Error in the Euclidean Preference Model

Luke Thorburn, **Maria Polukarov** and **Carmine Ventre**
King's College London
{luke.thorburn, maria.polukarov, carmine.ventre}@kcl.ac.uk

Abstract

Spatial models of preference, in the form of vector embeddings, are learned by many deep learning and multiagent systems, including recommender systems. Often these models are assumed to approximate a Euclidean structure, where an individual prefers alternatives positioned closer to their "ideal point", as measured by the Euclidean metric. However, previous work has shown there are ordinal preference profiles that cannot be represented with this structure if the Euclidean space has two fewer dimensions than there are individuals or alternatives. We extend this result, showing that there are situations in which almost all preference profiles cannot be represented with the Euclidean model, and derive a theoretical lower bound on the expected error when using the Euclidean model to approximate non-Euclidean preference profiles. Our results have implications for the interpretation and use of vector embeddings, because in some cases close approximation of arbitrary, true ordinal relationships can be expected only if the dimensionality of the embeddings is a substantial fraction of the number of entities represented.

1 Introduction

Accurate modelling of preferences is critical to the safe and efficient deployment of AI in multiagent systems [Christian, 2020]. Indeed, if the preference models used by AI agents are not sufficiently expressive to be capable of representing human preferences, they will be at least partially "mistaken" about what humans want and, consequently, may take actions that cause harm. For example, a recommender system built on inaccurate preference models may make recommendations that bias perceptions of politics [Huszár *et al.*, 2021], contribute to low self esteem [Faelens *et al.*, 2021], or encourage unsafe medical interventions [Johnson *et al.*, 2021].

Spatial models of preference, in the form of vector embeddings, are widely used in deep learning systems. In user-facing contexts such as recommender systems, user embeddings contain information that might be described literally as preferences [Pan and Ding, 2019]. More generally, embeddings that are intended to capture degrees of similarity between elements of a set can often be viewed as a spatial model of preference. For example, each word in a language could be considered to have a "preference" over all other words, "preferring" those with similar meanings. Spaces of word embeddings [Almeida and Xexéo, 2019] can thus be viewed as models of preference. Spatial models of (literal) preference are also used in the fields of political science [Hinich and Munger, 2008], social choice theory [Miller, 2015] and opinion dynamics [Aydogdu *et al.*, 2017], as well as in preference aggregation software such as *Polis* [Small *et al.*, 2021] and Twitter's *Community Notes* feature [Twitter, 2022].

A canonical spatial model is the Euclidean model, where both individuals and alternatives are represented as points in Euclidean space, and each individual prefers alternatives positioned nearer to them, as measured by the standard Euclidean metric [Bogomolnaia and Laslier, 2007]. A preference profile of I individuals over A alternatives is said to be d-Euclidean if it can be represented with a d-dimensional Euclidean model.

The Euclidean model is often used explicitly, and even spatial preference models that are not strictly Euclidean are often assumed to have an approximately Euclidean structure. For example, embeddings in deep learning systems are usually compared using cosine similarity [Jurafsky and Martin, 2022, Ch. 6.4], which induces the same ordinal relationships as the Euclidean metric when applied to normalized vectors. Given the importance of accurately representing human preferences and the prevalence of Euclidean preference models, it is important to understand their limitations.

In this paper, we assume a "ground truth" preference structure where each individual's preference is a strict order over the available alternatives. Consider the expressiveness of the Euclidean preference model relative to this ordinal model. There are three questions one might ask, of increasing informativeness. For fixed positive integers I, A and d:

1. Are there any preference profiles of I individuals over A alternatives that are not d-Euclidean?
2. What proportion of such profiles are not d-Euclidean?
3. How large is the expected error when approximating arbitrary preferences with a d-dimensional Euclidean model?

Question 1 was answered by [Bogomolnaia and Laslier, 2007], who showed that when $d < \min\{I-1, A-1\}$, there exist profiles of I preferences over A alternatives that cannot be represented in a d-dimensional Euclidean model. In this work, we address questions 2 and 3.

1.1 Contributions

We extend the results of [Bogomolnaia and Laslier, 2007] by proving a series of theoretical bounds. Intuitively, these bounds indicate that when the dimensionality of the Euclidean model is small relative to the number of individuals and the number of alternatives, it is possible that almost all preference profiles cannot be represented. Further, we describe conditions under which only a minute proportion of all possible preferences can be simultaneously represented, unless the dimensionality of the Euclidean model is almost as large as the number of individuals or the number of alternatives. Our final bound is on the expected error when approximating a randomly chosen preference in a Euclidean model of given dimensionality, where we formalize error as the number of adjacent swaps required to transform the nearest representable preference into the true preference that is being approximated.

Our results have implications for the interpretation and use of vector embeddings, because there are situations in which close approximation of arbitrary, true preferences (or preference-like data) is possible only if the dimensionality of the embeddings is a substantial fraction of the number of individuals or alternatives. In these situations, our theoretical bounds can inform the choice of the dimensionality of vector embeddings by quantifying the expected error when the dimensionality is too small.

1.2 Related Work

This paper builds on the work of [Bogomolnaia and Laslier, 2007]; we state the relevant results in Section 3. Important context is also provided by [Peters, 2017] who showed that the problem of determining whether a preference profile is d-Euclidean is, in general, NP-hard, and that some ordinal preference profiles require exponentially many bits to be represented in the Euclidean model. These hardness results rule out some of the most obvious approaches to evaluating the expressiveness of the Euclidean model, because for a given ordinal profile it is usually not feasible to check whether it is d-Euclidean, or to compute its best approximation with a d-dimensional Euclidean model [Tydrichová, 2023]. That said, the computational task of approximating an ordinal preference profile with a d-dimensional Euclidean model is known as *multidimensional unfolding*, and has a number of proposed algorithms [Bennett and Hays, 1960; Elkind and Faliszewski, 2014; Luaces *et al.*, 2015].

A related line of work discusses the limitations of the Euclidean preference model from the perspective of measurement theory and psychometric validity. For example, [Eguia, 2013] questions the validity of the utility functions, indifference curves and separability of preferences that are implied by the Euclidean model. [Henry and Mourifié, 2013] present analysis suggesting that the Euclidean model is not consistent with real world voting data. A number of works suggest the L^1 metric may be more appropriate than the L^2 metric in spatial preference models [Humphreys and Laver, 2010; Rodríguez, 2011; Eguia, 2011; Eguia, 2013].

