Information Loss in Euclidean Preference Models

Luke Thorburn, Maria Polukarov, Carmine Ventre

King’s College London
{luke.thorburn, maria.polukarov, carmine.ventre}@kcl.ac.uk

Abstract
Spatial models of preferences, in the form of vector embeddings, are learned by many deep learning 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, Bogomolnaia and Laslier(2007) showed that there exist 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 realistic situations in which almost all preference profiles cannot be represented with the Euclidean model, and derive a theoretical lower bound on the information lost when approximating non-representable preferences with the Euclidean model. Our results have implications for the interpretation and use of vector embeddings, because in some cases close approximation of arbitrary, true preferences is possible only if the dimensionality of the embeddings is a substantial fraction of the number of individuals or alternatives.

1 Introduction
Accurate modelling of preferences is critical to developing and governing AI safely and efficiently (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, whenever embeddings are intended to capture degrees of similarity between elements of a set, they can be viewed as spatial models 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 Xexéq 2019) can thus be viewed as models of preference. Spatial models of (literal) preferences are also used in the fields of political science (Hinich and Munger 2008), social choice theory (Miller 2015) and opinion dynamics (Aydogdu, Mcquade, and Duteil 2017), as well as in preference aggregation software such as Polis (Small et al. 2021) and Twitter Birdwatch (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. On average, how much information is lost 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.

Contributions. We extend the results of Bogomolnaia and Laslier (2007) by proving a series of theoretical bounds. In-
tuitively, these bounds indicate that when the dimensionality of the Euclidean model is small relative to the number of individuals and the number of alternatives, 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 information lost when approximating a randomly chosen preference in a Euclidean model of given dimensionality, where we formalize information loss 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 information lost when the dimensionality is too small.

**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 [2016] 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. 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, Egia [2013] questions the validity of the utility functions, indifference curves and separability of preferences that are implied by the Euclidean model. Henry and Mourifie [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, Rodriguez 2011, Egia 2011, 2013].

A general review of structured preference models is given by Elkind, Lackner, and Peters [2017].

**Outline.** Section 2 introduces our notation and definitions. Section 3 presents answers to the three questions listed above, including our theoretical bounds, which are then proved in Section 4. Implications are discussed in Section 5.

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 $\pi_i$ (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 set 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 $a \in \{1, \ldots, A\}$, and $w^i \in \mathbb{R}^d$ for all $i \in \{1, \ldots, I\}$ such that, for all alternatives $a, b$ and individuals $i$, $a >_i b \iff \|x^a - w^i\| < \|x^b - w^i\|$ (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 $\mathcal{P}_{A,I}$.

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 $I$.

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, this probability would be the proportion of profiles that exhibit that pathology, which is a lower bound on the proportion of profiles that are not $d$-Euclidean. Here, we consider one class of such pathological sub-profiles.

**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

$$a_1 >_1 a_2 >_2 \cdots >_k a_{k-1} >_1 a_k$$

$$a_2 >_2 a_3 >_3 \cdots >_k a_{k-1} >_2 a_1$$

$$\vdots$$

$$a_k >_k a_1 >_1 \cdots >_k a_{k-2} >_k a_{k-1}.$$  

**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 $\Pi$ 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 to be beyond the current capabilities of combinatorics, but there is some related work. For example, consider a particular circulant pathology constructed using fixed subset of all $A$ alternatives, with a fixed reference order for individual 1. 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 preferences used in the pathology. The probability that this specific pathology arises is the probability that when choosing $I$ preferences 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 (Buch and Ma 2014, Houdré and Xu 2018, Houdré and Islak 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 $P(C)$ be the probability that a profile chosen uniformly from $\mathcal{P}_{A,I}$ contains a circulant pathology of size $k \geq d + 2$. Then,

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

where

$$B_k = \binom{I}{k} \frac{k}{d+2} (d+2)! \left( \frac{1}{(d+2)!} \right)^k \left(1 - \frac{d+2}{(d+2)!} \right)^{I-k}$$

and $\{\binom{k}{d+2}\}$ denotes a Stirling number of the second kind.

