Statistics and Computing
https://doi.org/10.1007/s11222-019-09859-z

# Antithetic and Monte Carlo kernel estimators for partial rankings

M. Lomelí[1] · M. Rowland[2] · A. Gretton[3] · Z. Ghahramani[1,4]

Received: 25 July 2018 / Accepted: 2 February 2019


**Abstract**
In the modern age, rankings data are ubiquitous and they are useful for a variety of applications such as recommender systems, multi-object tracking and preference learning. However, most rankings data encountered in the real world are incomplete, which prevent the direct application of existing modelling tools for complete rankings. Our contribution is a novel way to extend kernel methods for complete rankings to partial rankings, via consistent Monte Carlo estimators for Gram matrices: matrices of kernel values between pairs of observations. We also present a novel variance-reduction scheme based on an antithetic variate construction between permutations to obtain an improved estimator for the Mallows kernel. The corresponding antithetic kernel estimator has lower variance, and we demonstrate empirically that it has a better performance in a variety of machine learning tasks. Both kernel estimators are based on extending kernel mean embeddings to the embedding of a set of full rankings consistent with an observed partial ranking. They form a computationally tractable alternative to previous approaches for partial rankings data. An overview of the existing kernels and metrics for permutations is also provided.

**Keywords** Reproducing kernel Hilbert space · Partial rankings · Monte Carlo · Antithetic variates · Gram matrix

## 1 Motivation

Permutations play a fundamental role in statistical modelling and machine learning applications involving rankings and preference data. A ranking over a set of objects can be encoded as a permutation; hence, kernels for permutations are useful in a variety of machine learning applications involving rankings such as recommender systems, multi-object tracking and preference learning. It is of interest to construct a kernel in the space of the data in order to capture similarities between datapoints and thereby influence the pattern of generalisation. A kernel input is required for the maximum mean discrepancy (MMD) two-sample test (Gretton et al. 2012), kernel principal component analysis (kPCA) (Schölkopf et al. 1999), support vector machines (Boser et al. 1992; Cortes and Vapnik 1995), Gaussian processes (GPs) (Rasmussen and Williams 2006) and agglomerative clustering (Duda and Hart 1973), among others.

Our main contributions are: (1) a novel and computationally tractable way to deal with incomplete or partial rankings by first representing the marginalised kernel (Haussler 1999) as a kernel mean embedding of a set of full rankings consistent with an observed partial ranking. We then propose two estimators that can be represented as the corresponding empirical mean embeddings; (2) a Monte Carlo kernel estimator that is based on sampling independent and identically distributed rankings from the set of consistent full rankings given an observed partial ranking; (3) an antithetic variate construction for the marginalised Mallows kernel that gives a lower variance estimator for the kernel Gram matrix. The Mallows kernel has been shown to be an expressive kernel; in

**Electronic supplementary material** The online version of this article (https://doi.org/10.1007/s11222-019-09859-z) contains supplementary material, which is available to authorized users.

✉ M. Lomelí
maria.lomeli@eng.cam.ac.uk

M. Rowland
mr504@cam.ac.uk

A. Gretton
arthur.gretton@gmail.com

Z. Ghahramani
zoubin@eng.cam.ac.uk

[1] Computational and Biological Learning Lab, University of Cambridge, Cambridge, UK

[2] Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Cambridge, UK

[3] Gatsby Computational Neuroscience Unit, University College London, London, UK

[4] Uber AI Labs, San Francisco, USA

Published online: 21 February 2019

particular, Mania et al. (2016) show that the Mallows kernel is an example of a universal and characteristic kernel, and hence, it is a useful tool to distinguish samples from two different distributions, it achieves the Bayes risk when used in kernel-based classification/regression (Sriperumbudur et al. 2011). Jiao and Vert (2015) have proposed a fast approach for computing the Kendall marginalised kernel; however, this kernel is not characteristic (Mania et al. 2016) and hence has limited expressive power.

The resulting estimators are used for a variety of kernel machine learning algorithms in the Experiments section. In particular, we present comparative simulation results demonstrating the efficacy of the proposed estimators for an agglomerative clustering task, a hypothesis test task using the maximum mean discrepancy (MMD) (Gretton et al. 2012) and a Gaussian process classification task. For the latter, we extend some of the existing methods in the software library GPy (GPy 2012).

Since the space of permutations is an example of a discrete space, with a non-commutative group structure, the corresponding reproducing kernel Hilbert spaces (RKHS) have only recently being investigated; see Kondor et al. (2007), Fukumizu et al. (2009), Kondor and Barbosa (2010), Jiao and Vert (2015) and Mania et al. (2016). First, we provide an overview of the connection between kernels and certain semimetrics when working on the space of permutations. This connection allows us to obtain kernels from given semimetrics or semimetrics from existing kernels. We can combine these semimetric-based kernels to obtain novel, more expressive kernels which can be used for the proposed Monte Carlo kernel estimator.

## 2 Definitions

We first briefly introduce the theory of permutation groups. A particular application of permutations is to use them to represent rankings; in fact, there is a natural one-to-one relationship between rankings of $n$ items and permutations (Stanley 2000). For this reason, we sometimes use ranking and permutation interchangeably. In this section, we state some mathematical definitions to formalise the problem in terms of the space of permutations.

Let $[n] = \{1, 2, \ldots, n\}$ be a set of indices for $n$ items, for some $n \in \mathbb{N}$. Given a ranking of these $n$ items, we use the notation $\succ$ to denote the ordering of the items induced by the ranking, so that for distinct $i, j \in [n]$, if $i$ is preferred to $j$, we will write $i \succ j$. Note that for a full ranking, the corresponding relation $\succ$ is a total order on $\{1, \ldots, n\}$.

We now outline the correspondence between rankings on $[n]$ and the permutation group $S_n$ that we use throughout the paper. In words, given a full ranking of $[n]$, we will associate it with the permutation $\sigma \in S_n$ that maps each ranking position $1, \ldots, n$ to the correct object under the ranking. More mathematically, given a ranking $a_1 \succ \cdots \succ a_n$ of $[n]$, we may associate it with the permutation $\sigma \in S_n$ given by $\sigma(j) = a_j$ for all $j = 1, \ldots, n$. For example, the permutation corresponding to the ranking on [3] given by $2 \succ 3 \succ 1$ corresponds to the permutation $\sigma \in S_3$ given by $\sigma(1) = 2, \sigma(2) = 3, \sigma(3) = 1$. This correspondence allows the literature relating to kernels on permutations to be leveraged for problems involving the modelling of ranking data.

In the next section, we first review some semimetrics on $S_n$ because of the existing relationship between semimetrics with an additional property and kernels. We state such relationship in Theorem 1.

### 2.1 Metrics for permutations and properties

**Definition 1** Let $\mathcal{X}$ be any set and $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a function, which we write $d(x, y)$ for every $x, y \in \mathcal{X}$. Then $d$ is a *semimetric* if it satisfies the following conditions, for every $x, y \in \mathcal{X}$ (Dudley 2002):

(i) $d(x, y) = d(y, x)$, that is, $d$ is a symmetric function.
(ii) $d(x, y) = 0$ if and only if $x = y$.
A *semimetric* is a *metric* if it satifies:
(iii) $d(x, z) \leq d(x, y) + d(y, z)$ for every $x, y, z \in \mathcal{X}$, that is, $d$ satisfies the triangle inequality.

The following are some examples of semimetrics on the space of permutations $S_n$ (Diaconis 1988). All semimetrics in bold have the additional property of being of negative type. Theorem 1 shows that negative-type semimetrics are closely related to kernels. This is because the semimetric can be written as the Hilbert space norm of a feature embedding and the kernel is the inner product for such feature embedding.

(1) *Spearman's footrule*

$$d_f(\sigma, \sigma') = \sum_{i=1}^{n} |\sigma(i) - \sigma'(i)| = \|\sigma - \sigma'\|_1.$$

(2) ***Spearman's rank correlation***

$$d_\rho(\sigma, \sigma') = \sum_{i=1}^{n} (\sigma(i) - \sigma'(i))^2 = \|\sigma - \sigma'\|_2^2.$$

(3) ***Hamming distance***

$$d_H(\sigma, \sigma') = \#\{i | \sigma(i) \neq \sigma'(i)\}.$$

It can also be defined as the minimum number of substitutions required to change one permutation into the other.

(4) *Cayley distance*

$$d_C(\sigma, \sigma') = \sum_{j=1}^{n-1} X_j(\sigma \circ (\sigma')^{-1}),$$

where the composition operation of the permutation group $S_n$ is denoted by $\circ$ and $X_j(\sigma \circ (\sigma')^{-1}) = 0$ if $j$ is the largest item in its cycle and is equal to 1 otherwise (Irurozki et al. 2016b). It is also equal to the minimum number of pairwise transpositions taking $\sigma$ to $\sigma'$. Finally, it can also be shown to be equal to $n - C(\sigma \circ (\sigma')^{-1})$ where $C(\eta)$ is the number of cycles in $\eta$.

(5) ***Kendall distance***

$$d_\tau(\sigma, \sigma') = n_d(\sigma, \sigma'),$$

where $n_d(\sigma, \sigma')$ is the number of discordant pairs for the permutation pair $(\sigma, \sigma')$. It can also be defined as the minimum number of pairwise adjacent transpositions taking $\sigma^{-1}$ to $(\sigma')^{-1}$.