A general review of structured preference models is given by [Elkind *et al.*, 2017].

Outline. Section 2 introduces our notation and definitions. Section 3 presents answers to the three questions listed above, including our theoretical bounds. The main paper contains proof sketches for each of the results, and full proofs are in the extended preprint[1]. Implications are discussed in Section 4.

2 Preliminaries

Let $A \in \mathbb{N}$ be the number of alternatives and $I \in \mathbb{N}$ be the number of individuals. Each individual is assumed to have a *preference*: a strict order (ranking without ties) over all A alternatives. The preference of individual i is denoted π_i (a ranked list) or $>_i$ (the corresponding order relation), so if a and b are alternatives, $a >_i b$ means individual i prefers a to b. The list of all preferences represented in the population of I individuals is called a *profile*. For given values of A and I, the set of all possible profiles is denoted $\mathcal{P}_{A,I}$. The number of *unique* preferences in a profile is denoted I^*.

In the d-dimensional Euclidean preference model, both alternatives and individuals are represented as points in d-dimensional Euclidean space, denoted $\mathbb{R}^d$. We refer to these points as the *location* of each alternative and the *ideal point* for each individual. Each individual's preference is determined by the distance between their ideal point and the location of each alternative. Nearer alternatives are preferred, as measured by the standard Euclidean metric.

Definition 1 (Euclidean preferences)**.** *A profile $\Pi \in \mathcal{P}_{A,I}$ is d-*Euclidean *if there exist points $x^a \in \mathbb{R}^d$ for all $a \in \{1, \dots, A\}$, and $w^i \in \mathbb{R}^d$ for all $i \in \{1, \dots, I\}$ such that, for all alternatives a, b and individuals i, $a >_i b \Leftrightarrow \|x^a - w^i\| < \|x^b - w^i\|$ (using the standard Euclidean norm).*

Note that while some versions of the Euclidean preference model are used to represent utilities or *cardinal* preferences, in our definition the Euclidean model only describes *ordinal* preferences. We focus on strict orders for simplicity, and also because true indifference in most random d-Euclidean preference profiles will occur with probability zero.

Definition 2. *If all profiles $\Pi \in \mathcal{P}_{A,I}$ are Euclidean of dimension d, then we say d is* sufficient *for $\mathcal{P}_{A,I}$.*

For a profile Π, we use $\text{Euclidean}(\Pi)$ to denote a Euclidean preference model that optimally approximates Π.

3 Three Questions

3.1 Are All Profiles Euclidean?

The first question was answered by [Bogomolnaia and Laslier, 2007], who identified the minimum dimensionality required to represent all possible profiles of a given size.

Theorem 1 (Bogomolnaia and Laslier 2007)**.** *Dimensionality d is sufficient for A, I if and only if $d \geq M$ where either $M = \min\{I - 1, A - 1\}$ or $M = \min\{I, A - 1\}$, depending on the values of A and I.*

[1] See: doi.org/10.48550/arXiv.2208.08160.

Thus, the answer to the first question is no, not all profiles are d-Euclidean. If $d < \min\{I-1, A-1\}$, then there exists at least one profile $\Pi \in \mathcal{P}_{A,I}$ which cannot be losslessly represented with a d-Euclidean model.

Is this really that big a deal? Maybe the profiles that are not d-Euclidean are only a small number of pathological edge cases that are unlikely to be encountered in the real world. The next question asks whether this is the case.

3.2 How Common Are Non-Euclidean Profiles?

For a given $d < \min\{I-1, A-1\}$, what proportion of preference profiles in $\mathcal{P}_{A,I}$ are not d-Euclidean? Given the NP-hardness of recognizing whether a given profile is Euclidean, a precise answer to this question is likely intractable. However, [Bogomolnaia and Laslier, 2007] defined three classes of pathological sub-profiles that, if present, cause a profile to not be d-Euclidean for some d. If we calculate the probability that a pathological sub-profile arises in a profile constructed uniformly at random—that is, in an *impartial culture* [Eğecioğlu and Giritligil, 2013]—this probability is precisely the proportion of profiles that exhibit that pathology (and hence are not d-Euclidean), which is a lower bound on the proportion of profiles that are not d-Euclidean. It is a lower bound because the set of all non d-Euclidean profiles is a superset of the set of profiles that contain this pathological sub-profile, and it is *only* a lower bound because we are—to the best of our knowledge—considering only a subset of the pathologies that cause a profile to be non d-Euclidean.

It is not known whether the three classes of pathological sub-profiles identified by [Bogomolnaia and Laslier, 2007] exhaustively characterize the ways in which a profile can fail to be d-Euclidean—there may be other pathologies. In this work, we focus on one of these classes (which we call the *circulant pathology*), as of the two other classes they consider, one requires the possibility of ties between alternatives [Bogomolnaia and Laslier, 2007, Ex. 6], (which is a different "ground truth" preference model to that which we consider) and the other requires that $A \geq 2^I$ [Bogomolnaia and Laslier, 2007, Ex. 14] (which we deemed implausible for the settings we are interested in, such as recommender systems, where $I \gg 100$). Further, while we were able to derive or bound the probability of each class occurring in isolation, calculation of the joint probabilities did not appear tractable, and considering the circulant pathology alone is sufficient to produce non-trivial results. We emphasize that the bounds we derive apply more broadly to the phenomenon of non d-Euclidean profiles, because the set of such profiles is a superset of those that contain the circulant pathology.

Definition 3 (circulant pathology)**.** *A circulant pathology of size k is a preference sub-profile consisting of k alternatives $a_1, \ldots, a_k$ and k individuals $1, \ldots, k$ such that*

$$\begin{array}{lllllllll}
a_1 & >_1 & a_2 & >_1 & \cdots & >_1 & a_{k-1} & >_1 & a_k \\
a_2 & >_2 & a_3 & >_2 & \cdots & >_2 & a_k & >_2 & a_1 \\
\vdots & & \vdots & & & & \vdots & & \vdots \\
a_k & >_k & a_1 & >_k & \cdots & >_k & a_{k-2} & >_k & a_{k-1}.
\end{array}$$

Note that for each instance of the circulant pathology, there is a unique *circular permutation* of k alternatives that is consistent with all the sub-preferences involved in the pathology. (A circular permutation is a unique ordering of the alternatives on a circle.) We refer to each of these sub-preferences as *necessary sub-preferences*.