The quantity $B_k$ is the probability that a particular version of pathology arises. 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? Do we actually lose that much information by approximating them with a Euclidean preference model?
3.3 How much information is lost?
In this section, we consider the question from the perspective of individual preferences, rather than complete profiles. Specifically, we ask: if one already possesses a profile consisting of \( I^* \) unique preferences, and then observes a new preference generated uniformly from among all \( A! \) possible preferences, on average how closely will one be able to approximate this new preference within a \( d \)-dimensional Euclidean model, if the model must also approximate the existing \( I^* \) preferences? Formally, let \( \pi \) be the newly generated, \((I^* + 1)\)th preference and \( \hat{\pi} \) be the closest possible approximation to it in such a Euclidean preference model. As an interpretable measure of information loss, we use the metric

\[
d(\pi, \hat{\pi}) = \# \text{ adjacent swaps required to transform } \pi \text{ to } \hat{\pi}.
\]

We are interested in both \( \mathbb{E}[d(\pi, \hat{\pi})] \) and \( \mathbb{E}[d(\pi, \hat{\pi})]/\binom{A}{2} \), which is the expected number of adjacent swaps required as a proportion of the maximum possible number of swaps between any two preferences. Given the lack of efficient algorithms 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 \( d \). For all following results, we will abbreviate \( r_{A,d} \) to \( r \) (the dependence on \( A \) and \( d \) is implied), and will assume that \( I^* \geq r \) because, in this case, the probability that \( \pi \) is representable is fixed. 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 \( \pi \) 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 \( \mathbb{E}[d(\pi, \hat{\pi})] \). 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). Our overestimate for \( \hat{r} \) is given by the following lemma.

**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

\[
r \leq \hat{r} = (1 - \mathbb{P}(E)) A!,
\]

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

\[
\mathbb{P}(E) = \mathbb{P}(D_1) + \mathbb{P}(D_2) \mathbb{P}(E | \bar{D}_2),
\]

where the conditional probability is defined recursively, as follows. For \( N < M - 2 \),

\[
\mathbb{P} \left( E \mid \bigcup_{n=1}^{N} \bar{D}_n \right) = \mathbb{P} \left( D_{N+1} \mid \bigcup_{n=1}^{N} \bar{D}_n \right) \mathbb{P} \left( E \mid \bigcup_{n=1}^{N} \bar{D}_n \right) + \mathbb{P} \left( \bar{D}_{N+1} \mid \bigcup_{n=1}^{N} \bar{D}_n \right) \mathbb{P} \left( E \mid \bigcup_{n=1}^{N} \bar{D}_n \right),
\]

and for \( N = M - 2 \),

\[
\mathbb{P} \left( E \mid \bigcup_{n=1}^{N} \bar{D}_n \right) = \mathbb{P} \left( D_{M-1} \mid \bigcup_{n=1}^{N} \bar{D}_n \right).
\]

How small is \( \hat{r} \), in real terms? Figure 2 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 information do we lose when approximating such preferences in a Euclidean preference model?

**Theorem 4** (lower bound on information loss). Let \( A, d \) be fixed positive integers such that \( d < A - 1 \), \( \pi \) be a preference chosen uniformly at random from the set of \( A! \) possible preferences, \( \hat{\pi} \) be the nearest representable preference (that is, the 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

\[
\mathbb{E}[d(\pi, \hat{\pi})] \geq \sum_{k=0}^{K} \binom{A! - n_{k,A} A!}{(A!)} \mathbb{I}(\hat{r} < A! - n_{k,A}),
\]

where \( ()_r \) denotes a falling factorial, \( \mathbb{I}(\cdot) \) the indicator function, and \( n_{k,A} = \min\{(A - 1)^k, A!\} \).
How high is the bound in real terms? Figure 3(a) plots some values of the lower bound on $E[d(\pi, \hat{\pi})]$, and Figure 3(b) plots the equivalent scaled values, $E[d(\pi, \hat{\pi})]/(\frac{A}{2})$. In both cases, we can see that expected information loss becomes more severe as $d/A \to 0$, but that the bound on the scaled loss appears less severe for larger values of $A$.

4 Proofs

Proof of Theorem 3. We build up the bound step by step. 

First, consider the case where a fixed subset of $d + 2$ alternatives $a_1, \ldots, a_{d+2}$ is used to construct the pathology. Further assume that the particular permutation of these alternatives implied by the sub-preference of individual 1 in the definition of the circulant pathology is fixed. Thus, we are considering only versions of the pathology that arise from a specific set of alternatives, and imply a specific permutation of those alternatives (modulo the circulant operation). For this version of the pathology to arise, we need at least one individual to have each of the following sub-preferences.

\[
\begin{align*}
    a_1 & > a_2 > \ldots > a_{d+2} \\
    a_2 & > a_3 > \ldots > a_1 \\
    \vdots & \vdots \\
    a_{d+2} & > a_1 > \ldots > a_{d+1}
\end{align*}
\]

We will call these the necessary sub-preferences. The probability that this version of the pathology arises in a profile chosen uniformly at random from $P_{A,I}$ is

\[
\begin{align*}
\left( \frac{I}{k} \right) \binom{k}{d+2} (d+2)! & \times \frac{1}{(d+2)!} \times \left( 1 - \frac{d+2}{(d+2)!} \right)^{I-k} \\
\end{align*}
\]

For brevity, we will use $B_k$ to refer to this binomial-like expression. The number of individuals involved in the pathology, $k$, must be at least $d+2$, and can be as high as $I$. Further, because we explicitly require that the $I-k$ individuals not involved in the pathology do not have any of the necessary sub-preferences, 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 $k$.