(6) $l_p$ *distances*

$$d_p(\sigma, \sigma') = \left(\sum_{i=1}^{n} |\sigma(i) - \sigma'(i)|^p\right)^{\frac{1}{p}} = \|\sigma - \sigma'\|_p,$$

with $p \geq 1$.

(7) $l_\infty$ *distances*

$$d_\infty(\sigma, \sigma') = \max_{1 \leq i \leq n} |\sigma(i) - \sigma'(i)| = \|\sigma - \sigma'\|_\infty.$$

**Definition 2** A *semimetric* is said to be of *negative type* if for all $n \geq 2$, $x_1, \ldots, x_n \in \mathcal{X}$ and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ with $\sum_{i=1}^n \alpha_i = 0$, we have

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j d(x_i, x_j) \leq 0. \tag{1}$$

In general, if we start with a Mercer kernel for permutations, that is, a symmetric and positive-definite function $k : S_n \times S_n \to \mathbb{R}$, the following expression gives a semimetric $d$ that is of negative type

$$d_k(\sigma, \sigma')^2 = k(\sigma, \sigma) + k(\sigma', \sigma') - 2k(\sigma, \sigma'). \tag{2}$$

Berlinet and Thomas-Agnan (2004) and Shawer-Taylor and Cristianini (2004) provide in-depth treatments about Mercer kernels and reproducing kernel Hilbert spaces (RKHS); see "Appendix A" for a short overview. A useful characterisation of semimetrics of negative type is given by the following theorem, which states a connection between negative-type metrics and a Hilbert space feature representation or feature map $\Phi$.

**Theorem 1** (Berg et al. 1984) *A semimetric $d$ is of negative type if and only if there exists a Hilbert space $\mathcal{H}$ and an injective map $\Phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, x' \in \mathcal{X}$, $d(x, x') = \|\Phi(x) - \Phi(x')\|^2_{\mathcal{H}}$.*

Once the feature map from Theorem 1 is found, we can directly take its inner product to construct a kernel. For instance, Jiao and Vert (2015) propose an explicit feature representation for Kendall kernel given by

$$\Phi(\sigma) = \left(\frac{1}{\sqrt{\binom{n}{2}}}\left[\mathbb{I}_{\{\sigma(i) > \sigma(j)\}} - \mathbb{I}_{\{\sigma(i) < \sigma(j)\}}\right]\right)_{1 \leq i < j \leq n}.$$

They show that the inner product between two such features is a positive-definite kernel. The corresponding metric, given by Kendall distance, can be shown to be the square of the norm of the difference of feature vectors. Hence, by Theorem 1, it is of negative type.

Analogously, Mania et al. (2016) propose an explicit feature representation for the Mallows kernel, given by

$$\Phi(\sigma) = \left(\frac{1 - \exp(-v)}{2}\right)^{\frac{1}{2}\binom{n}{2}} \left(\frac{1 - \exp(-v)}{1 + \exp(-v)}\right)^{\frac{r}{2}} \prod_{i=1}^{r} \bar{\Phi}(\sigma)_{s_i}$$

where $\bar{\Phi}(\sigma)_{s_i} = 2\mathbb{I}_{\{\sigma(a_i) < \sigma(b_i)\}-1}$ when $s_i = (a_i, b_i)$ and $\bar{\Phi}(\sigma)_\emptyset = 2^{\frac{1}{2}\binom{n}{2}}(1 + \exp(-v))^{\frac{1}{2}\binom{n}{2}}$.

In the following proposition, an explicit feature representation for the Hamming distance is introduced and we show that it is a distance of negative type.

**Proposition 1** *The Hamming distance is of negative type with*

$$d_H(\sigma, \sigma') = \frac{1}{2} Trace\left[\left(\Phi(\sigma) - \Phi(\sigma')\right)\left(\Phi(\sigma) - \Phi(\sigma')\right)^T\right] \tag{3}$$

*where the corresponding feature representation is a matrix given by*

$$\Phi(\sigma) = \begin{pmatrix} \mathbb{I}_{\{\sigma(1)=1\}} & \cdots & \mathbb{I}_{\{\sigma(n)=1\}} \\ \mathbb{I}_{\{\sigma(1)=2\}} & \cdots & \mathbb{I}_{\{\sigma(n)=2\}} \\ \vdots & \cdots & \vdots \\ \mathbb{I}_{\{\sigma(1)=n\}} & \cdots & \mathbb{I}_{\{\sigma(n)=n\}} \end{pmatrix}.$$

***Proof*** The Hamming distance can be written as a square difference of indicator functions in the following way

$$d_H(\sigma, \sigma') = \#\{i | \sigma(i) \neq \sigma'(i)\} = \frac{1}{2} \sum_{i=1}^{n} \sum_{\ell=1}^{n} \left(\mathbb{I}_{\{\sigma(i)=\ell\}} - \mathbb{I}_{\{\sigma'(i)=\ell\}}\right)^2$$

where each indicator is one whenever the given entry of the permutation is equal to the corresponding element of the identity element of the group. Let the $\ell$th feature vector be $\phi_\ell(\sigma) = \left(\mathbb{I}_{\{\sigma(1)=\ell\}}, \ldots, \mathbb{I}_{\{\sigma(n)=\ell\}}\right)$, then

$$
\begin{aligned}
&= \frac{1}{2}\sum_{\ell=1}^{n}(\phi_\ell(\sigma) - \phi_\ell(\sigma'))^T(\phi_\ell(\sigma) - \phi_\ell(\sigma') \\
&= \frac{1}{2}\sum_{\ell=1}^{n}\|\phi_\ell(\sigma) - \phi_\ell(\sigma')\|^2 \\
&= \frac{1}{2}\text{Trace}\left[\left(\Phi(\sigma) - \Phi(\sigma')\right)\left(\Phi(\sigma) - \Phi(\sigma')\right)^T\right].
\end{aligned}
$$

This is the trace of the difference of the product of the feature matrices $\Phi(\sigma) - \Phi(\sigma')$, where the difference of feature matrices is given by

$$
\begin{pmatrix}
\mathbb{I}_{\{\sigma(1)=1\}} - \mathbb{I}_{\{\sigma'(1)=1\}} & \cdots & \mathbb{I}_{\{\sigma(n)=1\}} - \mathbb{I}_{\{\sigma'(n)=1\}} \\
\mathbb{I}_{\{\sigma(1)=2\}} - \mathbb{I}_{\{\sigma'(1)=2\}} & \cdots & \mathbb{I}_{\{\sigma(n)=2\}} - \mathbb{I}_{\{\sigma'(n)=2\}} \\
\vdots & \vdots & \vdots \\
\mathbb{I}_{\{\sigma(1)=n\}} - \mathbb{I}_{\{\sigma'(1)=n\}} & \cdots & \mathbb{I}_{\{\sigma(n)=n\}} - \mathbb{I}_{\{\sigma'(n)=n\}}
\end{pmatrix}.
$$

This is the square of the usual Frobenius norm for matrices, by Theorem 1, and the Hamming distance is of negative type. □

Another example is Spearman's rank correlation, which is a semimetric of negative type since it is the square of the usual Euclidean distance (Berg et al. 1984).

The two alternative definitions given for some of the distances in the previous examples are handy from different perspectives. One is an expression in terms of either an injective or non-injective feature representation, whilst the other is in terms of the minimum number of operations to change one permutation to the other. Other distances can be defined in terms of this minimum number of operations, and they are called *editing metrics* (Deza and Deza 2009). Editing metrics are useful from an algorithmic point of view, whereas metrics defined in terms of feature embeddings are useful from a theoretical point of view. Ideally, having a particular metric in terms of both algorithmic and theoretical descriptions gives a better picture of which are the relevant characteristics of the permutation that the metric takes into account (Fig. 1). For instance, Kendall and Cayley distances algorithmic descriptions correspond to the bubble and quick sort algorithms, respectively (Knuth 1998).

Another property shared by most of the semimetrics in the examples is the following

**Definition 3** Let $\sigma_1, \sigma_2 \in S_n$, $(S_n, \circ)$ denote the symmetric group of degree n with the composition operation, a *right-invariant* semimetric (Diaconis 1988) satisfies

$$
d(\sigma_1, \sigma_2) = d(\sigma_1 \circ \eta, \sigma_2 \circ \eta) \quad \forall\ \sigma_1, \sigma_2, \eta \in S_n. \tag{4}
$$

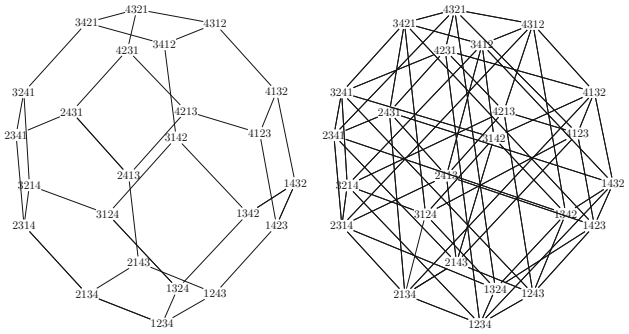


**Fig. 1** Kendall and Cayley distances for permutations of $n = 4$. There is an edge between two permutations in the graph if they differ by one adjacent or non-adjacent transposition, respectively

In particular, if we take $\eta = \sigma_1^{-1}$, then $d(\sigma_1, \sigma_2) = d(e, \sigma_2 \circ \sigma_1^{-1})$, where $e$ corresponds to the identity element of the permutation group.

