voting rules cannot, in general, guarantee the selection of the candidate (committee) with maximum social welfare. This challenge is reminiscent of settings in *approximation algorithms* [74] and *online algorithms* [23], where the goal is to approximate optimal solutions.

Distortion, introduced by Procaccia and Rosenschein [69], is used to quantify how well a voting rule approximates the optimal outcome in terms of social welfare. It is defined as the worst-case ratio between the maximum possible social welfare and the social welfare of the candidate selected by the voting rule. This metric has become a standard tool for analyzing the performance of ordinal mechanisms, motivating both theoretical investigations of existing rules and the design of new rules that aim to minimize distortion. Unfortunately, strong impossibility results show that distortion can be large in the general case [30], even when using randomized mechanisms [26].

To obtain more meaningful guarantees, Anshelevich et al. [8] proposed a more structured model in which both voters and candidates are embedded in a metric space. In this setting, the distance between a voter and a candidate represents the cost to the voter if that candidate is elected, and voters are assumed to prefer candidates that are closer to them. This model is well-motivated in many real-world applications—such as political elections—where a voter's preferences are often determined by an *ideological distance* between themselves and the candidates, capturing how closely a candidate's positions align with their own beliefs. Under this interpretation, the objective naturally shifts from maximizing social welfare to minimizing *social cost*, defined as the sum of distances between all voters and the selected candidate.

Although the metric setting allows for significantly better approximation guarantees compared to the unrestricted case, strong lower bounds still apply even here. For instance, no deterministic ordinal

mechanism can achieve distortion better than 3 in single-winner elections [8]. The situation becomes even more severe in the multi-winner setting: it is known that for $k \geq 3$, the distortion of purely ordinal mechanisms can be unbounded [33], meaning that such rules may select committees with arbitrarily high social cost compared to the optimum. These lower bounds highlight a fundamental gap between what can be achieved using only ordinal information and what is attainable with access to cardinal utilities. To overcome these limitations, we consider augmenting the algorithmic model with limited access to *cardinal information*, in the form of distance queries. This enhancement enables the design of voting rules that retain the simplicity of ordinal inputs while strategically using a small number of metric queries to substantially improve performance. The combination of a metric preference model with limited query access forms the central framework of this thesis, within which we study how structure and information can be jointly exploited to design mechanisms with provably low distortion.

This metric setting induces an underlying geometric structure that naturally defines a *clustering problem*: voters tend to form groups around their preferred candidates, and the objective of minimizing social cost aligns with identifying such candidate-centered clusters. From this perspective, selecting an approximately optimal candidate (or committee) can be viewed as a clustering task, where each cluster is centered around a candidate and consists of voters who are close to them in the metric space.

In light of this connection, it becomes natural to consider structural assumptions that reflect realistic features of voter distributions. One such assumption is that of *perturbation stability* [21, 20], which states that the optimal solution remains invariant under small changes to the distances. In the context of social choice, this captures the idea that voters are meaningfully and consistently grouped around candidates in a way that is resilient to small changes in perception or positioning. That is, the support bases of candidates are well-separated, and the relative proximity of voters to their favored candidates is robust.

The goal of this thesis is to explore how stability assumptions can be leveraged to match or even improve the performance of existing voting rules while requiring significantly less access to distance information, particularly in the multi-winner setting.

2.1 Voting Rules and Distortion

Traditional models in social choice theory represent each voter's input as a strict ranking over the set of candidates, and the role of a voting rule is to process these rankings and return a winning candidate. This formulation reflects how people naturally convey their preferences—by ordering alternatives—rather than assigning them precise numerical values. In the absence of cardinal utilities, one classical method for assessing the quality of voting rules is the *axiomatic approach*, where desirable principles are formalized as axioms. Voting rules are then judged based on which of these axioms they satisfy. Seminal contributions to this framework include the impossibility theorems of Arrow [14] and Gibbard [49], the characterization by May [64], the result of Satterthwaite [72], and the work of Young [78]. A broader overview of this axiomatic tradition is provided in the survey by Zwicker [79].

In contrast to the axiomatic framework, this thesis adopts the *utilitarian perspective*, a viewpoint rooted in game theory [76] and algorithmic mechanism design [68]. According to this approach, each voter's preference is modeled as a real-valued *utility function* over the set of candidates, and the collective goal is to select the outcome that maximizes the aggregate utility—commonly referred to as *social welfare*. Although this model does not suit all voting scenarios—particularly those in which utility values are not comparable across individuals—it proves highly relevant in many practical domains. For example, applications such as recommender systems and e-commerce platforms often rely on agents or users who internally evaluate options using cardinal utility, even if this information is not explicitly reported. As Boutilier et al. [27] point out, while these utilities typically remain

hidden, agents are still able to provide ordinal rankings that align with their underlying preferences. Moreover, behavioral research supports the idea that individuals generally find it difficult to assign exact numerical values to choices, reinforcing the practical necessity of working with ordinal data in many real-world settings.

When only ordinal rankings are available, it is generally impossible for a voting rule to always identify the candidate that maximizes social welfare. This observation motivates an algorithmic interpretation of voting rules: they can be viewed as *approximation algorithms* that aim to select outcomes that are near-optimal despite limited information. This perspective was introduced by Procaccia and Rosenschein [69], who proposed the concept of *distortion* to quantify the effectiveness of a voting rule. Distortion is defined as the worst-case ratio between the social welfare of the optimal candidate and that of the candidate selected by the rule. This framework enables a rigorous, numerical comparison of voting rules—where smaller distortion indicates better performance. A detailed overview of major developments in this line of work can be found in the survey by Anshelevich et al. [9].

Let V be the set of n voters and C the set of m candidates. In the most general setting—where voters have arbitrary utilities—Procaccia and Rosenschein [69] showed that distortion can be unbounded, even for simple and widely-used voting rules. To mitigate this, they introduced a normalization assumption, such as requiring that each voter's total utility over all candidates sums to one. Even under this assumption, distortion remains high: for deterministic rules, the best achievable distortion is $\Theta(m^2)$ [32, 31], while randomized rules can reduce this to $\tilde{O}(\sqrt{m})$ [27]. These results illustrate the limitations of ordinal voting rules in unrestricted environments and motivate the shift toward more structured models such as the metric setting considered in this thesis.For completeness, we note that distortion has also been studied in a variety of other settings, including communication-bounded protocols [62, 63], distributed environments [46], partial preferences [24], and fairness-oriented objectives [39].

2.2 Metric distortion of Purely Ordinal Rules

This thesis focuses on the framework of *metric distortion*, introduced by Anshelevich et al. [8], which models both voters and candidates as points in an abstract metric space. The core assumption is that voters favor candidates who are closer to them in this space—reflecting the intuitive idea that proximity corresponds to alignment of preferences. In this context, a voter's objective of minimizing the distance to the elected candidate naturally parallels the utilitarian goal of maximizing utility. This view aligns with the tradition of spatial voting models widely studied in political science [42, 13, 65, 38, 73], where a voter's stance on an issue can be mapped to a point, such as a position along a left-right ideological spectrum. While such models often assume a one-dimensional Euclidean structure, the metric spaces considered in this work are more general, allowing for a richer and more flexible representation of preferences.

We begin by formally introducing the main components and definitions used throughout this thesis. Let V and C be finite sets representing the set of *voters* and *candidates*, respectively, with $n = |V|$ and $m = |C|$. Voters are typically denoted by $u, v \in V$, and candidates by $c, x, y \in C$.

We assume that all voters and candidates are located in a metric space (X, d), where $d : (V \cup C) \times (V \cup C) \to \mathbb{R}_{\geq 0}$ is a distance function satisfying the metric properties. This distance represents a disutility or cost that a voter experiences when a particular candidate is elected.

Given a subset $W \subseteq C$ of size k, called a *committee*, the *social cost* of W under metric d is defined as

$$\mathrm{SC}(W, d) = \sum_{v \in V} \min_{c \in W} d(v, c),$$

that is, each voter is assigned to their nearest committee member. We will write $\mathrm{SC}(W)$ when the metric d is clear from the context.

A triplet (V, C, d) is called an *instance*. For any two candidates $c, c' \in C$, we say that voter $v \in V$ *prefers* c over c', denoted $c \succ_v c'$, if $d(v, c) < d(v, c')$.

Definition 2.2.1 (Preference Profile)**.** A *preference profile* $\succ := (\succ_v)_{v \in V}$ is an n-tuple of strict total orders, one for each voter. That is, for each $v \in V$, $\succ_v$ is a ranking over the candidates C, where $c \succ_v c'$ means that v strictly prefers c to c'.

We say that a metric d is *aligned* with a preference profile $\succ$, and write $d \rhd \succ$, if for all voters $v \in V$, $c \succ_v c'$ implies $d(v, c) < d(v, c')$. That is, the metric reflects the preference orders.

We now define the central problem of interest. An algorithm ALG, also referred to as a *voting rule*, receives as input a preference profile $\succ$ that is consistent with an unknown metric d. The algorithm must select a committee $W \subseteq C$ of size k with the aim of minimizing the social cost $\mathrm{SC}(W, d)$, even though it has no access to the actual distances in d.

The *distortion* of ALG is defined as the worst-case approximation ratio it may incur over all preference profiles and all metrics aligned with them:

$$\text{distortion}(\mathrm{ALG}) = \sup_{\succ} \sup_{d \rhd \succ} \frac{\mathrm{SC}(\mathrm{ALG}(\succ), d)}{\mathrm{SC}(W^*(d), d)},$$

where $W^*(d)$ denotes an optimal committee of size k minimizing the social cost under the metric d.

In chapter 3, we analyze the power and limitations of voting mechanisms that operate solely on ordinal preference information, within the metric distortion framework introduced by Anshelevich et al. [8]. While the algorithm does not have access to the underlying distance function, it must still strive to select a high-quality committee—one whose total distance to the electorate is close to the optimum. The gap between the performance of a voting rule and the true optimum is captured by the notion of *distortion*, a central metric studied in this work.

We begin with the single-winner setting ($k = 1$), where a tight characterization of deterministic distortion is known. A seminal result of Anshelevich et al.[8] shows that no deterministic rule can achieve distortion better than 3, even in the simplest possible setting with two candidates. This lower bound is tight: Gkatzelis, Halpern, and Shah [50] construct a deterministic algorithm achieving distortion exactly 3, relying on deep combinatorial structure. Remarkably, a significantly simpler voting rule known as the *Plurality-Veto rule*, introduced by Kızılkaya and Kempe [59], also matches this bound and is presented in detail, along with an elementary proof of its performance guarantee.

We now turn to the more general and practically significant case of *multi-winner elections* ($k \geq 2$). Unlike the single-winner setting, there are several natural ways to define a voter's cost for a committee, each giving rise to distinct models of representation with different normative properties [41, 44].

One widely studied approach, considered by Goel et al. [52] and Chen et al. [37], defines a voter's cost as the sum of her distances to all members of the committee. In this framework, Goel et al. [52] showed that applying a single-winner rule with distortion α independently k times yields a multi-winner rule with the same distortion bound α. Consequently, the optimal distortion of 3 established for the single-winner case [50] can also be achieved under this additive cost model. However, such rules tend to select committees that reflect the preferences of majority voters, often at the expense of diversity.

In contrast, our focus lies on the setting where a voter's cost is determined by her distance to the *closest* committee member. This formulation aligns with the objectives of the Chamberlin–Courant [34] and Monroe [66] voting rules, which seek to elect representative and diverse committees that reflect

the distribution of the entire electorate. Under this cost model, the problem becomes significantly more challenging, both in terms of algorithm design and in establishing tight distortion bounds.

Caragiannis, Shah, and Voudouris [33] provide a structural framework revealing a *trichotomy* in distortion depending on how voter costs are defined. Adapting their insights to the standard model where each voter is assigned to their nearest committee member, we derive the following lower and upper bounds.

Specifically, we prove that:

- For $k \geq 3$, the distortion of every deterministic voting rule is *unbounded*.
- For $k = 2$, the distortion can be bounded but grows *linearly* with the number of voters. We describe and analyze the PolarOpposites algorithm, which achieves distortion $O(n)$.
- Finally, we show that even *randomized* rules cannot circumvent this limitation: for $k = 2$, the distortion of every (possibly randomized) rule remains $\Omega(n)$ in the worst case.

These findings highlight a sharp contrast between the single-winner and multi-winner cases, and illustrate fundamental limits of ordinal-only decision making. They motivate further exploration of enhanced models incorporating cardinal feedback or randomization, which are taken up in subsequent chapters.

Following the resolution of the optimal distortion bound for deterministic voting rules, a key open problem remains: determining the best possible distortion achievable by *randomized* algorithms. Anshelevich and Postl [11] established a foundational lower bound of 2, and showed that the Random Dictatorship rule achieves distortion strictly less than 3, although this value converges to 3 as the number of voters increases. Subsequent work by Fain et al. [43] and Kempe [58] introduced alternative randomized mechanisms whose distortion also approaches 3, but in this case as the number of candidates grows.

More recently, Charikar and Ramakrishnan [35] improved the known lower bound to 2.1126, and Charikar et al. [36] designed a randomized rule with distortion at most 2.753, thereby breaking the long-standing barrier of 3 for the first time. Nonetheless, the exact optimal distortion for randomized single-winner rules remains unresolved.

The results above are summarized in the following Table.

Committee Size k	**Setting**	**Best Distortion**
$k = 1$	Deterministic	3 (tight)
$k = 1$	Randomized	$[2, 3 - \frac{2}{n}]$
$k = 2$	Deterministic	$\Theta(n)$
$k = 2$	Any randomized rule	$\Omega(n)$
$k \geq 3$	Deterministic	Unbounded
$k \geq 3$	Any randomized rule	Unbounded

Table 2.1: Distortion bounds for purely ordinal voting algorithms under metric preferences.

Additional lines of research have explored connections and generalizations of the metric distortion framework. For example, Goel, Lee, and Shah [51] consider hybrid models that combine metric distortion with utilitarian objectives, where agents' valuations are arbitrary normalized utilities. Other works examine enriched informational models, such as [1], which allows access to more than just

ordinal preferences, and [58], which investigates the trade-off between achievable distortion and the communication complexity of voting rules.

Metric distortion has also been studied in more constrained settings: in distributed environments [10], under the requirement of truthfulness [45], or when preference information is even more limited than standard rankings [11, 55, 43, 58, 25, 6]. These extensions highlight the rich interplay between informational assumptions, algorithmic complexity, and efficiency guarantees in social choice.

2.3 Metric Distortion of Algorithms with access to cardinal queries

The strong impossibility results established for purely ordinal algorithms—particularly the unbounded distortion for $k \geq 3$ raise a natural question: can limited access to the underlying metric information restore meaningful guarantees in multi-winner elections. Motivated by recent progress on query-based models [3, 5, 4], we study the impact of augmenting voting algorithms with a small number of *distance queries* to voters. This enhancement allows the algorithm to occasionally access exact distances, while still operating primarily on ordinal preferences.

In Chapter 4, we review recent work by Fotakis et al. [47], which investigates the trade-off between distortion and query complexity in the 1-dimensional Euclidean setting, where the set of voters need not coincide with the set of candidates. The *egalitarian cost* of a committee is defined as the maximum cost incurred by any voter, i.e., the distance from the worst-off voter to her closest committee member.

Their first result establishes a strong lower bound: any deterministic algorithm that uses fewer than $k-2$ distance queries might incur *unbounded* distortion, both for the social cost and the egalitarian cost. This result highlights the inherent limitations of purely ordinal mechanisms, even in highly structured metric spaces.

The authors then demonstrate that this barrier can be overcome with a modest number of queries. They adapt a greedy algorithm inspired by Gonzalez's 2-approximation for k-center, showing that it achieves a distortion of at most $5n$ for the social cost and at most 5 for the egalitarian cost, using only $O(k)$ distance queries. These bounds include additive factors of $3n$ and 3, respectively, which arise from the fact that the voter and candidate sets do not coincide.

Building on this, they introduce a more sophisticated method based on the concept of (ℓ, β)-bicriteria solutions, small subsets of candidates whose structure approximates that of an optimal committee. Through a hierarchical partitioning procedure, they construct an $(O(k \log n), 2)$-bicriteria solution using only $O(k \log n)$ queries. This reduction enables the use of dynamic programming over a restricted domain, yielding a committee with constant distortion (at most 5) while maintaining polynomial-time complexity and sublinear query usage.

In Chapter 5, we review recent work by Burkhardt et al. [29], which investigates the trade-off between distortion and query complexity in general metric spaces, where no structural assumptions (such as Euclidean geometry) can be made. In this setting, the lack of total order forces the candidate set to coincide with the set of voters, and all inter-agent distances must be obtained explicitly through queries.

While we do not detail their lower bounds in this thesis, it is worth noting that they establish strong impossibility results: no algorithm using fewer than $O(k)$ distance queries can guarantee bounded distortion for any (k, z)-clustering objective. Moreover, achieving constant distortion with respect to the social cost requires at least $\Omega(k + \log \log n)$ queries when k is variable, and at least $\Omega(k \cdot 2^{\log^\star n})$ when k is fixed. These results reinforce the necessity of limited metric access, even when paired with full ordinal information.

On the positive side, and of particular relevance to this thesis, they show that $O(k)$ distance queries suffice to achieve bounded distortion. In particular, a carefully designed greedy procedure, inspired

by Gonzalez's algorithm for k-center, yields distortion guarantees of at most 4 for the egalitarian cost and $4n$ for the social cost, while issuing only $2k$ queries.

Building on this scaffold, the authors propose a more sophisticated algorithm for the k-median objective, leveraging the concept of (ℓ, β)-bicriteria solutions and a ring-based hierarchical partitioning scheme. The central contribution is an algorithm that achieves constant expected distortion using only $O(k^4 \log^5 n)$ distance queries. The key idea is to emulate the k-median++ sampling procedure by approximating the distance-based distribution over points via ring decompositions, where each ring requires only one distance query. A boosted sampling scheme ensures that high-cost regions are likely to be sampled, and a geometric decay analysis of the uncovered cost guarantees rapid convergence. Finally, a bicriteria-to-true reduction step converts the resulting set into a valid committee of size k, preserving the constant-factor approximation. The resulting algorithm is robust, query-efficient, and operates entirely under ordinal access with limited metric information.

Together, these results demonstrate that strategic use of a few distance queries suffices to overcome the unbounded distortion barrier in multi-winner elections. They highlight the power of hybrid models that combine ordinal preferences with limited cardinal access, and provide a concrete path forward for designing voting rules that are both informative and efficient.

2.4 Clustering and Stability

In light of the geometric interpretation of metric preferences, the task of selecting a socially optimal committee can be naturally viewed as a clustering problem: each voter is assigned to their closest committee member, and the goal is to minimize the total assignment cost. This perspective reveals a strong connection between multi-winner voting and classical clustering objectives such as k-median and k-center, and it motivates the use of algorithmic tools from clustering theory in the design and analysis of voting rules.

Unfortunately, these objectives are NP-hard to optimize in general metric spaces, and distortion can be unbounded in the worst case—even with full access to distance information. To go beyond these limitations, we adopt a *beyond worst-case* viewpoint grounded in structural assumptions. In particular, we focus on *perturbation stability*, a widely studied condition in clustering which assumes that the optimal solution remains unchanged under bounded multiplicative perturbations of the metric. This assumption reflects the intuition that many real-world instances possess a strong underlying structure that can be algorithmically exploited.

In Chapter 6, we explore the implications of perturbation stability for $k - median$ clustering. We begin by formalizing the notion of γ-perturbation stability and studying the structural properties it implies, such as:

- γ-*center proximity*, which guarantees that each point is significantly closer to its own cluster center than to any other;
- *weak center proximity*, a relaxation that holds for most points but not necessarily all;
- *cluster separation*, ensuring that clusters are well-isolated;
- and the *min-stability* property, which implies that the optimal clustering corresponds to a pruning of the single-linkage tree.

We show that γ-center proximity with $\gamma \geq 2 + \sqrt{3}$ implies min-stability, thereby connecting local distance conditions to global structural guarantees.

These structural insights enable the design of efficient algorithms that recover the optimal committee under perturbation stability. In particular, we present two algorithmic frameworks that succeed on stable instances:

- Single-Link++: This algorithm constructs a minimum spanning tree (MST) over the metric space and evaluates all k-clusterings formed by removing $k-1$ edges. Among these, it selects the one minimizing the k-median objective. We show that for instances satisfying 2-perturbation stability, the optimal clustering corresponds to such a partition, allowing the algorithm to recover it in polynomial time.

- Hierarchical Clustering with Dynamic Programming: For instances with $\gamma \geq 2+\sqrt{3}$, we construct a hierarchical clustering tree using single-linkage. The optimal clustering is guaranteed to appear as a pruning of this tree. A dynamic programming routine efficiently searches for the optimal pruning, yielding an exact solution with runtime $O(nK^2 + nT(n))$, where $T(n)$ denotes the cost of evaluating the clustering objective on subtrees.

Both algorithms exploit the combinatorial structure implied by perturbation stability to overcome worst-case hardness barriers. In the context of voting, this demonstrates that stable preference profiles allow for efficient and query-efficient selection of high-quality committees, even when metric information is only partially accessible.

2.5 Contribution

In chapter 7, we investigate how perturbation stability can be leveraged to design query-efficient algorithms for committee selection in metric spaces. Building on existing structural insights from the clustering literature, we describe a simple yet general reduction framework that identifies a small candidate set—referred to as the *frontier*—which is guaranteed to contain the optimal solution under suitable stability assumptions.

The key observation is that in γ-stable instances—particularly when $\gamma \geq 2+\sqrt{3}$—the optimal clustering exhibits strong separation and laminarity properties. These properties allow us to identify the frontier using only ordinal information, without relying on any distance queries. As a result, we can construct a compact representation of the solution space whose size depends only on the committee size k, rather than the total number of candidates n.

We illustrate this framework in two settings:

- In one-dimensional Euclidean spaces, we show that the frontier-based reduction yields a $(2^k - 1, 1)$-bicriteria solution. Applying a known dynamic programming algorithm on this reduced instance leads to a deterministic algorithm with query complexity $O(2^k)$ and constant distortion.

- In general metric spaces, when voters and candidates coincide, we use the frontier to obtain a $(2^k - 1, 3)$-bicriteria solution. We then combine this with a standard approximation algorithm, resulting in a constant-factor approximation using $O(4^k)$ distance queries.

While the algorithms we present are adaptations of existing methods, our goal is to demonstrate how the structural properties of stable instances can guide the design of query-efficient algorithms in voting. These results offer a modest step toward bridging the gap between worst-case hardness and practical structure, and they point to broader opportunities for leveraging stability in metric social choice problems.

CHAPTER 3

Metric distortion of Purely Ordinal Algorithms

In this chapter, we study what can be achieved by voting mechanisms under metric preferences, assuming access only to ordinal information. Voters and candidates are modeled as points in a metric space, where each voter prefers candidates that are closer to them over those that are farther away. The goal is to select a set of k candidates that minimizes the social cost, defined as the total distance from all voters to the chosen committee.

A key concept for evaluating voting mechanisms under metric preferences is **metric distortion**. This measures the worst-case ratio between the social cost of the committee selected by a voting rule and the social cost of the optimal committee, which minimizes the total distance from all voters. In other words, distortion quantifies how far a rule's outcome can be from the best possible, assuming only access to ordinal information—that is, rankings over candidates—rather than the exact distances.

This concept is especially relevant in real-world scenarios where voters may find it difficult to assign exact numerical values to their preferences but can still rank candidates in order of desirability. Metric distortion helps us evaluate how much efficiency is lost when decisions are made using only rankings instead of full (cardinal) preference data.

We begin by presenting a result from Anshelevich, Bhardwaj, Elkind, and Postl [8], which establishes a lower bound of 3 on the distortion of any deterministic voting rule in the single-winner setting. The authors also conjectured that the bound of 3 is tight. This conjecture was later confirmed by Gkatzelis, Halpern, and Shah [50], who presented a polynomial-time deterministic algorithm achieving distortion 3. Their approach builds on a structural result known as the Ranking-Matching Lemma, whose proof relies on a combinatorial conjecture originally posed by Munagala [67]. Although the algorithm of Gkatzelis et al. is more involved, a significantly simpler rule achieving the same distortion was subsequently proposed and analyzed by Kızılkaya and Kempe [59]. In this chapter, we present their version, known as the Plurality-Veto rule, along with a simple proof of its distortion guarantee.

We then discuss results by Caragiannis, Shah, and Voudouris [33], who generalize this framework to the multi-winner case. They characterize the best possible distortion guarantees for both deterministic and randomized algorithms that rely solely on ordinal information.

3.1 Definitions and preliminaries

We now formalize the key components of our model, including the notion of metric spaces, voter preferences, and the definition of social cost. These definitions lay the foundation for the distortion framework discussed in later sections.

Let V and C be two finite sets, representing the set of *voters* and the set of *candidates*, respectively. We denote $n = |V|$ as the number of voters and $m = |C|$ as the number of candidates. Throughout this work, individual voters are typically denoted by u or v, and candidates by c, x, or y.

We now recall the formal definition of a *metric*. Given a non-empty set X, a function $d : X \times X \to \mathbb{R}^+$ is said to be a metric if it satisfies the following conditions:

(i) **Non-negativity and identity of indiscernibles:** For all $a, b \in X$, $d(a, b) \geq 0$, and $d(a, b) = 0$ if and only if $a = b$.

(ii) **Symmetry:** For all $a, b \in X$, $d(a, b) = d(b, a)$.

(iii) **Triangle inequality:** For all $a, b, c \in X$, $d(a, c) \leq d(a, b) + d(b, c)$.

A pair (X, d) satisfying these properties is called a *metric space*.

In our setting, we assume that the sets V and C are embedded in a common metric space (X, d), meaning that the distance $d(a, b)$ is well-defined for all pairs $a, b \in V \cup C$.

Definition 3.1.1 (Social cost)**.** For a voter $u \in V$ and a set $S \subseteq C$ of candidates, we define the *cost* experienced by u from the set S, denoted by $\text{cost}_u(S)$, as:

$$\text{cost}_u(S) = \min_{c \in S} d(u, c),$$

which represents the distance from u to her closest representative in S, under the metric d.

The *social cost* of the set S, denoted $\text{SC}(S)$, is then defined as the sum of the individual costs over all voters:

$$\text{SC}(S) = \sum_{v \in V} \text{cost}_v(S).$$

Definition 3.1.2 (Preference Ranking)**.** A triplet (V, C, d) as described above is referred to as an *instance*. The value $d(v, c)$ represents how much a voter $v \in V$ prefers a candidate $c \in C$; smaller distances indicate stronger preferences. We say that voter v prefers candidate c over candidate c' if $d(v, c) < d(v, c')$, and denote this by

$$c \succ_v c'.$$

Each instance (V, C, d) induces a preference ranking for every voter, defined as a strict total order $\succ_v$ over the candidates C. We write $c \succ_v c'$ if v ranks c strictly above c'.

Definition 3.1.3 (Preference Profile)**.** A *preference profile* $\succ := (\succ_v)_{v \in V}$ is an n-tuple of strict total orders, one for each voter. That is, for each $v \in V$, $\succ_v$ is a ranking over the candidates C, where $c \succ_v c'$ means that v strictly prefers c to c'.

We are now ready to describe the k**-Committee Election Problem**. An algorithm ALG receives as input a preference profile $\succ$, which is induced by an instance (V, C, d) and a positive number k. The algorithm does not have access to the underlying distance function d. The goal is to select a subset $S \subseteq C$ of candidates, called a *committee*, of size $|S| = k \leq m - 1$, that minimizes the total social cost:

$$\text{SC}(S) = \sum_{v \in V} \text{cost}_v(S).$$

We refer to such an algorithm as a *voting rule*, and to its output as the *winning committee* (or winner, in the single-candidate case) under that rule. The quality of a voting rule is evaluated using the notion of *distortion*, introduced by Procaccia and Rosenschein (2006).[69]. Distortion measures how well a rule approximates the optimal solution using only ordinal information.

Given a preference profile $\succ$, a committee size k, the distortion of a rule R is defined as the worst-case ratio between the social cost of the committee selected by R and that of an optimal committee:

$$\text{dist}(R, \succ, k) = \sup \frac{\text{SC}(R(\succ, k))}{\min_{S:|S|=k} \text{SC}(S)}, \tag{1}$$

where the supremum is taken over all metric embeddings of voter and candidate locations that are consistent with the preference profile $\succ$. That is, the metric must satisfy $d(v, c) < d(v, c')$ whenever $c \succ_v c'$.

The distortion of a deterministic k-committee rule R is then defined as the maximum value that $\text{dist}(R, \succ, k)$ attains over all preference profiles $\succ$ with n voters and m candidates.

3.2 Electing a single candidate

We start our investigation with the simplest possible setting: choosing a single candidate ($k = 1$) to minimize total distance to the voters. Despite its simplicity, this case already exhibits strong impossibility results, as well as tight algorithmic guarantees.

3.2.1 Lower Bound for Deterministic Voting Rules

We now present a foundational result from [8], which establishes that no deterministic single-winner voting rule can achieve distortion strictly less than 3. While randomized mechanisms are known to achieve lower distortion in the single-winner setting, we primarily focus on deterministic algorithms in this work and therefore only mention this fact without analyzing it further.