Ideally, we would be able to generalize this expression to account for other versions of the pathology that use the same fixed set of alternatives, but imply a different reference permutation (modulo the circulant operation). This is not easy to do because we have already counted the necessary sub-preferences for these versions of the pathology among the non-necessary sub-preferences in the expression of $B_k$. Put another way, the occurrences of versions of the pathology with different reference permutations are not independent.

However, versions of the pathology constructed using entirely different subsets of alternatives do occur independently. This is because the subsets of alternatives can be permuted within a preference order without affecting the relative positions of alternatives that are not part of that subset.

Thus, 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. Using the independence property from the previous paragraph, the probability increases to

\[
1 - \left( 1 - \sum_{k=d+2}^{I} B_k \right)^{\left\lfloor \frac{A}{d+2} \right\rfloor}.
\]

This expression is the probability that at least 1 of each of $\left\lfloor \frac{A}{d+2} \right\rfloor$ versions of the pathology, each constructed using a disjoint set of alternatives, occurs. However, it does not account for all possible subsets of alternatives that could be used, or all possible ways that those subsets could be permuted. Thus, it is a lower bound on $P(C)$, the probability that any version of the pathology is contained in a randomly chosen profile. □

Proof of Lemma 7. The intuition for the proof is as follows. We want to start with the full set of $A!$ possible preferences, and then “ban” the minimum number of preferences required to guarantee that the circulant pathology does not arise. What is this minimum number?

We need to only review circulant pathologies of size $d+2$ (the smallest possible), as any circulant pathologies of size $k > d+2$ would contain a size $d+2$ pathology as a sub-profile.

For every combination of $d+2$ alternatives, there are $(d+1)!$ circular permutations of those alternatives. (A circular permutation is a unique ordering of the alternatives on a circle.) Each of these circular permutations characterizes a way in which the circulant pathology might arise.

For each circular permutation of $d+2$ alternatives, there are $d+2$ unique sub-preferences over those alternatives that are consistent with that circular permutation. To avoid the pathology, we need to ban one of these necessary sub-preferences for every combination and circular permutation of $d+2$ alternatives.

At first, determining how many (complete) preferences are affected by this set of sub-preference bans appears difficult, because a preference may be consistent with more than one of the banned sub-preferences, and these banned
sub-preferences may involve some of the same alternatives. However, it turns out that the task is equivalent to computing the probability of a particular event related to the positions of some specific alternatives within a random permutation of all alternatives, and that this computation is tractable.

Consider each set of \( d + 2 \) sub-preferences (corresponding to each combination and circular permutation of \( d + 2 \) alternatives) that could form the circular pathology. We must ban one of these \( d + 2 \) sub-preferences. Without loss of generality, we choose to ban the sub-preference that ranks most highly the alternative with the minimum index. For example, of the set of sub-preferences

\[
\begin{align*}
  a_6 &> a_{13} > a_2 \\
a_{13} &> a_2 > a_6 \\
a_2 &> a_6 > a_{13}
\end{align*}
\]

we would ban the third sub-preference. 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. If (and only if) \( E \) occurs, then there is a low-index alternative ranked sufficiently highly to ensure that there will be enough high-index alternatives ranked below it to induce at least one of the banned sub-preferences.

It now remains to compute \( P(E) \). There is no closed form expression, but it can be written down using a recursive formula that corresponds to repeated application of the law of total probability. Specifically, if \( M = A - d \), then

\[
P(E) = P(D_1) + P(D_0)P(E \mid D_0)
\]

where the conditional probability is defined recursively, as follows. For \( N < M - 2 \),

\[
P \left( E \mid \bigcup_{n=1}^{N} D_n \right) = P \left( D_{N+1} \mid \bigcup_{n=1}^{N} D_n \right) + P \left( D_{N+1} \bigcap \bigcup_{n=1}^{N} D_n \right) P \left( E \mid \bigcup_{n=1}^{N+1} D_n \right),
\]

and for \( N = M - 2 \),

\[
P \left( E \mid \bigcup_{n=1}^{N} D_n \right) = P \left( D_{M-1} \bigcup_{n=1}^{M-2} D_n \right).
\]