This property is inherited by the *distance-induced kernel* from Sect. 2.2, Example 7. This symmetry is analogous to translation invariance for kernels defined in Euclidean spaces.

## 2.2 Kernels for $S_n$

If we specify a symmetric and positive-definite function or kernel $k$, it corresponds to defining an implicit feature space representation of a ranking data point. The well-known *kernel trick* exploits the implicit nature of this representation by performing computations with the kernel function explicitly, rather than using inner products between feature vectors in high or even infinite-dimensional space. Any symmetric and positive-definite function uniquely defines an underlying Reproducing Kernel Hilbert Space (RKHS); see the supplementary material Appendix A for a brief overview about the RKHS. Some examples of kernels for permutations are the following

1. The *Kendall kernel* (Jiao and Vert 2015) is given by

$$
k_\tau(\sigma, \sigma') = \frac{n_c(\sigma, \sigma') - n_d(\sigma, \sigma')}{\binom{d}{2}},
$$

   where $n_c(\sigma, \sigma')$ and $n_d(\sigma, \sigma')$ denote the number of concordant and discordant pairs between $\sigma$ and $\sigma'$, respectively.
2. The *Mallows kernel* (Jiao and Vert 2015) is given by

$$
k_\lambda(\sigma, \sigma') = \exp(-\lambda n_d(\sigma, \sigma')).
$$

3. The *Polynomial kernel of degree m* (Mania et al. 2016) is given by

$$k_P^{(m)}(\sigma, \sigma') = (1 + k_\tau(\sigma, \sigma'))^m.$$

4. The *Hamming kernel* is given by

$$k_H(\sigma, \sigma') = \text{Trace}\left[\left(\Phi(\sigma)\Phi(\sigma')\right)^T\right].$$

5. An *exponential semimetric kernel* is given by

$$k_{\exp}(\sigma, \sigma') = \exp\left\{-\lambda d(\sigma, \sigma')\right\},$$

where $d$ is a semimetric of negative type.

6. The *diffusion kernel* (Kondor and Barbosa 2010) is given by

$$k_\beta(\sigma, \sigma') = \exp\left\{\beta q(\sigma \circ \sigma')\right\},$$

where $\beta \in \mathbb{R}$ and $q$ is a function that must satisfy $q(\pi) = q(\pi^{-1})$ and $\sum_\pi q(\pi) = 0$. A particular case is $q(\sigma, \sigma') = 1$ if $\sigma$ and $\sigma'$ are connected by an edge in some Cayley graph representation of $S_n$, and $q(\sigma, \sigma') = -\text{degree}_\sigma$ if $\sigma = \sigma'$ or $q(\sigma, \sigma') = 0$ otherwise.

7. The *semimetric or distance-induced kernel* (Sejdinovic et al. 2013): if the semimetric $d$ is of negative type, then, a family of kernels $k$, parameterised by a central permutation $\sigma_0$, is given by

$$k_d(\sigma, \sigma') = \frac{1}{2}\left[d(\sigma, \sigma_0) + d(\sigma', \sigma_0) - d(\sigma, \sigma')\right].$$

If we choose any of the above kernels by itself, it will generally not be complex enough to represent the ranking data's generating mechanism. However, we can benefit from the allowable operations for kernels to combine kernels and still obtain a valid kernel. Some of the operations which render a valid kernel are the following: sum, multiplication by a positive constant, product, polynomial and exponential (Berlinet and Thomas-Agnan 2004).

In the case of the symmetric group of degree $n$, $S_n$, there exist kernels that are *right invariant*, as defined in Equation (4). This invariance property is useful because it is possible to write down the kernel as a function of a single argument and then obtain a Fourier representation. The caveat is that this Fourier representation is given in terms of certain matrix unitary representations due to the non-Abelian structure of the group (James 1978). Even though the space is finite, and every irreducible representation is finite-dimensional (Fukumizu et al. 2009), these Fourier representations do not have closed-form expressions. For this reason, it is difficult to work on the spectral domain in contrast to the $\mathbb{R}^n$ case. There is also no natural measure to sample from such as the one provided by Bochner's theorem in Euclidean spaces (Wendland 2005). In the next section, we will present a novel Monte Carlo kernel estimator for the case of partial rankings data.

## 3 Partial rankings

Having provided an overview of kernels for permutations, and reviewed the link between permutations and rankings of objects, we now turn to the practical issue that in real data sets, we typically have access only to partial ranking information, such as pairwise preferences and top-$k$ rankings. Partial rankings can be obtained from pairwise comparisons data given certain assumptions. For instance, a classic generative model for pairwise comparisons that can be used to obtain top$k$ rankings is the Bradley–Terry model (Bradley and Terry 1952) and its extension to multiple comparisons, the Plackett–Luce model (Luce 1959; Plackett 1974). See (Chen et al. 2017) for details on how to obtain a top$k$ partial ranking given pairwise comparisons from the Bradley–Terry model and (Caron et al. 2014) for a nonparametric Bayesian extension of the Plackett–Luce model and references therein. In the following, as Jiao and Vert (2015), we assume that our data are partial rankings of the following types

**Definition 4** (*Exhaustive partial rankings, top-k rankings*) Let $n \in \mathbb{N}$. A partial ranking on the set $[n]$ is specified by an ordered collection $\Omega_1 \succ \cdots \succ \Omega_l$ of disjoint non-empty subsets $\Omega_1, \ldots, \Omega_l \subseteq [n]$, for any $1 \leq l \leq n$. The partial ranking $\Omega_1 \succ \cdots \succ \Omega_l$ encodes the fact that the items in $\Omega_i$ are preferred to those in $\Omega_{i+1}$, for $i = 1, \ldots, l-1$. A partial ranking $\Omega_1 \succ \cdots \succ \Omega_l$ with $\cup_{i=1}^l \Omega_i = [n]$ termed *exhaustive*, as all items in $[n]$ are included within the preference information. A top-$k$ partial ranking is a particular type of exhaustive ranking $\Omega_1 \succ \cdots \succ \Omega_l$, with $|\Omega_1| = \cdots = |\Omega_{l-1}| = 1$, and $\Omega_l = [n] \setminus \cup_{i=1}^{l-1} \Omega_i$. We will frequently identify a partial ranking $\Omega_1 \succ \cdots \succ \Omega_l$ with the set $R(\Omega_1, \ldots, \Omega_l) \subseteq S_n$ of full rankings consistent with the partial ranking. Thus, $\sigma \in R(\Omega_1, \ldots, \Omega_l)$ iff for all $1 \leq i < j \leq l$, and for all $x \in \Omega_i$, $y \in \Omega_j$, we have $\sigma^{-1}(x) < \sigma^{-1}(y)$. When there is potential for confusion, we will use the term "subset partial ranking" when referring to a partial ranking as a subset of $S_n$, and "preference partial ranking" when referring to a partial ranking with the notation $\Omega_1 \succ \cdots \succ \Omega_l$.

Several interpretations are compatible with this definition; for instance, scenarios in which no preference information is known about items within a particular $\Omega_i$ are possible, as well as are scenarios where the preferences of all items in a particular $\Omega_i$ are ties. Thus, for many practical problems,

we require definitions of kernels between subsets of partial rankings rather than between full rankings, to be able to handle data sets containing only partial ranking information. A common approach (Tsuda et al. 2002) is to take a kernel $K$ defined on $S_n$, and use the *marginalised kernel*, defined on subsets of partial rankings by

$$K(R, R') = \sum_{\sigma \in R} \sum_{\sigma' \in R'} K(\sigma, \sigma') p(\sigma|R) p(\sigma'|R') \tag{5}$$

for all $R, R' \subseteq S_n$, for some probability distribution $p \in \mathscr{P}(S_n)$. Here, $p(\cdot|R)$ denotes the conditioning of $p$ to the set $R \subseteq S_n$. If some prior information about the distribution of complete rankings is available, it can be used to define a kernel over partial rankings with a non-uniform distribution. For instance, let $\sigma \in S_n$ be a permutation, its probability mass function under a *Mallows distribution* (Mallows 1957), given a metric $d : S_n \times S_n \to \mathbb{R}$, a location parameter $\sigma_0 \in S_n$, and a scale parameter $\theta > 0$, is

$$p(\sigma \mid R) = \frac{\exp\{-\theta d(\sigma, \sigma_0)\}}{\psi(\theta)} \mathbb{I}_{\{\sigma \in R\}},$$

with normalising constant $\psi(\theta) = \sum_{\sigma \in S_n} \exp\{-\theta d(\sigma, \sigma_0)\} \times \mathbb{I}_{\{\sigma \in R\}}$. This family of probability distributions has been extensively studied for the full rankings case, when $R = S_n$; see Fligner and Verducci (1986), Mukherjee (2016) and Busse et al. (2007) for mixtures of Mallows distributions. The R package "PerMallows" (Irurozki et al. 2016) provides random number generators based on different algorithms to sample from a Mallows distribution parameterised by different distance functions. These sampling procedures are not straightforwardly applicable to the partial rankings case. There have been various extensions for top$k$ partial rankings such as Lebanon and Mao (2008), who propose a nonparametric estimator based on kernel smoothing; Chierichetti et al. (2018), who extended the Mallows model by defining a distance measure directly over top$k$ rankings; and Vitelli et al. (2017), who developed a Bayesian framework for inference using a Metropolis–Hastings algorithm, among others. We assume that we do not have any prior information about the generative process of full rankings; hence, we only deal with the case of the marginalised kernel from Equation (5), in which we take the probability mass function to be uniform over each of the partial rankings denoted by $R, R'$. The corresponding kernel is given by

$$K(R, R') = \frac{1}{|R||R'|} \sum_{\sigma \in R} \sum_{\sigma' \in R'} K(\sigma, \sigma'). \tag{6}$$