Theorem 3.2.1 (Anshelevich-Bhardwaj-Elkind-Postl-Skowron)**.** *Any deterministic algorithm has worst-case distortion at least* 3 *for the social cost.*

Proof. We analyze a scenario with exactly two candidates, x and w. Suppose the voter population is evenly split: half prefer x over w, and the other half prefer w over x. In other words, the voter set V can be partitioned into two equal subsets, V_1 and V_2, with $|V_1| = |V_2| = \frac{n}{2}$. The preference profile $\succ$ is defined as follows:

- For every voter $v \in V_1$, their ranking is $\succ_v (x) = 1$ and $\succ_v (w) = 2$.
- For every voter $v \in V_2$, their ranking is $\succ_v (x) = 2$ and $\succ_v (w) = 1$.

Now, consider any algorithm ALG, and assume without loss of generality that the algorithm selects w as the winner on this profile. We construct a distance function d that is consistent with the preference profile $\succ$ as follows:

Each voter in V_1 (who prefers x) is located at zero distance from x, i.e., $d(v, x) = 0$, and at distance 2 from w, i.e., $d(v, w) = 2$.

Each voter in V_2 (who prefers w) is located approximately halfway between the two candidates, with $d(v, x) = 1 + \epsilon$ and $d(v, w) = 1 - \epsilon$, for some small $\epsilon > 0$.

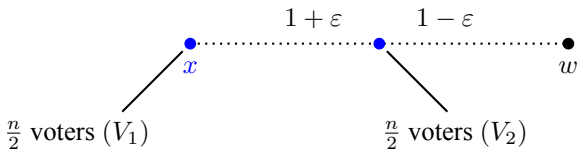


Figure 3.1: Example for Theorem 3.2.1.

We compute

$$\mathrm{SC}(x,d) = \sum_{v \in V} d(v,x) = (1+\epsilon)\,\frac{n}{2}$$

and

$$\mathrm{SC}(w,d) = \sum_{v \in V} d(v,w) = 2\,\frac{n}{2} + (1-\epsilon)\,\frac{n}{2}.$$

This shows that

$$\text{distortion}(\mathrm{ALG}) \geq \frac{2n + (1-\epsilon)n}{(1+\epsilon)n} = \frac{3-\epsilon}{1+\epsilon},$$

which tends to 3 as $\epsilon \to 0$. □

Remark: In the example used in the proof of Theorem 3.2.1, we set $d(x,v) = 0$ even though $x \neq v$, meaning that the function d is technically a pseudometric rather than a true metric. However, this is not a significant violation, as the value 0 could just as well be replaced with any arbitrarily small positive number without affecting the core argument. Throughout the rest of the thesis, we will adopt this convention when convenient.

Having established that no deterministic algorithm can achieve distortion better than 3 in the single-winner setting, a natural question is whether this bound is tight. This conjecture, posed by Anshelevich et al.[8], was later resolved by Gkatzelis, Halpern, and Shah[50], who presented a deterministic algorithm with distortion exactly 3. Their construction relies on a structural result known as the *Ranking-Matching Lemma*, whose proof builds on a combinatorial conjecture originally formulated by Munagala [67].

While this result provides a tight upper bound, the algorithm itself is relatively complex. Fortunately, Kizilkaya and Kempe [59] later showed that an extremely simple rule—called the *Plurality-Veto rule*—also achieves distortion 3. In the remainder of this section, we describe this rule and present their concise and elegant proof of its distortion guarantee.

3.2.2 A Simple Rule with Optimal Distortion: The Plurality-Veto Rule

Algorithm 1 PLURALITY VETO (Kizilkaya, Kempe)

Input: An election $\mathcal{E} = (V, C, \succ)$

Output: A winning candidate $c \in C$

1: **Initialize** $\text{score}(c) \leftarrow \text{plu}(c)$ for each $c \in C$

2: **Let** $(v_1, \ldots, v_n)$ be an arbitrary ordering of V

3: **For** $i = 1, 2, \ldots, n$:

4: $\quad A_i \leftarrow \{c \in C \mid \text{score}(c) > 0\}$

5: $\quad c_i \leftarrow \text{bottom}_{A_i}(v_i)$

6: $\quad$ **decrement** $\text{score}(c_i)$ **by** 1

7: **return** c_n {the candidate remaining at the end}

We now show that this rule achieves distortion at most 3, matching the lower bound. The argument follows a careful use of triangle inequalities and veto mechanics.

Proof. Let j_u denote the candidate vetoed by voter u, and let j^* be the final chosen candidate. Furthermore, for each candidate $j \in C$, let P_j be the set of voters who rank j in first place, and define $\text{plu}(j) = |P_j|$.

Since j^* maintains a strictly positive score until the final step of the algorithm, it must be that $j^* \succcurlyeq_u j_u$ for every voter $u \in V$; in other words, each voter weakly prefers j^* to the candidate j_u that they vetoed.

Now consider any candidate $i \in C$. We will compare the total distance cost of the selected candidate j^* to that of i.

$$
\begin{aligned}
\sum_{v \in V} d(j^*, v) &\leq \sum_{v \in V} d(j_v, v) && (j^* \succcurlyeq_v j_v) \\
&\leq \sum_{v \in V} \left(d(i, v) + d(i, j_v)\right) && \text{(triangle inequality)} \\
&= \sum_{v \in V} d(i, v) + \sum_{j \in C} \text{plu}(j) \cdot d(i, j) && (j \text{ is vetoed } \text{plu}(j) \text{ times}) \\
&= \sum_{v \in V} d(i, v) + \sum_{j \in C} \sum_{v \in P_j} d(i, j) \\
&\leq \sum_{v \in V} d(i, v) + \sum_{j \in C} \sum_{v \in P_j} \left(d(i, v) + d(j, v)\right) && \text{(triangle inequality)} \\
&\leq \sum_{v \in V} d(i, v) + \sum_{j \in C} \sum_{v \in P_j} 2d(i, v) && (v \in P_j \text{ implies } j \succcurlyeq_v i \Leftrightarrow d(j, v) < d(i, v)) \\
&= 3 \sum_{v \in V} d(i, v)
\end{aligned}
$$

Since this holds for any candidate, it must also hold for the optimal candidate. □

The Plurality-Veto rule thus provides a remarkably simple yet provably optimal deterministic so-

lution in the single-winner setting, achieving the minimum possible distortion of 3. However, many real-world applications—such as parliamentary elections, committee formation, or recommendation systems—require selecting not just a single candidate, but a set of representatives. This naturally leads us to the *multi-winner* setting, where the goal is to select a committee of size $k > 1$ that minimizes the total social cost. In the following subsection, we extend our focus to this more general and practically relevant problem, and explore how ordinal algorithms perform in the multi-winner case.

3.3 Multi-winner Voting

To address the general multi-winner setting, Caragiannis, Shah, and Voudouris [33] utilized a model in which each voter's cost is determined by the distance to their q-th closest candidate in the elected committee. Within this framework, they established a striking *trichotomy* in the distortion of multi-winner voting rules, depending on the relationship between the committee size k and the parameter q:

- When $q \leq \frac{k}{3}$, the distortion is unbounded;
- When $\frac{k}{3} < q \leq \frac{k}{2}$, the distortion grows asymptotically linearly with the number of voters;
- When $q > \frac{k}{2}$, the distortion is bounded by a constant.

In the following section, we instantiate the general framework of Caragiannis et al. [33] to the special case where each voter's cost is defined as the distance to their *closest* representative in the committee, that is, we set $q = 1$. Since this is a direct specialization of their model, all of their structural insights and distortion bounds apply without modification. Having already established the tight distortion bounds for the single-winner case ($k = 1$), we now present their results for the multi-winner setting with $k = 2$ and $k \geq 3$.

Unbounded distortion for $k \geq 3$

Theorem 3.3.1 (Caragiannis, Shah, Voudouris)**.** *For every deterministic multi-winner voting rule, the worst-case distortion is unbounded when $k \geq 3$.*

Proof. Let $k \geq 3$ be the committee size, and let f be a deterministic multi-winner voting rule. Define

$$L = k + 1 \geq 4.$$

We set $L = k + 1$, so that the number of voters and candidates exceeds the committee size by one. This ensures that at least one candidate must be excluded from any selected committee, which is crucial for our construction.

We construct an instance with $n = L$ agents, partitioned into two groups:

$$V = \{v_1, \ldots, v_{\lfloor L/2 \rfloor}\}, \quad U = \{u_1, \ldots, u_{\lceil L/2 \rceil}\}.$$

The set of candidates consists of $m = L$ alternatives, divided into:

$$X = \{x_1, x_2, \ldots, x_{\lfloor L/2 \rfloor}\}, \quad Y = \{y_1, y_2, \ldots, y_{\lceil L/2 \rceil}\}.$$

We now define a preference profile consistent with the following:

- Every agent in V ranks all candidates in X above those in Y.

- Every agent in U ranks all candidates in Y above those in X.
- For each $\ell \in [\lfloor L/2 \rfloor]$, agent v_ℓ ranks $x_i \succ x_j$ whenever $|\ell - i| < |\ell - j|$, and ranks the Y-candidates in the fixed order $y_1 \succ y_2 \succ \cdots \succ y_{\lceil L/2 \rceil}$.
- For each $\ell \in [\lceil L/2 \rceil]$, agent u_ℓ ranks $y_i \succ y_j$ whenever $|\ell - i| < |\ell - j|$, and ranks the X-candidates in reverse order: $x_{\lfloor L/2 \rfloor} \succ \cdots \succ x_1$.

Since $m = L > k$, not all alternatives can be included in the committee. We distinguish between two exhaustive cases:

Voter	Candidate ranking
v_1	$x_1 \succ x_2 \succ y_1 \succ y_2$
v_2	$x_2 \succ x_1 \succ y_1 \succ y_2$
u_1	$y_1 \succ y_2 \succ x_2 \succ x_1$
u_2	$y_2 \succ y_1 \succ x_2 \succ x_1$

Table 3.1: Example preference profile for Theorem 3.3.1, instantiated with committee size $k = 3$ and four alternatives.

(a) The metric space in Case 1.

(b) The metric space in Case 2.

Figure 3.2: The two metric spaces illustrated correspond to the construction used in the proof of Theorem 3.3.1, for the case $k = 3$. Both metrics are consistent with the ordinal preferences given in Table 3.1. In the first metric, if the committee fails to include both alternatives in $Y = \{y_1, y_2\}$, then either u_1 or u_2 will incur a positive social cost. However, selecting $\{y_1, y_2\}$ along with any one alternative from $X = \{x_1, x_2\}$ results in a total social cost of zero. On the other hand, if all alternatives from Y are included in the committee, then only one alternative from X can be selected. This means that in the second metric, either v_1 or v_2 will incur a positive cost. Since in this case, the optimal committee $\{x_1, x_2\}$ together with any single alternative from Y also achieves zero cost, the distortion remains unbounded.

Case 1: *At least one candidate in Y is excluded from the committee.*

Suppose that some alternative $y_{\ell^*} \in Y$ is not selected, for some $\ell^* \in [\lceil L/2 \rceil]$. Consider the following one-dimensional Euclidean embedding that respects the given rankings:

- All agents in V and all candidates in X are located at position 0.
- For each $\ell \in [\lceil L/2 \rceil]$, agent u_ℓ and candidate y_ℓ are placed at position $\lceil L/2 \rceil + \ell$.

Since y_{ℓ^*} is excluded from the committee, agent u_{ℓ^*} must be matched to a farther candidate, incurring strictly positive cost. Which implies that the total social cost under f is also strictly positive.

However, we can achieve zero social cost by selecting all alternatives in Y, and filling the remaining $k - \lceil L/2 \rceil$ committee spots with any subset of candidates from X. This is feasible under the

assumption $k \geq 3$, since it guarantees $k - \lceil L/2 \rceil \geq 1$. Hence, in this case, the distortion of f is unbounded.

Case 2: *All candidates in Y are included in the committee.*

Now suppose that f selects all candidates in Y. Consider a different one-dimensional Euclidean embedding consistent with the same rankings:

- For each $\ell \in [\lfloor L/2 \rfloor]$, agent v_ℓ and candidate x_ℓ are placed at position $-L + \ell$.
- All agents in U and all candidates in Y are placed at position 0.

Since the committee includes all of Y, at least one alternative from X must be omitted (as $|Y| = \lceil L/2 \rceil > k - 1$). Therefore, some candidate $x_\ell \in X$ is excluded, and the corresponding agent v_ℓ incurs a cost of at least 1. Consequently, the total social cost is at least 1.

On the other hand, selecting all alternatives in X, along with any $k - \lfloor L/2 \rfloor$ candidates from Y, yields zero social cost. Again, this is feasible since $k - \lfloor L/2 \rfloor \geq 1$ under the assumption $k \geq 3$. Thus, the distortion of f is unbounded in this case as well.

We conclude that for any deterministic voting rule f, if $k \geq 3$, the worst-case distortion can be made arbitrarily large. □

Remark. While the proof above is tailored to deterministic voting rules, the lower bound remains valid even for randomized algorithms. In particular, if a randomized rule fails to include some alternative from Y with non-zero probability (as in Case 1), or fails to include some alternative from X with non-zero probability (as in Case 2), then there exists a consistent metric embedding in which the expected social cost is strictly positive—while the optimal solution achieves cost zero. Hence, the distortion remains unbounded in expectation.

Linear Distortion for $k = 2$

We now focus on the case where $k = 2$. In this setting, Caragiannis, Shah, and Voudouris [33] demonstrated that although the distortion can be bounded, it remains linear in the number of agents, which may be very large in practice.

To address this, they introduced a deterministic multi-winner voting rule called *PolarOpposites*, which runs in polynomial time and achieves a distortion of $O(n)$. In what follows, we present the PolarOpposites algorithm—a conceptually simple yet effective rule—and provide an analysis of its distortion guarantee. While the algorithm itself is relatively straightforward, the upper bound analysis requires a more delicate argument.

Algorithm 2 Constructing the set S for structural guarantee

Input: Voters V and optimal committee O

Output: A subset $S \subseteq V$

1: **Initialize** $S \leftarrow \emptyset$

2: **Sort** the voters in V in non-decreasing order of $c_i(O)$

3: **For each** voter i in this sorted order:

4: **If** $\nexists\, j \in S$ such that $\text{top}_j(O) = \text{top}_i(O)$:

5: **Add** i to S

6: **return** S

To that end we first present a structural lemma of [33] which is useful to the proof of the upper bound.

Lemma 3.3.2. *Let $I = (V, C, d, k)$ be an instance with a set of voters V, candidates C, a metric d, and a desired committee size k. Let $O \subseteq C$ be an optimal committee of size k, minimizing the total cost SC(O). Then, there exists a subset of voters $S \subseteq V$ with $|S| \leq k$, such that for every voter $i \in V$, there exists a voter $j \in S$ satisfying:*

- *top$_i(O)$ = top$_j(O)$, and*
- *$c_j(O) \leq c_i(O)$,*

where top$_i(A)$ is the most-preferred candidate by voter i among a subset $A \subseteq C$, and define the corresponding cost as $c_i(A) = d(i, \text{top}_i(A))$. When the subset under consideration is the entire set of candidates C, we simplify the notation and write top$_i$ instead of top$_i(C)$, and similarly $c_i = d(i, \text{top}_i)$.

Furthermore, for any committee $C \supseteq \{\text{top}_j(A) : j \in S\}$, it holds that

$$c_i(C) \leq 3 \cdot c_i(O), \quad \textit{for all } i \in V.$$

Proof. We construct the set $S \subseteq V$ using Algorithm 2. Since we are only interested in proving the existence of such a set, we assume access to the underlying cost values.

By construction, for each voter $i \in V$, either $i \in S$, in which case the condition in the lemma holds trivially for $j = i$, or there exists a voter $j \in S$ who was considered before i in Algorithm 2 and satisfies $\text{top}_j(O) = \text{top}_i(O)$ and $c_j(O) \leq c_i(O)$.

Since each voter $j \in S$ contributes a distinct top choice from the optimal committee O, and each such choice belongs to O, we must have $|S| \leq |O| = k$.

For the second claim, consider any committee $C' \supseteq \{\text{top}_i(C) : i \in S\}$. Clearly, $c_i(C') \leq c_i(O)$ for every $i \in S$, since $\text{top}_i(C) \in C'$ and C is the candidate set.

By the property of S established above, for any voter $i \in V \setminus S$, there exists a voter $j \in S$ such that $\text{top}_i(O) = \text{top}_j(O)$ and $c_j(O) \leq c_i(O)$. Let $x = \text{top}_i(O) = \text{top}_j(O)$, and let $y = \text{top}_j(C') \in C'$ be the most-preferred candidate of j in the committee C'. We make the following observations:

- Since $x = \text{top}_i(O)$, it follows that $d(i, x) = c_i(O)$,
- Since $x = \text{top}_j(O)$, we have $d(j, x) = c_j(O) \leq c_i(O)$,
- Since $y = \text{top}_j(C') \in C'$ and $\text{top}_j(C) \subseteq C'$, it follows that $d(j, y) \leq c_j(C') \leq c_j(O) \leq c_i(O)$.

By the triangle inequality:

$$c_i(C') \leq d(i, y) \leq d(i, x) + d(j, x) + d(j, y) \leq 3 \cdot c_i(O).$$

This concludes the proof. □

Algorithm 3 PolarOpposites for $k = 2$

Choose an arbitrary voter $i \in V$
Choose an agent $j \in \arg\max_{\ell \in V \setminus \{i\}} c_i(\{\text{top}_\ell\})$
If $\text{top}_i \neq \text{top}_j$, **then**
    $W \leftarrow \{\text{top}_i, \text{top}_j\}$
Else
    Choose an arbitrary candidate $a \in C \setminus \{\text{top}_i\}$
    $W \leftarrow \{\text{top}_i, a\}$
return W

Theorem 3.3.3 (Caragiannis, Shah, Voudouris)**.** *The distortion of PolarOpposites for $k = 2$ is $O(n)$, where n is the number of voters.*

Proof. Let $I = (V, C, d, k = 2)$ be an instance. Let i and j be the agents chosen by PolarOpposites on I, let W be the committee returned by it, and let $O \in \arg\min_{C':|C'|=2} \text{SC}(C')$ be an optimal committee for I.

We will show that for every agent $\ell \in V$, it holds that

$$c_\ell(W) \leq c_\ell(O) + 4 \cdot \text{SC}(O).$$

By summing over all agents, we obtain that

$$\text{SC}(W) \leq (4n + 1) \cdot \text{SC}(O),$$

thus implying an upper bound of $4n + 1$ on the distortion of PolarOpposites.

We distinguish between the following two cases:

Case 1: $top_i(O) = top_j(O) = x$.

For any agent $\ell \in V$, since $top_j \in W$, we have

$$c_\ell(W) \leq d(\ell, top_j).$$

Using the triangle inequality,

$$d(\ell, top_j) \leq d(\ell, top_\ell) + d(i, top_\ell) + d(i, x) + d(j, x) + d(j, top_j).$$

Now observe:

- $d(\ell, top_\ell) = c_\ell \leq c_\ell(O)$.
- $d(i, top_\ell) \leq d(i, top_j)$, since j was chosen to maximize $c_i(top_j)$.
- $d(i, x) = c_i(O)$, and similarly $d(j, x) = c_j(O)$.
- $d(j, top_j) = c_j \leq c_j(O)$.

So

$$c_\ell(W) \leq c_\ell(O) + d(i, top_j) + c_i(O) + 2\, c_j(O).$$

Again by the triangle inequality,

$$d(i, top_j) \leq d(i, x) + d(j, x) + d(j, top_j) \leq c_i(O) + 2\, c_j(O),$$

and hence

$$c_\ell(W) \;\leq\; c_\ell(O) + 2\, c_i(O) + 4\, c_j(O).$$

Finally, since $\mathrm{SC}(O) \geq c_i(O) + c_j(O)$, we get

$$c_\ell(W) \leq c_\ell(O) + 4\, \mathrm{SC}(O).$$

Case 2: $\mathrm{top}_i(O) \neq \mathrm{top}_j(O)$.

Consider the set S guaranteed to exist by Lemma 3.3.2. Since $k = 2$ we have that $|S| \leq 2$.

If $|S| = 1$, then there exists a single agent $u \in S$ such that for every voter $\ell \in V$, it holds that $\mathrm{top}_\ell(O) = \mathrm{top}_u(O)$. In this case, the proof is immediate.

If $|S| = 2$, we claim that there exists a function $g : V \to S$ such that for every agent $\ell \in V$, it holds that $\mathrm{top}_\ell(O) = \mathrm{top}_{g(\ell)}(O)$, and moreover, that $S = \{g(i), g(j)\}$.

Lemma 3.3.2 guarantees the existence of a function g satisfying the first condition. Suppose for the sake of contradiction that there exists an agent $u \in S$ such that $\mathrm{top}_u(O) \notin \{\mathrm{top}_i(O), \mathrm{top}_j(O)\}$. Then, $\mathrm{top}_i(O)$, $\mathrm{top}_j(O)$, and $\mathrm{top}_u(O)$ would all be distinct elements of O, contradicting the fact that $|O| = 2$.

Therefore, such a function g exists and maps every agent $\ell \in V$ to either $g(i)$ or $g(j)$, with $\mathrm{top}_\ell(O)$ equal to one of $\mathrm{top}_i(O)$ or $\mathrm{top}_j(O)$.

Now consider any agent $\ell \in V$, and suppose that $g(\ell) = g(j) = u \in S$. (The case where $g(\ell) = g(i)$ is analogous and handled similarly.)

By the properties of S, there exists an alternative x such that

$$x = \mathrm{top}_\ell(O) = \mathrm{top}_u(O) = \mathrm{top}_j(O).$$

Since $\mathrm{top}_j \in W$, it follows that

$$c_\ell(W) \leq d(\ell, \mathrm{top}_j).$$

Applying the triangle inequality, we have:

$$d(\ell, \mathrm{top}_j) \leq d(\ell, x) + d(x, j) + d(j, \mathrm{top}_j).$$

Next, we bound each term individually:

- $d(\ell, x) = c_\ell(O)$,
- $d(x, j) \leq c_j(O)$,
- $d(j, \mathrm{top}_j) = c_j \leq c_j(O)$.

Combining these inequalities, we obtain:

$$c_\ell(W) \leq c_\ell(O) + 2c_j(O).$$

Since $SC(O) \geq c_j(O)$, it follows that:

$$c_\ell(W) \leq c_\ell(O) + 2SC(O).$$

Finally, noting that $2SC(O) \leq 4SC(O)$, we obtain:

$$c_\ell(W) \leq c_\ell(O) + 4SC(O),$$

as desired.

□

(a) The metric space in Case 1. **(b)** The metric space in Case 2.

The two metric spaces considered in the proof of Theorem 3.3.1.

Theorem 3.3.4 (Caragiannis, Shah, Voudouris)**.** *For k=2 the distortion of every (even randomized) multi-winner voting rule is* $\Omega(n)$.

Proof. Let f be an arbitrary multi-winner voting rule. We consider instances with $n = 2x + 1$ agents, partitioned into two sets V and U of size x each, and a singleton set $\{w\}$.

There are 3 alternatives, named x, y, and z.

Consider any preference profile subject to the following rules:

- Every agent in V has the ranking $x \succ y \succ z$.
- Every agent in U has the ranking $y \succ x \succ z$.
- Agent w has the ranking $z \succ y \succ x$.

Since $m = 3 > k$, the committee returned by f cannot include every alternative.

We now distinguish between the following two cases. Figure 2 depicts the two metric spaces considered in these cases.

Case 1: The committee chosen by f does not include alternative z.

Consider the following metric, which is consistent (up to tie-breaking) with the preference profile defined above:

- The agents in $V \cup U$ and the alternatives x and y are all located at 0.
- Agent w and the alternative z are located at 1.

Since alternative z is not included in the chosen committee, the expected cost of agent w, and thus the social cost under f, is strictly positive.

However, the committee that includes z and any of the alternatives x and y has social cost 0. Therefore, the distortion of f is unbounded in this case.

Case 2. The committee chosen by f includes alternative z.

Consider the following metric, which is again consistent (up to tie-breaking) with the preference profile defined above:

- The agents in V and the alternative x are located at 0.
- The agents in U and the alternative y are located at 1.
- Agent w and the alternative z are located at 2.

Since f selects a committee that includes alternative z, any such committee can include at most one of $\{x, y\}$.

If it does not include alternative x, then the cost of every agent in V is at least 1. Conversely, if it does not include alternative y, then the cost of every agent in U is at least 1.

Either way, f selects a committee with social cost at least n.

On the other hand, the committee that includes both alternatives x and y social cost 1, since only agent w incurs a cost of 1.

Thus, the distortion of f is at least $x = \Omega(n)$ in this case.

□

In summary, this chapter established tight distortion bounds for deterministic ordinal voting mechanisms under metric preferences. While deterministic rules achieve optimal constant distortion in the single-winner scenario, our results reveal inherent limitations in the multi-winner context, including linear and even unbounded distortion. These findings naturally motivate an exploration of randomized mechanisms or mechanisms enhanced with limited cardinal information, which we explore in subsequent chapters

CHAPTER 4

Committee Selection with Metric Preferences on the real line and Query Access

The previous chapter highlighted the limitations of purely ordinal algorithms in the context of committee selection under metric preferences. While constant distortion is achievable in the single-winner setting, the multi-winner case suffers from inherent gaps between ordinal information and social cost, leading to linear or even unbounded distortion.

In this chapter, we explore how access to partial metric information, through selective distance queries, can overcome these limitations. Specifically, we focus on the one-dimensional Euclidean setting, where voters and candidates lie on the real line, and distances correspond to absolute differences in position.

We present a sequence of results by Fotakis, Gourvès, and Patsilinakos [48], who study the trade-off between distortion and query complexity in this setting. They first show that any deterministic algorithm achieving bounded distortion must make at least $\Omega(k)$ distance queries, where k is the desired committee size. They then match this lower bound with a simple greedy algorithm that achieves linear distortion using only $O(k)$ queries. Finally, they propose an improved algorithm that increases the query complexity to $O(k \log n)$, where n is the number of voters, but guarantees constant distortion.

These results demonstrate that even limited access to distance information, when carefully leveraged, enables the design of algorithms with strong performance guarantees, significantly improving over what is possible in the purely ordinal model.

4.1 Model and Preliminaries

We consider a set $C = \{c_1, \ldots, c_m\}$ of m candidates and a set $V = \{v_1, \ldots, v_n\}$ of n voters. As in the previous chapter, we assume that voters and candidates are located in a metric space, and each voter prefers candidates that are closer to her. Here, we enrich this model by assuming a one-dimensional Euclidean structure and allowing limited access to cardinal information through distance queries.

Formally, each candidate $c \in C$ and each voter $v \in V$ is associated with a location $x(c), x(v) \in \mathbb{R}$ on the real line. For simplicity, we often identify each agent with her location. Candidates are indexed so that $x(c_1) < x(c_2) < \cdots < x(c_m)$, which defines the *candidate axis*, a fixed left-to-right ordering consistent with the embedding.

The cost for a voter $v \in V$ to be represented by a candidate $c \in C$ is given by their distance: $\text{cost}_v(c) = d(v, c) = |x(v) - x(c)|$. For a set $S \subseteq C$, we define $\text{cost}_v(S) = \min_{c \in S} d(v, c)$, and the *social cost* of a committee $S \subseteq C$ of size k is

$$\text{SC}(S) = \sum_{v \in V} \text{cost}_v(S).$$

We also consider the *egalitarian cost* $\text{EC}(S) = \max_{v \in V} \text{cost}_v(S)$.

Each voter provides only a strict total order $\succ_v$ over the candidates, consistent with her costs; that is, $c \succ_v c'$ if and only if $d(v, c) < d(v, c')$. The profile $\succ = (\succ_1, \ldots, \succ_n)$ is called a *1-Euclidean ranking profile*, and is assumed to arise from some fixed—but unknown—embedding of voters and candidates on the real line. As is standard, we assume all rankings are strict (no ties).

In addition to the ranking profile $\succ$, the algorithm has access to a distance oracle: a query of the form $(v, c) \in V \times C$ reveals the true distance $d(v, c)$. A deterministic algorithm receives the profile $\succ$, the committee size k, and a query budget q, and may adaptively issue up to q distance queries. The algorithm must then return a committee $S \subseteq C$ of size k.

We assume that the candidate axis (i.e., the ordering of candidates on the real line) is known. This assumption is without loss of generality, since the axis can be reconstructed in polynomial time from the ranking profile and is unique up to reflection as shown by Elkind and Faliszewski [40]

Distortion. We evaluate the performance of a committee election rule R (also referred to as an algorithm or mechanism) in terms of its *distortion*, i.e., the worst-case approximation ratio it achieves with respect to the social cost under limited metric access. Given a 1-Euclidean ranking profile $\succ$, committee size k, and a query budget q, the distortion of R is defined as

$$\text{dist}(R, \succ, k, q) = \sup \frac{\text{SC}(R(\succ, k, q))}{\min_{S \subseteq C, |S|=k} \text{SC}(S)},$$

where the supremum is taken over all collections of voter and candidate locations on the real line that are consistent with $\succ$ and with the answers to the q distance queries made by R.

The distortion of a deterministic k-committee rule is the maximum of $\text{dist}(R, \succ, k, q)$ over all 1-Euclidean ranking profiles $\succ$ with n voters and m candidates. We also consider the distortion with respect to the *egalitarian cost* $\text{EC}(S)$ by explicitly referring to it when needed.