Theorem 2 (Bogomolnaia and Laslier 2007)**.** *If a profile $\Pi \in \mathcal{P}_{A,I}$ contains a circulant pathology of size k as a sub-profile, then Π is not d-Euclidean for any $d \leq k-2$.*

Deriving the exact probability that a circulant pathology of size k arises in a randomly generated profile appears difficult, but there is some related work. For example, consider a particular circulant pathology constructed using fixed subset of all A alternatives, consistent with a fixed circular permutation of those alternatives. The set of all $A!$ possible preferences can be partitioned into those that could play the role of individual 1, those that could be individual 2 etc., and those that cannot be part of the pathology because the order in which the k alternatives appear does not match any of the necessary sub-preferences. The probability that this specific pathology arises is the probability that when choosing I preferences independently and uniformly at random, we choose at least one from all but the last part in this partition. This can be framed as the probability of completing a particular row on a bingo card within I draws (with replacement). It is also closely related to the *coupon collector problem* [Neal, 2008] in probability theory. However, results for these problems are not easily generalized to the situation in which every possible subset of alternatives of size $k \geq d+2$ might be used to construct the pathology.

From another angle, we might start from the probability that k randomly chosen preferences contain a common sub-preference over any subset of k alternatives, and then hope to adjust the probability to account for the circulant offset. The closest work to this is a set of papers on the longest common subsequences in random words or permutations [Bukh and Ma, 2014; Houdré and Xu, 2018; Houdré and Işlak, 2022], however so far this work is limited to the setting of $k = 2$.

Whilst we cannot compute it exactly, we can bound from below the probability by restricting ourselves to a subset of the ways in which the pathology might be constructed. This brings the problem within reach of an approach similar to the bingo framing described above.

Theorem 3 (lower bound on probability of circulant pathology)**.** *Let A, I, and d be fixed positive integers such that $d < \min\{I, A-1\}$, and $\mathbf{P}(C)$ be the probability that a profile chosen uniformly from $\mathcal{P}_{A,I}$ contains a circulant pathology of size $k \geq d+2$. Then,*

$$\mathbf{P}(C) \geq 1 - \left(1 - \sum_{k=d+2}^{I} B_k\right)^{\left\lfloor \frac{A}{d+2} \right\rfloor},$$

where $B_k = \binom{I}{k} \left\{ {k \atop d+2} \right\} (d+2)! \left(\frac{1}{(d+2)!}\right)^k \left(1 - \frac{d+2}{(d+2)!}\right)^{I-k}$ and $\left\{ {k \atop d+2} \right\}$ denotes a Stirling number of the second kind.

Proof sketch. Build up the bound step by step. First, consider the version of the pathology that involves a specific subset of $d+2$ alternatives $a_1, \ldots, a_{d+2}$, and is consistent with a specific circular permutation of those $d+2$ alternatives. The probability that this version of the pathology arises in a profile chosen uniformly at random from $\mathcal{P}_{A,I}$ is

$$\underbrace{\binom{I}{k}\left\{\begin{matrix}k\\ d+2\end{matrix}\right\}(d+2)!}_{\begin{array}{c}\text{\# ways to choose } k \text{ individuals from } I,\\ \text{partition them into } d+2 \text{ non-empty parts,}\\ \text{and assign each part to a necessary}\\ \text{sub-preference}\end{array}}$$

$$\times \underbrace{\left(\frac{1}{(d+2)!}\right)^{k}}_{\begin{array}{c}\text{probability those } k \text{ individuals}\\ \text{randomly get assigned}\\ \text{preferences that have those}\\ \text{necessary sub-preferences}\end{array}} \times \underbrace{\left(1-\frac{d+2}{(d+2)!}\right)^{I-k}}_{\begin{array}{c}\text{probability the other } I-k\\ \text{individuals randomly get assigned}\\ \text{preferences that lack those}\\ \text{necessary sub-preferences}\end{array}}$$

Call this binomial-like expression B_k. The number of individuals involved in the pathology, k, must be at least $d{+}2$, and can be as high as I, and the events where different numbers of individuals are involved are mutually exclusive. Thus, we can sum over the alternate choices of k to get $\sum_{k=d+2}^{I} B_k$, which is the probability that this particular version of the pathology arises for any k. Note also that versions of the pathology constructed using disjoint subsets of alternatives occur independently, so we can generalize the expression to allow for the fact that there are multiple disjoint subsets of alternatives from which this pathology can arise. Each subset must be of size $d{+}2$, so at most we could specify $\left\lfloor \frac{A}{d+2} \right\rfloor$ disjoint subsets. The probability thus increases to

$$1-\underbrace{\left(1-\sum_{k=d+2}^{I} B_k\right)^{\left\lfloor \frac{A}{d+2} \right\rfloor}}_{\begin{array}{c}\text{probability that none of these}\\ \lfloor A/(d+2) \rfloor \text{ versions of the}\\ \text{pathology occur}\end{array}}.$$

This expression is the probability that at least 1 of each of $\lfloor A/(d+2) \rfloor$ versions of the pathology, each constructed using a disjoint set of alternatives, occurs. It does not account for all possible subsets of alternatives, or all possible circular permutations thereof, so it is a lower bound on $\mathbf{P}(C)$. □

We can now numerically evaluate this expression for a range of values of A, I and d to see what the probability (or proportion) is in real terms (Figure 1). When d is a lot smaller than A and I, almost all profiles are not d-Euclidean, and for any fixed d and A this proportion appears to approach 1 as $I \to \infty$. Increasing d appears to be quite effective at reducing the proportion of non-Euclidean profiles (there is a lot of room in a high-dimensional space).

We conclude that in some circumstances, almost all profiles are not d-Euclidean. So what? If we *approximate* them with a Euclidean preference model, is the approximation error big enough to worry about?