Jiao and Vert (2015) also use this kernel and called it the *convolution kernel* (Haussler 1999) between partial rankings. In general, the use of a marginalised kernel quickly becomes computationally intractable, with the number of terms in the right-hand side of Eq. (5) growing super-exponentially with $n$, for a fixed number of items in the partial rankings $R$ and $R'$; see "Appendix E" for a table that illustrates such growth. An exception is the Kendall kernel case for two interleaving partial rankings of $k$ and $m$ items or a top-$k$ and top-$m$ ranking. In this case, the sum can be tractably computed and it can be done in $\mathcal{O}(k \log k + m \log m)$ time (Jiao and Vert 2015).

We propose a variety of Monte Carlo methods to estimate the marginalised kernel of Eq. (5) for the general case, where direct calculation is intractable.

**Definition 5** The Monte Carlo estimator approximating the marginalised kernel of Eq. (5) is defined for a collection of partial rankings $(R_i)_{i=1}^I$, given by

$$\widehat{K}(R_i, R_j) = \frac{1}{M_i M_j} \sum_{l=1}^{M_i} \sum_{m=1}^{M_j} w_l^{(i)} w_m^{(j)} K(\sigma_l^{(i)}, \sigma_m^{(j)}) \tag{7}$$

for $i, j = 1, \ldots, I$, where $((\sigma_n^{(i)})_{m=1}^{M_i})_{i=1}^I$ are random permutations and $\left((w_m^{(i)})_{m=1}^{M_i}\right)_{i=1}^I$ are random weights. Note that this general setup allows for several possibilities:

- For each $i = 1 \ldots, I$, the permutations $(\sigma_m^{(i)})_{m=1}^{M_i}$ are drawn exactly from the distribution $p(\cdot|R_i)$. In this case, the weights are simply $w_n^{(i)} = 1$ for $m = 1, \ldots, M_i$.
- For each $i = 1, \ldots, I$, the permutations $(\sigma_m^{(i)})_{m=1}^{M_i}$ drawn from some proposal distribution $q(\cdot|R_i)$ with the weights given by the corresponding *importance weights* $w_n^{(i)} = p(\sigma_n^{(i)}|R)/q(\sigma_n^{(i)}|R)$ for $m = 1, \ldots, M_i$.

An alternative perspective on the estimator defined in Eq. (7), more in line with the literature on random feature approximations of kernels, is to define a random feature embedding for each of the partial rankings $(R_i)_{i=1}^I$.

More precisely, let $\mathcal{H}_K$ be the (finite-dimensional) Hilbert space associated with the kernel $K$ on the space $S_n$, and let $\boldsymbol{\Phi}$ be the associated feature map, so that $\Phi(\sigma) = K(\sigma, \cdot) \in \mathcal{H}_K$ for each $\sigma \in S_n$. Then observe that we have $K(\sigma, \sigma') = \langle \boldsymbol{\Phi}(\sigma), \boldsymbol{\Phi}(\sigma') \rangle$ for all $\sigma, \sigma' \in S_n$. We now extend this feature embedding to partial rankings as follows. Given a partial ranking $R \subseteq S_n$, we define the feature embedding of $R$ by

$$\boldsymbol{\Phi}(R) = \frac{1}{|R|} \sum_{\sigma \in R} K(\sigma, \cdot) \in \mathcal{H}_K$$

With this extension of $\boldsymbol{\Phi}$ to partial rankings, we may now directly express the marginalised kernel of Eq. (5) as an inner product in the same Hilbert space $\mathcal{H}_K$

$$K(R, R') = \langle \boldsymbol{\Phi}(R), \boldsymbol{\Phi}(R') \rangle$$

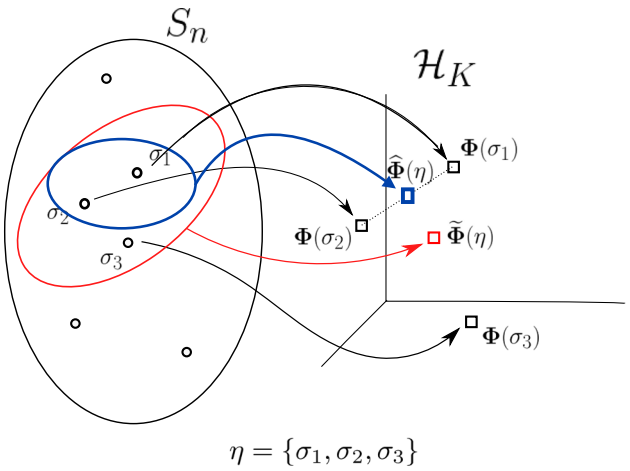


**Fig. 2** Visualisation of the various embeddings discussed in the proof of Theorem 3. $\sigma_1, \sigma_2 and \sigma_3$ are permutations in $S_n$, which are mapped into the RKHS $\mathcal{H}_K$ by the embedding $\boldsymbol{\Phi}$. $\eta$ is a partial ranking subset which contains $\sigma_1, \sigma_2, \sigma_3$, and its embedding $\widetilde{\boldsymbol{\Phi}}(\eta)$ is given as the average of the embeddings of its full rankings. The Monte Carlo embedding $\widehat{\boldsymbol{\Phi}}(\eta)$ induced by Equation (7) is computed by taking the average of a randomly sampled collection of consistent full rankings from $\eta$

for all partial rankings $R, R' \subseteq S_n$. If we define a random feature embedding of the partial rankings $(R_i)_{i=1}^I$ by

$$\widehat{\boldsymbol{\Phi}}(R_i) = \sum_{m=1}^{M_i} w_m^{(i)} \boldsymbol{\Phi}(\sigma_m^{(i)}),$$

then the Monte Carlo kernel estimator of Eq. (7) can be expressed directly as

$$\begin{aligned}
\widehat{K}(R_i, R_j) &= \frac{1}{M_i M_j} \sum_{l=1}^{M_i} \sum_{m=1}^{M_j} w_l^{(i)} w_m^{(j)} K(\sigma_l^{(i)}, \sigma_m^{(j)}) \\
&= \frac{1}{M_i M_j} \sum_{l=1}^{M_i} \sum_{m=1}^{M_j} w_l^{(i)} w_m^{(j)} \langle \boldsymbol{\Phi}(\sigma_l^{(i)}), \boldsymbol{\Phi}(\sigma_m^{(j)}) \rangle \\
&= \left\langle \frac{1}{M_i} \sum_{l=1}^{M_i} w_l^{(i)} \boldsymbol{\Phi}(\sigma_l^{(i)}), \frac{1}{M_j} \sum_{m=1}^{M_j} w_m^{(j)} \boldsymbol{\Phi}(\sigma_m^{(j)}) \right\rangle \\
&= \langle \widehat{\boldsymbol{\Phi}}(R_i), \widehat{\boldsymbol{\Phi}}(R_j) \rangle \qquad (8)
\end{aligned}$$

for each $i, j \in \{1, \ldots, I\}$. This expression of the estimator as an inner product between randomised embeddings will be useful in the sequel.

We provide an illustration of the various RKHS embeddings at play in Fig. 2, using the notation of the proof of Theorem 3. In this figure, $\eta$ is a partial ranking, with three consistent full rankings $\sigma_1, \sigma_2, and \sigma_3$. The extended embedding $\widetilde{\boldsymbol{\Phi}}$ applied to $\eta$ is the barycentre in the RKHS of the embeddings of the consistent full rankings, and a Monte Carlo approximation $\widehat{\boldsymbol{\Phi}}$ to this embedding is also displayed.