Additional Notation. Recall that for a voter $v \in V$, $\text{top}(v)$ denotes the candidate ranked first in $\succ_v$. The *cluster* of a candidate $c \in C$, denoted $\text{Cluster}(c)$, is the set of all voters who rank c as their top choice. A candidate is said to be *active* if $\text{Cluster}(c) \neq \emptyset$, i.e., if some voter ranks c first.

We assume that algorithms operate only on the set of *active candidates*, i.e., those who are the top choice of at least one voter. An instance is said to be *candidate-restricted* if every candidate is active and each voter is placed at the location of her top-ranked candidate. Such instances can be compactly represented by m pairs (c_i, n_i), where n_i denotes the number of voters co-located with candidate c_i. This simplification is justified by Fotakis et al. [47] (see Proposition 3), who show that for candidate-restricted instances inactive candidates can be eliminated without increasing the social cost. Additionally, relocating each voter to her top candidate increases the distortion by at most a factor of 3 for the social cost (see Theorem 4.3.2), and a similar bound holds for the egalitarian cost (see Theorem 4.2.4).

Although our analysis sometimes refers to candidate-restricted instances for simplicity, the algorithms work for general 1-Euclidean inputs and do not rely on any such structural assumptions. The

distortion guarantees we prove always compare against the optimal committee in the original (possibly unrestricted) instance.

Ranking at Candidate Locations. A subtle but important technical challenge arises when using rankings in algorithm design: for a voter v, the ranking $\succ_v$ may differ from the ranking $\succ_{\text{top}(v)}$, which orders candidates by increasing distance from top(v) instead of v. This discrepancy—due to the fact that v and top(v) may be at different locations—limits the ability to infer rankings at other points on the line from a single voter's perspective. This issue marks a key difference from the clustering models studied in related work [29], where such location-based discrepancies are not present.

4.2 Bounded Distortion

4.2.1 Lower bound for the number of queries required for bounded distortion

First we note that there are 3 types of queries:

Regular queries: Given a voter $v \in V$ and a candidate $c \in C$, we ask for the distance $d(v, c) = |v - c|$.

Candidate queries: Given two candidates $c, c' \in C$, we ask for the distance $d(c, c') = |c - c'|$.

Voter queries: Given two voters $v, v' \in V$, we ask for the distance $d(v, v') = |v - v'|$.

Following the approach of Fotakis et al. [47], we focus primarily on *candidate queries* when designing and analyzing committee election rules.It was shown in Appendix E of [47] that both candidate queries and voter queries can be simulated using a small, constant number of regular queries in the 1-Euclidean setting (at most 6 and 2, respectively). Thus, from an asymptotic perspective, these query types are interchangeable. For clarity and simplicity, we therefore assume access to candidate queries, while noting that equivalent results can be obtained using regular queries with only a constant-factor overhead.

Before designing algorithms with limited metric access, it is important to understand the minimal amount of information required to guarantee reasonable performance. The following theorem, due to Fotakis, Gourvès, and Patsilinakos [48], provides a tight lower bound on the number of distance queries needed to ensure bounded distortion.

Theorem 4.2.1 (Fotakis-Gourvés-Patsilinakos)**.** *For any $k \geq 3$, the distortion of any deterministic k-committee election rule that uses at most $k-3$ distance queries and selects k out of at least $2(k-1)$ candidates on the real line cannot be bounded by any function of n, m, and k (for both the social cost and the egalitarian cost).*

Proof. Let $k \geq 3$ and consider $m = 2(k-1)$ candidates $c_1 < c_2 < \cdots < c_{2k-2}$ located on the real line. Fix a constant D large enough such that $D^2 \gg \max\{2D+1, k\}$, and a small $\epsilon \in (0, 1/k)$.

We construct a *basic instance* as follows. For each $i \in [k-1]$, set:

$$d(c_{2i-1}, c_{2i}) = 1,$$

and for each $i \in [k-2]$, define:

$$d(c_{2i}, c_{2i+1}) = D^2 + (i-1)\epsilon.$$

Let $n = m$, and assign each of the n voters to be co-located with a distinct candidate (their top choice). Thus, all voters have unique top preferences, and the ordinal rankings are identical across all instances.

We now define $2(k-1)$ *variants* of the basic instance. In the j-th variant:

- If j is odd, move candidate c_j left by D, so that $d(c_j, c_{j+1}) = D + 1$.
- If j is even, move candidate c_j right by D, so that $d(c_{j-1}, c_j) = D + 1$.

All other candidates remain fixed. Thus, each variant introduces exactly one *distant pair* of candidates (at distance $D + 1$), while all other original pairs (c_{2i-1}, c_{2i}) remain close (at distance 1).

In each variant, the optimal committee (with respect to both social and egalitarian cost) includes the two candidates in the distant pair:

- If j is odd, the optimal committee includes $\{c_j, c_{j+1}\}$;
- If j is even, the optimal committee includes $\{c_{j-1}, c_j\}$.

The remaining $k-2$ candidates can be selected arbitrarily, one from each of the remaining close pairs. This yields:

- Social cost: $k - 2$,
- Egalitarian cost: 1.

Any committee that does not include the distant pair causes at least one voter to be at distance $\geq D$ from the committee, resulting in a social and egalitarian cost of at least D, which is arbitrarily worse due to our choice of $D \gg k$.

Crucially, the ordinal rankings of the voters are identical across all variants, since each voter remains co-located with their top candidate. Therefore, no deterministic rule can distinguish between the variants based on ordinal information alone.

Now suppose that the algorithm uses at most $k - 3$ distance queries. Observe that:

- Each variant differs from the basic instance in only the distances involving one candidate c_j.
- Any distance query not involving c_j confirms the basic structure and can eliminate at most two variants.

Thus, in order to identify which of the $2(k-1)$ variants is the true instance, the algorithm must use enough queries to rule out all others. Since each query rules out at most 2 variants, it must make at least:

$$\left\lceil \frac{2(k-1)-1}{2} \right\rceil = k - 2$$

queries in the worst case. But this contradicts the assumption that at most $k - 3$ queries are allowed.

Therefore, any deterministic rule making at most $k-3$ distance queries cant not guarantee bounded distortion in both social and egalitarian cost. □

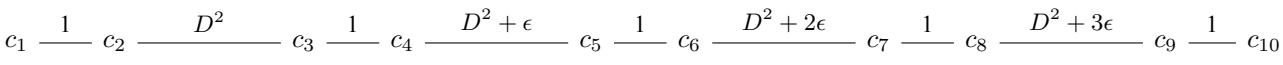


Figure 4.1: The basic instance used in theorem 4.2.1 for $k = 6$

$c_1 \overset{D+1}{\text{—}} c_2 \overset{D^2}{\text{———}} c_3 \overset{1}{\text{—}} c_4 \overset{D^2+\epsilon}{\text{———}} c_5 \overset{1}{\text{—}} c_6 \overset{D^2+2\epsilon}{\text{———}} c_7 \overset{1}{\text{—}} c_8 \overset{D^2+3\epsilon}{\text{———}} c_9 \overset{1}{\text{—}} c_{10}$

$c_1 \overset{D+1}{\text{—}} c_2 \overset{D^2-D}{\text{———}} c_3 \overset{1}{\text{—}} c_4 \overset{D^2+\epsilon}{\text{———}} c_5 \overset{1}{\text{—}} c_6 \overset{D^2+2\epsilon}{\text{———}} c_7 \overset{1}{\text{—}} c_8 \overset{D^2+3\epsilon}{\text{———}} c_9 \overset{1}{\text{—}} c_{10}$

$c_1 \overset{1}{\text{—}} c_2 \overset{D^2-D}{\text{———}} c_3 \overset{D+1}{\text{—}} c_4 \overset{D^2+\epsilon}{\text{———}} c_5 \overset{1}{\text{—}} c_6 \overset{D^2+2\epsilon}{\text{———}} c_7 \overset{1}{\text{—}} c_8 \overset{D^2+3\epsilon}{\text{———}} c_9 \overset{1}{\text{—}} c_{10}$

$c_1 \overset{1}{\text{—}} c_2 \overset{D^2}{\text{———}} c_3 \overset{D+1}{\text{—}} c_4 \overset{D^2-D+\epsilon}{\text{———}} c_5 \overset{1}{\text{—}} c_6 \overset{D^2+2\epsilon}{\text{———}} c_7 \overset{1}{\text{—}} c_8 \overset{D^2+3\epsilon}{\text{———}} c_9 \overset{1}{\text{—}} c_{10}$

$c_1 \overset{1}{\text{—}} c_2 \overset{D^2}{\text{———}} c_3 \overset{1}{\text{—}} c_4 \overset{D^2-D+\epsilon}{\text{———}} c_5 \overset{D+1}{\text{—}} c_6 \overset{D^2+2\epsilon}{\text{———}} c_7 \overset{1}{\text{—}} c_8 \overset{D^2+3\epsilon}{\text{———}} c_9 \overset{1}{\text{—}} c_{10}$

$c_1 \overset{1}{\text{—}} c_2 \overset{D^2}{\text{———}} c_3 \overset{1}{\text{—}} c_4 \overset{D^2+\epsilon}{\text{———}} c_5 \overset{D+1}{\text{—}} c_6 \overset{D^2-D+2\epsilon}{\text{———}} c_7 \overset{1}{\text{—}} c_8 \overset{D^2+3\epsilon}{\text{———}} c_9 \overset{1}{\text{—}} c_{10}$

$c_1 \overset{1}{\text{—}} c_2 \overset{D^2}{\text{———}} c_3 \overset{1}{\text{—}} c_4 \overset{D^2+\epsilon}{\text{———}} c_5 \overset{1}{\text{—}} c_6 \overset{D^2-D+2\epsilon}{\text{———}} c_7 \overset{D+1}{\text{—}} c_8 \overset{D^2+3\epsilon}{\text{———}} c_9 \overset{1}{\text{—}} c_{10}$

$c_1 \overset{1}{\text{—}} c_2 \overset{D^2}{\text{———}} c_3 \overset{1}{\text{—}} c_4 \overset{D^2+\epsilon}{\text{———}} c_5 \overset{1}{\text{—}} c_6 \overset{D^2+2\epsilon}{\text{———}} c_7 \overset{D+1}{\text{—}} c_8 \overset{D^2-D+3\epsilon}{\text{———}} c_9 \overset{1}{\text{—}} c_{10}$

$c_1 \overset{1}{\text{—}} c_2 \overset{D^2}{\text{———}} c_3 \overset{1}{\text{—}} c_4 \overset{D^2+\epsilon}{\text{———}} c_5 \overset{1}{\text{—}} c_6 \overset{D^2+2\epsilon}{\text{———}} c_7 \overset{1}{\text{—}} c_8 \overset{D^2-D+3\epsilon}{\text{———}} c_9 \overset{D+1}{\text{—}} c_{10}$

$c_1 \overset{1}{\text{—}} c_2 \overset{D^2}{\text{———}} c_3 \overset{1}{\text{—}} c_4 \overset{D^2+\epsilon}{\text{———}} c_5 \overset{1}{\text{—}} c_6 \overset{D^2+2\epsilon}{\text{———}} c_7 \overset{1}{\text{—}} c_8 \overset{D^2+3\epsilon}{\text{———}} c_9 \overset{D+1}{\text{—}} c_{10}$

Figure 4.2:The $2(k-1) = 10$ variants obtained from the basic instance used in the lower bound of Theorem 4.2.1 for $k = 6$

4.2.2 Bounded distortion with $\Theta(k)$ queries

In this section, we show that bounded distortion can be achieved using only $\Theta(k)$ distance queries, thereby asymptotically matching the lower bound established in Theorem 4.2.1. This result is accomplished through a query-efficient implementation of the classical 2-approximation algorithm for the k-center problem, originally introduced by Gonzalez [53] and later presented in a simplified form by Williamson and Shmoys [77].

Specifically, we show that the well-known greedy 2-approximation algorithm for k-center can be executed using only a small number of distance queries. The greedy algorithm iteratively constructs a set S of centers: it starts by choosing any candidate (typically at random), and then, in each subsequent iteration, adds the candidate c that maximizes the minimum distance $d(c, S)$ to the current set S. For linear instances, the algorithm can be further optimized by initializing with the leftmost candidate c_1 and the rightmost candidate c_m, then repeatedly adding the candidate $c \in C$ with the maximum distance to S, until k centers have been selected.

Algorithm 4 Query-efficient implementation of the greedy k-Center algorithm

Input: Candidates $C = \{c_1, \dots, c_m\}$, integer $k \in \{2, \dots, m-1\}$, ranking profile $\succ = (\succ_1, \dots, \succ_n)$
Output: A set $S \subseteq C$ of size k

Initialize $S \leftarrow \{c_1, c_m\}$ {pick leftmost and rightmost}
$\hat{C} \leftarrow$ Distant-Candidate$\big(C[c_1, c_m]\big)$
while $|S| < k$ **do**
  Let c be s.t. $(c, \delta) \in \hat{C}$ and $\delta \geq \delta'$ for all $(c', \delta') \in \hat{C}$
  $S \leftarrow S \cup \{c\}$
  $\hat{C} \leftarrow \hat{C} \setminus \{(c, \delta)\}$
  if $|S| < k$ **then**
    Let c_i be the rightmost candidate in S to the left of c
    Let c_{i+1} be the leftmost candidate in S to the right of c
    $\hat{C} \leftarrow \hat{C} \cup \{\text{Distant} - \text{Candidate}(C[c_i, c])\} \cup \{\text{Distant} - \text{Candidate}(C[c, c_{i+1}])\}$
  endif
end while
return S =0

To implement the greedy algorithm in this model (Algorithm 4), we need an efficient method to identify the candidate that is farthest from the current set of selected centers $S = \{c_1, \dots, c_\ell\} \subseteq C$, where the candidates are indexed from left to right along the candidate axis. Here, c_1 and c_ℓ denote the leftmost and rightmost candidates in S and C, respectively.

For each $1 \leq i \leq \ell - 1$, we define $\hat{c}_i$ as the candidate in the interval $C[c_i, c_{i+1}]$ that is maximally distant from its two endpoints c_i and $c_{i+1} \in S$. Formally, we define

$$\hat{c}_i = \arg \max_{c \in C[c_i, c_{i+1}]} d(c, \{c_i, c_{i+1}\}) \tag{4.1}$$

and we let

$$\delta_i = d(\hat{c}_i, \{c_i, c_{i+1}\}).$$

the distance of $\hat{c}_i$ to the endpoints of the interval $C[c_i, c_{i+1}]$. This information is provided by the Distant-Candidate algorithm (Algorithm 5). We now show that the candidate in S that is farthest from the rest can be identified as the $\hat{c}_i$ with the maximum distance δ_i.

Proposition 4.2.2. *Let S be the current set of selected candidates in Algorithm 4, and let $\hat{c}_1, \dots, \hat{c}_{\ell-1}$ be defined as in Equation 4.1. Then,*

$$\max_{c \in C} d(c, S) = \max_{1 \leq i \leq \ell - 1} d(\hat{c}_i, \{c_i, c_{i+1}\}).$$

Proof. Since c_1 is the leftmost candidate and c_ℓ is the rightmost candidate in C, every candidate in $C \setminus S$ lies within one of the intervals $C[c_1, c_2], \dots, C[c_{\ell-1}, c_\ell]$. Suppose the farthest candidate c from S lies in the interval $C[c_i, c_{i+1}]$. By construction, the candidate in this interval that is farthest from both endpoints is $\hat{c}_i$, and therefore:

$$d(c, S) = d(c, \{c_i, c_{i+1}\}) \leq d(\hat{c}_i, \{c_i, c_{i+1}\}) = d(\hat{c}_i, S).$$

Hence, the maximum distance from any candidate to S is achieved by some $\hat{c}_i$, and the claim follows. □

Algorithm 5 The Distant-Candidate algorithm

Input: Candidate interval $C[c, c']$, a voter $v \in \text{Cluster}(c'')$ for every $c'' \in C[c, c']$
Output: Candidate $\hat{c} \in C[c, c']$ with maximum $d(\hat{c}, \{c, c'\})$

1: **if** $|C[c, c']| = 3$ **then**
2: $c'' \leftarrow C[c, c'] \setminus \{c, c'\}$
3: **return** $(c'', \min\{d(c'', c), d(c'', c')\})$
4:Let c'' be the leftmost candidate in $C[c, c'] \setminus \{c\}$
5:**while** $c'' \in C[c, c']$ **do**
6: Let $\succ_{c''}$ be the ranking $\succ_v$ of any $v \in \text{Cluster}(c'')$
7: **if** $c' \succ_{c''} c$ **then**
8: Let c_r be c'' and c_ℓ be next candidate on c'''s left {c_ℓ and c_r found, while-loop terminates}
9: **break**
10: **else**
11: $c'' \leftarrow$ the next candidate on c'''s right {proceed to the next candidate on the right}
12: **if** $d(c, c_\ell) \geq d(c_r, c')$ **then**
13: **return** $(c_\ell, \min\{d(c, c_\ell), d(c', c_\ell)\})$
14:**else**
15: **return** $(c_r, \min\{d(c_r, c), d(c_r, c')\})$

Consider any interval $C[c, c']$ with $|C[c, c']| \geq 3$, defined by two consecutive centers $c, c' \in S$. Algorithm 3 then computes $\hat{c} = \arg\max_{x \in C[c,c']} d\big(x, \{c, c'\}\big)$, and obtains $d\big(\hat{c}, \{c, c'\}\big) = d(\hat{c}, S)$. Moreover, each invocation of Algorithm 5 uses at most three distance queries.

Lemma 4.2.3. *Let $c, c' \in C$ with $c < c'$. Then Algorithm 5 correctly returns the candidate $\hat{c} \in C[c, c']$ satisfying $\hat{c} = \arg\max_{c'' \in C[c,c']} d\big(c'', \{c, c'\}\big)$, and also provides the distance $d\big(\hat{c}, \{c, c'\}\big) = d(\hat{c}, S)$.*

Proof. **Base case.** If $|C[c, c']| = 3$, the unique interior candidate is returned and its distance to $\{c, c'\}$ is obtained with two queries.

General case $|C[c, c']| \geq 4$**.** Let $m := (c + c')/2$ and define

$$f(x) := \min\{d(x, c), d(x, c')\}, \qquad \Delta(x) := d(x, c) - d(x, c').$$

On the line, Δ is strictly monotone non-decreasing, with $\Delta(c) < 0 < \Delta(c')$; hence it crosses 0 exactly once. Set

$$c_\ell := \max\{x \in C[c, c'] : \Delta(x) < 0\}, \qquad c_r := \min\{x \in C[c, c'] : \Delta(x) > 0\}.$$

Then $c_\ell \leq m \leq c_r$ (with at least one of the two inequalities being strict) and no candidate lies strictly between them. Moreover,

$$x \leq c_\ell \Rightarrow f(x) \leq f(c_\ell), \qquad x \geq c_r \Rightarrow f(x) \leq f(c_r), \tag{4.2}$$

so $\hat{c} \in \{c_\ell, c_r\}$ and $f(\hat{c}) = \max\{f(c_\ell), f(c_r)\}$.

Locating the border via the first flipping voter. Traverse candidates from left to right. For each candidate x, fix one voter u_x whose *top candidate* is x (so $u_x \in \text{Cluster}(x)$). Let u_r be the first encountered with $c' \succ_{u_r} c$, and let t be the top candidate of u_r. Write p and s for the predecessor and successor of t (when they exist). Exactly one of the following holds:

(i) $t \le m < s$ (i.e., t is the rightmost candidate on or before m);

(ii) $p \le m < t$ (i.e., t is the leftmost candidate after m).

Indeed, a voter prefers c' to c iff the voter lies to the right or on top of the perpendicular bisector of $\{c, c'\}$, i.e., to the right or on top of m. By minimality of u_r in the left-to-right scan over candidates, the first cluster that contains such a voter must be anchored either at the rightmost candidate $\le m$ or at the leftmost $> m$.

Case (i): $t \le m < s$. Because u_r ranks t above s and u_r lies to the right of m, we must have $m < u_r < s$ (if $u_r \ge s$, then $d(u_r, s) < d(u_r, t)$, a contradiction). Then

$$d(u_r, t) = d(u_r, m) + d(m, t), \qquad d(u_r, s) = d(m, s) - d(u_r, m),$$

and $d(u_r, t) < d(u_r, s)$ implies

$$d(m, t) + 2\, d(u_r, m) < d(m, s) \;\Rightarrow\; d(m, t) < d(m, s).$$

Thus t is strictly closer to the midpoint than its right neighbour s. Since t is the rightmost candidate on/before m, every candidate left of t is at least as far from m as t, and every candidate right of s is at least as far from m as s (and therefore farther than t). Hence t is the unique closest-to-midpoint candidate in $C[c, c']$, equivalently $\hat{c} = t$.

In the algorithm we name $c_r := t$ and $c_\ell := p$ (the predecessor of t). Note that

$$d(c_r, c') - d(c_r, c) = 2(m - t) \ge 0, \qquad d(c_r, c) - d(c_\ell, c) = t - p > 0,$$

so

$$d(c_r, c') \ge d(c_r, c) > d(c_\ell, c).$$

Therefore Line 12's comparison $d(c, c_\ell)$ versus $d(c_r, c')$ selects $c_r = t = \hat{c}$, and the returned value equals $d(\hat{c}, \{c, c'\})$.

Case (ii): $p \le m < t$. Here the only candidates that can maximise f are the adjacent pair $\{p, t\}$: any candidate right of t is farther from m than t, and any candidate left of p is farther from m than p. The algorithm sets $c_r := t$ and $c_\ell := p$, and then compares

$$d(c, c_\ell) = f(p) \quad \text{and} \quad d(c_r, c') = f(t).$$

Thus Line 12 computes $\max\{f(p), f(t)\} = f(\hat{c})$ and returns $\hat{c} \in \{p, t\}$ with the correct distance $d(\hat{c}, \{c, c'\})$.

Combining the base case with the two cases above, Algorithm 5 always returns the farthest candidate and its distance. □

Theorem 4.2.4 (Fotakis-Gourvés-Patsilinakos)**.** *Let (C, V) be an instance of the k-committee election. Let $S \subseteq C$ (respectively, $S^* \subseteq C$) be a β-approximate (respectively, an optimal) k-committee with respect to the egalitarian cost for the candidate-restricted instance (respectively, the original instance). Then,*

$$\mathrm{EC}(S) \;\le\; (1 + 2\beta)\, \mathrm{EC}(S^*).$$

Proof. Recall that for each voter $v \in V$, $\mathrm{top}(v)$ denotes v's top-ranked candidate in C. By the triangle inequality,

$$d(v, S) \;\le\; d\big(v, \mathrm{top}(v)\big) \;+\; d\big(\mathrm{top}(v), S\big).$$

Taking the maximum over all $v \in V$ yields

$$\mathrm{EC}(S) \;\le\; \mathrm{EC}(C) \;+\; \mathrm{EC}\big(C_{\mathrm{cr}}, S\big), \tag{4.3}$$

where

$$\mathrm{EC}(C) \;=\; \max_{v \in V} d\big(v, \mathrm{top}(v)\big) \;=\; \max_{v \in V} d(v, C), \quad \mathrm{EC}\big(C_{\mathrm{cr}}, S\big) \;=\; \max_{v \in V} d\big(\mathrm{top}(v), S\big).$$

$\mathrm{EC}(C_{\mathrm{cr}}, S)$ denotes the egalitarian cost of S on the candidate-restricted instance C_{cr} induced by C. Since S is a β-approximate k-committee for the candidate-restricted instance C_{cr},

$$\mathrm{EC}\big(C_{\mathrm{cr}}, S\big) \;\le\; \beta\,\mathrm{EC}\big(C_{\mathrm{cr}}, S^\sharp\big) \;\le\; \beta\,\mathrm{EC}\big(C_{\mathrm{cr}}, S^*\big), \tag{4.4}$$

where $S^\sharp$ is an optimal solution on C_{cr}. For the second inequality we use the fact that S^* is also an alternative to $S^\sharp$ as optimal solution for C_{cr}. Therefore, using the optimality of $S^\sharp$ for the C_{cr} instance with respect to the egalitarian cost we get $\beta\,\mathrm{EC}\big(C_{\mathrm{cr}}, S^\sharp\big) \;\le\; \beta\,\mathrm{EC}\big(C_{\mathrm{cr}}, S^*\big)$

On the other hand, applying the triangle inequality to each $\mathrm{top}(v)$ and S^* gives

$$\mathrm{EC}\big(C_{\mathrm{cr}}, S^*\big) \;\le\; \mathrm{EC}(C) \;+\; \mathrm{EC}(S^*).$$

Substituting (4.4) and this bound into (4.3) yields

$$\mathrm{EC}(S) \;\le\; (1+\beta)\,\mathrm{EC}(C) \;+\; \beta\,\mathrm{EC}(S^*) \;\le\; (1+2\beta)\,\mathrm{EC}(S^*),$$

as required. □

Theorem 4.2.5 (Fotakis-Gourvés-Patsilinakos)**.** *For any $k \ge 3$, Algorithm 4 achieves a distortion of at most $5n$ for the social cost, and at most 5 for the egalitarian cost, for k-Committee Election on the real line using at most $6k - 15$ candidate distance queries.*

Proof. We first bound the number of distance queries. In Algorithm 4, the Distant-Candidate subroutine is invoked once in Step 2, and then twice in each iteration of the **while**-loop from $|S| = 3$ up to $|S| = k - 1$. Hence there are $1 \;+\; 2(k-3)$ calls in total, and since each call uses at most three distance queries, the overall query complexity is

$$3\big(1 + 2(k-3)\big) \;=\; 6(k-3) + 3 \;=\; 6k - 15.$$

Correctness follows from Lemma 4.2.3 and Proposition 4.2.2, which guarantee that each iteration indeed adds the candidate c maximizing $d(c, S)$. The Williamson-Shmoys greedy algorithm is a 2-approximation for the egalitarian cost on candidate-restricted instances (see [77, Theorem 2.3]), and thus by Theorem 4.2.4 the resulting distortion on the original instance is at most 5.

Finally, since for any committee $S \subseteq C$ we have

$$\mathrm{EC}(S) \;\le\; \mathrm{SC}(S) \;\le\; n\,\mathrm{EC}(S),$$

it follows that if S^* is optimal for the egalitarian cost and S^{**} is optimal for the social cost, then

$$\frac{\mathrm{SC}(S)}{\mathrm{SC}(S^{**})} \;\le\; n\,\frac{\mathrm{EC}(S)}{\mathrm{EC}(S^{**})} \;=\; n\,\frac{\mathrm{EC}(S)}{\mathrm{EC}(S^{*})} \cdot \frac{\mathrm{EC}(S^{*})}{\mathrm{EC}(S^{**})} \;\le\; 5n,$$

where we used $\mathrm{EC}(S^*) \le \mathrm{EC}(S^{**})$ and the bound $\mathrm{EC}(S)/\mathrm{EC}(S^*) \le 5$. □

Thus, the greedy center selection strategy achieves strong distortion guarantees for both objectives with only $O(k)$ queries, matching the information-theoretic lower bound. While the distortion for the social cost scales linearly with the number of voters, this is unavoidable without additional access to distances or additional assumptions about the instance.

4.3 Constant Distortion with $O(k \log n)$ queries

4.3.1 Constant Distortion with $O(k \log n)$ queries

We now turn our attention to the complementary result of Fotakis, Gourvés and Patsilinakos, who give an algorithm that guarantees a *constant* distortion bound in the one dimensional model. Their method incorporates a careful selection and merging strategy to control the worst-case distance error. In what follows, we outline the main steps of this constant-distortion algorithm, explain how it leverages the structure of (ℓ, β)-bicriteria solutions, and sketch the proof of its constant-factor performance.

Definition 4.3.1 ((ℓ, β)-bicriteria solution)**.** *Let C be a set of candidates and let S^* be an optimal k-committee that minimizes the social cost $SC(S^*)$. A subset $C' \subseteq C$ is said to be* (ℓ, β)-good*, for some $\ell \geq k$ and $\beta \geq 1$, if the following conditions hold:*

(i) *$|C'| = \ell$, and*

(ii) *The social cost incurred by representing each voter by her top candidate in C' satisfies:*

$$SC(C') \leq \beta \cdot SC(S^*)$$

In other words:

- *C' is ℓ-sparse, meaning it contains only ℓ candidates (ideally $\ell \ll |C|$), and*
- *C' is β-good, achieving a social cost within a factor β of the optimal.*

Note:

- *The original candidate set C is trivially $(m, 1)$-good, where $m = |C|$.*
- *Any k-committee with distortion β is (k, β)-good.*

This notion formalizes the idea of a small representative subset of candidates that approximates the performance of the optimal committee. By focusing on such subsets, we reduce the complexity of the original instance while preserving enough structure to enable efficient optimization.

Given an (ℓ, β)-good subset of candidates C', we define the *candidate-restricted instance* induced by C' as

$$C'_{\text{cr}} = \{(c_1, n_1), \ldots, (c_\ell, n_\ell)\}$$

where $c_1 < \cdots < c_\ell$ denote the positions of the candidates in C' along the real line, and $n_i = |\text{Cluster}(c_i)|$ is the number of voters who rank c_i as their top choice in C'. By construction, we have $n_1 + \cdots + n_\ell = n$, and each $n_i > 0$, since we discard any inactive candidates from C'.

We will show that an optimal k-committee for the restricted instance C'_{cr} achieves a distortion of at most $1 + 2\beta$ when measured against the original instance.