The above expressions look a bit intimidating but the basic idea—the law of total probability—is quite simple. We compute the probability that \( a_1 \) is in the top \( A - d - 1 \) positions, or if not, that \( a_2 \) is in the top \( A - d - 2 \) positions, or if not, that \( a_3 \) is in the top \( A - d - 3 \) positions, and so on. The relevant probabilities are analytically simple to compute. Explicitly,

\[
P \left( D_n \mid \bigcup_{n=1}^{N} D_n \right) = 1 - \frac{\prod_{k=d+n+1}^{A} (k-n)}{\prod_{k=d+n+1}^{A} (k-n+1)}
\]

The second equality comes from a combinatorial argument: for both the numerator and the denominator, write out how many alternatives are "available" to populate each position in the permutation, given the conditioning constraints, and multiply them together.

**Proof of Theorem 4.** The distribution of the \( \hat{r} \) "representable" preferences over the vertices of the permutohedron that would minimize the expectation \( E[d(\pi, \hat{\pi})] \) is the uniform distribution. Thus, to lower bound the expectation, we 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 \( d(\pi, \hat{\pi}) \). For any non-negative integer \( k \),

\[
F(k) = 1 - \frac{(A! - n_kA)!}{(A!)_{\hat{r}}} 1(\hat{r} < A! - n_kA),
\]

where the probability that none of the \( r \) "representable" preferences fall on the \( n_kA \) "reachable" vertices.
where $(\cdot)_!$ denotes a falling factorial, $1(\cdot)$ 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 formula for $n_{k,A}$ arises from the fact that each additional swap makes accessible at most a factor of $A - 1$ additional unique preferences, though this is a very loose upper bound because many of those will be reachable with fewer swaps. The minimum is used because there are only $A!$ vertices on the permutohedron. The form of $F$ is analogous to the exact distribution function of $d(\pi, \hat{\pi})$ 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 vertices that are reachable within $k$ swaps (Figure 4). Because of these substitutions, as well as the assumption of the uniform distribution, we have

$$P(d(\pi, \hat{\pi}) \leq k) \leq F(k).$$

Now, using the well known formula that expresses the expectation of a non-negative random variable in terms of its distribution function, we can turn this into a bound on the expectation. For any non-negative integer $K < \left(\frac{A}{2}\right)$,

$$E[d(\pi, \hat{\pi})] = \sum_{k=0}^{\infty} P(d(\pi, \hat{\pi}) > k)$$

$$= \sum_{k=0}^{\infty} 1 - P(d(\pi, \hat{\pi}) \leq k)$$

$$\geq \sum_{k=0}^{\infty} 1 - F(k) \geq K \sum_{k=0}^{\infty} 1 - F(k),$$

which is the bound in the theorem.

## 5 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 information loss 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, in some cases, the expected error when approximating an additional, arbitrary preference in the Euclidean model is at least 7% of the maximum possible variation between any two preferences.

### Implications

Our lower bound on the expected information loss has been proven for $I^* \geq r$, and is largest when $d \ll \min\{I, A\}$. These conditions can be met, for example, in social choice models of national elections, or in models of reputation systems where there are a large number of Sybils. In such settings, our bound shows that close approximation of true underlying preference profiles cannot be expected. The amount of information lost will depend on the particular profile, and can be reduced by increasing the dimensionality, $d$. Our bound provides a means by which to quantify a 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 $\approx 10^3$ 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 (e.g. 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 those embeddings are compared using cosine similarity or Euclidean distance.

### Future Work

Our results pave the way for a research agenda useful to inform the choice of the dimensionality of embeddings of Euclidean preference models. As discussed, the bound on the expected information loss 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 information loss 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, 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 information loss 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 also to derive meaningful upper bounds.
6 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).

References

Almeida, F.; and Xexó, G. 2019. Word Embeddings: A Survey. arXiv:1901.09069.

Aydogdu, A.; Mcquade, S. T.; and Duteil, N. P. 2017. Opinion Dynamics on a General Compact Riemannian Manifold. Networks and Heterogeneous Media, 12(3): 489 – 523.

Bennett, J. F.; and Hays, W. L. 1960. Multidimensional Unfolding: Determining the Dimensionality of Ranked Preference Data. Psychometrika, 25(1): 27–43.

Bogomolnaia, A.; and Laslier, J.-F. 2007. Euclidean Preferences. Journal of Mathematical Economics, 43(2): 87–98.

Bukh, B.; and Ma, J. 2014. Longest Common Subsequences in Sets of Words. arXiv:1406.7017.

Christian, B. 2020. The Alignment Problem: Machine Learning and Human Values. WW Norton & Company.