3.3 How Large Is the Expected Error?

There are different ways to quantify the error of a d-Euclidean approximation to an arbitrary preference profile. The most natural approach may be to count the minimal number of swaps needed to change the given profile into one that is d-Euclidean. However, this intuitive approach seems difficult given the findings of [Peters, 2017], who proved that the task of finding such best-approximations is, in general, NP-hard. Instead, we consider the question from the perspective of individual preferences, rather than complete profiles, and use the following setup:

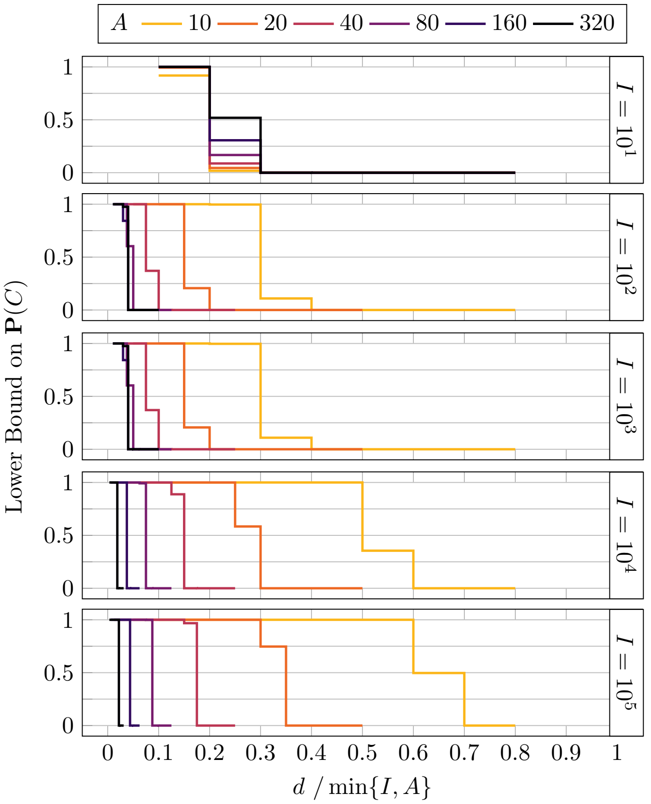

Figure 1: Lower bound on $\mathbf{P}(C)$, and hence the probability that a profile chosen uniformly at random is not d-Euclidean, for various d, A and I.

1. Take an arbitrary profile $\Pi \in \mathcal{P}_{A,I^*}$ consisting of I^* unique preferences. (It doesn't matter which profile it is — we assume we don't know.)
2. Approximate Π as well as possible in a d-dimensional Euclidean model. Call this model Euclidean(Π). (It doesn't matter how good the approximation algorithm is, our bound assumes it finds the best possible fit.)
3. Observe a new preference π generated uniformly from among all $A!$ preferences. (It could duplicate one of the existing I^* preferences. The new preference π is generated *after* the model Euclidean(Π) is fit. Knowledge of π cannot inform Euclidean(Π), otherwise one could trivially choose a model in which π was representable.)
4. Let $\hat{\pi}$ be the preference representable in Euclidean(Π) that minimizes $m(\hat{\pi}, \pi)$ for some error measure m.

That is, $\hat{\pi}$ is the closest possible approximation to π in such a Euclidean preference model. As our measure of error m, we

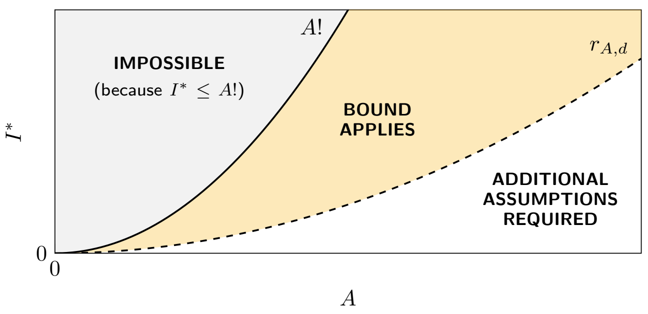

Figure 2: A (not to scale) diagram indicating the values of I^* and A for which the bound on the expected error given in Theorem 4 applies. When $I^* < r_{A,d}$, additional assumptions would be required to produce a non-trivial lower bound, because the expected error could be as low as zero depending on the profile Π.

use the Kendall tau distance (or bubble-sort distance):

$$m(\pi, \pi') = \text{\# pairwise disagreements between } \pi \text{ and } \pi'.$$

Equivalently, $m(\pi, \pi')$ can be defined as the minimum number of adjacent swaps required to transform π to π'. The Kendall tau distance is well-established [Kumar and Vassilvitskii, 2010] and has been widely used as a metric between preferences in the social choice literature [Obraztsova and Elkind, 2012; Obraztsova *et al.*, 2013; Anand and Dey, 2021]. We are interested in both $\mathbf{E}[m(\pi, \hat{\pi})]$ and $\mathbf{E}[m(\pi, \hat{\pi})]/\binom{A}{2}$, which is the expected number of adjacent swaps required as a proportion of the maximum possible number of swaps between any two preferences. Intuitively, $\mathbf{E}[m(\pi, \hat{\pi})]/\binom{A}{2}$ can also be interpreted as the probability that two alternatives chosen uniformly at random will be ranked differently by π and $\hat{\pi}$. Given the lack of tractable algorithms or formulae for optimally approximating arbitrary profiles with Euclidean models (see Section 1), we do not compute these expectations exactly, and instead derive lower bounds.

The intuition for our approach is as follows. For any choice of A and d, there will be some maximum number of unique preferences (or permutations) $r_{A,d} \leq A!$ that can be simultaneously represented in Euclidean space of dimension d. In what follows, we will abbreviate $r_{A,d}$ to r (the dependence on A and d is implied) and in order to bound the expected error will assume that $I^* \geq r$ because, in this case, the probability that π is representable is fixed (Figure 2). Assume that we have an overestimate of r, say $\hat{r}$. Then, if we "distribute" these $\hat{r}$ "representable" preferences "evenly" throughout the set of all possible preferences, this will minimize the expected distance between π and $\hat{\pi}$, the nearest of the $\hat{r}$ "representable" preferences. We use this minimized distance (or more precisely, distribution of distances) to produce a lower bound for $\mathbf{E}[m(\pi, \hat{\pi})]$. It is only a lower bound because $\hat{r}$ may overestimate r.