Theorem 4.3.2 (Fotakis-Gourvés-Patsilinakos)**.** *Let (C, V) be an instance of the k-committee election, let $C' \subseteq C$ be an (ℓ, β)-bicriteria solution, and let C'_{cr} be the candidate-restricted instance induced by C'. Let S (resp. S^*) be an optimal k-committee for C'_{cr} (resp. for (C, V)). Then,* $\text{SC}(S) \leq (1 + 2\beta)\,\text{SC}(S^*)$.

Proof. For each voter $v \in V$, let $\text{top}'(v) \in C'$ denote their top-ranked candidate within the set C'. By the triangle inequality, we have

$$d(v, S) \leq d\big(v, \text{top}'(v)\big) + d\big(\text{top}'(v), S\big).$$

Summing over all voters gives:

$$\text{SC}(S) \leq \text{SC}(C') + \text{SC}\big(C'_{\text{cr}}, S\big), \tag{4.5}$$

where:

- $\text{SC}(C') = \sum_{v \in V} d\big(v, \text{top}'(v)\big) = \sum_{v \in V} d(v, C')$, and
- $\text{SC}(C'_{\text{cr}}, S) = \sum_{v \in V} d\big(\text{top}'(v), S\big)$ is the social cost of S in the candidate-restricted instance C'_{cr} induced by C'.

Since S is the optimal k-committee for C'_{cr}, we have: $\text{SC}(C'_{\text{cr}}, S) \leq \text{SC}(C'_{\text{cr}}, S^\sharp)$. for any feasible solution $S^\sharp$ on the candidate-restricted instance.

However, S^* may not be feasible for C'_{cr}, as it can include candidates not in C'. To address this, we argue that S^* can be transformed into a valid committee $S^\sharp \subseteq C'$ without increasing the social cost.

If S^* includes an inactive candidate $c \notin C'$, and no voter is assigned to c, we can safely remove it without affecting the cost and replace it by any active candidate. If c does have assigned voters, we divide them into two groups: those to the left of c (V_{left}) and those to the right (V_{right}). Since all voters in C'_{cr} are collocated with their top candidates, we can find a candidate in C' who is collocated with a voter in the larger group. Specifically, if $|V_{\text{left}}| > |V_{\text{right}}|$, we replace c with the candidate collocated with the rightmost voter in V_{left}; otherwise, we use the one collocated with the leftmost voter in V_{right}. This replacement ensures that the maximum increase in distance (if any) affects fewer voters and that the total cost does not increase. Repeating this process for all inactive candidates in S^* yields a feasible committee $S^\sharp \subseteq C'$ with:

$$\text{SC}(C'_{\text{cr}}, S^\sharp) \leq \text{SC}(C'_{\text{cr}}, S^*).$$

Combining this with the optimality of S, we conclude:

$$\text{SC}(C'_{\text{cr}}, S) \leq \text{SC}(C'_{\text{cr}}, S^\sharp) \leq \text{SC}(C'_{\text{cr}}, S^*).$$

Finally, by the triangle inequality again,

$$d\big(\text{top}'(v), S^*\big) \leq d\big(\text{top}'(v), v\big) + d\big(v, S^*\big),$$

so summing over all voters gives:

$$\text{SC}(C'_{\text{cr}}, S^*) \leq \text{SC}(C') + \text{SC}(S^*).$$

Substituting this into (4.5), we obtain:

$$\text{SC}(S) \leq \text{SC}(C') + \text{SC}(C'_{\text{cr}}, S) \leq \text{SC}(C') + \text{SC}(C') + \text{SC}(S^*) = 2\,\text{SC}(C') + \text{SC}(S^*).$$

Finally, since C' is an (ℓ, β)-bicriteria solution, we have $\text{SC}(C') \leq \beta\,\text{SC}(S^*)$, and thus:

$$\text{SC}(S) \leq (1 + 2\beta)\,\text{SC}(S^*).$$

□

This theorem justifies our overall strategy: if we can find a good representative subset C' then we can solve a much smaller problem and still retain strong approximation guarantees.

As soon as we have the distances between all candidates in an (ℓ, β)-bicriteria solution C' which requires $l-1$ candidate queries an optimal k-committee for the C'_{cr} instance can be computed in polynomial time by the dynamic programming algorithm of [56] as mentioned by Fotakis et al. [47]

The next challenge is to efficiently construct such a good subset using limited distance information. To this end, we now describe how to construct an $(O(k \log n), O(1))$-bicriteria solution of candidates using hierarchical partitioning of the candidate axis. This is implemented by Algorithm 6, which computes such a set using only $O(k \log n)$ distance queries.

In this approach, we maintain a collection $\mathcal{I}$ of intervals over the candidate axis $C = \{c_1, \ldots, c_m\}$. Each interval $C[c_a, c_b] \in \mathcal{I}$ is annotated with: (i) the number n_{ab} of voters whose top-ranked candidate lies in $C[c_a, c_b]$, and (ii) the length $d(c_a, c_b)$ of the interval. We define the *weight* of an interval as $\text{wt}(c_a, c_b) = n_{ab} \cdot d(c_a, c_b)$.

Algorithm 6 begins with a partitioning of C into k regions based on a reference committee $S = \{c^1, \ldots, c^k\}$, computed via Algorithm 4. For each $i \in [k]$, we define an interval $C[c^i_a, c^i_b]$ containing all candidates closer to c^i than to any other member of S. Let n_i denote the number of voters whose top choice lies in $C[c^i_a, c^i_b]$, and let $d(c^i_a, c^i_b)$ be the interval's length. We define

$$\delta^* = \max_{i \in [k]} \{d(c^i_a, c^i_b)\}.$$

Since the committee S has distortion at most 5 for the egalitarian cost and all candidates are assumed to be active, we obtain the lower bound $\text{SC}(S^*) \geq \delta^*/5$, where S^* is an optimal k-committee for the original instance.

The partition $\mathcal{I}$ is then refined iteratively. In each step, the interval in $\mathcal{I}$ with the largest weight and at least four candidates is split into two subintervals. This is done using a subroutine called Partitioning (Algorithm 7), which extends the Distant-Candidate approach from Section 6. For an interval $C[c_a, c_b]$, the algorithm identifies the midpoint $m = (c_a + c_b)/2$, and splits the interval into two: $C[c_a, c_\ell]$ and $C[c_r, c_b]$, where c_ℓ is the rightmost candidate to the left of m, and c_r is the leftmost candidate to the right. The algorithm uses at most 4 distance queries per split. Notably, the union of these subintervals exactly covers the original interval, i.e.,

$$C[c_a, c_\ell] \cup C[c_r, c_b] = C[c_a, c_b],$$

so the invariant that $\mathcal{I}$ is a partition of C is preserved throughout.

After at most $O(k \log n)$ such splits, the algorithm terminates with a partition $\mathcal{I}$ of size at most $7k(\log_2(5nk) + 2)$. From this partition, we extract a candidate set $C'(\mathcal{I}) \subseteq C$ as follows: for each interval $C[c_a, c_b] \in \mathcal{I}$,

- If the interval contains more than 3 candidates, include only its endpoints c_a and c_b in $C'(\mathcal{I})$,
- Otherwise, include all candidates in the interval.

Since $|\mathcal{I}| = O(k \log n)$, the size of $C'(\mathcal{I})$ is also $O(k \log n)$. We will nextshow that this set satisfies

$$\text{SC}(C'(\mathcal{I})) \leq 2\,\text{SC}(S^*),$$

where S^* is the optimal committee with respect to the original instance. Hence, $C'(\mathcal{I})$ is an $(O(k \log n), O(1))$-bicriteria solution.

Algorithm 6 Hierarchical partitioning of $C = [c_1, c_m]$

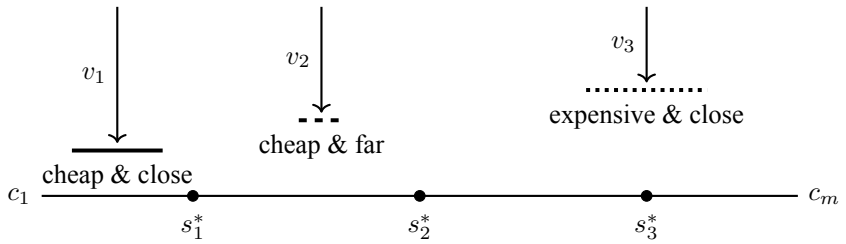
Input: Candidates $\mathcal{C}= \{C_1, ..., C_m\}$, $k \in \{2, ..., m-1\}$, voter ranking profile $\succ= (\succ_1, \ldots, \succ_n)$
Output: Partiotioning $\mathcal{I}$ of $\mathcal{C}$ into $O(klogn)$ intervals

1: Let $S = \{c^1, ..., c^k\}$ be the output of Algorithm 4
2: $\mathcal{I} \leftarrow \{(\mathcal{C}[c_a^1, c_b^1], n_1, d(c_a^1, c_b^1)), ..., \mathcal{C}[c_a^k, c_b^k], n_k, d(c_a^k, c_b^k))\}$ {Start with the partiotioning of $\mathcal{C}, \mathcal{V}$ induced by S
3: $\delta^* \leftarrow max_{i \in [k]}\{d(c_a^i, c_b^i)\}$
4:**while** $|\mathcal{I}| \leq 7k(log_2(5nk) + 2)$ **do**
5: Let $(\mathcal{C}[c_a, c_b], n_{ab}, d(c_a, c_b) \in \mathcal{I}$ with $|\mathcal{C}[c_a, c_b]| \geq 4$ and maximum weight $wt(c_a, c_b) = n_{ab} d(c_a, c_b)$
6: **if** $wt(c_a, c_b) \leq \delta^*/(5k)$ **then break-while-loop**
7: $\mathcal{I} \leftarrow (\mathcal{I} \setminus \{(\mathcal{C}[c_a, c_b], n_{ab}, d(c_a, c_b))\}) \cup$ Partitioning$(\mathcal{C}[c_a, c_b])$
8:**end while**
9:**return** $\mathcal{I}$

Theorem 4.3.3 (Fotakis-Gourvés-Patsilinakos)**.** *Let $\mathcal{I}$ be the partition of the candidate set C computed by Algorithm 6. Then, the resulting set of candidates $C'(\mathcal{I}) \subseteq C$ is an $(O(k \log n), 2)$-bicriteria solution.*

Proof. Let $S^* \subseteq \mathcal{C}$ denote an optimal committee of size k, and let $\mathcal{I}$ be the final partition of candidates produced by Algorithm 6. Let $C'(\mathcal{I})$ be the set of endpoints of all intervals in $\mathcal{I}$. We aim to show that $C'(\mathcal{I})$ is an $(O(k \log n), 2)$-bicriteria solution.

We organize the proof into three parts: bounding the social cost, bounding the number of intervals, and bounding the query complexity.



Light: wt$(I) \leq$ SC$(S^*)/k$
Heavy: wt$(I) >$ SC$(S^*)/k$

Interval classification with regards to whether they are cheap or expensive and far or close. We note that far/close term classifies heavy intervals

1. Bounding the social cost. Each voter v belongs to a unique interval $I = [c_a, c_b] \in \mathcal{I}$, determined by the position of their top-ranked candidate. We define two types of intervals(see Figure 4.3):

- An interval I is *cheap* if it contains no candidate from S^*,
- Otherwise, it is *expensive*.

In a cheap interval, the closest candidate from S^* lies outside the interval. Since $C'(\mathcal{I})$ contains both endpoints of I, we have:

$$d(v, C'(\mathcal{I})) \leq d(v, S^*).$$

In an expensive interval, the distance from v to $C'(\mathcal{I})$ is at most:

$$d(v, C'(\mathcal{I})) \leq d(v, S^*) + d(c_a, c_b).$$

Let $\text{wt}(I) = n_I \cdot d(c_a, c_b)$ denote the *weight* of interval I, where n_I is the number of voters with top-ranked candidate in I. Then:

$$\text{SC}(C'(\mathcal{I})) \leq \text{SC}(S^*) + \sum_{\text{expensive } I} \text{wt}(I).$$

We now show that the total additional cost is at most $\text{SC}(S^*)$, once enough intervals have been created.

2. Bounding the number and weight of expensive intervals. To bound this cost, we analyze the structure of the partitioning process.

Interval levels. We group intervals by *level*, based on their diameter. A level-i interval satisfies:

$$2^{i-1}\delta^* < d(c_a, c_b) \leq 2^i \delta^*,$$

where δ^* is the maximum diameter among the initial k intervals.

As the algorithm recursively splits intervals, the length of each resulting subinterval is halved. The smallest level produced corresponds to intervals of diameter at most:

$$\frac{\delta^*}{5nk},$$

which occurs at level $i = -\log_2(5nk)$.

Each such interval contains at most n voters, so its weight is bounded by:

$$\text{wt}(I) \leq n \cdot \frac{\delta^*}{5nk} = \frac{\delta^*}{5k} \leq \frac{\text{SC}(S^*)}{k}.$$

Therefore, at level $-\log_2(5nk)$, all intervals are light, and further splitting is not permitted. Since levels range from $i = 0$ to $i = -\log_2(5nk)$, the total number of levels is at most $\log_2(5nk) + 1$.

3. Bounding the number of heavy intervals. We distinguish two types of heavy intervals (see Figure 4.3) :

- A heavy interval is *far* if $d(\{c_a, c_b\}, S^*) \geq d(c_a, c_b)$,
- Otherwise, it is *close*.

Far heavy intervals. Let v be a voter in a far heavy interval $I = [c_a, c_b]$, and let $c_v^* \in S^*$ be the closest candidate to v. Using the triangle inequality, one can show:

$$d(v, c_v^*) \geq \frac{1}{2} d(c_a, c_b).$$

Since $\text{wt}(I) > \text{SC}(S^*)/k$, voters in I contribute at least $\text{SC}(S^*)/(2k)$ to the optimal cost. Therefore, there can be at most $2k$ such intervals per level, and at most:

$$2k(\log_2(5nk) + 1)$$

in total.

Close heavy intervals. Fix $c^* \in S^*$ and a level i. A close heavy interval $I = [c_a, c_b]$ satisfies:

$$d(\{c_a, c_b\}, c^*) < d(c_a, c_b).$$

This implies that I lies within a ball of radius $2^i\delta^*$ centered at c^*, since both endpoints are closer to c^* than they are to each other. Also, the length of each level-i interval is at least $2^{i-1}\delta^*$, by definition of levels.

Hence,as shown also in the figure 4.4 the number of such disjoint intervals that can fit within a ball of radius $2^i\delta^*$ is at most:

$$\left\lfloor \frac{2 \cdot 2^i\delta^*}{2^{i-1}\delta^*} \right\rfloor = \left\lfloor \frac{4\delta^*}{\delta^*} \right\rfloor = 4.$$

Allowing for overlaps and rounding, we conservatively upper bound this number by 5. Thus, at most 5 disjoint close heavy intervals at level i can be associated with any candidate $c^* \in S^*$.

Therefore, the total number of close heavy intervals is at most:

$$5k(\log_2(5nk) + 1).$$

4. Bounding the total number of intervals. Only heavy intervals are split. The total number of heavy intervals (close and far) encountered is:

$$(2k + 5k)(\log_2(5nk) + 1) = 7k(\log_2(5nk) + 1).$$

Starting with k intervals and adding one per split, the total number of intervals is at most:

$$7k(\log_2(5nk) + 1) + k + 1 < 7k(\log_2(5nk) + 2) = O(k \log n).$$

Since $C'(\mathcal{I})$ includes two candidates per interval, its size is $O(k \log n)$.

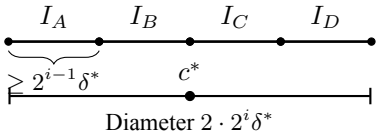
Moreover, once only light intervals remain, each expensive interval has weight at most $\text{SC}(S^*)/k$, and there can be at most k such intervals, since there are k candidates on the optimal solution. Therefore, the total additive cost from expensive intervals is at most $\text{SC}(S^*)$, yielding:

$$\text{SC}(C'(\mathcal{I})) \leq 2 \cdot \text{SC}(S^*).$$

□

Algorithm 6 performs $O(k \log n)$ splits throughout its execution. Since each invocation of the Partitioning algorithm requires at most 4 distance queries, the total number of distance queries used by Algorithm 6 is bounded by $O(k \log n)$. Combining Theorem 4.3.2 (and the accompanying discussion) with the analysis of Algorithm 6 in Theorem 4.3.3, we obtain the following result:

Theorem 4.3.4 (Fotakis-Gourvés-Patsilinakos)**.** *There exists a deterministic polynomial-time rule for the k-Committee Election problem that uses $O(k \log n)$ distance queries and achieves distortion at most 5.*



Packing argument for close heavy intervals. A 1D "ball" of radius $2^i\delta^*$ (the horizontal segment) centered at c^* can contain at most $\frac{2\cdot 2^i\delta^*}{2^{i-1}\delta^*} = 4$ disjoint level-i intervals of length $\geq 2^{i-1}\delta^*$. We conservatively bound the number by 5 to account for slight variations in interval lengths and ensure robustness.

To conclude with this chapter we are left to state the Partitioning algorithm and verify its properties used by Algorithm 6

Algorithm 7 Partitioning of interval $C[c, c']$

Input: Candidate interval $(C[c, c'], n, d(c, c'))$, and for each $c'' \in C[c, c']$ a ranking $\succ_v$ of any voter $v \in \text{Cluster}(c'')$
Output: Two subintervals $(\mathcal{C}[c, c_\ell], n_\ell, d(c, c_\ell))$ and $(\mathcal{C}[c_r, c'], n_r, d(c_r, c'))$ subdividing $(\mathcal{C}[c, c'], n, d(c, c'))$

Let c'' be the leftmost candidate in $\mathcal{C}[c, c'] \setminus \{c\}$
while $c'' \in \mathcal{C}[c, c']$ **do**
  Let $\succ_{c''}$ be the ranking of any $v \in \text{Cluster}(c'')$
  if $c' \succ_{c''} c$ **then**
    $c_r \leftarrow c''$, $c_\ell \leftarrow$ candidate immediately left of c'' {c_l and c_r found, while-loop terminates}
    break-while-loop
  else
    $c'' \leftarrow$ next candidate to the right of c'' {proceed to the next candidate on the right }
  end if
end while
if $d(c, c_\ell) \geq d(c_r, c')$ **then**
  {c_l is the most distant candidate to $\{c, c'\}$}
  if $d(c, c_\ell) > d(c_\ell, c')$ **then**
    $c_r \leftarrow c_\ell$ {c_ℓ is the first candidate on the right of $(c + c')/2$
    $c_\ell \leftarrow$ candidate immediately left of c_r
  end if
else
  {c_r is the most distant candidate to $\{c, c'\}$}
  if $d(c, c_r) < d(c_r, c')$ **then**
    $c_\ell \leftarrow c_r$ {c_r is the first candidate on the left of $(c + c')/2$
    $c_r \leftarrow$ candidate immediately right of c_ℓ
  end if
end if
$n_\ell \leftarrow \sum_{\tilde{c} \in C[c, c_\ell]} |\text{Cluster}(\tilde{c})|$
$n_r \leftarrow \sum_{\tilde{c} \in C[c_r, c']} |\text{Cluster}(\tilde{c})|$
return $(C[c, c_\ell], n_\ell, d(c, c_\ell)), (C[c_r, c'], n_r, d(c_r, c'))$

Lemma 4.3.5. *Let $c, c' \in C$ with $c < c'$ and $|C[c, c']| \geq 4$. Algorithm 7 returns two candidates c_ℓ and c_r that partition $C[c, c']$ into two disjoint intervals $C[c, c_\ell]$ and $C[c_r, c']$, with the following properties:*

- *If no candidate lies exactly at the midpoint $\mu = \frac{c+c'}{2}$, then:*

  - *c_ℓ is the rightmost candidate in $C[c, c']$ strictly to the left of μ,*
  - *c_r is the leftmost candidate in $C[c, c']$ strictly to the right of μ.*

- *If there exists a candidate $\hat{c} \in C[c, c']$ located exactly at the midpoint μ, then the algorithm assigns $\hat{c}$ as follows:*
  - *If the voter $v \in$ Cluster$(\hat{c})$ given as part of the input ranks $c' \succ_v c$, then $\hat{c}$ is assigned to the right interval (i.e., becomes c_r),*
  - *If $c \succ_v c'$, then $\hat{c}$ is assigned to the left interval (i.e., becomes c_ℓ).*

Thus, the midpoint candidate is consistently assigned based on voter rankings, ensuring a valid partition of the interval.

Proof. We note that if $|C[c, c']| \geq 4$, the first ten steps of Algorithm 5 7 are identical to the first ten steps of Algorithm 3 (i.e., steps 5 to 14, applied to this case). Therefore, by the proof of Lemma 4.2.3, when Algorithm 7 reaches step 10, either c_l or c_r is the candidate $\hat{c} \in C[c, c']$ with the largest distance to $\{c, c'\}$.

Then, by the proof of Lemma 4.2.3:

- If $d(c, c_l) \geq d(c_r, c')$, then $\hat{c} = c_l$.
- Otherwise, $\hat{c} = c_r$.

In both cases, $\hat{c}$ is the candidate in $C[c, c']$ closest to or on to the midpoint $\mu = \frac{c+c'}{2}$. In each case (i.e., either if $\hat{c} = c_l$, where steps 13–16 are executed, or if $\hat{c} = c_r$, where steps 19–22 are executed), Algorithm 5 distinguishes two subcases depending on whether $\hat{c}$ is on the left or on the right of μ.

Case 1: $d(c, c_l) \geq d(c_r, c')$ and $\hat{c} = c_l$

- If $d(c, c_l) > d(c_l, c')$, then c_l is on the right of the midpoint μ. In this case, c_l is in fact c_r (i.e., the leftmost candidate on the right of μ; so the value of the algorithm's variable c_r is set to c_l in step 14), and c_l (i.e., the rightmost candidate on the left of μ) is the first candidate on the left of c_r on the candidate axis (step 15).
- Otherwise (i.e., if $d(c, c_l) \leq d(c_l, c')$), since $\hat{c} = c_l$ and c_l and c_r are consecutive on the candidate axis, c_l is indeed the rightmost candidate on the left of μ and c_r is the leftmost candidate on the right of μ. Thus, the values of the corresponding algorithm's variables are set correctly.

Case 2: $d(c, c_l) < d(c_r, c')$ and $\hat{c} = c_r$

- If $d(c, c_r) < d(c_r, c')$, then c_r is on the left of the midpoint μ. In this case, c_r is in fact c_l (i.e., the rightmost candidate on the left of μ; so the value of the algorithm's variable c_l is set to c_r in step 20), and c_r (i.e., the leftmost candidate on the right of μ) is the first candidate on the right of c_l on the candidate axis (step 21).
- Otherwise (i.e., if $d(c, c_r) \geq d(c_r, c')$), since $\hat{c} = c_r$ and c_l and c_r are consecutive on the candidate axis, c_r is indeed the leftmost candidate on the right of μ and c_l is the rightmost candidate on the left of μ. Thus, the values of the corresponding algorithm's variables are set correctly.

Therefore, when Algorithm 7 reaches step 23, the value of the variable c_l corresponds to the rightmost candidate on the left of (or exactly at) the midpoint $\mu = \frac{c+c'}{2}$, and the value of the variable c_r corresponds to the leftmost candidate on the right of (or exactly at) the midpoint $\mu = \frac{c+c'}{2}$ of the interval $C[c, c']$. In the case where there exists a candidate exactly at μ, it is assigned consistently to either c_l or c_r according to the algorithm's tie-breaking rule (e.g., axis ordering), ensuring that the partition remains correct.

□

The number of voters n_l and n_r associated with the subintervals $C[c, c_l]$ and $C[c_r, c']$ are correctly computed in steps 24 and 25 using the voters' preference profile $\succ$.

Regarding the distances $d(c, c_l)$ and $d(c_r, c')$, which represent the lengths of the two subintervals, one additional distance query may be needed depending on whether steps 14–15 or 20–21 are executed:

- If steps 14–15 are executed, then $d(c_r, c')$ is already known from step 13 (since $c_r = c_l$ at that point), and the algorithm performs an extra query to compute $d(c, c_l)$.
- If steps 20–21 are executed, then $d(c, c_l)$ is already known from step 19 (since $c_l = c_r$ at that point), and an additional query is made to obtain $d(c_r, c')$.

In every case, Algorithm 7 successfully partitions the interval $C[c, c']$ into $C[c, c_l]$ and $C[c_r, c']$, where c_l (respectively, c_r) is the rightmost (respectively, leftmost) candidate strictly to the left (respectively, right) of the midpoint $\mu = \frac{c+c'}{2}$. If a candidate lies exactly at μ, it is assigned to either c_l or c_r based on the algorithm's selection logic, but never to both. Thus, at most one of c_l or c_r can coincide with the midpoint. The algorithm ensures that n_l, n_r, $d(c, c_l)$, and $d(c_r, c')$ are computed correctly using no more than four distance queries in total.

Conclusion. In this chapter, we investigated the power of limited distance information in the committee election problem under metric preferences. Building on the candidate-restricted model, we established that strong guarantees on social cost distortion can be achieved even with a sublinear number of queries. Specifically, we presented a deterministic algorithm that uses only $O(k \log n)$ distance queries and attains constant worst-case distortion. This result demonstrates that carefully selected queries can significantly enrich the ordinal model, enabling near-optimal outcomes while respecting communication constraints. These findings complement the purely ordinal results of the previous chapter and highlight the benefits of modest access to metric structure in collective decision-making.

CHAPTER 5

Committee Selection with Metric Preferences in General Metric Spaces and Query Access

In this chapter, we consider the problem of committee selection under metric preferences in general metric spaces, where access to exact distances is restricted and ordinal information is freely available. We focus on the recent work of Burkhardt et al. [29], who study this setting in the context of clustering and propose algorithms with provable guarantees under limited query access.

Their results address both the k-center problem and the broader (k, z)-clustering framework, which generalizes several classical clustering objectives. Of particular relevance to our setting is the case $z = 1$, corresponding to the k-median objective and the committee election problem studied throughout this thesis.

A key contribution of their work is an algorithm that achieves constant distortion using only $O(k^4 \log^5 n)$ distance queries, where n is the number of agents. This query complexity is sublinear in n, highlighting the power of combining ordinal information with a small number of carefully chosen distance queries. While we do not detail their lower bounds in this thesis, it is worth noting that they establish strong impossibility results: no algorithm using fewer than $O(k)$ distance queries can guarantee bounded distortion for any (k, z)-clustering objective. Moreover, achieving constant distortion with respect to the social cost requires at least $\Omega(k + \log \log n)$ queries when k is variable, and at least $\Omega(k \cdot 2^{\log^\star n})$ when k is fixed.

We now formalize the model and present the algorithmic framework and analysis.

5.1 Preliminaries and Specifics of this model

In general metric spaces, we cannot infer distances *between candidates* by querying voters about their distances *to* candidates, unlike in the one-dimensional Euclidean setting where a common total order enables such reconstruction. Ordinal rankings $\{\pi_x\}_{x \in X}$ are inherently *local* to each voter x and need not embed into a single global order over candidates. Consequently, we adopt the standard variant in which the committee is selected from the ground set X itself (candidates coincide with voters), so any required inter-point distance can be queried directly.

Formally, let (X, d) be a finite metric space with $|X| = n$, where $d : X \times X \to \mathbb{R}_{\geq 0}$ satisfies the metric axioms. The algorithm has *query access* to d: a query on (x, y) returns $d(x, y)$ at unit cost, and

the total number of queries is bounded by a budget. In contrast, ordinal information remains free: for each $x \in X$, we are given a ranking $\pi_x : [n] \to X$ such that for all $i < j$,

$$d\big(x, \pi_x(i)\big) \leq d\big(x, \pi_x(j)\big),$$

with ties broken arbitrarily. We refer to $P = \{\pi_x\}_{x \in X}$ as the *ordinal profile*, and write $P(d)$ for the collection of ordinal profiles consistent with d.

Definition 5.1.1 (Ordinal (k, z)-Clustering Problem)**.** *Given positive integers k and z, and a set X of n points forming a metric space (X, d), the goal is to select a subset $C \subseteq X$ of k centers that minimizes a cost function. Each point $x \in X$ provides a ranking π_x over all points in X, where the ranking is consistent with the metric d, i.e., $d(x, \pi_x(i)) \leq d(x, \pi_x(j))$ for all $i < j$.*

Let $P = \{\pi_x\}_{x \in X}$ denote the set of all such rankings (the ordinal profile). For any subset $S \subseteq X$ and candidate solution $C \subseteq X$, the cost of serving S with centers in C is defined as:

$$\phi_C(S, d) = \sqrt[z]{\sum_{x \in S} d(x, C)^z},$$

*where $d(x, C) = \min_{c \in C} d(x, c)$ denotes the distance from x to its closest center in C. The **k-Committee Election Problem** is a special case of Ordinal (k, z)-Clustering Problem where the cost function is that of the well-studied k-median problem.*

Unified objective notation. For the special case where $S = X$, we drop explicit dependence on S and write the objective as

$$\phi_z(C) := \left(\sum_{x \in X} d(x, C)^z \right)^{1/z}, \quad \phi_\infty(C) := \max_{x \in X} d(x, C).$$

We use $\phi^{(z)}_{\mathrm{OPT}} := \min_{|C|=k} \phi_z(C)$ to denote the optimal cost.

The objective is to identify a subset $C \subseteq X$ of k centers that minimizes the relevant objective value $\phi_z(C)$ (or $\phi_\infty(C)$ for k-center). We drop the explicit dependence on d when clear from context and use $\phi^{(z)}_{\mathrm{OPT}}$ (resp. $\phi^{(\infty)}_{\mathrm{OPT}}$) for the optimal value.