Eguia, J. X. 2011. Foundations of Spatial Preferences. Journal of Mathematical Economics, 47(2): 200–205.

Eguia, J. X. 2013. Challenges to the Standard Euclidean Spatial Model. In Advances in Political Economy: Institutions, Modelling and Empirical Analysis, 169–180. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-35239-3.

Elkind, E.; and Faliszewski, P. 2014. Recognizing 1-Euclidean Preferences: An Alternative Approach. In Lavi, R., ed., Algorithmic Game Theory, Lecture Notes in Computer Science, 146–157. Berlin, Heidelberg: Springer. ISBN 978-3-662-44803-8.

Elkind, E.; Lackner, M.; and Peters, D. 2017. Structured Preferences. In Endriss, U., ed., Trends in Computational Social Choice, chapter 10, 187–207. AI Access.

Faelens, L.; Hoorelbeke, K.; Cambier, R.; van Put, J.; Van de Putte, E.; De Raedt, R.; and Koster, E. H. 2021. The Relationship Between Instagram Use and Indicators of Mental Health: A Systematic Review. Computers in Human Behavior, 100121.

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

Henry, M.; and Mourifié, I. 2013. Euclidean Revealed Preferences: Testing the Spatial Voting Model. Journal of Applied Econometrics, 28(4): 650–666.

Hinich, M. J.; and Munger, M. C. 2008. Spatial Theory, In Readings in Public Choice and Constitutional Political Economy, 295–304. Boston, MA: Springer US. ISBN 978-0-387-75870-1.

Houdré, C.; and Işlak, Ü. 2022. A Central Limit Theorem for the Length of the Longest Common Subsequences in Random Words. arXiv:1408.1559.

Houdré, C.; and Xu, C. 2018. A Note on the Expected Length of the Longest Common Subsequences of Two i.i.d. Random Permutations. arXiv:1703.07691.

Humphreys, M.; and Laver, M. 2010. Spatial Models, Cognitive Metrics, and Majority Rule Equilibria. British Journal of Political Science, 40(1): 11–30.

Huszár, F.; Ktena, S. I.; O’Brien, C.; Belli, L.; Schlaikjer, A.; and Hardt, M. 2021. Algorithmic Amplification of Politics on Twitter. Proceedings of the National Academy of Sciences, 119(1): e2025334119.

Johnson, S. B.; Parsons, M.; Dorff, T.; Moran, M. S.; Ward, J. H.; Cohen, S. A.; Akerley, W.; Bauman, J.; Hubbard, J.; Spratt, D. E.; Bylund, C. L.; Swire-Thompson, B.; Onega, T.; Scherer, L. D.; Tward, J.; and Fagerlin, A. 2021. Cancer Misinformation and Harmful Information on Facebook and Other Social Media: A Brief Report. JNCl: Journal of the National Cancer Institute, 114(7): 1036–1039.

Jurafsky, D.; and Martin, J. H. 2022. Speech and Language Processing. https://web.stanford.edu/~jurafsky/slp3/ed3book_jan122022.pdf Accessed: 2022-08-09.

Luaces, O.; Díez, J.; Joachims, T.; and Bahamonde, A. 2015. Mapping Preferences into Euclidean Space. Expert Systems with Applications, 42(22): 8588–8596.

Miller, N. 2015. The Spatial Model of Social Choice and Voting. In Heckelman, J. C.; and Miller, N. R., eds., Handbook of Social Choice and Voting, chapter 10. Cheltenham, UK: Edward Elgar Publishing. ISBN 9781783470723.

Nagy, W. E.; and Anderson, R. C. 1984. How Many Words Are There in Printed School English? Reading Research Quarterly, 19(3): 304–330.

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

Pan, S.; and Ding, T. 2019. Social Media-based User Embedding: A Literature Review. arXiv:1907.00725.

Peters, D. 2016. Recognising Multidimensional Euclidean Preferences. arXiv:1602.08109 [cs].

Rodríguez, G. R. 2011. Integrity 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.

Satuluri, V.; Wu, Y.; Zheng, X.; Qian, Y.; Wickers, B.; Dai, Q.; Tang, G. M.; Jiang, J.; and Lin, J. 2020. 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, 3183–3193. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379984.

Small, C.; Bjorkgren, M.; Erkkilä, T.; Shaw, L.; and Megill, C. 2021. Polis: Scaling Deliberation by Mapping High Dimensional Opinion Spaces. Recerca. Revista de Pensament i Ànàlisi, 26(2): 1–26.

Twitter. 2022. Birdwatch Guide: Note Ranking. https://twitter.github.io/birdwatch/ranking-notes/ Accessed: 2022-07-28.