Along with an upper bound $\hat{r}$, this approach requires a structure on the set of all $A!$ preferences that allows us to "evenly distribute" the $\hat{r}$ preferences that might be representable. For this, we use the *permutohedron* (Figure 3), a high dimensional polytope where each vertex corresponds to a permutation, and edges between vertices correspond to single swap operations between adjacent elements. (There is a bijection between permutations and preferences.) For example, the permutohedron of order 4 will have a vertex corresponding to the permutation `1234` that is connected to three other vertices: `2134`, `1324`, and `1243`, because they are the permutations that can be reached by applying a single adjacent swap to `1234` [Gaiha and Gupta, 1977]. Our overestimate for $\hat{r}$ is given by the following lemma.

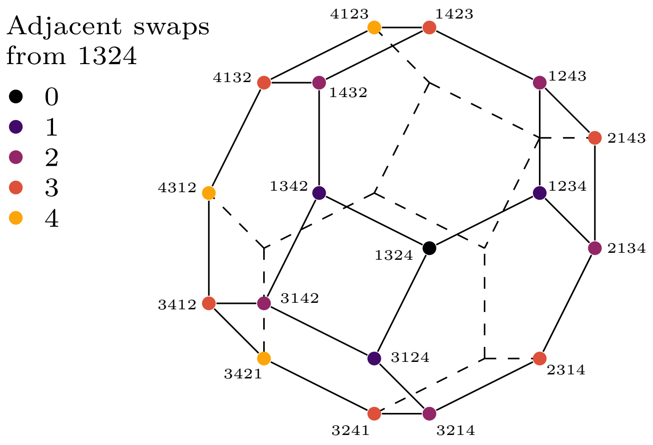

Figure 3: The permutohedron of order 4.

Lemma 1 (upper bound on r)**.** *Let r be the maximum number of $A!$ unique preferences over A alternatives that are simultaneously representable in a d-dimensional Euclidean preference model. If $I^* \geq r$, then*

$$r \leq \hat{r} = (1 - \mathbf{P}(E))\, A!,$$

where $\mathbf{P}(E)$ is the proportion of $A!$ unique preferences over A alternatives that cannot *be simultaneously represented due to the circulant pathology. Explicitly, let D_n be the event that a random permutation of integers $1, \ldots, A$ places the integer n ($n \leq A$) within the first $A - d - n$ positions, and $\overline{D_n}$denote the complement of D_n. Then $\mathbf{P}(E)$ can be written as*

$$\mathbf{P}(E) = \mathbf{P}(D_1) + \mathbf{P}(\overline{D_1})\mathbf{P}(E \mid \overline{D_1}),$$

where the conditional probability is defined recursively, as follows. For $N < A - d - 2$,

$$\mathbf{P}\left(E \mid \bigcap_{n=1}^{N} \overline{D_n}\right) = \mathbf{P}\left(D_{N+1} \mid \bigcap_{n=1}^{N} \overline{D_n}\right) + \mathbf{P}\left(\overline{D_{N+1}} \mid \bigcap_{n=1}^{N} \overline{D_n}\right) \mathbf{P}\left(E \mid \bigcap_{n=1}^{N+1} \overline{D_n}\right),$$

and for $N = A - d - 2$,

$$\mathbf{P}\left(E \mid \bigcap_{n=1}^{N} \overline{D_n}\right) = \mathbf{P}\left(D_{A-d-1} \mid \bigcap_{n=1}^{A-d-2} \overline{D_n}\right).$$

Proof sketch. For every combination of $d + 2$ alternatives, there are $(d + 1)!$ circular permutations of those alternatives. For each circular permutation, there are $d + 2$ necessary subpreferences over the $d+2$ alternatives that are consistent with the circular permutation. To avoid the pathology, we need to

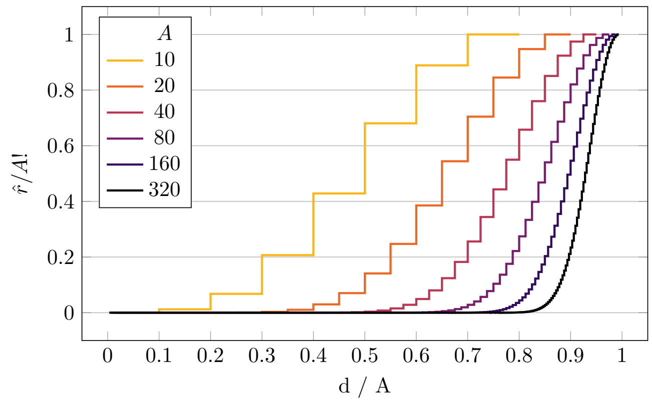

Figure 4: Upper bound on the proportion of all $A!$ possible preferences that can be simultaneously represented in a d-Euclidean model, for various A, d, and valid when $I^* \geq r$.

“ban” one of these necessary sub-preferences for every combination and circular permutation of $d + 2$ alternatives.

Without loss of generality, we choose to ban the sub-preference that ranks most highly the alternative with the minimum index. With this convention, the proportion of preferences that are banned is equal to the probability of the event E that a random preference over the alternatives $a_1, \ldots, a_A$ “positions a low index alternative sufficiently near the top”.

More precisely, $E = \bigcup_{n=1}^{A-d-1} D_n$, where D_n is the event that a random permutation of the alternatives $a_1, \ldots, a_A$ positions a_n within the first $A - d - n$ positions of the permutation. There is no closed form expression for $\mathbf{P}(E)$, but it can be written down using a recursive formula that corresponds to repeated application of the law of total probability. □

How small is $\hat{r}$, in real terms? Figure 4 plots $\hat{r}/A!$ for a variety of values of d and A. In most cases, only a small proportion of preferences can be simultaneously represented, and it is only as d nears A that the proportion increases. So how much error should we expect when approximating such preferences in a Euclidean preference model?