Terminology and conventions. Throughout this chapter we work on a single ground set X endowed with a metric d. To avoid ambiguity across clustering and committee-selection language, we fix the following usage.

- **Points / agents.** Elements of X. We use "point" and "agent" interchangeably.
- **Voters (committee view).** When we interpret the problem as committee selection, the elements of X are the voters. Each $x \in X$ provides an ordinal ranking π_x.
- **Clients (clustering view).** In clustering arguments we call the same elements "clients." Thus, *voters* and *clients* both refer to X.
- **Centers / representatives / committee members.** The selected set $C \subseteq X$ of size k is the set of *centers*; in the committee interpretation these are the *representatives* or *committee members*. We use these three terms interchangeably for elements of C. (Note: an element of X can be both a client/voter and a center/representative; if $x \in C$ then $d(x, C) = 0$.)

- **Candidate (avoid ambiguity).** We do *not* maintain a separate external candidate set. When the word "candidate" appears in this chapter it always means "*prospective center*" (i.e., an element of X being considered for addition to C). To prevent confusion, we otherwise prefer "center/representative."

5.2 Bounded Distortion with $O(k)$ Distance Queries

Before aiming for constant distortion with sublinear query complexity, it is useful to establish a strong baseline that uses *very few* distance queries. In this section we show that a carefully guided farthest-first procedure, steered by ordinal information and sparingly augmented with distance queries, yields a 4-approximation for ordinal k-center while using only $2k$ queries. This construction will serve as a scaffold for the more elaborate constant-distortion algorithm later on.

Theorem 5.2.1 (Deterministic Ordinal k-Center Approximation)**.** *There exists a deterministic algorithm for the ordinal k-center clustering problem that achieves a distortion of at most 4 with respect to the optimal solution and makes at most $2k$ distance queries.*

Proof sketch. At iteration i, let C_i be the current centers and $Q_i \subseteq C_i$ the *query set*. For each $y \in C_i$, write $S_{y,i} = \{x \in X : d(x, y) = \min_{c \in C_i} d(x, c)\}$ and let $z_i(y) \in \arg\max_{x \in S_{y,i}} d(y, x)$ be its farthest client (identified ordinally). In round i we query only the distances $\{d(y, z_i(y)) : y \in Q_i\}$, add the client attaining the maximum, and update Q_i via an ordinal dominance rule; no additional queries are made during this update.

Lemma 5.2.2 (stability) implies that $z_i(y)$ does not change unless it is picked as a center, so each $y \in Q_i$ is queried at most once per (re)appearance; this yields the $2k$ bound (Lemma 5.2.3). Moreover, the chosen point constitutes a $1/2$-approximate farthest-first step (Lemma 5.2.4). Applying the standard argument (Lemma 5.2.5) then gives a 4-approximation for k-center, proving the theorem. □

The procedure performs a greedy traversal in which preference rankings determine the evolving cluster structure, and exact distances are queried only for a small query set of centers in each round. The full algorithm is given below.

Algorithm 8 Ordinal k-Center with $2k$ Queries

Input: Point set X, metric d, ordinal profile $P = \{\pi_x\}_{x \in X}$, integer k
Output: A set $C \subseteq X$ of k centers

Pick an arbitrary point $z \in X$
$C \leftarrow \{z,\ \pi_z(n)\},\ \ Q \leftarrow \{z,\ \pi_z(n)\}$
Define $S_{z,0} = \{\, x \in X \mid x \text{ ranks } z \text{ above } \pi_z(n)\},\ S_{\pi_z(n),0} = X \setminus S_{z,0}$
for $i = 1, \ldots, k-2$ **do**
  $\delta_{\max} \leftarrow 0$
  for each $y \in Q$ **do**
    Let $z_i \leftarrow \arg\max_{x \in S_{y,i}} d(y, x)$
    Query $\delta \leftarrow d(y, z_i)$
    if $\delta \geq \delta_{\max}$ **then**
      $\delta_{\max} \leftarrow \delta,\ r \leftarrow z_i,\ v \leftarrow y$
    end if
  end for
  $C \leftarrow C \cup \{r\},\ \ Q \leftarrow Q \setminus \{v\}$
  $R \leftarrow C \setminus Q$
  for each $u \in R$ **do**
    add $\leftarrow$ **true**
    Let $w \leftarrow \arg\max_{x \in S_{u,i+1}} d(u, x)$
    for each $p \in Q$ **do**
      $q \leftarrow \arg\max_{x \in S_{p,i+1}} d(p, x)$
      if $d(p, q) \geq d(u, q)$ **then**
        add $\leftarrow$ **false**
      end if
    end for
    if add **then**
      $Q \leftarrow Q \cup \{u\}$
    end if
  end for
end for
return C

Implementation note. Cluster assignments $(S_{y,i})$ and the farthest clients $z_i(y)$ are read off from the ordinal profile. The update of Q_i (the "ordinal dominance" test) checks, for each $p \in Q_i$ with bottleneck client q_p, whether q_p ranks a candidate u ahead of p; if not, u is dominated and not inserted. No extra distance queries are needed for this check.

We now turn to the formal analysis, beginning with the structural invariant that keeps the number of queries small.

Stability of farthest clients. The next lemma states that once we have queried a center $y \in Q_i$ for its farthest client, that client remains farthest for y unless it is added as a center; thus we never need to re-query y while that client is unselected.

Lemma 5.2.2. *For any $y \in Q_i$, let $z_i = \arg\max_{x \in S_{y,i}} d(y, x)$, and suppose $z_i \notin C_{i+1}$. Then* $\arg\max_{x \in S_{y,i}} d(y, x) = \arg\max_{x \in S_{y,i+1}} d(y, x)$.

Proof. We proceed by induction on i. The base case $i = 0$ is immediate. Now suppose the claim holds for iteration i. Let w be the new center added, and u the center whose cluster contained w.

Consider any $y \in Q_i$. If y was already in Q_i before u was added, then

$$d(z_i, w) > d(y, z_i) \Rightarrow \arg\max_{x \in S_{y,i+1}} d(y, x) = z_i.$$

If instead y joined Q_i after u, we again conclude z_i remains the farthest due to the relative distances, completing the proof. □

Bounding the number of queries. Using Lemma 5.2.2, each removal from Q_i corresponds to a single past query, and each insertion leads to at most one future query. Since there are at most k removals and k insertions, the total number of queries is at most $2k$.

Lemma 5.2.3. *The total number of distance queries performed by the algorithm is at most* $2k$.

Proof. By Lemma 5.2.2, each time a center is removed from Q_i it must have been queried once, and there are k such removals. Each addition to Q also incurs a query, and at most k such additions occur. Thus, the total number of queries is at most $2k$. □

We now argue that the algorithm performs an approximate farthest-first traversal: in each iteration, the selected point is not much closer than the true farthest point in the space.

Lemma 5.2.4. *At iteration i, let z be the newly chosen center (so $z \in S_{y,i}$ for some $y \in Q_i$), and let $u \in C_i$, $w \in S_{u,i}$. Then*

$$d(y, z) \geq \frac{1}{2}\, d(u, w).$$

Proof. Since z maximizes $d(y, x)$ over clusters in Q_i, for any $u \notin Q_i$ there exists some $y' \in Q_i$ such that

$$d(y', z') \geq d(u, w), \quad \text{with } z' = \arg\max_{x \in S_{y',i}} d(y', x).$$

Applying the triangle inequality gives

$$d(u, w) \leq d(y', z') + d(z', w) \leq 2d(y', z') \leq 2d(y, z),$$

so $d(y, z) \geq \frac{1}{2}\, d(u, w)$. □

Finally, we show that an approximate farthest-first traversal suffices to guarantee a constant-factor approximation to the optimal cost.

Lemma 5.2.5. *Suppose we iteratively select points such that, in each iteration, the newly selected point $z \in X$ satisfies*

$$d(z, C_i) \geq \alpha \cdot \max_{x \in X} d(x, C_i),$$

for some $\alpha \in (0, 1]$, where C_i is the set of centers selected so far. Then, after k-2 iterations, the set C_k satisfies

$$\max_{x \in X} d(x, C_{k-1}) \leq \frac{2}{\alpha} \cdot \phi^{(\infty)}_{\text{OPT}},$$

where $\phi^{(\infty)}_{\text{OPT}} := \min_{|C|=k} \max_{x \in X} d(x, C)$ denotes the optimal cost for the k-center objective.

Proof. Let $C^* = \{A_1, \ldots, A_k\}$ be the optimal clustering of X

We consider two cases:

Case 1: Each optimal cluster A_j contains at least one point from C_{k-2}. In this case, for every point $x \in X$, the triangle inequality yields:

$$d(x, C_{k-2}) \leq d(x, c_j^*) + d(c_j^*, C_{k-2}) \leq \phi_{\text{OPT}}^{(\infty)} + \phi_{\text{OPT}}^{(\infty)} = 2 \cdot \phi_{\text{OPT}}^{(\infty)}.$$

So we obtain a 2-approximation.

Case 2: Now consider the case where C_{k-2} includes two points $x_1, x_2 \in A_j$ from the same optimal cluster. By the triangle inequality and the definition of the optimal clustering, we have:

$$d(x_2, x_1) \leq d(x_2, c_j^*) + d(c_j^*, x_1) \leq 2 \cdot \phi_{\text{OPT}}^{(\infty)}.$$

Moreover, since x_2 was selected by the greedy rule,

$$d(x_2, C_i) \geq \alpha \cdot \max_{x \in X} d(x, C_i) \Rightarrow \max_{x \in X} d(x, C_i) \leq \frac{1}{\alpha} d(x_2, C_i).$$

But $x_1 \in C_i$, so:

$$d(x_2, C_i) \leq d(x_2, x_1) \leq 2 \cdot \phi_{\text{OPT}}^{(\infty)}.$$

Combining,

$$\max_{x \in X} d(x, C_i) \leq \frac{2}{\alpha} \cdot \phi_{\text{OPT}}^{(\infty)}.$$

□

This lemma formalizes a key intuition: as long as each new center reaches sufficiently far into unserved areas (i.e., is not too close to the existing centers), the algorithm makes meaningful progress in reducing the maximum client distance. Even though we never know the true distances for all points, using approximate farthest-first steps based on sparse queries still ensures that we don't miss large uncovered regions. This is why a small number of well-guided queries can lead to strong global guarantees.

With Lemma 5.2.5 we obtain a k-center bound: the α-approximate farthest-first procedure returns a set C_k with

$$\max_{x \in X} \min_{c \in C_k} d(x, c) \;\leq\; \frac{2}{\alpha} \min_{|C|=k} \max_{x \in X} \min_{c \in C} d(x, c),$$

i.e., a $(2/\alpha)$-approximation for k-center (setting $\alpha = \frac{1}{2}$ yields the 4-approximation we use as a scaffold). To leverage this for other clustering objectives, we relate the k-center objective to the (k, z) cost via norm inequalities: for any center set C and $z \geq 1$,

$$\Big(\sum_{x \in X} d(x, C)^z \Big)^{1/z} \;\leq\; n^{1/z} \max_{x \in X} d(x, C), \qquad \text{and} \qquad \max_{x \in X} d(x, C) \;\leq\; \Big(\sum_{x \in X} d(x, C)^z \Big)^{1/z}.$$

Thus an approximation for k-center immediately translates into a corresponding bound for (k, z), yielding the following translation from k-center to (k, z).

Lemma 5.2.6 (From k-center to (k, z)-clustering)**.** *Fix $z \geq 1$. Let $S \subseteq X$ with $|S| = k$. If*

$$\max_{x \in X} d(x, S) \;\leq\; \alpha \cdot \min_{|C|=k} \max_{x \in X} d(x, C),$$

then

$$\phi_z(S) \;\leq\; \alpha\, n^{1/z}\, \phi_{\text{OPT}}^{(z)}.$$

In particular, for $z = 1$ (the k-median objective), $\phi_1(S) \le \alpha n \, \phi_{\mathrm{OPT}}^{(1)}$.

Proof. First, for any S,

$$\phi_z(S)^z = \sum_{x \in X} d(x,S)^z \;\le\; \sum_{x \in X} \Big(\max_{y \in X} d(y,S) \Big)^z = n \Big(\max_{y \in X} d(y,S) \Big)^z,$$

so taking z-th roots yields

$$\phi_z(S) \;\le\; n^{1/z} \max_{x \in X} d(x,S).$$

By the hypothesis on S,

$$\max_{x \in X} d(x,S) \;\le\; \alpha \cdot \min_{|C|=k} \max_{x \in X} d(x,C).$$

Moreover, for every C we have $\max_x d(x,C) \le \phi_z(C)$ (since $\|\cdot\|_\infty \le \|\cdot\|_z$), hence

$$\min_{|C|=k} \max_{x} d(x,C) \;\le\; \min_{|C|=k} \phi_z(C) \;=\; \phi_{\mathrm{OPT}}^{(z)}.$$

Combining the above gives $\phi_z(S) \le \alpha \, n^{1/z} \, \phi_{\mathrm{OPT}}^{(z)}$. For $z = 1$, this is $\phi_1(S) \le \alpha n \, \phi_{\mathrm{OPT}}^{(1)}$. □

The power of this reduction lies in the relationship between different clustering norms: while the k-center objective minimizes the worst-case distance, k-median minimizes the average. Bounding the maximum distance immediately bounds the average — up to an $n^{1/z}$ factor — which means that any approximate k-center solution can serve as a useful (though loose) approximation for k-median. This justifies our use of the low-query k-center scaffold as a safe starting point for further refinement.

Combining Lemma 5.2.4 with Lemma 5.2.5 (with $\alpha = \frac{1}{2}$) gives a 4-approximate k-center solution using only $2k$ queries. By Lemma 5.2.6, this immediately yields a $4n$-approximation for k-median. This k-center scaffold provides a bounded-cost starting point with *very few* distance queries, which we leverage next to obtain a constant-distortion algorithm using only $O(k^4 \log^5 n)$ queries.

5.3 Constant Distortion with $O(k^4 \log^5 n)$ queries

To achieve constant distortion with a sublinear number of distance queries, Burkhardt et al. [29] emulate the distance-proportional sampling of k-median++ [15] without access to the full distance matrix. Their method partitions each current cluster into $O(\log n)$ ordinal rings and issues just one distance query per ring, enabling an approximation of the true sampling distribution up to constant factors. This allows them to sample new centers with probabilities that remain within a constant factor of those in the original full-information scheme, ensuring that each optimal cluster is hit with sufficiently high probability in expectation. Iterating this process for $T = \Theta(k \log n)$ rounds leads to geometric decay of the uncovered cost, culminating in a solution with constant distortion and polylogarithmic query complexity However, unlike the aforementioned works, they cannot guarantee an upper bound on the cost when relying solely on the sampling guarantee provided by Claim 5.3.3: there remains a nonzero probability of repeatedly sampling from the same clusters, which may lead to arbitrarily large distortion. To sidestep this issue, they employ the deterministic k-center solution developed in the previous section, which achieves a 4-approximation using only $2k$ distance queries. This initialization ensures that the total cost is at most $O(n)$ times the optimal (k, z)-cost. As a result, the low-probability event that previously implied unbounded distortion now only results in an $O(n)$-approximation, significantly strengthening the robustness of the overall algorithm.

First, we present some new notation and definitions that are going to be necessary for the understanding of their algorithm that follows.

Definition 5.3.1. *For any set of points $S \subseteq X$ and any point $c \notin S$, we define a partition of $S \cup \{c\}$ into disjoint subsets*

$$\{S_{c,1}, \ldots, S_{c,\ell}\}, \quad \textit{where} \quad \ell = \lfloor \log |S| \rfloor.$$

This partition is constructed recursively, starting from $S_{c,\ell}$. We let $S_{c,\ell}$ be the singleton set containing the point in S that is farthest from c. Then, for each $\ell > j > 1$, we define $S_{c,j}$ to consist of the $2^{\ell-j}$ farthest points from c in the set $S \setminus \bigcup_{i=j+1}^{\ell} S_{c,i}$. Finally, we set

$$S_{c,1} = S \setminus \bigcup_{j=2}^{\ell} S_{c,j}.$$

Furthermore, given a current set of centers C, we define an estimated cost for each ring $S_{i,j}$ in the hierarchical partition as $\widehat{\phi}_C(S_{i,j}) = |S_{i,j}| \cdot \min_{x \in S_{i,j-1}} d(x, c_i)$.To evaluate this estimate, it suffices to determine the distance between c_i and the top-ranked point in its preference list π_{c_i} that lies in $S_{i,j-1}$. We then emulate the k-median++ algorithm by defining a probability distribution $\mathcal{D}$ over candidate points. For any $c \in S_{r,j}$, its sampling probability is given by: $\widehat{p}(c) := \frac{1}{|S_{r,j}|} \cdot \frac{\widehat{\phi}_C(S_{r,j})}{\sum_{i,j} \widehat{\phi}_C(S_{i,j})}$, which forms a valid distribution, since the total probability mass sums to one. While the distribution $\widehat{p}$ defines a valid sampling probability over candidate points, in the algorithm, they amplify the probability mass by a factor of T, and sample each point c with adjusted probability $\widehat{p}_T(c) := \min\{1,\ T \cdot \widehat{p}(c)\}$. This boosted sampling scheme increases the likelihood of selecting points from high-cost regions, effectively emulating the geometric convergence of the original k-median++ sampling process, while enabling them to control the number of queries performed in each round.

Definition 5.3.2 (Covered Optimal Cluster)**.** *Let $C^* = \{A_1, \ldots, A_k\}$ denote the optimal clustering, and let C be the clustering induced by the algorithm under consideration. For each $i \in [k]$, we say that the optimal cluster A_i is* covered *if*

$$\phi_C(A_i) \leq 10 \cdot \phi_{C^*}(A_i),$$

and uncovered *otherwise. Throughout, $\phi_C(A_i)$ denotes the cost of serving the subset A_i using the centers in C, that is,*

$$\phi_C(A_i) = \left(\sum_{x \in A_i} d(x, C)^z \right)^{1/z}.$$

For ease of notation, we let Uncovered *(U) denote the set of all points contained in uncovered clusters.*

Having defined the key components — the hierarchical partitioning, cost estimates, and a method to evaluate coverage — we now present the full algorithm. This algorithm iteratively samples candidate centers using the approximate distribution $\widehat{p}$, while initializing with a robust k-center solution to ensure bounded cost even in worst-case sampling scenarios.

Algorithm 9 k-median with $O(k^4 \log^5 n)$ queries

Input: Point set X, ordinal profile $P = \{\pi_x\}_{x \in X}$, and $k \in \mathbb{N}$
Output: A set $C \subseteq X$ of size k

1: Initialize $C \leftarrow C_0$ where C_0 is the output of Algorithm 8
2: Sample a point c uniformly at random from X; set $C \leftarrow C \cup \{c\}$
3: **for** $t = 1$ to T **do**
4: **for each** $c \in C$ **do**
(a) $S_c \leftarrow \{x \in X : c = \arg\min_{c' \in C} \pi_x(c')\}$
(b) Partition S_c per Definition 5.3.1
(c) Compute $\widehat{\phi}_C(S_{c,j})$ for all j
end for
5: Sample $c \in S_{xj}$ with probability:

$$\widehat{p} \leftarrow \min \left\{ 1,\ \frac{T}{|S_{x,j}|} \cdot \frac{\widehat{\phi}_C(S_{x,j})}{\sum_{i,j} \widehat{\phi}_C(S_{i,j})} \right\}$$

6: set $C \leftarrow C \cup \{c\}$
7: **endfor**
8: Approximate C with a committee C' of size k
9: **return** C'

We now analyze the performance of Algorithm 9. Our goal is to show that the uncovered cost decreases geometrically across iterations, culminating in a constant distortion approximation. To do so, we first establish that the emulated sampling distribution retains a constant fraction of the probability mass of the ideal k-median++ distribution. Then, we show that uncovered optimal clusters are likely to be hit in each round, and finally, we prove that the expected uncovered cost shrinks by a constant factor in each iteration.

Claim 5.3.3. *Given a current set of centers C, let $p(c)$ denote the probability that the standard k-median++ algorithm adds point c to C, and let $\widehat{p}(c)$ denote the probability that c is sampled under the distribution $\mathcal{D}$. Then,*

$$\widehat{p}(c) \geq \frac{1}{2} \cdot p(c).$$

Proof. We begin by recalling the expressions for the sampling probabilities in the standard k-median++ algorithm and in our modified distribution $\mathcal{D}$.

In the standard k-median++ algorithm, the probability of adding point c to the set C is

$$p(c) = \frac{d(c, C)}{\sum_{x \in X} d(x, C)} = \frac{d(c, C)}{\phi_C(X)},$$

where $\phi_C(X) = \sum_{x \in X} d(x, C)$ denotes the total cost with respect to the current centers C.

Recall that the estimated cost, used in Algorithm 9 is defined as

$$\widehat{\phi}_C(S_{x,j}) = |S_{x,j}| \cdot \min_{q \in S_{x,j-1}} d(q, C).$$

By construction, for every $c \in S_{x,j}$, we have

$$\min_{q \in S_{x,j-1}} d(q, C) \geq d(c, C),$$

which implies

$$\widehat{\phi}_C(S_{x,j}) \geq |S_{x,j}| \cdot d(c, C).$$

Substituting this into the expression for $\widehat{p}(c)$, we obtain:

$$\widehat{p}(c) \geq \frac{1}{|S_{x,j}|} \cdot \frac{|S_{x,j}| \cdot d(c, C)}{\sum_{i,j} \widehat{\phi}_C(S_{i,j})} = \frac{d(c, C)}{\sum_{i,j} \widehat{\phi}_C(S_{i,j})}.$$

Thus, to prove that $\widehat{p}(c) \geq \frac{1}{2} \cdot p(c)$, it suffices to show that

$$\sum_{i,j} \widehat{\phi}_C(S_{i,j}) \leq 2 \cdot \phi_C(X).$$

To see this, observe that for all i, j,

$$\widehat{\phi}_C(S_{i,j}) = |S_{i,j}| \cdot \min_{q \in S_{i,j-1}} d(q, C) \leq 2 \cdot |S_{i,j-1}| \cdot \min_{q \in S_{i,j-1}} d(q, C) \leq 2 \cdot \phi_C(S_{i,j-1}),$$

where the inequality follows from the fact that each ring $S_{i,j}$ contains at most twice as many points as $S_{i,j-1}$ by construction.

Summing over all i, j, we obtain:

$$\sum_{i,j} \widehat{\phi}_C(S_{i,j}) \leq 2 \cdot \sum_{i,j} \phi_C(S_{i,j-1}) \leq 2 \cdot \phi_C(X),$$

which completes the proof. □

Having established that our sampling distribution $\widehat{p}(\cdot)$ approximates the standard k-median++ probabilities up to a constant factor, we can now leverage this bound to argue about the likelihood of covering uncovered optimal clusters.

Lemma 5.3.4. *Let C be the current set of centers and let A be an optimal cluster from the optimal solution C^* that is not yet covered by C. Then, after sampling a new center (according to the distribution described in Algorithm 9), the probability that A remains uncovered is at most*

$$\Pr[A \textit{ remains uncovered}] \leq \exp\left(-\frac{T \cdot \phi_C(A)}{10 \cdot \phi_C(X)}\right),$$

where $\phi_C(A)$ denotes the total cost of points in A under the current center set C, and $\phi_C(X)$ is the total cost over all points.

Proof. Burkhardt et al. [29] (Claim B.3) first prove that for the k-median objective, sampling a point $c \in A$ according to the distribution D^{++} (induced by the standard k-median++ algorithm) leads to

$$\mathbb{E}_{D^{++}}[\phi_{C \cup \{c\}}(A)] \leq 4 \cdot \phi_{C^*}(A).$$

By Markov's inequality, this implies

$$\Pr_{D^{++}}\left[\phi_{C \cup \{c\}}(A) \geq 5 \cdot \phi_{C^*}(A)\right] \leq \frac{4}{5}.$$

Therefore, with probability at least $\frac{1}{5}$, the cluster A becomes covered when a center is sampled from A using D^{++}.

This means there exists a subset $A' \subseteq A$ such that:

(i) Sampling any point $c \in A'$ makes A covered, and

(ii) The total cost of A' under C satisfies $\phi_C(A') \geq \frac{1}{5} \cdot \phi_C(A)$.

Now since we sample a center according to distribution D, amplified T times, using the boosted lower bounds from Claim 5.3.3, we have:

$$\Pr[A \text{ remains uncovered}] \leq \prod_{c \in A'} (1 - T \cdot \widehat{p}(c)) \leq \prod_{c \in A'} \left(1 - \frac{T \cdot p(c)}{2}\right).$$

Applying the inequality $1 - x \leq e^{-x}$, we get:

$$\Pr[A \text{ remains uncovered}] \leq \exp\left(-\sum_{c \in A'} \frac{T \cdot p(c)}{2}\right).$$

Recalling that $\sum_{c \in A'} p(c) = \frac{\phi_C(A')}{\phi_C(X)} \geq \frac{1}{5} \cdot \frac{\phi_C(A)}{\phi_C(X)}$, it follows that

$$\Pr[A \text{ remains uncovered}] \leq \exp\left(-\frac{T}{2} \cdot \frac{\phi_C(A')}{\phi_C(X)}\right) \leq \exp\left(-\frac{T \cdot \phi_C(A)}{10 \cdot \phi_C(X)}\right),$$

which completes the proof. □

This lemma shows that, in expectation, uncovered optimal clusters—especially those with non-negligible cost—are likely to be covered in each iteration. This insight is the key to showing geometric decay in the uncovered cost, which we pursue next.

To formalize this decay, we now partition the uncovered optimal clusters into those that contribute significantly to the cost (*heavy*) and those that do not (*light*). This separation enables us to argue that, even if some low-cost clusters remain uncovered, the bulk of the cost decreases sharply.

Lemma 5.3.5. *Let ϕ_{OPT} be the cost of an optimal k-median clustering, and let t be a round such that the cost of the current solution C_t satisfies*

$$\phi_{C_t}(X) \geq 20 \cdot \phi_{OPT}.$$

Then, the expected cost of the uncovered clusters after the next iteration satisfies:

$$\mathbb{E}[\phi_{C_{t+1}}(U)] \leq \frac{1 + \exp\left(-\frac{T}{40k}\right)}{2} \cdot \phi_{C_t}(U)$$

Proof. We begin by observing that the assumption $\phi_{C_t}(X) \geq 20 \cdot \phi_{\text{OPT}}$ implies

$$\phi_{C_t}(U) \geq \frac{1}{2} \cdot \phi_{C_t}(X),$$

since otherwise the total cost of the covered clusters would exceed $\frac{1}{2} \cdot \phi_{C_t}(X)$, implying that

$$\phi_{C_t}(X) < 2 \cdot \phi_{C_t}(U) < 2 \cdot 10 \cdot \phi_{\text{OPT}} = 20 \cdot \phi_{\text{OPT}},$$

which contradicts the assumption.

We now partition the uncovered optimal clusters U into two collections:

- The *heavy* clusters:
$$\mathcal{H}_t := \left\{ A \subseteq U \mid \phi_{C_t}(A) \geq \frac{\phi_{C_t}(U)}{2k} \right\},$$
- The *light* clusters: $\mathcal{L}_t := U \setminus \mathcal{H}_t$.

Let $A \in \mathcal{H}_t$. By Lemma 5.3.4, the probability that such a heavy cluster A remains uncovered in round $t+1$ is at most

$$\exp\left(-\frac{T \cdot \phi_{C_t}(A)}{10 \cdot \phi_{C_t}(X)}\right) \leq \exp\left(-\frac{T \cdot \phi_{C_t}(U)}{20k \cdot \phi_{C_t}(X)}\right) \leq \exp\left(-\frac{T}{40k}\right),$$

where the last inequality uses $\phi_{C_t}(U) \geq \frac{1}{2} \cdot \phi_{C_t}(X)$.