**Theorem 2** *Let $R \subseteq S_n$ be a partial ranking, and let $(\sigma_m)_{m=1}^M$ independent and identically distributed samples from $p(\cdot \mid R)$. The kernel Monte Carlo mean embedding,*

$$\widehat{\Phi}(R) = \frac{1}{M} \sum_{m=1}^{M} K(\sigma_m, \cdot) \qquad (9)$$

*is an unbiased estimator of the marginalised kernel embedding*

$$\widetilde{\Phi}(R) = \frac{1}{|R|} \sum_{\sigma \in R} K(\sigma, \cdot).$$

***Proof*** Note that the RKHS in which these embeddings take values is finite-dimensional, and the Monte Carlo estimator is the average of iid terms, each of which is equal to the true embedding in expectation. Thus, we immediately obtain unbiasedness of the Monte Carlo embedding. □

**Theorem 3** *The Monte Carlo kernel estimator from Eq. (7) does define a positive-definite kernel; further, it yields unbiased estimates of the off-diagonal elements and consistent for the diagonal elements of the kernel matrix.*

***Proof*** We first deal with the positive-definiteness claim. Let $R_1, \ldots, R_I \subseteq S_n$ be a collection of partial rankings, and for each $i = 1, \ldots, I$, let $(\sigma_m^{(i)}, w_m^{(i)})_{m=1}^{M_i}$ be an i.i.d. weighted collection of complete rankings distributed according to $p(\cdot|R_i)$. To show that the Monte Carlo kernel estimator $\widehat{K}$ is positive definite, we observe that by Eq. (8), the $I \times I$ matrix with $(i, j)$th element given by $\widehat{K}(R_i, R_j)$ is the Gram matrix of the vectors $(\widehat{\boldsymbol{\Phi}}(R_i))_{i=1}^I$ with respect to the inner product of the Hilbert space $\mathcal{H}_K$. We therefore immediately deduce that the matrix is positive semidefinite. Furthermore, the Monte Carlo kernel estimator is unbiased for the off-diagonal elements and consistent for the diagonal elements of the kernel matrix; see Appendix C in the supplementary material for the proof. □

We highlight that whilst the mean embedding estimator in Eq. (9) is unbiased, the corresponding kernel estimator is consistent for the diagonal elements of the kernel matrix and unbiased for the off-diagonal elements. Having established that the Monte Carlo estimator $\widehat{K}$ is itself a kernel, we note that when it is evaluated at two partial rankings $R, R' \subseteq S_n$, the resulting expression is *not* a sum of iid terms; the following result quantifies the quality of the estimator through its variance.

**Theorem 4** *The variance of the Monte Carlo kernel estimator evaluated at a pair of partial rankings $R_i, R_j$, with $M_i, N_j$ Monte Carlo samples, respectively, is given by*

$$
\begin{aligned}
&\mathrm{Var}\left(\widehat{K}(R_i, R_j)\right) \\
&= \frac{1}{M_i} \sum_{\sigma^{(i)} \in R_i} p(\sigma^{(i)}|R_i) \left( \sum_{\sigma^{(j)} \in R_j} p(\sigma^{(j)}|R_j) K(\sigma^{(i)}, \sigma^{(j)}) \right)^2 \\
&\times \frac{-1}{M_i} \left( \sum_{\substack{\sigma^{(i)} \in R_i \\ \sigma^{(j)} \in R_j}} K(\sigma^{(i)}, \sigma^{(j)}) p(\sigma^{(i)}|R_i) p(\sigma^{(j)}|R_j) \right)^2 \\
&\frac{-1}{M_i N_j} \sum_{\sigma^{(i)} \in R_i} p(\sigma^{(i)}|R_i) \left( \sum_{\sigma^{(j)} \in R_j} p(\sigma^{(j)}|R_j) K(\sigma^{(i)}, \sigma^{(j)}) \right)^2 \\
&+ \frac{1}{M_i N_j} \sum_{\substack{\sigma^{(i)} \in R_i \\ \sigma^{(j)} \in R_j}} K(\sigma^{(i)}, \sigma^{(j)})^2 p(\sigma^{(i)}|R_i) p(\sigma^{(j)}|R_j).
\end{aligned}
$$

The proof is given in the supplementary material, "Appendix D". We have presented some theoretical properties of the embedding corresponding to the Monte Carlo kernel estimator which confirm that it is a sensible embedding. In the next section, we present a lower variance estimator based on a novel antithetic variates construction.

## 4 Antithetic random variates for permutations

A common, computationally cheap variance-reduction technique in Monte Carlo estimation of expectations of a given function is to use antithetic variates (Hammersley and Morton 1956), the purpose of which is to introduce negative correlation between samples without affecting their marginal distribution, resulting in a lower variance estimator. Antithetic samples have been used when sampling from Euclidean vector spaces, for which antithetic samples are straightforward to define. Ross (2006) defines the antithetic of a full ranking by reversing the order of the original permutation. We give a definition of antithetic permutations for partial rankings in terms of distance maximisation and show that this coincides with the definition of Ross (2006) in the case of full rankings. We begin with a preliminary lemma, before giving the full definition of antithetic permutations given a fixed partial ranking.

**Lemma 1** *Let $R \subseteq S_n$ be a top-k partial ranking, let $\sigma \in R$. Then, there exists a unique solution to the problem*

$$\arg\max_{\sigma' \in R} d_\tau(\sigma, \sigma').$$

*Moreover, it can be calculated directly; if the preference partial ranking corresponding to R is given by $a_1 \succ \cdots \succ a_k$, so that the full ranking $\sigma \in R$ satisfies $\sigma(1) = a_1, \ldots, \sigma(k) = a_k$, then the unique distance-maximising permutation $\sigma'$ is given by*

$$
\begin{aligned}
&\sigma'(i) = a_i && \text{for } i = 1, \ldots, k, \\
&\sigma'(k+j) = \sigma(n+1-j) && \text{for } j = 1, \ldots, n-k.
\end{aligned}
$$

*In this case, we have $d_\tau(\sigma, \sigma') = \binom{n-k}{2}$.*

See "Appendix B" for the proof.

**Definition 6** (*Antithetic permutations*) Let $R \subseteq S_n$ be a top-$k$ partial ranking. The antithetic operator $A_R : R \to R$ maps each permutation $\sigma \in R$ to the permutation in $R$ of maximal Kendall distance from $\sigma$. $A_R(\sigma)$ is said to be antithetic to $\sigma$.

This definition of antithetic samples for permutations has parallels with the standard notion of antithetic samples in vector spaces, in which typically a sampled vector $x \in \mathbb{R}^d$ is negated to form $-x$, its antithetic sample; $-x$ is the vector maximising the Euclidean distance from $x$, under the restrictions of fixed norm. We note here also that the computational cost of generating an antithetic permutation via the method described in Lemma 1 is no greater than the cost associated with generating an independent permutation.

**Proposition 1** *Let $R$ be a partial ranking and $\{\sigma, A_R(\sigma)\}$ be an antithetic pair from $R$, $\sigma$ is distributed uniformly in the region $R$. Let $d_\tau : S_n \to \mathbb{R}^+$ be the Kendall distance and $\sigma_0 \in R$ a fixed permutation, let $X = d_\tau(\sigma, \sigma_0)$ and $Y = d_\tau(A_R(\sigma), \sigma_0)$, then $X$ and $Y$ have negative covariance.*

Proposition 1 is useful because one of the main tasks in statistical inference is to compute expectations of a function of interest, denoted by $h$. Once the antithetic variates are constructed, the functional form of $h$ determines whether or not the antithetic variate construction effectively produces a lower variance estimator for its expectation. The proof of this proposition is presented after the relevant lemmas are proved. If $h$ is a monotone function, we have the following corollary.

**Corollary 2** *Let $h$ be a monotone increasing (decreasing) function. Then, the random variables $h(X)$ and $h(Y)$ have negative covariance.*

***Proof*** The random variable $Y$ from Proposition 1 is equal in distribution to $Y \overset{d}{=} C - X$, where $C$ is a constant which specialises depending on whether $\sigma$ is a full ranking or an exhaustive partial ranking; see the proof of Proposition 1 in the next section for the specific form of the constant for each case. By Chebyshev's integral inequality (Fink and Jodeit 1984), the covariance between a monotone increasing (decreasing) and a monotone decreasing (increasing) functions is negative. □

The next theorem presents the antithetic empirical feature embedding and corresponding antithetic kernel estimator.

Indeed, if we take the inner product between two embeddings, this yields the kernel antithetic estimator which is a function of a pair of partial rankings subsets. In this case, the $h$ function from above is the kernel evaluated in each pair, and this is an example of a $U$-statistic (Serfling 1980, Chapter 5).

**Theorem 5** *Let $R_i \subseteq S_n$ be a partial ranking, $S_n$ denotes the space of permutations of $n \in \mathbb{N}$, $(\sigma_m^{(i)}, A_{R_i}(\sigma_m^{(i)}))_{m=1}^{M_i}$ are antithetic pairs of i.i.d. samples from the region $R_i$. The kernel antithetic Monte Carlo mean embedding*

$$\widehat{\phi}(R_i) = \frac{1}{M_i} \sum_{m=1}^{M_i} \left[ \frac{K(\sigma_m^{(i)}, \cdot) + K(A_{R_i}(\sigma_m^{(i)}), \cdot)}{2} \right]$$

*is a unbiased estimator of the embedding that corresponds to the marginalised kernel. The corresponding antithetic kernel estimator is*

$$\begin{aligned} \widehat{K}(R_i, R_j) = \frac{1}{4MN} \sum_{m=1}^{M} \sum_{n=1}^{N} \big( & K(\sigma_m^{(i)}, \sigma_n^{(j)}) \\ & + K(A_{R_i}(\sigma_m^{(i)}), \sigma_n^{(j)}) + K(\sigma_m^{(i)}, A_{R_j}(\sigma_n^{(j)}) \\ & + K(A_{R_i}(\sigma_m^{(i)}), A_{R_j}(\sigma_n^{(j)})\big) \end{aligned} \tag{10}$$

*using $M$ antithetic pairs of samples $(\sigma_m^{(i)}, A_{R_i}(\sigma_m^{(i)}))_{m=1}^{M}$ from region $R_i$ and $N$ antithetic pairs of samples $(\sigma_n^{(j)}, A_{R_j}(\sigma_n^{(j)}))_{n=1}^{N}$, from $R_j$.*

***Proof*** Since the antithetic kernel embedding is a convex combination of the Monte Carlo kernel embedding, unbiasedness follows. □

In the next section, we present the main result about the kernel estimator from Eq. (10), namely, that it has lower asymptotic variance than the Monte Carlo kernel estimator from Eq. 7 if we use the Mallows kernel.