Theorem 4 (lower bound on expected error)**.** *Let A, d be fixed positive integers such that $d < A-1$, $\Pi \in \mathcal{P}_{A,I^*}$ consist of I^* unique preferences, π be a preference chosen uniformly at random from the set of $A!$ possible preferences, $\hat{\pi}$ be the nearest preference to π that is representable in Euclidean(Π) (that is, the representable preference that can be reached in the fewest number of adjacent swaps), and K be a positive integer such that $K \leq \binom{A}{2}$. If $I^* \geq r$, then*

$$\mathbf{E}[m(\pi, \hat{\pi})] \geq \sum_{k=0}^{K} \frac{(A! - n_{k,A})_{\hat{r}}}{(A!)_{\hat{r}}} \mathbf{1}(\hat{r} < A! - n_{k,A}),$$

where $(\cdot)_{\hat{r}}$ denotes a falling factorial, $\mathbf{1}(\cdot)$ the indicator function, and $n_{k,A} = \min\{(A-1)^k,\ A!\}$.

Proof sketch. The distribution of the $\hat{r}$ “representable” preferences over the vertices of the permutohedron that would minimize the expectation $\mathbf{E}[m(\pi, \hat{\pi})]$ is the uniform distribution. Thus, to lower bound the expectation, select $\hat{r}$ vertices uniformly at random (without replacement) at which to position the “representable” preferences. This gives a function F that bounds from above the distribution function of $m(\pi, \hat{\pi})$. For any non-negative integer k,

$$F(k) = 1 - \underbrace{\frac{(A! - n_{k,A})_{\hat{r}}}{(A!)_{\hat{r}}} \mathbf{1}(\hat{r} < A! - n_{k,A})}_{\substack{\text{probability that none of the } \hat{r} \text{ “representable”} \\ \text{preferences fall on the } n_{k,A} \text{ “reachable” vertices}}},$$

where $(\cdot)_{\hat{r}}$ denotes a falling factorial, $\mathbf{1}(\cdot)$ the indicator function, and $n_{k,A} = \min\{(A-1)^k,\ A!\}$ is an upper bound on the number of unique preferences that are reachable within k swaps. The form of F is analogous to the exact distribution function of $m(\pi, \hat{\pi})$ under our assumed uniform distribution on the positions of the representable preferences, but uses $\hat{r}$ in place of r and $n_{k,A}$ in place of the true number of reachable vertices. Because of these substitutions and the uniform assumption, $\mathbf{P}(m(\pi, \hat{\pi}) \leq k) \leq F(k)$. We can then turn this into a bound on the expectation. For any non-negative integer $K < \binom{A}{2}$,

$$\mathbf{E}[m(\pi, \hat{\pi})] = \sum_{k=0}^{\infty} \mathbf{P}(m(\pi, \hat{\pi}) > k) = \sum_{k=0}^{\infty} 1 - \mathbf{P}(m(\pi, \hat{\pi}) \leq k)$$
$$\geq \sum_{k=0}^{\infty} 1 - F(k) \geq \sum_{k=0}^{K} 1 - F(k). \quad \square$$

How high is the bound in real terms? Figure 5(a) plots some values of the lower bound on $\mathbf{E}[m(\pi, \hat{\pi})]$, and Figure 5(b) plots the equivalent scaled values, $\mathbf{E}[m(\pi, \hat{\pi})] / \binom{A}{2}$. In both cases, we can see that expected error becomes more severe as $d/A \to 0$, but that the bound on the scaled loss appears less severe for larger values of A.

We reiterate that while the bounds in Lemma 1 and Theorems 3 and 4 are derived using only one class of pathological sub-profiles (the circulant pathology), they apply more broadly to the general phenomenon of non d-Euclidean profiles, because the set of such profiles is a superset of those that contain a circulant pathology.

4 Conclusions

We have proven theoretical lower bounds on the proportion of preference profiles of I individuals over A alternatives that are not d-Euclidean, and on the expected error when representing an arbitrary preference in a Euclidean model. Ultimately, these bounds show that when d is small relative to I and A ($d \ll \min\{I, A\}$), almost all preference profiles are not d-Euclidean and, with preferences from an impartial culture, the expected error when approximating an additional, arbitrary preference in the Euclidean model can in some cases be at least 7% of the maximum possible variation between any two preferences, as measured by the Kendall tau distance.

4.1 Implications

Our lower bound on the expected error has been proven for $I^* \geq r$, and is largest when $d \ll \min\{I, A\}$. These conditions may be met, for example, in some social choice models of national elections, or in models of reputation systems where there are a large number of Sybils. In such settings,

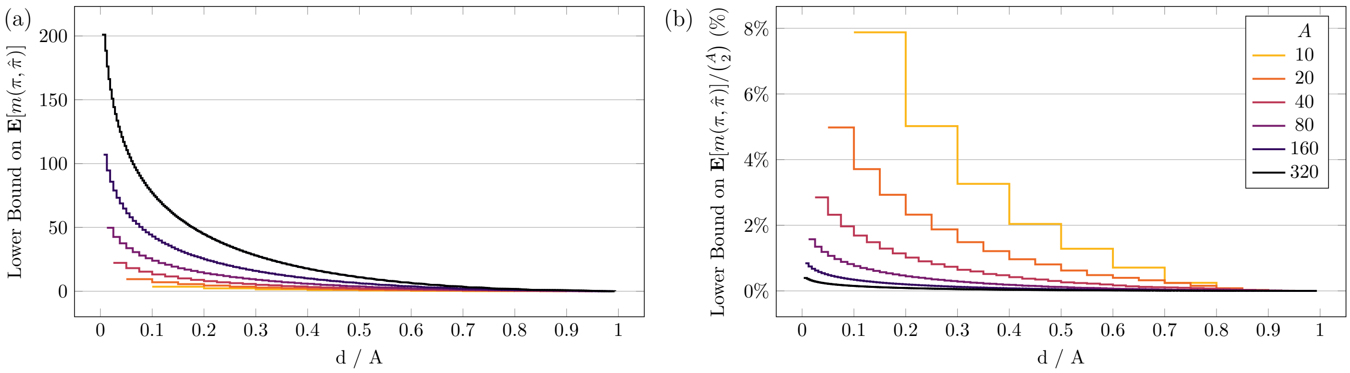

Figure 5: (a) The lower bound on $\mathbf{E}[m(\pi, \hat{\pi})]$ and (b) the lower bound on $\mathbf{E}[m(\pi, \hat{\pi})]/\binom{A}{2}$, both for $I^* \geq r$ and various A, d.

our bound shows that close approximation of true underlying preference profiles may not always be expected. The magnitude of error will depend on the particular profile, and can be reduced by increasing the dimensionality, d. Our bound provides a way to quantify the trade-off between dimensionality and accuracy, and hence inform the choice of d.