Therefore, the probability that a heavy cluster gets covered is at least $1 - \exp\left(-\frac{T}{40k}\right)$. This implies that the expected decrease in total uncovered cost satisfies:

$$\phi_{C_t}(U) - \mathbb{E}[\phi_{C_{t+1}}(U)] \geq \left(1 - \exp\left(-\frac{T}{40k}\right)\right) \cdot \sum_{A \in \mathcal{H}_t} \phi_{C_t}(A).$$

But the total cost of the light clusters is at most:

$$\sum_{A \in \mathcal{L}_t} \phi_{C_t}(A) \leq k \cdot \frac{\phi_{C_t}(U)}{2k} = \frac{1}{2} \cdot \phi_{C_t}(U),$$

so the heavy clusters contribute at least:

$$\sum_{A \in \mathcal{H}_t} \phi_{C_t}(A) \geq \phi_{C_t}(U) - \frac{1}{2} \cdot \phi_{C_t}(U) = \frac{1}{2} \cdot \phi_{C_t}(U).$$

Hence,

$$\mathbb{E}[\phi_{C_{t+1}}(U)] \leq \underbrace{\sum_{A \in \mathcal{L}_t} \phi_{C_t}(A)}_{\leq \frac{1}{2} \cdot \phi_{C_t}(U)} + \underbrace{\sum_{A \in \mathcal{H}_t} \Pr[A \text{ remains uncovered}] \cdot \phi_{C_t}(A)}_{\leq \exp\left(-\frac{T}{40k}\right) \cdot \sum_{A \in \mathcal{H}_t} \phi_{C_t}(A)}.$$

Since $\sum_{A \in \mathcal{L}_t} \phi_{C_t}(A) \leq \frac{1}{2} \cdot \phi_{C_t}(U)$ and $\sum_{A \in \mathcal{H}_t} \phi_{C_t}(A) \leq \frac{1}{2} \cdot \phi_{C_t}(U)$, we obtain:

$$\mathbb{E}[\phi_{C_{t+1}}(U)] \leq \left(\frac{1}{2} + \frac{1}{2} \cdot \exp\left(-\frac{T}{40k}\right)\right) \cdot \phi_{C_t}(U),$$

which simplifies to

$$\mathbb{E}[\phi_{C_{t+1}}(U)] \leq \frac{1 + \exp\left(-\frac{T}{40k}\right)}{2} \cdot \phi_{C_t}(U).$$

This completes the proof. □

Theorem 5.3.6 (Burkhardt et al.)**.** *Algorithm 9 achieves an expected $O(1)$-distortion for the committee election problem using $O(k^4 \log^5 n)$ distance queries.*

Proof. Algorithm 9 invokes Algorithm 8 as a subroutine, which yields a 4-approximation to the optimal k-center cost. By Lemma 5.2.6, this implies:

$$\mathbb{E}[\phi_{C_0}(U)] \leq 4n \cdot \phi_{\text{OPT}}.$$

Then, applying Lemma 5.3.5 iteratively, we obtain the recurrence:

$$\mathbb{E}[\phi_{C_{t+1}}(U)] \leq 20 \cdot \phi_{\text{OPT}} + \alpha \cdot \mathbb{E}[\phi_{C_t}(U)], \quad \text{where } \alpha := \frac{1 + \exp\left(-\frac{T}{40k}\right)}{2}.$$

Unfolding this recurrence for T rounds gives:

$$\mathbb{E}[\phi_{C_T}(U)] \leq \alpha^T \cdot 4n \cdot \phi_{\text{OPT}} + 20 \cdot \phi_{\text{OPT}} \cdot \sum_{t=0}^{T-1} \alpha^t.$$

Letting $T \geq 40k \log n$, we ensure that

$$\exp\left(-\frac{T}{40k}\right) \leq \frac{1}{n}, \quad \text{which implies} \quad \alpha = \frac{1 + \exp\left(-\frac{T}{40k}\right)}{2} \leq \frac{n+1}{2n}.$$

which yields:

$$\mathbb{E}[\phi_{C_T}(U)] \leq \frac{n+1}{2n}^{40k \log n} \cdot 4n \cdot \phi_{\text{OPT}} + 20 \cdot \phi_{\text{OPT}} \cdot \frac{2n}{n-1} \lesssim 2 \cdot \phi_{\text{OPT}} + 40 \cdot \phi_{\text{OPT}} = 42 \cdot \phi_{\text{OPT}}.$$

Thus,

$$\mathbb{E}[\phi_{C_T}(U)] \leq 42 \cdot \phi_{\text{OPT}}.$$

which then yields:

$$\mathbb{E}[\phi_{C_T}(X)] \leq \mathbb{E}[\phi_{C_T}(U)] + 10 \cdot \phi_{\text{OPT}} \leq 52 \cdot \phi_{\text{OPT}}.$$

Since the algorithm runs for T rounds and adds $\Theta(T)$,in expectation, centers per round, the total number of centers opened is $O(k^2 \log^2 n)$.

Therefore Algorithm 9 at the end of step 7 has created an $(O(k^2 \log^2 n), 52)$-bicriteria solution, meaning it uses $O(k^2 \log^2 n)$ as many centers and achieves an approximation (distortion) of 52

Finally, applying any constant-factor approximation algorithm to select k centers from the committee C produced by Algorithm 9 (e.g., the 2.613-approximation of Gowda et al. [54]), and leveraging claim 5.3.7 (Stated below) which incurs an additional factor of 4, we get a final distortion bound of:

$$4 \cdot 52 \cdot 2.613 < 544.$$

The algorithm runs for $T = O(k \log n)$ rounds. In each round, it adds $\Theta(T)$ new centers in expectation, resulting in a total of $T^2 = O(k^2 \log^2 n)$ candidate centers over the course of the algorithm.

To select each new center, the algorithm samples one client from every ring of the current center set. Each center defines $O(\log n)$ rings (as per Definition 5.3.1), so at each round, it samples $O(T \log n)$ clients. Across all T rounds, this leads to a total of

$$O(T^2 \log n) = O(k^2 \log^3 n)$$

sampled clients.

For the final reduction step, which transforms the bicriteria solution into a true k-clustering, we need to know the distances between all sampled clients and all candidate centers. This results in

$$O(k^2 \log^3 n) \cdot O(k^2 \log^2 n) = O(k^4 \log^5 n)$$

distance queries in total. These queries suffice both for computing the sampling probabilities during the algorithm and for executing the reduction step.

□

Claim 5.3.7 (Approximation from bicriteria solution to true k-clustering Solution)**.** *Let X be a point set in a metric space, and let $C' \subseteq X$ be an (α, β)-bicriteria solution for the (k, z)-clustering problem. Construct a multiset X' by assigning each point $x \in X$ to its closest center $c_x \in C'$, and replicating c_x once per assigned point. Let $C \subseteq C'$ be any γ-approximate solution for the (k, z)-clustering problem on X'. Then C is a $4\alpha\gamma$-approximate solution for X.*

Proof. Let $C_{\text{OPT}} \subseteq X$ be an optimal solution of size k, and for each $x \in X$, let $c_x = \arg\min_{c \in C'} d(x, c)$.

We begin by bounding the cost of clustering X' with respect to C_{OPT}:

$$\sum_{x \in X} d^z(c_x, C_{\text{OPT}}) \leq 2^{z-1} \sum_{x \in X} \left(d^z(c_x, x) + d^z(x, C_{\text{OPT}})\right) = 2^{z-1} \left(\phi_z(C') + \phi_z(C_{\text{OPT}})\right) \leq 2^{z-1}(\alpha^z + 1) \cdot \phi_z(C_{\text{OPT}}),$$

where the first inequality follows from the Minkowski inequality for $z \geq 1$, and the last from the fact that C' is an α-approximate solution.

Now, since C is a γ-approximate solution on X', we have:

$$\sum_{x \in X} d^z(c_x, C) \leq \gamma^z \cdot \sum_{x \in X} d^z(c_x, C_{\text{OPT}}) \leq \gamma^z \cdot 2^{z-1}(\alpha^z + 1) \cdot \phi_z(C_{\text{OPT}}).$$

Finally, using the triangle inequality again, we bound the total cost of clustering X with respect to C:

$$\begin{aligned}
\sum_{x \in X} d^z(x, C) &\leq 2^{z-1} \sum_{x \in X} \left(d^z(x, c_x) + d^z(c_x, C)\right) \\
&= 2^{z-1} \left(\phi_z(C') + \sum_{x \in X} d^z(c_x, C)\right) \\
&\leq 2^{z-1} \cdot \alpha^z \cdot \phi_z(C_{\text{OPT}}) + 2^{z-1} \cdot \gamma^z \cdot 2^{z-1}(\alpha^z + 1) \cdot \phi_z(C_{\text{OPT}}).
\end{aligned}$$

Combining both terms:

$$\phi_z(C) \leq \left(2^{z-1}\alpha^z + 2^{2z-2}\gamma^z(\alpha^z + 1)\right) \cdot \phi_z(C_{\text{OPT}}) \leq 2^{2z} \cdot \gamma^z \cdot \alpha^z \cdot \phi_z(C_{\text{OPT}}),$$

where the last inequality holds for $\alpha, \gamma \geq 1$.

Taking the zth root yields:

$$\phi(C) \leq 2^2 \cdot \gamma \cdot \alpha \cdot \phi(C_{\text{OPT}}) = 4\gamma\alpha \cdot \phi(C_{\text{OPT}}).$$

□

CHAPTER 6

Stability on Clustering and Voting Mechanism

A standard approach in the design and analysis of computational problems is *worst-case analysis*, and *voting mechanisms* is no exception. While this framework provides a comprehensive measure of a problem's computational complexity, it imposes a constraint: we must consider a single algorithm to handle all possible instances, even when our interest lies primarily in certain "special" or structured cases that might admit more efficient solutions.

This issue is particularly relevant in problems like *clustering* and the *k-committee election problem*, both of which involve optimization tasks that are generally *NP-Hard* in arbitrary metric spaces [2, 57]. That is, there is no known polynomial-time algorithm that can solve all instances of these problems exactly. As a result, a significant body of work has focused on developing *approximation algorithms* that compute near-optimal solutions with provable guarantees, particularly for problems like k-median, k-means and facility location and other objective functions [12, 61, 75]. However, in real-world applications, our concern is rarely with all theoretical inputs, but rather with those that reflect meaningful, structured data.

In the context of clustering, this distinction is elegantly summarized by Bilu, Daniely, Linial, and Saks, who argue that "clustering is either easy or pointless" [22], and echoed by Roughgarden, who observes that "clustering is hard when it doesn't matter" [70] [71]. These perspectives suggest that hard instances may be of limited practical relevance, whereas instances encountered in the real world often exhibit features that make them tractable.

Clustering aims to partition a dataset into groups such that elements within the same group are "similar," while those in different groups are "dissimilar." In real-world scenarios, it is usually assumed that such clusters are well-defined—that is, similar items are located close together and are clearly separated from other clusters. This intuition carries over to k-committee election, where voters or agents often form communities with clear boundaries. These well-separated structures simplify the identification of cohesive groups and their representative candidates.

In the following sections, we refer to such instances—where the problem exhibits sufficient internal structure—as *stable instances*, and we explore the interesting properties of stability.

6.1 Stable Clustering

6.1.1 Definitions and Preliminaries

First of all, we will define the components of the classic clustering problem.

Definition 6.1.1 (Clustering Problem)**.** *A clustering problem is defined by a tuple* $((X, d), \mathcal{H}, k)$*, where:*

- (X, d) *is a metric space, with* X *being a set of data points and* d *a distance function defined on pairs of points in* X*,*
- $\mathcal{H}$ *is an objective function that assigns a nonnegative real-valued cost to any partition of* X *into* k *subsets* $C_1, \ldots, C_k$*, based on the metric* d*,*
- $k > 1$ *is the number of clusters.*

The goal is to find a partition $\{C_1, \ldots, C_k\}$ *of* X *that minimizes the cost given by* $\mathcal{H}$*.*

Definition 6.1.2 (Center-based and Separable Objectives)**.** *A clustering objective is center-based if the optimal solution can be defined by k points* $c_1, ..., c_k$ *in the metric space, called centers, such that every data point is assigned to its nearest center. Such a clustering objective is separable if it further satisfies the following two conditions:*

- *The objective function value of a given clustering is either a (weighted) sum or the maximum of the individual cluster scores.*
- *Given a proposed single cluster, its score can be computed in polynomial time.*

The most well-studied and, perhaps, most interesting clustering objectives are k*-means*, k*-median*, and k*-center*. These objectives are defined as follows. Given a clustering $C_1, \ldots, C_k$, the objective is the minimum over all choices of centers $c_1 \in C_1, \ldots, c_k \in C_k$ of the following functions:

$$\mathcal{H}_{\text{means}}(C_1, \ldots, C_k; d) = \sum_{i=1}^{k} \sum_{u \in C_i} d(u, c_i)^2$$

$$\mathcal{H}_{\text{median}}(C_1, \ldots, C_k; d) = \sum_{i=1}^{k} \sum_{u \in C_i} d(u, c_i)$$

$$\mathcal{H}_{\text{center}}(C_1, \ldots, C_k; d) = \max_{i \in \{1,\ldots,k\}} \left\{ \max_{u \in C_i} d(u, c_i) \right\}$$

It is evident that the $\mathcal{H}_{\text{means}}$ objective corresponds to the *social cost*, and the $\mathcal{H}_{\text{center}}$ objective corresponds to the *maximum cost* for any voter in the k-committee election problem.

Several studies have explored different formalizations of *stability* in the context of clustering. One such notion, *approximation stability*, was introduced by Balcan, Blum, and Gupta [17]. In their framework, a clustering instance is said to exhibit approximation stability if any solution that approximates the objective function well is also close to a desired ground-truth clustering. More precisely, a k-median instance is said to be (c, ϵ)-approximation stable if every c-approximate k-clustering is ϵ-accurate, meaning that it agrees with the target clustering on at least a $1 - \epsilon$ fraction of the data points.

In this work, however, we focus on a different notion known as *perturbation stability*, originally proposed by Bilu and Linial and subsequently studied in [21], [16], [19], and [7]. The motivation behind perturbation stability arises from the observation that, in practical applications, pairwise distances between data points are often determined using heuristics such as Euclidean distance. As such, if the optimal clustering under a given distance function is meaningful, then it should remain optimal under small perturbations of the input distances, unless the correct solution is obtained merely by chance.

Definition 6.1.3 (γ-perturbation)**.** *Given a metric space (X, d) and a parameter $\gamma \geq 1$, we say that a function $d' : X \times X \to \mathbb{R}_{>0}$ is a γ-perturbation of d if, for all $x, y \in X$, the following holds:*

$$\frac{d(x,y)}{\gamma} \leq d'(x,y) \leq d(x,y).$$

Definition 6.1.4 (γ-stability)**.** *Suppose we have a clustering instance composed of n points residing in a metric (X, d) and an objective function $\mathcal{H}$ we wish to optimize. We call the clustering instance γ-perturbation stable for $\mathcal{H}$ if for any d' which is a γ-perturbation of d, the (only) optimal clustering of (X, d') under $\mathcal{H}$ is identical, as a partition of points into subsets, to the optimal clustering of (X, d) under $\mathcal{H}$.*

A related but weaker notion of stability is known as γ-*metric perturbation stability*. In contrast to general perturbation stability—where the perturbed distance function d' need not be a metric (i.e., it may violate the triangle inequality)—γ-metric perturbation stability restricts attention to perturbations that do preserve the metric properties. Specifically, an instance is said to be γ-stable if it admits the same optimal solution under every γ-perturbation of the original distance function. For γ-metric stability, this requirement is relaxed: the optimal clustering must remain unchanged only under γ-*metric* perturbations, which form a subset of all γ-perturbations. Consequently, the set of γ-metric stable instances strictly contains the set of γ-stable instances. We refer to γ-metric perturbation stability as a weaker notion because it imposes less stringent conditions, making it applicable to a broader class of instances.

Definition 6.1.5 (γ-metric perturbation and γ-metric stability)**.** *Let (X, d) be a metric space and let $\gamma \geq 1$. A metric d' is called a γ-metric perturbation of d if for all $u, v \in X$, it holds that*

$$\frac{d(u,v)}{\gamma} \leq d'(u,v) \leq d(u,v),$$

and d' satisfies the properties of a metric (including the triangle inequality).

An instance $((X, d), \mathcal{H}, k)$ is said to be γ-metric perturbation stable if, for every γ-metric perturbation d' of d, the optimal clustering for $((X, d'), \mathcal{H}, k)$ is identical to that of $((X, d), \mathcal{H}, k)$.

6.2 Properties Of Perturbation Stable Instances

In any γ-stable instance, the optimal clustering satisfies the *center proximity property*, which ensures that every point is at least γ times closer to its assigned center than to any other center in the optimal solution. This captures the idea that points are most strongly associated with their own cluster.

Definition 6.2.1 (γ-center proximity)**.** *Let $\gamma \geq 1$, and let $((X, d), \mathcal{H}, k)$ be a γ-stable clustering instance with unique optimal clustering $\{C_1, \ldots, C_k\}$ and corresponding optimal centers $\{c_1, \ldots, c_k\}$.*

Then the instance satisfies the γ-center proximity property: for all $i \neq j$ and for every point $x_i \in C_i$, it holds that

$$d(x_i, c_j) > \gamma \cdot d(x_i, c_i).$$

Proof. Let $\gamma \geq 1$, and let C_i and C_j be any two clusters in the optimal clustering, with centers c_i and c_j respectively. Let $x \in C_i$ be an arbitrary point.

Define a perturbed distance function d' such that:

- $d'(x, c_i) = d(x, c_i)$,
- All other distances are scaled down by a factor of γ: for all other $u, v \in X$, $d'(u, v) = \frac{1}{\gamma} \cdot d(u, v)$.

Since the instance is γ-stable, the optimal clustering must remain unchanged under any γ-perturbation, including this one. In particular, x must remain assigned to C_i, so:

$$d'(x, c_i) < d'(x, c_j).$$

Substituting in the values from d', we have:

$$d(x, c_i) < \frac{1}{\gamma} \cdot d(x, c_j),$$

which implies:

$$d(x, c_j) > \gamma \cdot d(x, c_i).$$

Thus, the instance satisfies the γ-center proximity property. □

An immediate consequence of the γ-center proximity property is that any γ-stable instance with $\gamma \geq 2$ satisfies the *weak* γ-center proximity condition. Specifically, for any pair of clusters C_i and C_j with $i \neq j$, and for any points $x \in C_i$, $y \in C_j$, it holds that:

$$d(x, y) > (\gamma - 1) \cdot d(x, c_i),$$

where c_i is the center of cluster C_i in the optimal clustering.

Proposition 6.2.2 (Weak γ-center proximity)**.** *Let $\gamma \geq 2$, and let $((X, d), \mathcal{H}, k)$ be a γ-stable instance with unique optimal clustering $\{C_1, \ldots, C_k\}$ and corresponding optimal centers $\{c_1, \ldots, c_k\}$. Then, for all $i \neq j$, for all points $x \in C_i$ and $y \in C_j$, it holds that:*

$$d(x, y) > (\gamma - 1) \cdot d(x, c_i).$$

Proof. Let $x \in C_i$ and $y \in C_j$ for $i \neq j$, and let c_i and c_j be the centers of clusters C_i and C_j, respectively. We consider two cases based on the relative distances from x and y to their respective centers.

- **Case (a):** $d(y, c_j) \geq d(x, c_i)$.
  By the triangle inequality, we have:

  $$d(x, y) \geq d(y, c_i) - d(x, c_i).$$

  Since the instance is γ-stable, it satisfies the γ-center proximity property, which implies:

  $$d(y, c_i) > \gamma \cdot d(y, c_j).$$

Substituting, we get:

$$d(x, y) > \gamma \cdot d(y, c_j) - d(x, c_i) \geq \gamma \cdot d(y, c_j) - d(y, c_j) = (\gamma - 1) \cdot d(y, c_j).$$

- **Case (b):** $d(y, c_j) < d(x, c_i)$.
  Again, by the triangle inequality:

  $$d(x, y) \geq d(x, c_j) - d(y, c_j).$$

  Using γ-center proximity for x, we know:

  $$d(x, c_j) > \gamma \cdot d(x, c_i),$$

  and since $d(y, c_j) < d(x, c_i)$, it follows that:

  $$d(x, y) > \gamma \cdot d(x, c_i) - d(y, c_j) > \gamma \cdot d(x, c_i) - d(x, c_i) = (\gamma - 1) \cdot d(x, c_i).$$

In both cases, the inequality $d(x, y) > (\gamma - 1) \cdot d(x, c_i)$ holds. □

We now state a central implication of stability: the *cluster separation property*. In any γ-stable instance, the distance between any two points assigned to different clusters is bounded below by a function of γ, ensuring that inter-cluster distances are sufficiently large.

Lemma 6.2.3 (Cluster separation property)**.** *Let $\gamma \geq 2$, and let $((X, d), \mathcal{H}, k)$ be a γ-stable instance with unique optimal clustering $\{C_1, \ldots, C_k\}$ and corresponding centers $\{c_1, \ldots, c_k\}$. Let $x_i, x_i' \in C_k$ and $x_j \in C_{k'}$ for $k \neq k'$. Then:*

$$d(x_i, x_j) > \frac{(\gamma - 1)^2}{2\gamma} \cdot d(x_i, x_i').$$

Proof. Let c_k and $c_{k'}$ denote the centers of clusters C_k and $C_{k'}$, respectively. Since the instance is γ-stable, it satisfies the γ-center proximity property:

$$d(x_i, c_{k'}) > \gamma \cdot d(x_i, c_k).$$

Applying the triangle inequality:

$$\begin{aligned} d(x_i, c_k) + d(c_k, c_{k'}) &> \gamma \cdot d(x_i, c_k) \\ \Rightarrow \quad d(c_k, c_{k'}) &> (\gamma - 1) \cdot d(x_i, c_k). \end{aligned}$$

We also have, from the triangle inequality:

$$d(c_k, c_{k'}) < d(c_k, x_i) + d(x_i, x_j) + d(x_j, c_{k'}).$$

Next, by the weak γ-center proximity property (which holds for $\gamma \geq 2$), we know:

$$d(x_j, c_{k'}) > \frac{1}{\gamma - 1} \cdot d(x_i, x_j), \quad \text{and} \quad d(x_i, c_{k'}) > \frac{1}{\gamma - 1} \cdot d(x_i, x_j).$$

Substituting into the earlier inequality, we obtain:

$$d(c_k, c_{k'}) < \frac{1}{\gamma - 1} \cdot d(x_i, x_j) + d(x_i, x_j) + \frac{1}{\gamma - 1} \cdot d(x_i, x_j) = \frac{\gamma + 1}{\gamma - 1} \cdot d(x_i, x_j).$$

Finally, apply the triangle inequality again to relate $d(x_i, x_i')$:

$$\begin{aligned} d(x_i, x_i') &< d(x_i, c_k) + d(c_k, x_i') \leq 2 \cdot d(x_i, c_k) \\ &< \frac{2}{\gamma - 1} \cdot d(x_i, x_j) + \frac{2}{\gamma - 1} \cdot d(c_k, c_{k'}) \\ &< \frac{2}{\gamma - 1} \cdot d(x_i, x_j) + \frac{2(\gamma + 1)}{(\gamma - 1)^2} \cdot d(x_i, x_j) = \frac{2\gamma}{(\gamma - 1)^2} \cdot d(x_i, x_j). \end{aligned}$$

Rearranging gives:

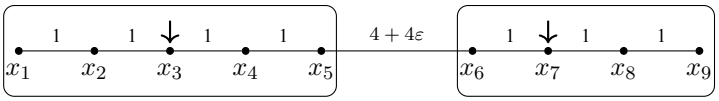
$$d(x_i, x_j) > \frac{(\gamma - 1)^2}{2\gamma} \cdot d(x_i, x_i'),$$

which concludes the proof. □

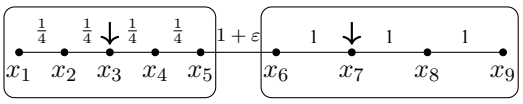
The three properties—Definition 6.2.1, Proposition 6.2.2, and Lemma 6.2.3—provide useful inequalities that characterize the inter- and intra-cluster distances of a γ-stable instance.It is important to note that the three properties presented above are *necessary* for γ-stability but not *sufficient*. That is, an instance may satisfy all three properties and still fail to be γ-stable.

As illustrated in Figure 6.1, we can construct such a counterexample. In this instance, all the inequalities corresponding to the γ-center proximity, weak γ-center proximity, and cluster separation properties hold for $\gamma = 4$ and any $\varepsilon > 0$. However, if we apply a γ-perturbation that scales down the distances between agents $x_1 - x_2$, $x_2 - x_3$, $x_3 - x_4$ and $x_4 - x_5$ by a factor of γ, the optimal clustering changes.

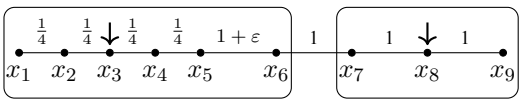
In the middle configuration (Figure 6.1b), the clustering remains the same as in the original instance, and its cost is $\frac{6}{4} + 4 = 5.5$. In contrast, the configuration on the bottom (Figure 6.1c) has a lower cost of $5 + \varepsilon$, due to the new center placements. If the original instance were truly γ-stable, the optimal clustering would remain unchanged under all valid γ-perturbations. Therefore, this instance is not 4-stable.



(a) Original instance.



(b) Perturbed Instance with original Clustering



(c) Perturbed Instance with different Clustering

1.1 An instance where the three geometric properties hold but the instance is not γ-stable.

We now formally introduce the *Min-stability* property, first defined by [18] and later applied in the context of perturbation-stable clustering by [16]. This property plays a central role in our approach,

as it underlies the hierarchical structure we exploit to design efficient algorithms for the k-committee election problem under perturbation stability.

Lemma 6.2.4 (Min-Stability [16])**.** *Let $\gamma \geq 2 + \sqrt{3}$, and let $((X, d), \mathcal{H}, k)$ be a γ-stable instance with unique optimal clustering $\{C_1, \ldots, C_k\}$. Let $C, C' \in \{C_1, \ldots, C_k\}$ be two distinct clusters, and let $A \subset C$ be a proper subset. Then:*

$$\min_{x \in A,\, y \in C \setminus A} d(x, y) \leq \min_{x \in A,\, z \in C'} d(x, z).$$

Proof. Let C_i^* and C_j^* be any two distinct clusters in the optimal clustering, and let $A \subset C_i^*$, $A' \subseteq C_j^*$. Let $p \in A$ and $p' \in A'$ be the pair realizing the minimum inter-set distance $d(p, p') = \min_{x \in A,\, z \in A'} d(x, z)$. Let $q \in C_i^* \setminus A$ be the point in the rest of C_i^* closest to p, and let c_i^*, c_j^* be the centers of C_i^* and C_j^*, respectively.

By the γ-center proximity property:

$$d(p, p') + d(p', c_j^*) > \gamma \cdot d(p, c_i^*) \tag{1}$$
$$d(p, p') + d(p, c_i^*) > \gamma \cdot d(p', c_j^*) \tag{2}$$
$$d(p, p') + d(p', c_j^*) + d(p, q) > \gamma \cdot (d(q, p) - d(p, c_i^*)) \tag{3}$$

Multiplying first inequality by $1 - \frac{1}{\gamma+1} - \frac{1}{\gamma-1}$, the second one by $\frac{1}{\gamma+1}$, the last one by $\frac{1}{\gamma-1}$ and summing all three inequalities, we get:

$$d(p, p') > \frac{\gamma}{2 - 4\gamma + 1} \cdot d(p, c_i^*) + d(p, q)$$

For $\gamma \geq 2 + \sqrt{3}$, this implies $d(p, p') > d(p, q)$. Therefore:

$$\min_{x \in A,\, y \in C_i^* \setminus A} d(x, y) < \min_{x \in A,\, z \in C_j^*} d(x, z),$$

which concludes the proof. □

In words, the Min-stability property states that for any strict subset $A \subsetneq C$ of a cluster C in the optimal clustering, the point closest to A lies in $C \setminus A$, rather than in any other cluster.

6.3 Algorithms for Perturbation Stable Instances

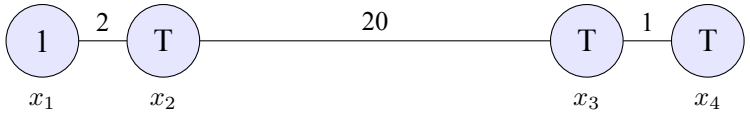
On this section, we present algorithmic frameworks that leverage the structural guarantees provided by perturbation-stable instances to recover optimal clusterings efficiently. Specifically, we show that under appropriate stability conditions—such as γ-perturbation stability for sufficiently large γ—it is possible to design polynomial-time algorithms that solve otherwise intractable clustering problems.

Single-link clustering is a classical and extensively studied hierarchical clustering algorithm. It models the input metric space (X, d) as a complete weighted graph, where vertices represent data points in X, and edge weights correspond to the pairwise distances $d(x, y)$. The algorithm proceeds by executing Kruskal's algorithm to construct a minimum spanning tree (MST) of the graph, but terminates once exactly k connected components have been formed.This process corresponds to halting the MST construction just before the final $k - 1$ edge insertions, thus producing a partition of the dataset into k clusters.

An alternative yet equivalent perspective is to view the algorithm as starting with each data point in its own singleton cluster. At each iteration, the algorithm merges the pair of clusters that are closest together—i.e., those with the smallest inter-cluster distance—until precisely k clusters remain.

At first glance, one might expect that single-link clustering would recover the optimal solution for sufficiently stable instances, as it proceeds by repeatedly merging the pair of clusters with the smallest inter-cluster distance. However, this intuition fails even in relatively simple instances that exhibit strong stability properties. The fundamental limitation lies in the algorithm's disregard for the underlying clustering objective—such as the H_{means} cost—during its execution. Since cluster merges are determined solely based on local pairwise distances, without considering their impact on the overall clustering cost, the resulting partition can be significantly suboptimal with respect to the intended objective function.

To illustrate this limitation, consider a simple instance consisting of a single data point located at $x_1 = 0$, along with three dense clusters of points: $T \gg 1$ data points located at $x_2 = 2$, T data points at $x_3 = 40$, and another T data points at $x_4 = 41$. Suppose the goal is to produce $k = 3$ clusters. In the optimal clustering, the centers are placed at positions x_2, x_3, and x_4, with the isolated point at x_1 assigned to the center at x_2, resulting in a total cost of 2. In contrast, single-link clustering merges the two closest dense clusters—those at x_3 and x_4—and chooses either x_3 or x_4 as the center of the merged cluster . This leads to a significantly higher cost, denoted by T, which can be made arbitrarily large relative to the optimal cost.



An example instance illustrating the failure of single-link clustering on a stable input.

It is important to note that the instance, by construction, satisfies γ-stability for arbitrarily large values of γ, provided that the cluster size parameter T is chosen appropriately.