### 4.1 Variance of the antithetic kernel estimator

We now establish some basic theoretical properties of antithetic samples in the context of marginalised kernel estimation. In order to do so, we require a series of lemmas to derive the main result in Theorem 6 that guarantees that the antithetic kernel estimator has lower asymptotic variance than the Monte Carlo kernel estimator for the marginalised Mallows kernel.

The following result shows that antithetic permutations may be used to achieve coupled samples which are marginally distributed uniformly on the subset of $S_n$ corresponding to a top-$k$ partial ranking.

**Lemma 2** *If $R \subseteq S_n$ is a top-$k$ partial ranking, then if $\sigma \sim Unif(R)$, then $A_R(\sigma) \sim Unif(R)$.*

See "Appendix B" for the proof. Lemma 2 establishes a base requirement of an antithetic sample—namely, that it has the correct marginal distribution. In the context of antithetic sampling in Euclidean spaces, this property is often trivial to establish, but the discrete geometry of $S_n$ makes this property less obvious. Indeed, we next demonstrate that the condition of exhaustiveness of the partial ranking in Lemma 2 is necessary.

***Example 1*** Let $n = 3$, and consider the partial ranking $2 \succ 1$. Note that this is not an exhaustive partial ranking, as the element 3 does not feature in the preference information. There are three full rankings consistent with this partial ranking, namely $3 \succ 2 \succ 1$, $2 \succ 3 \succ 1$, and $2 \succ 1 \succ 3$. Encoding these full rankings as permutations, as described in the correspondence outlined in Sect. 2, we obtain three permutations, which we, respectively, denote by $\sigma_A, \sigma_B, \sigma_C \in S_3$. Specifically, we have

$$\begin{aligned} &\sigma_A(1) = 3, \quad \sigma_A(2) = 2, \quad \sigma_A(3) = 1. \\ &\sigma_B(1) = 2, \quad \sigma_B(2) = 3, \quad \sigma_A(3) = 1. \\ &\sigma_C(1) = 2, \quad \sigma_C(2) = 1, \quad \sigma_A(3) = 3. \end{aligned}$$

Under the right-invariant Kendall distance, we obtain pairwise distances given by

$$\begin{aligned} &d_\tau(\sigma_A, \sigma_B) = 1, \\ &d_\tau(\sigma_A, \sigma_C) = 2, \\ &d_\tau(\sigma_B, \sigma_C) = 1. \end{aligned}$$

Thus, the marginal distribution of an antithetic sample for the partial ranking $2 \succ 1$ places no mass on $\sigma_B$, and half of its mass on each of $\sigma_A$ and $\sigma_C$, and is therefore not uniform over $R$.

We further show that the condition of right invariance of the metric $d$ is necessary in the next example.

***Example 2*** Let $n = 3$, and suppose $d$ is a distance on $S_3$ such that, with the notation introduced in Example 1, we have

$$\begin{aligned} &d(\sigma_A, \sigma_B) = 1, \\ &d(\sigma_A, \sigma_C) = 0.5, \\ &d(\sigma_B, \sigma_C) = 1. \end{aligned}$$

Note that $d$ is not right invariant, since

$$\begin{aligned} &d((\sigma_A, \sigma_C) \\ &\quad = d(\sigma_B \nu, \sigma_A \nu) \\ &\quad \neq d(\sigma_B, \sigma_A), \end{aligned}$$

where $\nu \in S_3$ is given by $\nu(1) = 1, \nu(2) = 3, \nu(3) = 2$. Then, note that an antithetic sample for the kernel associated

with this distance and the partial ranking $1 \succ 2$ is equal to $\sigma_B$ with probability $2/3$ and the other two full rankings with probability $1/6$ each and therefore does not have a uniform distribution.

Examples 1 and 2 serve to illustrate the complexity of antithetic sampling constructions in discrete spaces. Finally, we remark that an alternative phrasing of Lemma 2 is that the pushforward of the distribution $\text{Unif}(R)$ through the function $A_R$ is again $\text{Unif}(R)$. Whilst it may be possible to design distributions such that $p(\cdot|R)$ has this property for each top-$k$ ranking $R \subseteq S_n$, many commonly used non-uniform distributions over permutations, such as Mallows models, do not satisfy this property.

We now begin direct calculation with antithetic permutations and partial rankings. We primarily focus on the case of top-$k$ rankings, as calculation turns out to be particularly tractable in this case and also due to the fact that top-$k$ rankings feature in many applications of interest. The following two lemmas state some useful relationships between the distance between two permutations $(\sigma, \nu)$ and the corresponding pair $(A_R(\sigma), \nu)$ in both the unconstrained and constrained cases which correspond to not having any partial ranking information and having partial ranking information, respectively.

**Lemma 3** *Let $\sigma, \nu \in S_n$. Then, $d_\tau(\sigma, \nu) = \binom{n}{2} - d_\tau(A_{S_n}(\sigma), \nu)$.*

***Proof*** This is immediate from the interpretation of the Kendall distance as the number of discordant pairs between two permutations; a distinct pair $i, j \in [n]$ is discordant for $\sigma, \nu$ iff they are concordant for $A_{S_n}(\sigma), \nu$. □

In fact, Lemma 3 generalises in the following manner.

**Lemma 4** *Let $R$ be a top-$k$ ranking $a_1 \succ \cdots \succ a_l \succ [n] \setminus \{a_1, \ldots, a_l\}$, and let $\sigma, \nu \in R$. Then $d_\tau(\sigma, \nu) = \binom{n-l}{2} - d_\tau(A_R(\sigma), \nu)$.*

See "Appendix B" for the proof. Next, we show that it is possible to obtain a unique closest element in a given partial ranking set $R$, denoted by $\Pi_R(\nu)$, with respect to any given permutation $\nu \in S_n, \nu \notin R$. This is based on the usual generalisation of a distance between a set and a point (Dudley 2002). We then use such closest element in Lemmas 6 and 7 to obtain useful decompositions of distances identities. Finally, in Lemma 8 we verify that the closest element is also distributed uniformly on a subset of the original set $R$.

**Lemma 5** *Let $R \subseteq S_n$ be a top-$k$ partial ranking, let $\nu \in S_n$ be arbitrary. There is a unique closest element in $R$ to $\nu$. In other words, $\arg\min_{\sigma \in R} d_\tau(\sigma, \nu)$ is a set of size 1.*

See "Appendix B" for the proof.

**Definition 7** Let $R \subseteq S_n$ be a top-$k$ partial ranking. Let $\Pi_R : S_n \to R$ be the map that takes a permutation to the corresponding Kendall-closest permutation in $R$; by Lemma 5, this is well defined.

**Lemma 6** *Let $\sigma \in R$, and $\nu \in S_n$. We have the following decomposition of the distance $d(\sigma, \nu)$*

$$d_\tau(\sigma, \nu) = d_\tau(\sigma, \Pi_R(\nu)) + d_\tau(\Pi_R(\nu), \nu).$$

See "Appendix B" for the proof.

**Lemma 7** *Let $\sigma \in R$, and let $\nu \in R'$. We have the following relationship between $d_\tau(A_R(\sigma), \nu)$ and $d_\tau(\sigma, \nu)$*

$$d_\tau(A_R(\sigma), \nu) = d_\tau(\sigma, \nu) + \binom{n-k}{2} - 2d_\tau(\sigma, \Pi_R(\nu)). \tag{11}$$

See "Appendix B" for the proof.

**Lemma 8** *Let $R, R' \subseteq S_n$ be top-$k$ rankings, in preference notation given by*

$$R : a_1 \succ \cdots \succ a_l \succ [n] \setminus \{a_1, \ldots, a_l\},$$
$$R' : b_1 \succ \cdots \succ b_m \succ [n] \setminus \{b_1, \ldots, b_m\}.$$

*If $\nu \sim \text{Unif}(R')$, then $\Pi_R(\nu)$ is a full ranking with distribution $\text{Unif}(R'')$, where $R'' \subseteq R$ is the partial ranking given by*

$$R'' : a_1 \succ \cdots \succ a_l \succ b_{i_1} \succ \cdots \succ b_{i_q} \succ [n] \setminus \{a_1, \ldots, a_l, b_1, \ldots, b_m\},$$

*where $\{b_{i_1}, \ldots, b_{i_q}\} = \{b_1, \ldots, b_m\} \setminus \{a_1, \ldots, a_l\}$, and $i_j < i_{j+1}$ for all $j = 1, \ldots, q-1$.*

See "Appendix B" for the proof.

Having introduced the antithetic operator for a top-$k$ partial ranking $R$, $A_R : R \to R$ and the projection map $\Pi_R : S_n \to R$, we next study how these operations interact with one another.

**Lemma 9** *Let $R'' \subseteq R \subseteq S_n$ be top-$k$ partial rankings. Then for $\sigma \in R$, we have*

$$A_{R''}(\Pi_{R''}(\sigma)) = \Pi_{R''}(A_R(\sigma)).$$

See "Appendix B" for the proof.