Settings where $d \ll \min\{I, A\}$ are also common in deep learning. For example, text embeddings used for natural language processing commonly have $d \approx 10^3$ dimensions [Jurafsky and Martin, 2022, Ch. 6.8], but the English language has $\approx 10^5$ unique words [Nagy and Anderson, 1984], and the number of sentences or paragraphs will be orders of magnitude larger. Similarly, embeddings in recommender systems have up to $d \approx 10^5$ dimensions, but large platforms can serve $I \approx 10^9$ users and host $A \approx 10^{11}$ items of content [Satuluri *et al.*, 2020]. To the extent that there are true, underlying preferences (or more generally, ordinal relationships) in such domains, our results suggest that these relationships may not be able to be closely approximated by such relatively low dimensional embeddings if the use of those embeddings assumes the Euclidean structure.

It is important that we are precise about the uses of embeddings in deep learning systems to which these bounds apply. There are at least two common use cases potentially affected.

Proximity-based ranking. Embeddings are commonly compared using cosine similarity, which induces the same ordinal relationships as the Euclidean metric when applied to normalized vectors (see Section 1). Thus when cosine similarity or Euclidean distance is used to rank the similarity of embeddings to an ideal point, such as when ranking the top-k most preferred items for a user in a recommender system, the bounds suggest a limit on how accurately these ordinal relationships can be recovered.

SoftMax. Embeddings are also used to recover a probability distribution over a set of alternatives. Often this is done by taking the inner product between the embedding of each alternative with a fixed reference vector, and then converting these values into probabilities using the SoftMax function [Goodfellow *et al.*, 2016, Sec. 6.2.2.3]. Geometrically, this corresponds to projecting the embeddings onto the reference vector, so the ordering of the alternatives from most probable to least probable in the resulting distribution is the same as the ordering implied by the preferences of an individual in a Euclidean preference model where the individual's ideal point is placed sufficiently distant in the direction of the reference vector. In this context, our bounds suggest that there will be pairs of alternatives for which the model is mistaken about which is more likely.

The extent to which these bounds apply to embeddings when they are used in their original context—that is, when used as inputs to a neural network with which they were jointly trained—remains an open question. It is possible that a trained network interprets the relationships between embeddings using an approximation to the Euclidean metric, in which case similar limits on accuracy may apply. But it is also possible that the network effectively memorizes the exceptions to the ordinal relations implied by the underlying spatial model, and is thus able to compensate for any error.

4.2 Future Work

Our results pave the way for a research agenda useful to inform the choice of the dimensionality of embeddings or Euclidean preference models. As discussed, the bound on the expected error has only been proven for cases where $I^* \geq r$, which cannot be guaranteed for all applications. Moreover, the (computational) complexity of the bounds prevents them from being evaluated for settings where both I and A are larger than about 10^3. Future work should thus prioritize extending the bound on the expected error to cases where $I^* < r$, and developing simpler approximations to the bounds that can be evaluated efficiently for large I and A.

Another important direction would be to explore the robustness of our results to different distributions over preferences, as well as to different "ground truth" models of preference. For example, our proofs assume a uniform distribution over preferences and profiles (an impartial culture), but in practice some will be more probable than others. It is not clear how this would affect the bounds.

Finally, we note that while our lower bounds are informative, they are not tight. It is possible that non d-Euclidean profiles are substantially more common, or that the error is substantially more severe, than the bounds themselves. Given the possible implications for the accuracy of a widely-used preference model, it would be valuable to improve the tightness of the lower bounds, and to derive meaningful upper bounds.

Acknowledgments

Luke Thorburn was supported by UK Research and Innovation [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (safeandtrustedai.org). Carmine Ventre acknowledges funding from the UKRI Trustworthy Autonomous Systems Hub [EP/V00784X/1].

References

[Almeida and Xexéo, 2019] Felipe Almeida and Geraldo Xexéo. Word Embeddings: A Survey. *arXiv preprint arXiv:1901.09069*, January 2019.

[Anand and Dey, 2021] Aditya Anand and Palash Dey. Distance restricted manipulation in voting. *Theoretical Computer Science*, 891:149–165, 2021.

[Aydogdu *et al.*, 2017] Aylin Aydogdu, Sean T Mcquade, and Nastassia Pouradier Duteil. Opinion Dynamics on a General Compact Riemannian Manifold. *Networks and Heterogeneous Media*, 12(3):489 – 523, September 2017.

[Bennett and Hays, 1960] Joseph F. Bennett and William L. Hays. Multidimensional unfolding: Determining the dimensionality of ranked preference data. *Psychometrika*, 25(1):27–43, Mar 1960.

[Bogomolnaia and Laslier, 2007] Anna Bogomolnaia and Jean-François Laslier. Euclidean preferences. *Journal of Mathematical Economics*, 43(2):87–98, February 2007.

[Bukh and Ma, 2014] Boris Bukh and Jie Ma. Longest common subsequences in sets of words. *arXiv preprint arXiv:1406.7017*, October 2014.

[Christian, 2020] Brian Christian. *The Alignment Problem: Machine Learning and Human Values*. WW Norton & Company, New York, US, 2020.

[Eguia, 2011] Jon X. Eguia. Foundations of spatial preferences. *Journal of Mathematical Economics*, 47(2):200–205, 2011.

[Eguia, 2013] Jon X. Eguia. Challenges to the standard euclidean spatial model. In *Advances in Political Economy: Institutions, Modelling and Empirical Analysis*, pages 169–180. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[Elkind and Faliszewski, 2014] Edith Elkind and Piotr Faliszewski. Recognizing 1-Euclidean Preferences: An Alternative Approach. In Ron Lavi, editor, *Algorithmic Game Theory*, Lecture Notes in Computer Science, pages 146–157, Berlin, Heidelberg, 2014. Springer.

[Elkind *et al.*, 2017] Edith Elkind, Martin Lackner, and Dominik Peters. Structured preferences. In Ulle Endriss, editor, *Trends in Computational Social Choice*, chapter 10, pages 187–207. AI Access, (online), 2017.