The issue of single-link Clustering was that it paid no mind to the objective cost function. *Single-link++* is a more sophisticated version of single-link clustering.

6.3.1 Single-link++

Having seen it's predecessor single-link Clustering, it is now easier to understand the motivation behind the following algorithm. *Single-link++* is a clustering algorithm designed for use on γ-stable instances with respect to the H_{means} objective. It is capable of recovering the optimal clustering in polynomial time.

Algorithm 10 Single-link++

Input: Metric space (X, d)
Output: The corresponding clustering

1: Create a complete graph with vertices X and edge weights given by d
2: Run Kruskal's algorithm to compute the minimum spanning tree T of the complete graph
3: Among all $\binom{n-1}{k-1}$ subsets of $k-1$ edges in T, consider the induced k-clusterings (one cluster per connected component)
4: Compute the clustering with the minimum k-median objective value

First, we need a way to verify that our algorithm not only has a way to validate the existence of an

optimal clustering but can also produce it as an output.

Lemma 6.3.1. *Single-link++ recovers the optimal solution of a k-median instance (X, d) if and only if every optimal cluster C_i^* induces a connected subgraph of the minimum spanning tree.*

Proof. The Single-link++ algorithm generates clusterings by removing $k-1$ edges from the minimum spanning tree T, resulting in k connected components. Consequently, any clustering it outputs must consist of clusters that form connected subgraphs of T. Therefore, if some optimal cluster C_i^* does not induce a connected subgraph in T, then Single-link++ cannot recover the optimal solution.

Conversely, any partition of X into k non-empty connected subgraphs of T can be realized by deleting $k-1$ edges from T, specifically the edges that connect points in different clusters. Since the algorithm considers all such possible edge removals and evaluates the clustering cost for each, it is guaranteed to find the optimal solution provided that the optimal clustering corresponds to such a partition. Thus, if the optimal clustering forms connected components in T, the Single-link++ algorithm will correctly identify it. □

We now have a method to distinguish optimal clusterings in our induced MST. We only have to apply our core stability properties on the induced instance to receive the following result:

Theorem 6.3.2. *In every 2-perturbation-stable k-median instance, the single-link++ algorithm recovers the optimal solution (in polynomial time).*

Proof. It is enough to show that the correctness condition in 6.3.1 holds—that is, in every 2-perturbation-stable H_{median} instance, every optimal cluster C_i^* induces a connected subgraph of T. We proceed by contradiction. If not, there is a point x $\in C_i^*$ such that the (unique) c_i-x path in T concludes with the edge (y, x) with y $\notin C_i^*$. At the time (y, x) was added by Kruskal's algorithm, x and c_i were in different connected components (otherwise the addition of (y, x) would have created a cycle). Thus, Kruskal's algorithm also had the option of including the edge (x, c_i) instead. Since the algorithm chose (y, x) over (x, c_i), d(x, y) $\leq$ d(x, c_i). But then x is as close to y $\notin C_i^*$ as its own center, contradicting the weak 2-center proximity property. □

To the extent that we believe that "real-world" clustering instances with "meaningful solutions" are 2-perturbation-stable, 6.3.2 gives a formal sense in which clustering is hard only when it does not matter. It is a largely open research direction to prove robust versions of Theorem 6.3.2, where perturbations can cause a small number of points to switch clusters, while still preserving the optimal clustering of the instance, a property called *approximation stability*.

6.3.2 Single-Linkage with Dynamic Programming

When clustering instances are assumed to be γ-perturbation stable for $\gamma \geq 2+\sqrt{3}$, it becomes possible to recover the optimal solution without exhaustively enumerating all k-clusterings derived from MST cuts. In this section, we describe an algorithmic framework that leverages this structural property, following the approach of Balcan et al [18] and Awasthi et al. [16]. This framework relies on a hierarchical clustering procedure based on Single-Linkage, followed by a dynamic programming (DP) routine that identifies the optimal pruning corresponding to the target number of clusters.

The first step is to construct a hierarchical clustering tree via Single-Linkage. The procedure is summarized in Algorithm 11.

Algorithm 11 Hierarchical Clustering via Single-Linkage

Input: Metric space (X, d)
Output: A hierarchical clustering tree

1: Initialize each point in X as a singleton cluster
2: **While** more than one cluster remains:
   (a) Find clusters C, C' minimizing $d_{\min}(C, C') = \min_{x \in C, y \in C'} d(x, y)$
   (b) Merge C and C' into a single cluster
3: Record each merge as an internal node in a binary tree
4: Return the full tree with original points as leaves

We now show that when the instance satisfies the min-stability property (Lemma 6.2.4), the tree returned by Algorithm 11 contains the optimal clustering as a pruning.

Theorem 6.3.3. *Let $\gamma \geq 2 + \sqrt{3}$, and let $((X, d), \mathcal{H}, k)$ be a γ-stable instance with unique optimal clustering $\{C_1, \ldots, C_k\}$. Then the hierarchical tree produced by Algorithm 11 contains $\{C_1, \ldots, C_k\}$ as a pruning.*

Proof. We prove that at every step of the single-linkage algorithm, the resulting clustering remains *laminar* with respect to the ground-truth clustering $\mathcal{C} = \{C_1, \ldots, C_k\}$. That is, each cluster formed during the execution is either a subset of some $C_r \in \mathcal{C}$, equal to C_r, or a union of such clusters.

Initially, the algorithm begins with n singleton clusters (one for each data point), and this collection is trivially laminar with $\mathcal{C}$, since every singleton is a subset of some C_r.

Assume inductively that at a given step the current clustering is laminar with $\mathcal{C}$. The algorithm chooses to merge the pair of clusters C and C' minimizing the minimum pairwise distance $d_{\min}(C, C')$, and creates a new node in the tree representing their union.

Suppose, without loss of generality, that C is a strict subset of some cluster $C_r \in \mathcal{C}$. By the *min-stability property*, the point in $X \setminus C$ that is closest to any point in C must lie in $C_r \setminus C$. Therefore, if the algorithm chooses to merge C with some C', it must be that $C' \subseteq C_r$ as well. Hence, $C \cup C'$ is still contained in C_r, and the resulting clustering remains laminar with $\mathcal{C}$.

By induction, this property holds throughout the execution of the algorithm. Thus, the final hierarchical clustering forms a tree where each node is either a subset, equal to, or union of clusters in $\mathcal{C}$. □

The existence of such a laminar structure enables an efficient search for the optimal k-clustering via dynamic programming. The DP routine operates on the binary tree and recursively evaluates the best way to partition each subtree.

Algorithm 12 Recovering Optimal Clustering via Tree Pruning

Input: Metric space (X, d), number of clusters k
Output: Optimal k-clustering

1: Run Algorithm 11 to construct a full binary tree T over the dataset
2: Use dynamic programming to find the optimal k-pruning of T using:
   best-k-pruning$(T) = \min_{0<k_0<k} \left\{\text{best-}k_0\text{-pruning}(T_{\text{left}}) + \text{best-}(k-k_0)\text{-pruning}(T_{\text{right}})\right\}$
3: Return the clustering induced by the selected k pruned subtrees

The correctness of the dynamic programming algorithm follows from the recursive structure of the binary tree and the assumption that the overall clustering cost can be computed in terms of the costs

of individual clusters. Specifically, for each node T in the tree and each integer $1 \leq k \leq K$, we define a table entry DP$[T][k]$, representing the optimal cost of partitioning the subtree rooted at T into k clusters. The computation proceeds in a bottom-up fashion. If $k = 1$, the base case simply treats the entire subtree as a single cluster and computes its associated cost, denoted Cost(T), which depends solely on the points within the subtree. If $k > 1$, the optimal k-clustering is obtained by distributing the clusters between the left and right children of node T, denoted T_{left} and T_{right}. For each valid split $k_1 + k_2 = k$, where $k_1, k_2 \geq 1$, the DP value DP$[T][k]$ is computed either as the sum of the two subproblems for sum-based objectives (such as k-median or k-means), or as their maximum in the case of max-based objectives (such as k-center).

Since the tree contains at most $2n$ nodes and up to K clustering options must be evaluated per node, the total number of table entries is $O(nK)$. Each DP entry for $k > 1$ involves considering $O(K)$ ways of splitting k clusters between the two children, resulting in $O(K^2)$ time per node. The base case DP$[T][1] =$ Cost(T) requires evaluating the cost of treating the subtree as a single cluster, which depends on the chosen objective. For instance, in the k-median problem over a finite metric space, this can be done by computing distances to all candidate centers in $O(n^2)$ time; in Euclidean k-means, the optimal center is the mean, allowing the cost to be computed in linear time using precomputed statistics; and in k-center, the cost is the maximum distance to a center, which can also be found in linear time. As a result, the total runtime of the algorithm is

$$O(nK^2 + nT(n)) = O(n(K^2 + T(n))),$$

where $T(n)$ denotes the time required to compute the cost of clustering any subtree of size n as a single cluster. This completes the description of the algorithm's correctness and computational efficiency.

CHAPTER 7

Metric Distortion on Stable Instances

In this chapter we study how to obtain *constant-distortion* solutions under γ-perturbation stability. As discussed in Chapter 6, stability lets us bypass worst-case hardness by exposing combinatorial structure in the metric. Our approach is to exploit this structure to build a *small candidate set* that is guaranteed to contain an optimal clustering, and then solve only on that reduced ground set.

Concretely, we identify (using only ordinal information) a *hierarchical decomposition* of well-separated groups that forms a *laminar family*. We then define the *frontier*: the deepest nodes that can still be reached with a budget of k centers. Intuitively, each node represents a feasible "cluster at some resolution," and the frontier captures the smallest clusters that could still appear in an optimal k-clustering. A key structural bound we prove is that the frontier size depends only on k (and not on n); in fact, it is at most 2^{k-1}. This gives us a compact candidate set to work with.

We instantiate this framework in two models:

1-Dimensional case. Following the model of [47], we use the frontier to form a reduced 1-D instance and run a Hassin–Tamir style dynamic program [56]. Crucially, we query distances only *along frontier nodes*, yielding a constant-distortion solution with just $O(2^k)$ distance queries.

General metric case. In the model of [29] (where candidates and agents coincide), we first pick one representative per frontier node using only ordinal information, which yields a $(2^{k-1}, \alpha)$-bicriteria solution for a constant α (we instantiate $\alpha = 3$). We then *restrict the instance to these* 2^{k-1} *centers*, query only their pairwise distances, and run an off-the-shelf k-median approximation on this β-point metric (e.g., the 2.613 bi-point rounding of [54]). This standard "bicriteria $\Rightarrow$ true" reduction converts the bicriteria solution into a *true* k-solution with constant distortion, using $O(4^k)$ distance queries—independent of n.

We begin with the necessary definitions and preliminaries (well-separated groups, laminarity), describe the frontier construction and its size bound, and then present the two instantiations above together with their query and distortion guarantees. We first start with some definitions and preliminaries that are necessarry for our latter theorems

Definition 7.0.1 (Well-Separated Groups)**.** *A set of clusters $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ in a metric space (X, d) is said to be* well-separated *if for every cluster $C_i \in \mathcal{C}$, the following holds:*

$$\operatorname{diam}(C_i) < \min_{x \in C_i, y \notin C_i} d(x, y),$$

where the diameter of a cluster is defined as

$$\text{diam}(C_i) = \max_{x,y \in C_i} d(x, y).$$

In other words, each cluster is tighter (smaller in diameter) than its closest distance to any point outside the cluster.

Lemma 7.0.2. *Let (X, d) be a metric space, and suppose the instance is γ-perturbation stable with $\gamma \geq 2 + \sqrt{3}$. Then, the clusters $\mathcal{C}^* = \{C_1^*, C_2^*, \ldots, C_k^*\}$ of the optimal clustering solution form well-separated groups.*

Proof. Let $\mathcal{C}^* = \{C_1^*, C_2^*, \ldots, C_k^*\}$ be the unique optimal clustering of the γ-perturbation stable instance with $\gamma \geq 2 + \sqrt{3}$. Fix any cluster C_i^*, and let $x, x' \in C_i^*$ and $y \notin C_i^*$.

By Lemma 6.2.3, for any $x, x' \in C_i^*$ and $y \in C_j^*$ with $j \neq i$, we have:

$$d(x, y) > \frac{(\gamma - 1)^2}{2\gamma} \cdot d(x, x').$$

Taking the maximum over all pairs in C_i^* for the right-hand side and the minimum over all $x \in C_i^*, y \notin C_i^*$ for the left-hand side, we obtain:

$$\min_{x \in C_i^*,\, y \notin C_i^*} d(x, y) > \frac{(\gamma - 1)^2}{2\gamma} \cdot \text{diam}(C_i^*).$$

When $\gamma \geq 2 + \sqrt{3}$, it can be verified that

$$\frac{(\gamma - 1)^2}{2\gamma} > 1.$$

Hence,

$$\min_{x \in C_i^*,\, y \notin C_i^*} d(x, y) > \text{diam}(C_i^*),$$

which shows that each cluster is well-separated from the rest of the dataset. Thus, the optimal clusters form well-separated groups. □

Definition 7.0.3 (Laminar Family)**.** *A laminar family on an underlying set X is a collection $\mathcal{F}$ of non-empty subsets of X such that for any pair of sets $S, S' \in \mathcal{F}$, one of the following holds:*

- *$S \subseteq S'$,*
- *$S' \subseteq S$, or*
- *$S \cap S' = \emptyset$.*

Claim 7.0.4. *Well-separated groups in a general metric space (X, d) form a laminar family.*

Proof. A subset $G \subseteq X$ is *well-separated* if and only if for every two distinct candidates $a, b \in G$ and every candidate $c \in X \setminus G$:

$$d(a, b) < d(a, c) \quad \text{and} \quad d(a, b) < d(b, c).$$

Suppose for the sake of contradiction that there exist two well-separated subsets $A, B \subseteq X$ which violate laminarity. This means that:

$$A \not\subseteq B, \quad B \not\subseteq A, \quad \text{and} \quad A \cap B \neq \emptyset.$$

Then, there exist candidates:

$$x \in A \cap B, \quad a \in A \setminus B, \quad b \in B \setminus A.$$

Since A is well-separated, it follows that:

$$d(x, a) < d(x, b).$$

Similarly, since B is well-separated, we must have:

$$d(x, b) < d(x, a).$$

These two inequalities clearly contradict each other. Thus, our initial assumption must be false, and there can be no two well-separated groups violating laminarity.

Therefore, the set of well-separated groups forms a laminar family. □

We next describe an algorithm that identifies well-separated groups using only the ordinal preferences of the candidates. Although our 1-Dimensional model assumes access to voters' rankings over candidates rather than candidates' rankings, this limitation can be bypassed using the same technique employed in Chapter 4 during the analysis of the Distant-Candidate Algorithm (Algorithm 5). In contrast, in the model of [29], the sets of voters and candidates coincide, and thus this issue does not arise.

Algorithm 13 Find well-separated groups from ordinal preferences

Input: Set of candidates $X = \{c_1, \ldots, c_n\}$, each with a strict ranking over $X \setminus \{c_i\}$
Output: Family $\mathcal{S}$ of well-separated groups

```
Initialize S ← ∅
for each a ∈ X do
    Let Pref_a = [p_1, ..., p_{n-1}] {a's preference list}
    for k ← 1 to n − 1 do
        Let G ← {a} ∪ {p_1, ..., p_k}
        Let valid ← true
        for each x ∈ G do
            Let r = min{rank_x(y) | y ∉ G}
            if r ≤ k then
                valid ← false
                break
            endif
        end for
        if valid then
            S ← S ∪ {G}
        endif
    end for
end for
Remove duplicates from S
return S
```

Algorithm 13 generates at most $O(n^2)$ candidate groups, since for each of the n candidates, it considers up to n top-k prefixes. Verifying whether a given group is well-separated requires $O(n^2)$

time in the worst case, as it involves checking the preference list of each member of the group against all other candidates.

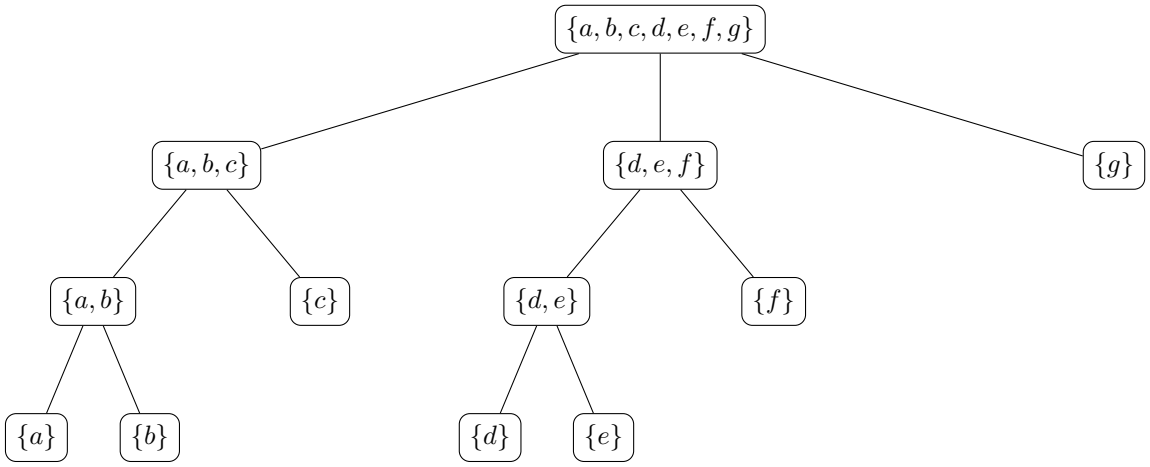
Therefore, the total complexity of the algorithm is $O(n^4)$, and it operates using only ordinal information.

Having computed the family of well-separated groups using only ordinal preferences via Algorithm 13, we now leverage their laminar structure (Claim 7.0.4) to organize them into a hierarchical tree. This tree captures the containment relationships among the well-separated groups and serves as a foundation for further algorithmic processing.

Definition 7.0.5 (Hierarchy Tree of Well-Separated Groups)**.** *Given a laminar family $\mathcal{S}$ of well-separated groups containing the full set X, we define a rooted tree $\mathcal{T}$ whose nodes correspond to the sets in $\mathcal{S}$, and where an edge from node G to node G' is present if and only if:*

- *$G \subset G'$, and*
- *there exists no $H \in \mathcal{S}$ such that $G \subset H \subset G'$.*

This construction naturally induces a tree rooted at the node corresponding to the full set X, where edges represent immediate containment between well-separated groups. Crucially, this hierarchical structure reflects the laminar nature of the family: any two groups are either nested or disjoint. Figure 7.1 illustrates an example of such a hierarchy built from a laminar family of well-separated groups.



Hierarchical tree corresponding to the laminar family of well-separated groups: $\big\{\{a\}, \{b\}\{a, b\}, \{c\}, \{a, b, c\}, \{d\}, \{e\}\{d, e\}, \{f\}, \{d, e, f\}, \{g\}, \{a, b, c, d, e, f, g\}\big\}$.

Claim 7.0.6. *For γ-stable instances with $\gamma \geq 2 + \sqrt{3}$, the optimal clusters correspond to nodes in the hierarchical tree described above.*

Proof. For such instances, all optimal clusters are well-separated. Algorithm 13 enumerates all possible groups that may form a well-separated group and identifies those that do. Consequently, any optimal cluster must be included in the output of the algorithm and thus appears as a node in the hierarchical tree constructed as described above. □

Now that we have established that each optimal cluster corresponds to a node in the hierarchical tree, our next goal is to understand how these nodes relate to one another within the tree. In particular,

we seek structural constraints that govern which combinations of nodes can be part of a valid clustering. To that end, we prove the following claim, which describes how optimal clusters may (or may not) overlap within the tree.

Claim 7.0.7. *Let G be a node in the hierarchical tree of well-separated groups with children $G_1, \ldots, G_m$. Then, in an optimal clustering solution:*

- *either all of G (i.e., all of its leaves) are clustered together in a single optimal cluster corresponding to G or one of its ancestors,*
- *or each optimal cluster intersects candidates from at most one child G_i.*

Proof. Since the instance is γ-stable with $\gamma \geq 2 + \sqrt{3}$, all optimal clusters are well-separated and, by construction, correspond to nodes in the hierarchical tree produced by Algorithm 13.

Let G be a node in the tree with children $G_1, G_2, \ldots, G_m$. Suppose, for contradiction, that there exists an optimal cluster C^* that contains candidates from more than one child — say, both G_i and G_j with $i \neq j$ — but not the entire set G, and not any of its ancestors.

We make three key observations:

1. Path coverage. Since the optimal clustering forms a partition of the ground set X, each candidate must belong to exactly one cluster. In the tree, this means that for every leaf (candidate), there is exactly one selected node along its path to the root that covers it. Thus, optimal clusters correspond to non-overlapping nodes in the tree that together cover all leaves.

2. Laminarity contradiction. Since $H = C^*$ intersects both G_i and G_j, and G_i, G_j are disjoint children of G, the laminar property implies that H must be a superset of both. Hence, $G_i \cup G_j \subseteq H$.

Now, if $H \subset G$, then H is strictly between G and its children in the tree — i.e., it lies *between* G and its descendants. In that case, the algorithm should have discovered H as a well-separated group during its bottom-up construction and added it to the tree. This contradicts the assumption that G_i and G_j are direct children of G, since they would have instead appeared as descendants of H.

On the other hand, if $H \supseteq G$, then H must be an ancestor of G, again contradicting the assumption that C^* is not equal to G or one of its ancestors.

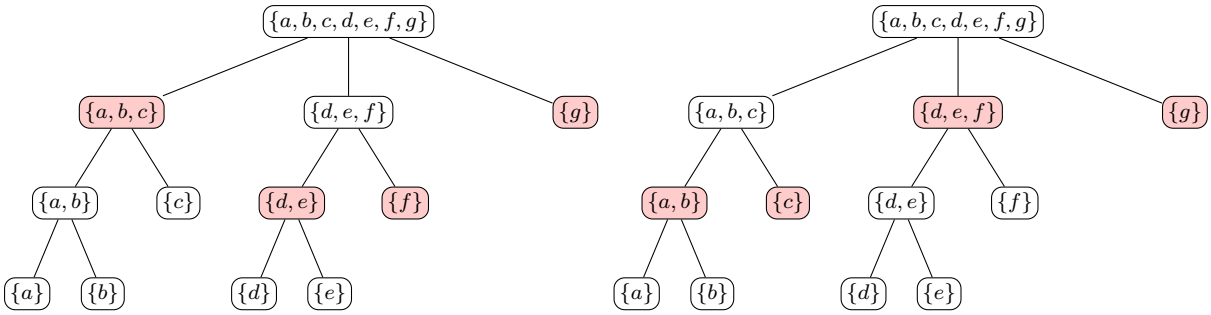
In either case, we reach a contradiction: the structure of the tree is inconsistent with the existence of a cluster C^* intersecting multiple children of G without fully containing G.

Therefore, in the optimal clustering, for any internal node G in the tree, one of the following must hold:

- All of G is assigned to a single optimal cluster corresponding to G or one of its ancestors, or
- Each optimal cluster intersects at most one of the children G_i, with no cluster spanning across multiple children.

This concludes the proof. See figure 7.2 below for examples □

This structural constraint implies that any optimal clustering corresponds to a selection of nodes in the tree that are pairwise non-overlapping (i.e., no node is an ancestor or descendant of another) and together cover all the leaves. This tree-based perspective allows us to significantly narrow down the search space for optimal clusters: rather than considering all possible subsets of the input, we can restrict our attention to certain nodes of the hierarchical decomposition tree. Our goal is to identify a small collection of such nodes that is guaranteed to contain all optimal clusters. In other words, we



(a) Highlighting $\{g\}$, $\{a, b, c\}$, $\{d, e\}$, $\{f\}$. **(b)** Highlighting $\{g\}$, $\{d, e, f\}$, $\{a, b\}$, $\{c\}$.

Two example selections of $k = 4$ clusters (four red nodes) on the hierarchical tree from Figure 7.1. Each subfigure highlights a valid clustering: exactly four non-overlapping nodes covering all leaves.

aim to construct a set of at most ℓ candidate clusters such that every optimal cluster intersects at most one of them — that is, an $(\ell, 1)$-bicriteria solution.

To achieve this, we ask the following question: *How many nodes in the hierarchical tree could possibly correspond to optimal clusters?* Intuitively, if the number of leaves n is much larger than the number of clusters k, then optimal clusters must be relatively "high" in the tree. Selecting clusters that are too deep in the tree would leave too many leaves uncovered, violating the requirement that the clustering covers all candidates using only k disjoint groups. We formalize this notion by introducing

the concept of the *frontier*, which intuitively captures the set of nodes that might serve as candidates for optimal clusters under a size constraint.

Definition 7.0.8. *Let T be a hierarchical decomposition tree. The* frontier *is the set of nodes in T that could potentially be selected as part of an optimal clustering of size at most k. Its size serves as an upper bound on the number of candidate clusters that must be considered.*

We now describe a recursive procedure for identifying the *frontier* — the set of nodes in the hierarchical clustering tree that are the deepest reachable under a clustering budget of k centers.

Intuitively, the procedure performs a depth-first traversal from the root of the tree. At each internal node with multiple children, we are allowed to use part of our budget to cover all but one of the children, and recursively explore the remaining child. This models the idea that to go deeper into the tree (i.e., to cluster smaller groups), we must "pay" at each branching point by assigning clusters to the sibling groups we choose not to explore.

Formally, the algorithm proceeds as follows:

Algorithm 14 Depth-first traversal to compute the frontier

Input: Rooted tree (V, E) with root r, integer k (number of centers)
Output: Frontier set $F \subseteq V$

```
Initialize F ← ∅
procedure DFS(v, t)
    Let children ← Children(v)
    Let m ← |children|
    if m = 0 or t < m then
        F ← F ∪ {v}      {add v to frontier}
        return
    end if
    for each u ∈ children do
        DFS(u, t − (m − 1))
    end for
end procedure
DFS(r, k)
return F
```

To analyze the size of the frontier computed by Algorithm 14, we define a recursive function that models its behavior. Specifically, for any node j in the tree and budget k, let $f_j(k)$ denote the number of nodes in the frontier of the subtree rooted at j when at most k centers are available. This function captures the structure of the recursive exploration performed by the algorithm: at each internal node j, we must allocate $k - |C_j| + 1$ centers to each of the $|C_j|$ children if we choose to explore them. If the budget is insufficient to do so, the node is added to the frontier and recursion terminates at that point. The following theorem formalizes this recurrence relation and shows that it accurately reflects the output of the algorithm.

Claim 7.0.9. *Let $f_j(k)$ denote the size of the frontier of the subtree rooted at node j when a clustering budget of at most k centers is available. Then, the function $f_j(k)$ computed by the depth-first traversal algorithm described in Algorithm 14 satisfies the following recurrence:*

$$f_j(k) = \begin{cases} \sum_{i \in C_j} f_i\left(k - |C_j| + 1\right) & \text{if } k \geq |C_j| \\ 1 & \text{otherwise} \end{cases}$$

where C_j denotes the set of children of node j, and $|C_j|$ its cardinality.

Although the algorithm proceeds in a top-down recursive manner, our proof will use structural induction on the tree, which is conceptually bottom-up. This allows us to reason about the correctness of the recurrence relation by assuming it holds for the children of a node and verifying it for the parent. The direction of the induction does not affect the validity of the proof, since it is a mathematical argument about the values computed by the algorithm, rather than the control flow itself. Note that although the algorithm only recurses into the children of a node j when $k \geq |C_j|$, the structural induction assumes that the recurrence holds for all subtrees in the tree structure, independently of whether they are visited in a particular run of the algorithm. In the inductive step, we only apply the hypothesis to the children that the algorithm actually recurses into, ensuring that the analysis precisely matches the algorithm's behavior.

Proof. We prove the theorem by structural induction on the subtree rooted at node j.

Base Case: Suppose node j is a leaf, i.e., it has no children ($|C_j| = 0$). Then the algorithm directly adds j to the frontier (line 6 of Algorithm 14), since the condition $m = 0$ is satisfied. No recursive calls are made. Therefore, the size of the frontier of the subtree rooted at j is exactly $f_j(k) = 1$ for all $k \geq 0$, which matches the second case of the recurrence.

Inductive Step: Suppose the recurrence holds for all subtrees rooted at the children of a node j, and let $|C_j|$ be the number of children of node j.

- If $k < |C_j|$, then the clustering budget is insufficient to cover $|C_j| - 1$ sibling groups and explore the remaining child. According to Algorithm 14, this triggers the base case (line 5), and node j is added directly to the frontier. No recursion is performed. Thus, $f_j(k) = 1$, which corresponds to the second case of the recurrence.

- If $k \geq |C_j|$, then the algorithm proceeds to explore all children $i \in C_j$, deducting $|C_j| - 1$ centers from the budget to conceptually cover all but one child at each branching point. Each recursive call to a child i receives the reduced budget $k' = k - |C_j| + 1$. By the inductive hypothesis, the size of the frontier in each child subtree is correctly given by $f_i(k - |C_j| + 1)$. Therefore, the total frontier size for the subtree rooted at j is:

$$f_j(k) = \sum_{i \in C_j} f_i(k - |C_j| + 1)$$

  which matches the first case of the recurrence.

By structural induction, we conclude that the size of the frontier computed by Algorithm 14 coincides with the value of $f_j(k)$ as given by the recurrence.