Finally, the last lemma states the most general identity for a distance, which involves the antithetic operator, the closest element map given a partial rankings set $R$ and a subset of it, denoted by $R''$.

**Lemma 10** *Let $R'' \subseteq R \subseteq S_n$ be top-k partial rankings, given in preference notation by*

$$R : a_1 \succ \cdots \succ a_l \succ [n] \setminus \{a_1, \ldots, a_l\},$$
$$R'' : a_1 \succ \cdots \succ a_l \succ a_{l+1} \succ \cdots a_m \succ [n] \setminus \{a_1, \ldots, a_m\}.$$

*Let $\alpha$ be the number of unranked elements under $R$, and let $\beta$ be the additional number of elements ranked under $R''$ relative to $R$. Then for $\sigma \in R$, we have*

$$d_\tau(\sigma, \Pi_{R''}(\sigma)) = ((n-l)-(m-l))(m-l) + \binom{m-l}{2} - d_\tau(A_R(\sigma), \Pi_{R''}(A_R(\sigma))).$$

See "Appendix B" for the proof.

***Proof of Proposition 1*** **Case** $\sigma_0 \in S_n$ be the fixed permutation, then

$$\mathrm{Cov}\,(d_\tau(\sigma, \sigma_0), d_\tau(A_R(\sigma), \sigma_0)) < 0.$$

This holds true since
$d_\tau(A_R(\sigma), \sigma_0) = \binom{n}{2} - d_\tau(\sigma, \sigma_0), \forall \sigma \in S_n, \forall n \in \mathbb{N}$ by Lemma 3.
**Case** $\emptyset \subset \mathbf{R}$: Let $\sigma_0 \in R$, we have that
$d_\tau(A_R(\sigma), \sigma_0) = \binom{n-k}{2} - d_\tau(\sigma, \sigma_0)\ \forall \sigma_0 \in R$ by Lemma 4. □

In general, if $\sigma_0 \notin R$, by Lemma 7, $d_\tau(A_R(\sigma), \sigma_0) = d_\tau(\sigma, \sigma_0) + \binom{n-k}{2} - 2d_\tau(\sigma, \Pi_{R_i}(\sigma_0))$.

After proving all the relevant Lemmas, we now present our main result regarding antithetic samples, namely, that this scheme provides negatively correlated pairs of samples.

**Theorem 6** *Consider the antithetic kernel estimator for the Mallows kernel evaluated on a pair of partial rankings $R_i, R_j$ using $M$ antithetic pairs of samples $(\sigma_m^{(i)}, A_{R_i}(\sigma_m^{(i)}))_{m=1}^M$ from region $R_i$ and $N$ antithetic pairs of samples $(\sigma_n^{(j)}, A_{R_j}(\sigma_n^{(j)}))_{n=1}^N$, from $R_j$. The asymptotic variance of this estimator is lower than the kernel estimator using $2M$ (respectively, $2N$) i.i.d. samples from $R_i$ (respectively, $R_j$).*

***Proof*** It has been shown previously that the antithetic kernel estimator is unbiased (in the off-diagonal case), so showing that it has lower MSE in the antithetic case which is equivalent to showing that its second moment is smaller in the antithetic case than in the i.i.d. case. The second moment is given by

$$\begin{aligned}
&\mathbb{E}\big[\widehat{K}(R_i, R_j)^2\big] \\
&= \mathbb{E}\Bigg[\Bigg(\frac{1}{4NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\big(K(\sigma_n, \nu_m) + K(\widetilde{\sigma}_n, \nu_m) + K(\sigma_n, \widetilde{\nu}_m) + K(\widetilde{\sigma}_n, \widetilde{\nu}_m)\big)\Bigg)^2\Bigg] \\
&= \frac{1}{16M^2N^2}\sum_{n,n'=1}^{N}\sum_{m,m'=1}^{M}\mathbb{E}\Big[\big(K(\sigma_n, \nu_m) + K(\widetilde{\sigma}_n, \nu_m) + K(\sigma_n, \widetilde{\nu}_m) + K(\widetilde{\sigma}_n, \widetilde{\nu}_m)\big) \times \big(K(\sigma_{n'}, \nu_{m'}) + K(\widetilde{\sigma}_{n'}, \nu_{m'}) + K(\sigma_{n'}, \widetilde{\nu}_{m'}) + K(\widetilde{\sigma}_{n'}, \widetilde{\nu}_{m'})\big)\Big].
\end{aligned}$$

We identify three types of terms in the above sum: (i) those where $n \neq n'$ and $m \neq m'$; (ii) those where $n = n'$ but $m \neq m'$, or $m = m'$ but $n \neq n'$; (iii) those where $n = n'$ and $m = m'$.

We remark that in case (i), the 16 terms that appear in the summand all have the same distribution in the antithetic and i.i.d. case, so terms of the form (i) contribute no difference between antithetic and i.i.d.. There are $\mathcal{O}(N^2M + M^2N)$ terms of the form (ii) and $\mathcal{O}(NM)$ terms of the form (iii). We thus refer to terms of the form (ii) as cubic terms and terms of the form (iii) as quadratic terms. We observe that due to the proportion of cubic terms to quadratic terms diverging as $N, M \to \infty$, it is sufficient to prove that each cubic term is less in the antithetic case than the i.i.d. case to establish the claim of lower MSE.

Thus, we focus on cubic terms. Let us consider a term with $n = n'$ and $m \neq m'$. The term has the form

$$\mathbb{E}\Bigg[\Big(K(\sigma_n, \nu_m) + K(\widetilde{\sigma}_n, \nu_m) + K(\sigma_n, \widetilde{\nu}_m) + K(\widetilde{\sigma}_n, \widetilde{\nu}_m)\Big) \times \Big(K(\sigma_n, \nu_{m'}) + K(\widetilde{\sigma}_n, \nu_{m'}) + K(\sigma_n, \widetilde{\nu}_{m'}) + K(\widetilde{\sigma}_n, \widetilde{\nu}_{m'})\Big)\Bigg].$$

Of the sixteen terms appearing in the expectation above, there are only two distinct distributions they may have. The two types of terms are given below:

$$\mathbb{E}\left[K(\sigma_n, \nu_m)K(\sigma_n, \nu_{m'})\right], \tag{12}$$

and

$$\mathbb{E}\left[K(\sigma_n, \nu_m)K(\widetilde{\sigma}_n, \nu_{m'})\right]. \tag{13}$$

Terms of the form in Eq. (12) have the same distribution in the antithetic and i.i.d. cases, so we can ignore these. However, terms of the form in Eq. (13) have differing distributions in these two cases, so we focus in on these. We deal specifically with the case where $K_\lambda(\sigma, \nu) = \exp(-\lambda d_\tau(\sigma, \nu))$, so we may rewrite the expression in Eq. (13) as

$$\mathbb{E}\left[\exp(-\lambda(d_\tau(\sigma_n, \nu_m) + d_\tau(\widetilde{\sigma}_n, \nu_{m'})))\right]. \tag{14}$$

We now decompose the distances $d_\tau(\sigma_n, \nu_m)$, $d_\tau(\widetilde{\sigma}_n, \nu_{m'})$ using the series of lemmas introduced before. First, we use

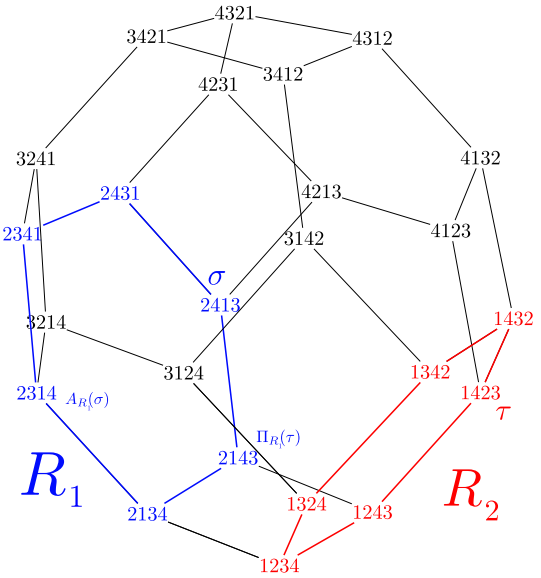


**Fig. 3** An example of the variables appearing in the decomposition in Eq. (15)

Lemma 6 to write

$$\begin{aligned} d_\tau(\sigma_n, \nu_m) &= d_\tau(\sigma_n, \Pi_{R_1}(\nu_m)) + d_\tau(\Pi_{R_1}(\nu_m), \nu_m), \\ d_\tau(\widetilde{\sigma}_n, \nu_{m'}) &= d_\tau(\widetilde{\sigma}_n, \Pi_{R_1}(\nu_{m'})) + d_\tau(\Pi_{R_1}(\nu_{m'}), \nu_{m'}). \end{aligned} \tag{15}$$

We give a small example illustrating some of the variables at play in this decomposition in Fig. 3.