[Eğecioğlu and Giritligil, 2013] Ömer Eğecioğlu and Ayça E. Giritligil. The impartial, anonymous, and neutral culture model: A probability model for sampling public preference structures. *The Journal of Mathematical Sociology*, 37(4):203–222, 2013.

[Faelens *et al.*, 2021] Lien Faelens, Kristof Hoorelbeke, Ruben Cambier, Jill van Put, Eowyn Van de Putte, Rudi De Raedt, and Ernst H.W. Koster. The relationship between instagram use and indicators of mental health: A systematic review. *Computers in Human Behavior Reports*, 4:100121, 2021.

[Gaiha and Gupta, 1977] P. Gaiha and S. K. Gupta. Adjacent Vertices on a Permutohedron. *SIAM Journal on Applied Mathematics*, 32(2):323–327, March 1977.

[Goodfellow *et al.*, 2016] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. http://www.deeplearningbook.org.

[Henry and Mourifié, 2013] Marc Henry and Ismael Mourifié. Euclidean revealed preferences: Testing the spatial voting model. *Journal of Applied Econometrics*, 28(4):650–666, 2013.

[Hinich and Munger, 2008] Melvin J. Hinich and Michael C. Munger. Spatial theory. In *Readings in Public Choice and Constitutional Political Economy*, pages 295–304. Springer US, Boston, MA, 2008.

[Houdré and Işlak, 2022] Christian Houdré and Ümit Işlak. A Central Limit Theorem for the Length of the Longest Common Subsequences in Random Words. *arXiv preprint arXiv:1408.1559*, March 2022.

[Houdré and Xu, 2018] Christian Houdré and Chen Xu. A Note on the Expected Length of the Longest Common Subsequences of two i.i.d. Random Permutations. *arXiv preprint arXiv:1703.07691*, June 2018.

[Humphreys and Laver, 2010] Macartan Humphreys and Michael Laver. Spatial models, cognitive metrics, and majority rule equilibria. *British Journal of Political Science*, 40(1):11–30, 2010.

[Huszár *et al.*, 2021] Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. Algorithmic amplification of politics on twitter. *Proceedings of the National Academy of Sciences*, 119(1):e2025334119, 2021.

[Johnson *et al.*, 2021] Skyler B Johnson, Matthew Parsons, Tanya Dorff, Meena S Moran, John H Ward, Stacey A Cohen, Wallace Akerley, Jessica Bauman, Joleen Hubbard, Daniel E Spratt, Carma L Bylund, Briony Swire-Thompson, Tracy Onega, Laura D Scherer, Jonathan Tward, and Angela Fagerlin. Cancer Misinformation and Harmful Information on Facebook and Other Social Media: A Brief Report. *JNCI: Journal of the National Cancer Institute*, 114(7):1036–1039, 07 2021.

[Jurafsky and Martin, 2022] Daniel Jurafsky and James H. Martin. Speech and language processing. https://web.stanford.edu/~jurafsky/slp3/ed3book_jan122022.pdf, 2022. Accessed: 2022-08-09.

[Kumar and Vassilvitskii, 2010] Ravi Kumar and Sergei Vassilvitskii. Generalized distances between rankings. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 571–580, New York, NY, USA, 2010. Association for Computing Machinery.

[Luaces *et al.*, 2015] Oscar Luaces, Jorge Díez, Thorsten Joachims, and Antonio Bahamonde. Mapping preferences into Euclidean space. *Expert Systems with Applications*, 42(22):8588–8596, December 2015.

[Miller, 2015] Nicholas R. Miller. *The spatial model of social choice and voting*, chapter 10, pages 163–181. Edward Elgar Publishing, Cheltenham, UK, 2015.

[Nagy and Anderson, 1984] William E. Nagy and Richard C. Anderson. How many words are there in printed school english? *Reading Research Quarterly*, 19(3):304–330, 1984.

[Neal, 2008] Peter Neal. The Generalised Coupon Collector Problem. *Journal of Applied Probability*, 45(3):621–629, 2008.

[Obraztsova and Elkind, 2012] Svetlana Obraztsova and Edith Elkind. Optimal manipulation of voting rules. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, page 2141–2147, Toronto, Ontario, Canada, 2012. AAAI Press.

[Obraztsova *et al.*, 2013] Svetlana Obraztsova, Edith Elkind, Piotr Faliszewski, and Arkadii Slinko. On swap-distance geometry of voting rules. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS '13, page 383–390, Richland, SC, 2013. International Foundation for Autonomous Agents and Multiagent Systems.

[Pan and Ding, 2019] Shimei Pan and Tao Ding. Social media-based user embedding: A literature review. *arXiv preprint arXiv:1907.00725*, 2019.

[Peters, 2017] Dominik Peters. Recognising multidimensional euclidean preferences. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 642–648, San Francisco, CA, USA, 2017. AAAI Press.

[Rodríguez, 2011] Gonzalo Rivero Rodríguez. Integrality and separability in multidimensional voting models: Ideology and nationalism in spanish regional elections. Technical Report 265, Juan March Institute Center for Advanced Study in the Social Sciences, 2011.

[Satuluri *et al.*, 2020] Venu Satuluri, Yao Wu, Xun Zheng, Yilei Qian, Brian Wichers, Qieyun Dai, Gui Ming Tang, Jerry Jiang, and Jimmy Lin. Simclusters: Community-based representations for heterogeneous recommendations at twitter. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3183–3193, New York, NY, USA, 2020. Association for Computing Machinery.

[Small *et al.*, 2021] Christopher Small, Michael Bjorkegren, Timo Erkkilä, Lynette Shaw, and Colin Megill. Polis: Scaling deliberation by mapping high dimensional opinion spaces. *Recerca. Revista de Pensament i Anàlisi*, 26(2):1–26, 2021.

[Twitter, 2022] Twitter. Community Notes Guide: Note Ranking. https://twitter.github.io/communitynotes/ranking-notes/, 2022. Accessed: 2023-01-17.

[Tydrichová, 2023] Magdaléna Tydrichová. *Structural and algorithmic aspects of preference domain restrictions in collective decision making: Contributions to the study of single-peaked and Euclidean preferences*. PhD thesis, Sorbonne Université, March 2023.