□

Next we define a global recursvive function $F(k)$ independent of the structure of the hierarhical tree that upper bounds the value of $f_j(k)$ for every node j of the hierarchical tree and any positive value of k.

Lemma 7.0.10. *Let $f_j(k)$ denote the size of the frontier of the subtree rooted at node j under a clustering budget of k, as defined in Claim 7.0.9.*

Define the function $F : \mathbb{N}^+ \to \mathbb{N}$ recursively by

$$F(k) = \begin{cases} 1 & \text{if } k < 2, \\ \max\limits_{2 \leq j \leq k} \{j \cdot F(k - j + 1)\} & \text{if } k \geq 2. \end{cases}$$

Then for any hierarchical decomposition tree and any node j, we have:

$$f_j(k) \leq F(k) \quad \text{for all } k \in \mathbb{N}^+.$$

Proof. Let T be an arbitrary rooted tree. For each node j in T, let C_j denote its set of children, with $|C_j|$ denoting the number of children.

Recall that $f_j(k)$ is defined recursively by:

$$f_j(k) = \begin{cases} \sum\limits_{i \in C_j} f_i(k - |C_j| + 1) & \text{if } k \geq |C_j|, \\ 1 & \text{otherwise.} \end{cases}$$

We prove the desired bound by induction on k.

Base Case: $k < 2$.
For any node j, we distinguish two cases: If $k < |C_j|$, then by definition $f_j(k) = 1 = F(k)$. If $k \geq |C_j|$, then since $|C_j| \geq 2$, this case cannot occur for $k < 2$.

Hence, $f_j(k) \leq F(k)$ holds for all $k < 2$.

Inductive Step: Suppose the inequality holds for all smaller values of k, i.e., for all $k' < k$ and all nodes j, we have $f_j(k') \leq F(k')$. We prove it for k.

Let j be an arbitrary node.

If $k < |C_j|$, then again $f_j(k) = 1 \leq F(k)$ by definition. If $k \geq |C_j|$, then by the recursive definition:

$$f_j(k) = \sum_{i \in C_j} f_i(k - |C_j| + 1).$$

By the inductive hypothesis, for each child $i \in C_j$,

$$f_i(k - |C_j| + 1) \leq F(k - |C_j| + 1).$$

Therefore,

$$f_j(k) \leq |C_j| \cdot F(k - |C_j| + 1).$$

By the definition of $F(k)$, we have:

$$F(k) = \max_{2 \leq j' \leq k} \left\{ j' \cdot F(k - j' + 1) \right\} \geq |C_j| \cdot F(k - |C_j| + 1).$$

Hence, $f_j(k) \leq F(k)$, as required.

Conclusion: By induction on k, the inequality holds for all $k \in \mathbb{N}^+$ and all nodes j in the tree. □

Although the function $F(k)$ provides a tree-independent upper bound on the size of the frontier, it is defined recursively and remains somewhat opaque. To better understand the asymptotic behavior of the algorithm, we now analyze the growth of $F(k)$ directly. In particular, we show that $F(k)$ grows at most exponentially with k, by proving that $F(k) \leq 2^k$ for all $k \in \mathbb{N}^+$. This explicit bound will allow us to derive a clean worst-case guarantee on the frontier size and will play a central role in the proof of our main result.

Lemma 7.0.11. *Let $F : \mathbb{N} \to \mathbb{N}$ be the function defined recursively by:*

$$F(k) = \begin{cases} 1 & \text{if } k < 2, \\ \max\limits_{2 \leq j \leq k} \{ j \cdot F(k - j + 1) \} & \text{if } k \geq 2. \end{cases}$$

Then, for all $k \geq 1$, we have:

$$F(k) \leq 2^{k-1}.$$

Proof. We prove by induction on $k \geq 1$ that

$$F(k) \leq 2^{k-1}.$$

Base case: For $k = 1$, we have

$$F(1) = 1 = 2^0,$$

so the base case holds.

Inductive hypothesis: Suppose that for all integers $k' < k$, we have

$$F(k') \leq 2^{k'-1}.$$

Inductive step: From the recurrence definition of $F(k)$,

$$F(k) = \max_{2 \leq j \leq k} \{j \cdot F(k-j+1)\}.$$

For each such j, we have $k' = k - j + 1 < k$, so by the inductive hypothesis:

$$F(k-j+1) \leq 2^{k-j}.$$

Thus,

$$F(k) \leq \max_{2 \leq j \leq k} \left\{ j \cdot 2^{k-j} \right\}.$$

Now define $g_k(j) := j \cdot 2^{k-j}$, and we want to show:

$$g_k(j) \leq 2^{k-1} \quad \text{for all } j \in [2, k].$$

We rewrite:

$$g_k(j) = 2^{k-1} \cdot \underbrace{j \cdot 2^{-j+1}}_{h(j)}.$$

To analyze the behavior of $g_k(j)$, we fix k and study its maximum over j. We define

$$h(j) := j \cdot 2^{-j+1},$$

and observe that this function controls the shape of $g_k(j)$. To show that $g_k(j)$ is maximized at $j = 2$, we analyze the monotonicity of $h(j)$ via its derivative.

This derivative is negative for $j > \frac{1}{\ln 2} \approx 1.44$, so $h(j)$ is strictly decreasing for $j \geq 2$. Thus, its maximum on $j \in [2, k]$ is attained at $j = 2$, giving:

$$h(j) \leq h(2) = 2 \cdot 2^{-1} = 1.$$

Therefore,

$$g_k(j) = 2^{k-1} \cdot h(j) \leq 2^{k-1},$$

and hence:

$$F(k) \leq \max_j g_k(j) \leq 2^{k-1}.$$

Conclusion: By induction, $F(k) \leq 2^{k-1}$ for all $k \geq 1$. □

Since the set of optimal clusters is guaranteed to be a subset of the frontier, the *Frontier Set* F forms a collection of at most 2^{k-1} clusters that contains the optimal ones. Thus, by selecting the median candidate from each cluster in F, we obtain a $(2^{k-1}, 3)$-bicriteria solution. Applying the dynamic programming algorithm of Hassin and Tamir [56] on the reduced candidate-restricted instance yields the following result:

Theorem 7.0.12. *There exists a polynomial-time deterministic algorithm for k-committee election in the one-dimensional Euclidean space that, under γ-perturbation stability with $\gamma \geq 2 + \sqrt{3}$, uses at most $O(2^{k-1})$ distance queries and achieves a distortion of at most 7.*

Proof. The distortion guarantee follows from Theorem 4.3.2, since the selected set of medians constitutes a $(2^{k-1}, 3)$-bicriteria solution. To reconstruct the candidate-restricted instance, it suffices to perform $2^{k-1} - 1$ distance queries to locate the medians of the clusters in F, on which the dynamic programming algorithm is then applied. □

Having established our main result in the one-dimensional Euclidean setting, it is natural to ask whether the frontier-based approach can also be applied in arbitrary metric spaces. While the lack of geometric structure prevents us from directly using the dynamic programming method, the crucial property that the frontier contains all optimal clusters still holds. In this setting, we follow the model and techniques of [29], where the sets of candidates and voters coincide. This framework suggests that by selecting suitable representatives from the frontier and subsequently applying known approximation techniques for the k-median problem, we can still achieve constant-distortion guarantees with a small number of distance queries. In the following, we outline how this adaptation can be carried out and present the resulting bounds on query complexity and distortion.

Lemma 7.0.13. *There exists a polynomial-time deterministic algorithm for k-committee election in general metric spaces where the sets of candidates and voters coincide, that under γ-perturbation stability with $\gamma \geq 2 + \sqrt{3}$ uses at most $O(4^k)$ distance queries and achieves a constant distortion of at most $3 \cdot \alpha$, where α is the approximation factor of the underlying k-median subroutine (e.g., $\alpha = 2.613$ from [54]).*

Proof. By Lemma 7.0.11, the frontier computed by Algorithm 14 contains all possible optimal clusters, and its size is at most 2^{k-1}. We can therefore select a single representative from each frontier cluster. Using the single-winner rules of either [50] or [60], each representative can be chosen so that it is within a factor of 3 of the optimal choice for that cluster. This yields a $(2^{k-1}, 3)$-bicriteria solution, which we denote by $C' \subseteq X$.

Following [29], we interpret C' as a *multiset*: for every point $x \in X$, we add to C' the representative $c \in C'$ assigned to x in the bicriteria solution. Thus, the multiplicity of each $c \in C'$ corresponds exactly to the number of points assigned to it. This interpretation produces a reduced k-median instance defined over C' in which all distances are inherited from the original metric space.

On this reduced instance of size at most 2^{k-1}, we run a constant-factor approximation algorithm for k-median, such as the 2.613 bi-point rounding of [54]. This converts the $(2^{k-1}, 3)$-bicriteria solution into a true k-median solution with constant distortion.

The approximation algorithm requires access to the pairwise distances between the points of C'. Since $|C'| \leq 2^{k-1}$, we need at most $\binom{2^{k-1}}{2} = O(4^k)$ distance queries. Therefore, the overall algorithm achieves constant distortion while using only $O(4^k)$ queries, completing the proof. □

Conclusion and Open Questions In both the one-dimensional and general metric settings, perturbation stability allowed us to design algorithms whose query complexity is independent of the number of candidates n and depends only on the committee size k. This stands in sharp contrast to the worst-case setting, where the number of required queries typically scales with n, and highlights the efficiency gains attainable under γ-perturbation stability. Several natural questions remain open. A first direction is to establish lower bounds on the query complexity under perturbation stability. Another, raised in [48], is to determine bounds and algorithms whose complexity depends on both n and k but remains significantly better than in the general case, thereby enabling a principled choice between n-dependent and n-independent approaches based on the values of n and k. Finally, an intriguing avenue is to explore learning-augmented algorithms that leverage predictions to further reduce query complexity while preserving robustness guarantees.

Bibliography

[1] B. Abramowitz, E. Anshelevich, and W. Zhu, "Awareness of voter passion greatly improves the distortion of metric social choice," in *Web and Internet Economics - 15th International Conference, WINE 2019, New York, NY, USA, December 10-12, 2019, Proceedings*, ser. Lecture Notes in Computer Science, I. Caragiannis, V. S. Mirrokni, and E. Nikolova, Eds., vol. 11920. Springer, 2019, pp. 3–16. [Online]. Available: https://doi.org/10.1007/978-3-030-35389-6_1

[2] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, "Np-hardness of euclidean sum-of-squares clustering," *Machine Learning*, vol. 75, pp. 245–248, 05 2009.

[3] G. Amanatidis, G. Birmpas, A. Filos-Ratsikas, and A. A. Voudouris, "Peeking behind the ordinal curtain: Improving distortion via cardinal queries," *Artif. Intell.*, vol. 296, p. 103488, 2021. [Online]. Available: https://doi.org/10.1016/j.artint.2021.103488

[4] ——, "Don't roll the dice, ask twice: The two-query distortion of matching problems and beyond," in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022. [Online]. Available: http://papers.nips.cc/paper_files/paper/2022/hash/c5ec22711f3a4a2f4a0a8ffd92167190-Abstract-Conference.html

[5] ——, "A few queries go a long way: Information-distortion tradeoffs in matching," *J. Artif. Intell. Res.*, vol. 74, 2022. [Online]. Available: https://doi.org/10.1613/jair.1.12690

[6] I. Anagnostides, D. Fotakis, and P. Patsilinakos, "Metric-distortion bounds under limited information," *J. Artif. Intell. Res.*, vol. 74, pp. 1449–1483, 2022. [Online]. Available: https://doi.org/10.1613/jair.1.13338

[7] H. Angelidakis, K. Makarychev, and Y. Makarychev, "Algorithms for stable and perturbation-resilient problems," in *Proc. of the 49th ACM Symposium on Theory of Computing (STOC 2017)*, 2017, pp. 438–451.

[8] E. Anshelevich, O. Bhardwaj, E. Elkind, J. Postl, and P. Skowron, "Approximating optimal social choice under metric preferences," *Artif. Intell.*, vol. 264, pp. 27–51, 2018. [Online]. Available: https://doi.org/10.1016/j.artint.2018.07.006

[9] E. Anshelevich, A. Filos-Ratsikas, N. Shah, and A. A. Voudouris, "Distortion in social choice problems: The first 15 years and beyond," *CoRR*, vol. abs/2103.00911, 2021. [Online]. Available: https://arxiv.org/abs/2103.00911

[10] E. Anshelevich, A. Filos-Ratsikas, and A. A. Voudouris, "The distortion of distributed metric social choice," *CoRR*, vol. abs/2107.05456, 2021. [Online]. Available: https://arxiv.org/abs/2107.05456

[11] E. Anshelevich and J. Postl, "Randomized social choice functions under metric preferences," *J. Artif. Intell. Res.*, vol. 58, pp. 797–827, 2017. [Online]. Available: https://doi.org/10.1613/jair.5340

[12] S. Arora, P. Raghavan, and S. Rao, "Approximation schemes for euclidean," 07 2000.

[13] K. Arrow, *Advances in the Spatial Theory of Voting*, J. M. Enelow and M. J. Hinich, Eds. Cambridge University Press, 1990.

[14] K. J. Arrow, *Social Choice and Individual Values*. Yale University Press, 2012. [Online]. Available: http://www.jstor.org/stable/j.ctt1nqb90

[15] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '07. USA: Society for Industrial and Applied Mathematics, 2007, p. 1027–1035.

[16] P. Awasthi, A. Blum, and O. Sheffet, "Center-based clustering under perturbation stability," *Inf. Process. Lett.*, vol. 112, no. 1-2, pp. 49–54, 2012.

[17] M.-F. Balcan, A. Blum, and A. Gupta, "Clustering under approximation stability," *Journal of the ACM*, vol. 60, no. 2, 2013.

[18] M. Balcan, A. Blum, and S. S. Vempala, "A discriminative framework for clustering via similarity functions," in *Proceedings of the 40th Annual ACM Symposium on Theory of Computing, Victoria, British Columbia, Canada, May 17-20, 2008*, C. Dwork, Ed. ACM, 2008, pp. 671–680. [Online]. Available: https://doi.org/10.1145/1374376.1374474

[19] M. Balcan and Y. Liang, "Clustering under perturbation resilience," *SIAM Journal on Computing*, vol. 45, no. 1, pp. 102–155, 2016.

[20] S. Ben-David and L. Reyzin, "Data stability in clustering: A closer look," 2014. [Online]. Available: https://arxiv.org/abs/1107.2379

[21] Y. Bilu and N. Linial, "Are Stable Instances Easy?" in *Proc. of the 1st Symposium on Innovations in Computer Science (ICS 2010)*. Tsinghua University Press, 2010, pp. 332–341.

[22] Y. Bilu, A. Daniely, N. Linial, and M. E. Saks, "On the practically interesting instances of MAX-CUT," in *Proceedings of the 30th International Symposium on Theoretical Aspects of Computer Science (STACS 2013)*, ser. LIPIcs, N. Portier and T. Wilke, Eds., vol. 20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2013, pp. 526–537.

[23] A. Borodin and R. El-Yaniv, *Online computation and competitive analysis*. Cambridge University Press, 1998.

[24] A. Borodin, D. Halpern, M. Latifian, and N. Shah, "Distortion in voting with top-t preferences," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, L. D. Raedt, Ed. International Joint Conferences on Artificial Intelligence Organization, 7 2022, pp. 116–122, main Track. [Online]. Available: https://doi.org/10.24963/ijcai.2022/17

[25] ——, "Distortion in voting with top-t preferences," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, ser. Proceedings of the International Joint Conference on Artificial Intelligence, L. De Raedt, Ed. IJCAI Organization, Jul. 2022, pp. 116–122, the 31st International Joint Conference on Artificial Intelligence and the 25th European Conference on Artificial Intelligence, IJCAI-ECAI 2022 ; Conference date: 23-07-2022 Through 29-07-2022. [Online]. Available: https://ijcai-22.org/

[26] C. Boutilier, I. Caragiannis, S. Haber, T. Lu, A. D. Procaccia, and O. Sheffet, "Optimal social choice functions: A utilitarian view," *Artif. Intell.*, vol. 227, pp. 190–213, 2015. [Online]. Available: https://doi.org/10.1016/j.artint.2015.06.003

[27] ——, "Optimal social choice functions: A utilitarian view," *Artif. Intell.*, vol. 227, pp. 190–213, 2015. [Online]. Available: https://doi.org/10.1016/j.artint.2015.06.003

[28] F. Brandt, V. Conitzer, U. Endriss, J. Lang, and A. D. Procaccia, Eds., *Handbook of Computational Social Choice*. Cambridge University Press, 2016. [Online]. Available: https://doi.org/10.1017/CBO9781107446984

[29] J. Burkhardt, I. Caragiannis, K. Fehrs, M. Russo, C. Schwiegelshohn, and S. Shyam, "Low-distortion clustering with ordinal and limited cardinal information," 2024. [Online]. Available: https://arxiv.org/abs/2402.04035

[30] I. Caragiannis, S. Nath, A. D. Procaccia, and N. Shah, "Subset selection via implicit utilitarian voting," *J. Artif. Intell. Res.*, vol. 58, pp. 123–152, 2017. [Online]. Available: https://doi.org/10.1613/jair.5282

[31] ——, "Subset selection via implicit utilitarian voting," *J. Artif. Intell. Res.*, vol. 58, pp. 123–152, 2017. [Online]. Available: https://doi.org/10.1613/jair.5282

[32] I. Caragiannis and A. D. Procaccia, "Voting almost maximizes social welfare despite limited communication," *Artif. Intell.*, vol. 175, no. 9-10, pp. 1655–1671, 2011. [Online]. Available: https://doi.org/10.1016/j.artint.2011.03.005

[33] I. Caragiannis, N. Shah, and A. A. Voudouris, "The metric distortion of multiwinner voting," *Artif. Intell.*, vol. 313, p. 103802, 2022. [Online]. Available: https://doi.org/10.1016/j.artint.2022.103802

[34] J. R. Chamberlin and P. N. Courant, "Representative deliberations and representative decisions: Proportional representation and the borda rule," *American Political Science Review*, vol. 77, no. 3, p. 718–733, 1983.

[35] M. Charikar and P. Ramakrishnan, "Metric distortion bounds for randomized social choice," in *Proceedings of the 2022 ACM-SIAM Symposium on Discrete Algorithms, SODA 2022, Virtual Conference / Alexandria, VA, USA, January 9 - 12, 2022*, J. S. Naor and N. Buchbinder, Eds. SIAM, 2022, pp. 2986–3004. [Online]. Available: https://doi.org/10.1137/1.9781611977073.116

[36] M. Charikar, K. Wang, P. Ramakrishnan, and H. Wu, "Breaking the metric voting distortion barrier," in *Proceedings of the 2024 ACM-SIAM Symposium on Discrete Algorithms, SODA 2024, Alexandria, VA, USA, January 7-10, 2024*, D. P. Woodruff, Ed. SIAM, 2024, pp. 1621–1640. [Online]. Available: https://doi.org/10.1137/1.9781611977912.65

[37] X. Chen, M. Li, and C. Wang, "Favorite-candidate voting for eliminating the least popular candidate in a metric space," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2020, pp. 1894–1901. [Online]. Available: https://doi.org/10.1609/aaai.v34i02.5558

[38] S. Cho and J. W. Endersby, "Issues, the spatial theory of voting, and british general elections: A comparison of proximity and directional models," *Public Choice*, vol. 114, no. 3/4, pp. 275–293, 2003. [Online]. Available: http://www.jstor.org/stable/30025956

[39] S. Ebadian, A. Kahng, D. Peters, and N. Shah, "Optimized distortion and proportional fairness in voting," *ACM Transactions on Economics and Computation*, vol. 12, no. 1, p. 1–39, Mar. 2024. [Online]. Available: http://dx.doi.org/10.1145/3640760

[40] E. Elkind and P. Faliszewski, "Recognizing 1-euclidean preferences: An alternative approach," in *Algorithmic Game Theory - 7th International Symposium, SAGT 2014, Haifa, Israel, September 30 - October 2, 2014. Proceedings*, ser. Lecture Notes in Computer Science, R. Lavi, Ed., vol. 8768. Springer, 2014, pp. 146–157. [Online]. Available: https://doi.org/10.1007/978-3-662-44803-8_13

[41] E. Elkind, P. Faliszewski, P. Skowron, and A. Slinko, "Properties of multiwinner voting rules," *Soc. Choice Welf.*, vol. 48, no. 3, pp. 599–632, 2017. [Online]. Available: https://doi.org/10.1007/s00355-017-1026-z

[42] J. M. Enelow and M. J. Hinich, *The Spatial Theory of Voting*, ser. Cambridge Books. Cambridge University Press, 1984, no. 9780521275156. [Online]. Available: https://ideas.repec.org/b/cup/cbooks/9780521275156.html

[43] B. Fain, A. Goel, K. Munagala, and N. Prabhu, "Random dictators with a random referee: constant sample complexity mechanisms for social choice," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. [Online]. Available: https://doi.org/10.1609/aaai.v33i01.33011893

[44] P. Faliszewski, P. Skowron, A. Slinko, and N. Talmon, *Multiwinner voting: A new challenge for social choice theory*, ser. Trends in computational social choice. Lulu Publisher, 2017, vol. 74, pp. 27–47.

[45] M. Feldman, A. Fiat, and I. Golomb, "On voting and facility location," *CoRR*, vol. abs/1512.05868, 2015. [Online]. Available: http://arxiv.org/abs/1512.05868

[46] A. Filos-Ratsikas, E. Micha, and A. A. Voudouris, "The distortion of distributed voting," *Artif. Intell.*, vol. 286, p. 103343, 2020. [Online]. Available: https://doi.org/10.1016/j.artint.2020.103343

[47] D. Fotakis, L. Gourvès, and P. Patsilinakos, "On the distortion of committee election with 1-euclidean preferences and few distance queries," *CoRR*, vol. abs/2408.11755, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2408.11755

[48] D. Fotakis, L. Gourvès, and P. Patsilinakos, "On the distortion of committee election with 1-euclidean preferences and few distance queries," 2024. [Online]. Available: https://arxiv.org/abs/2408.11755

[49] A. Gibbard, "Manipulation of voting schemes," *Econometrica*, vol. 41, pp. 587–602, 1973.

[50] V. Gkatzelis, D. Halpern, and N. Shah, "Resolving the optimal metric distortion conjecture," in *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020, Durham, NC, USA, November 16-19, 2020*, S. Irani, Ed. IEEE, 2020, pp. 1427–1438. [Online]. Available: https://doi.org/10.1109/FOCS46700.2020.00134

[51] V. Gkatzelis, M. Latifian, and N. Shah, "Best of both distortion worlds," in *Proceedings of the 24th ACM Conference on Economics and Computation, EC 2023, London, United Kingdom, July 9-12, 2023*, K. Leyton-Brown, J. D. Hartline, and L. Samuelson, Eds. ACM, 2023, pp. 738–758. [Online]. Available: https://doi.org/10.1145/3580507.3597739

[52] A. Goel, R. Hulett, and A. K. Krishnaswamy, "Relating metric distortion and fairness of social choice rules," in *Proceedings of the 13th Workshop on Economics of Networks, Systems and Computation, NetEcon@SIGMETRICS 2018, Irvine, CA, USA, June 18, 2018*. ACM, 2018, p. 4:1. [Online]. Available: https://doi.org/10.1145/3230654.3230658

[53] T. F. Gonzalez, "Clustering to minimize the maximum intercluster distance," *Theoretical Computer Science*, vol. 38, pp. 293–306, 1985. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0304397585902245

[54] K. N. Gowda, T. Pensyl, A. Srinivasan, and K. Trinh, "Improved bi-point rounding algorithms and a golden barrier for k-median," 2022. [Online]. Available: https://arxiv.org/abs/2210.13395

[55] S. Gross, E. Anshelevich, and L. Xia, "Vote until two of you agree: Mechanisms with small distortion and sample complexity," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, Feb. 2017. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/10587

[56] R. Hassin and A. Tamir, "Improved complexity bounds for location problems on the real line," *Oper. Res. Lett.*, vol. 10, no. 7, pp. 395–402, 1991. [Online]. Available: https://doi.org/10.1016/0167-6377(91)90041-M

[57] O. Kariv and S. L. Hakimi, "An algorithmic approach to network location problems. ii: The p-medians," *Siam Journal on Applied Mathematics*, vol. 37, pp. 539–560, 1979. [Online]. Available: https://api.semanticscholar.org/CorpusID:120247607

[58] D. Kempe, "Communication, distortion, and randomness in metric voting," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2020, pp. 2087–2094. [Online]. Available: https://doi.org/10.1609/aaai.v34i02.5582

[59] F. E. Kizilkaya and D. Kempe, "Plurality veto: A simple voting rule achieving optimal metric distortion," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, L. D. Raedt, Ed. ijcai.org, 2022, pp. 349–355. [Online]. Available: https://doi.org/10.24963/ijcai.2022/50

[60] ——, "Generalized veto core and a practical voting rule with optimal metric distortion," in *Proceedings of the 24th ACM Conference on Economics and Computation, EC 2023, London, United Kingdom, July 9-12, 2023*, K. Leyton-Brown, J. D. Hartline, and L. Samuelson, Eds. ACM, 2023, pp. 913–936. [Online]. Available: https://doi.org/10.1145/3580507.3597798

[61] A. Kumar, Y. Sabharwal, and S. Sen, "A simple linear time (1+ε)-approximation algorithm for k-means clustering in any dimensions." 01 2004, pp. 454–462.

[62] D. Mandal, A. D. Procaccia, N. Shah, and D. P. Woodruff, "Efficient and thrifty voting by any means necessary," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 7178–7189. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/c09f9caf5e08836d4673ccdd69bb041e-Abstract.html

[63] D. Mandal, N. Shah, and D. P. Woodruff, "Optimal communication-distortion tradeoff in voting," in *EC '20: The 21st ACM Conference on Economics and Computation, Virtual Event, Hungary, July 13-17, 2020*, P. Biró, J. D. Hartline, M. Ostrovsky, and A. D. Procaccia, Eds. ACM, 2020, pp. 795–813. [Online]. Available: https://doi.org/10.1145/3391403.3399510

[64] K. O. May, "A set of independent necessary and sufficient conditions for simple majority decision," *Econometrica*, vol. 20, p. 680, 1952. [Online]. Available: https://api.semanticscholar.org/CorpusID:14254458

[65] S. Merrill, III and B. Grofman, *A Unified Theory of Voting: Directional and Proximity Spatial Models*. Cambridge University Press, 1999.

[66] B. L. Monroe, "Fully proportional representation," *American Political Science Review*, vol. 89, no. 4, p. 925–940, 1995.

[67] K. Munagala and K. Wang, "Improved metric distortion for deterministic social choice rules," in *Proceedings of the 2019 ACM Conference on Economics and Computation, EC 2019, Phoenix, AZ, USA, June 24-28, 2019*, A. R. Karlin, N. Immorlica, and R. Johari, Eds. ACM, 2019, pp. 245–262. [Online]. Available: https://doi.org/10.1145/3328526.3329550

[68] N. Nisan, *Introduction to Mechanism Design (for Computer Scientists)*. Cambridge University Press, 2007, p. 209–242.

[69] A. D. Procaccia and J. S. Rosenschein, "The distortion of cardinal preferences in voting," in *Cooperative Information Agents X, 10th International Workshop, CIA 2006, Edinburgh, UK, September 11-13, 2006, Proceedings*, ser. Lecture Notes in Computer Science, M. Klusch, M. Rovatsos, and T. R. Payne, Eds., vol. 4149. Springer, 2006, pp. 317–331. [Online]. Available: https://doi.org/10.1007/11839354_23

[70] T. Roughgarden, *Beyond the Worst-Case Analysis of Algorithms*. Cambridge University Press, 2020.

[71] ——, "Lecture 6: Perturbation-stable clustering," *CS264: Beyond Worst-Case Analysis*, 2017. [Online]. Available: http://timroughgarden.org/w17/l/l6.pdf

[72] M. Satterthwaite, "Strategyproofness and arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions," *Journal of Economic Theory*, vol. 10, pp. 187–217, 1975.

[73] N. Schofield, *The Spatial Model of Politics*, ser. Routledge frontiers of political economy. Routledge, 2008. [Online]. Available: https://books.google.gr/books?id=rVSamKrb_LEC

[74] V. V. Vazirani, *Approximation algorithms*. Springer, 2001. [Online]. Available: http://www.springer.com/computer/theoretical+computer+science/book/978-3-540-65367-7

[75] D. Vega, M. Karpinski, and C. Kenyon, "Approximation schemes for clustering problems (extended abstract)," 04 2003.

[76] J. von Neumann, O. Morgenstern, and A. Rubinstein, *Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition)*. Princeton University Press, 1944. [Online]. Available: http://www.jstor.org/stable/j.ctt1r2gkx

[77] D. P. Williamson and D. B. Shmoys, *The Design of Approximation Algorithms*. Cambridge University Press, 2011. [Online]. Available: http://www.cambridge.org/de/knowledge/isbn/item5759340/?site_locale=de_DE

[78] H. P. Young, "Condorcet's theory of voting," *American Political Science Review*, vol. 82, no. 4, p. 1231–1244, 1988.

[79] W. S. Zwicker, "Introduction to the theory of voting," in *Handbook of Computational Social Choice*, F. Brandt, V. Conitzer, U. Endriss, J. Lang, and A. D. Procaccia, Eds. Cambridge University Press, 2016, pp. 23–56. [Online]. Available: https://doi.org/10.1017/CBO9781107446984.003