Now, writing $R_3 \subseteq R_1$ for the partial ranking described by Lemma 8, we have that $\Pi_{R_1}(\nu_m), \Pi_{R_1}(\nu_{m'}) \overset{\text{i.i.d.}}{\sim} \text{Unif}(R_3)$. Therefore, the distances in Eq. (15) may be decomposed further

$$\begin{aligned} d_\tau(\sigma_n, \nu_m) &= d_\tau(\sigma_n, \Pi_{R_3}(\sigma_n)) \\ &\quad + d_\tau(\Pi_{R_3}(\sigma_n), \Pi_{R_1}(\nu_m)) \\ &\quad + d_\tau(\Pi_{R_1}(\nu_m), \nu_m), \\ d_\tau(\widetilde{\sigma}_n, \nu_{m'}) &= d_\tau(\widetilde{\sigma}_n, \Pi_{R_3}(\widetilde{\sigma}_n)) \\ &\quad + d_\tau(\Pi_{R_3}(\widetilde{\sigma}_n), \Pi_{R_1}(\nu_{m'})) \\ &\quad + d_\tau(\Pi_{R_1}(\nu_{m'}), \nu_{m'}). \end{aligned} \tag{16}$$

We now consider each term and argue as to whether the distribution is different in the antithetic and i.i.d. cases, recalling that in the i.i.d. case, $\widetilde{\sigma}_n$ is drawn from $R_1$ independently from $\sigma_n$, whilst in the antithetic case, $\widetilde{\sigma}_n = A_{R_1}(\sigma_n)$.

- Each of the terms $d_\tau(\Pi_{R_1}(\nu_m), \nu_m)$ and $d_\tau(\Pi_{R_1}(\nu_{m'}), \nu_{m'})$ has the same distribution under the i.i.d. case and antithetic case. Further, in both cases, $d_\tau(\Pi_{R_1}(\nu_m), \nu_m)$ is independent of $\Pi_{R_1}(\nu_m)$, and $d_\tau(\Pi_{R_1}(\nu_{m'}), \nu_{m'})$ is independent of $\Pi_{R_1}(\nu_{m'})$, so these two terms are independent of all others appearing in the sum in both cases.
- Each of the terms $d_\tau(\Pi_{R_3}(\sigma_n), \Pi_{R_1}(\nu_m))$ and $d_\tau(\Pi_{R_3}(\widetilde{\sigma}_n), \Pi_{R_1}(\nu_{m'}))$ has the same distribution under the i.i.d. case and the antithetic case and is independent of all other terms in both cases.
- We deal with the terms $d_\tau(\sigma_n, \Pi_{R_3}(\sigma_n))$ and $d_\tau(\widetilde{\sigma}_n, \Pi_{R_3}(\widetilde{\sigma}_n))$ using Lemma 10. More specifically, under the i.i.d. case, these two distances are clearly i.i.d.. However, under the antithetic case, the lemma tells us that the sum of these two distances is equal to the mean under the distribution of the i.i.d. case almost surely. Thus, in the antithetic case, this random variable has the same mean as in the i.i.d. case, but is more concentrated (strictly so iff $d(\sigma_n, \Pi_{R_3}(\sigma_n))$ is not a constant almost surely, which is the case iff $R_1 \neq R_3$).

Thus, $d_\tau(\sigma_n, \nu_m) + d_\tau(\widetilde{\sigma}_n, \nu_{m'})$ has the same mean under the i.i.d. and antithetic cases, but is strictly more concentrated when $R_1 \neq R_3$ This holds true iff the partial rankings $R_1$ and $R_2$ do not concern exactly the same set of objects. Thus, by a conditional version of Jensen's inequality, since $\exp(-\lambda x)$ is strictly convex as a function of $x$, we obtain the variance result. □

### 4.2 The antithetic kernel estimator and kernel herding

In this section, having established the variance-reduction properties of antithetic samples in the context of Monte Carlo kernel estimation, we now explore connections to kernel herding (Chen et al. 2010). Kernel herding is a deterministic approach to numerical integration, in which quadrature points are selected according to a distance-minimisation algorithm taking place in a particular Hilbert space.

More precisely, given an integration problem of the form $\mathbb{E}_{X\sim\mu}[f(X)]$, for some domain $\mathcal{X}$, a function $f : \mathcal{X} \to \mathbb{R}$ and probability measure $\mu \in \mathscr{P}(\mathcal{X})$, kernel herding proceeds by first selecting a kernel $K : \mathcal{X}^2 \to \mathbb{R}$. Successively, the reproducing kernel Hilbert space is chosen $\mathcal{H}_K = \overline{\text{span}\{K(x, \cdot) | x \in \mathcal{X}\}}$ with inner product defined as the unique continuous linear extension of $\langle K(x, \cdot), K(y, \cdot)\rangle_{\mathcal{H}_K} = K(x, y)$ for all $x, y \in \mathcal{X}$, and with corresponding embedding $\phi_K : \mathcal{X} \to \mathcal{H}_K$ given by $\phi_K(x) = K(x, \cdot)$ for all $x \in \mathcal{X}$. An initial quadrature point $x_1 \in \mathcal{X}$ is then specified, and then, additional quadrature points are selected iteratively according to the following rule: given $m$ quadrature points $x_{1:m}$, the next quadrature point $x_{m+1}$ is selected by

$$x_{m+1} = \arg\min_{x\in\mathcal{X}} \left\| \mathbb{E}_{X\sim\mu}[K(X, \cdot)] - \frac{1}{m}\sum_{i=1}^{m} \phi_K(x_i) \right\|^2_{\mathcal{H}_K}.$$

Our main result in this section makes clear the connection between kernel herding and our antithetic construction.

**Theorem 7** *The antithetic variate construction of Theorem 5 is equivalent to the optimal solution for the first two steps of a kernel herding procedure in the space of permutations.*

***Proof*** Let $R$ be a partial ranking of $n$ elements. We calculate the sequence of herding samples from the uniform distribution $p(\cdot|R)$ over full rankings consistent with $R$ associated with the exponential semimetric kernel $K_{\exp}(\sigma, \sigma') = \exp(-\lambda d(\sigma, \sigma'))$, for a metric $d$ of negative definite type. Following Chen et al. (2010), we note that the herding samples from $p(\cdot|R)$ associated with the kernel $K$, with RKHS embedding $\phi : S_n \to \mathcal{H}$, are defined iteratively by

$$\sigma_T = \arg\min_{\sigma_T} \left\| \mu_p - \frac{1}{T}\sum_{t=1}^{T} \phi(\sigma_t) \right\|_{\mathcal{H}}^2 \quad \text{for } T = 1, \ldots,$$

where $\mu_p$ is the RKHS mean embedding of the distribution $p$. Since $p$ is uniform over its support, any ranking $\sigma$ in the support of $p(\cdot|R)$ is a valid choice as the first sample in a herding sequence. Given such an initial sample, we then calculate the second herding sample, by considering the herding objective as follows

$$\begin{aligned} \left\| \mu_p - \frac{1}{2}\sum_{t=1}^{2} \phi(\sigma_t) \right\|_{\mathcal{H}}^2 &= \|\mu_p\|_{\mathcal{H}}^2 - \sum_{t=1}^{2} \frac{1}{|R|} \sum_{\sigma \in R} K(\sigma_t, \sigma) \\ &\quad + \frac{1}{4}\big(K_{\exp}(\sigma_1, \sigma_1) + 2K_{\exp}(\sigma_1, \sigma_2) \\ &\quad + K_{\exp}(\sigma_2, \sigma_2)\big) \end{aligned} \tag{17}$$

which, as a function of $\sigma_2$, is equal to $2K_{\exp}(\sigma_1, \sigma_2) = 2\exp(-\lambda d(\sigma_1, \sigma_2))$, up to an additive constant. Thus, selecting $\sigma_2$ to minimise the herding objective is equivalent to maximising $d(\sigma_1, \sigma_2)$, which is exactly the definition of the antithetic sample to $\sigma_1$. □

After this result, one would like to do a herding procedure for more than two steps. However, the solution is not the same as picking $k$ herding samples simultaneously. Specifically, the following counterexample, illustrated in Fig. 4, clearly shows why. The left plot shows the result of solving the herding objective for 2 samples—the result is an antithetic pair of samples for the region $R$. If a third sample is selected greedily, with these first two samples fixed, it will yield a different result than if the herding objective is solved for 3 samples simultaneously, as illustrated in the right of the figure.

***Remark 3*** Theorem 7 says that if we first pick a point uniformly at random from $R$, then put it into the herding objective and then select the second deterministically to minimise the herding objective and this is equivalent to the antithetic variate construction of Definition 6. Alternatively, we could pick the second point uniformly at random from $R$, independently from the first point. This second scheme will produce a higher value of the herding objective on average.

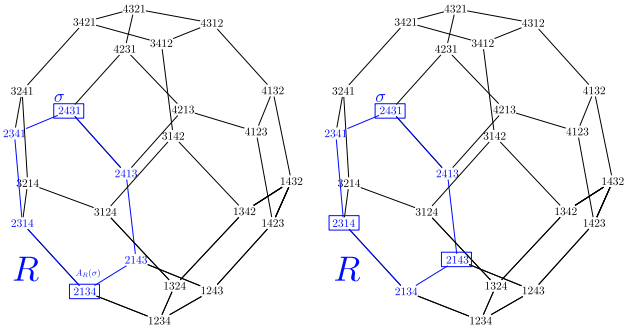


**Fig. 4** Samples from the region $R$, illustrating the difference between solving the herding objective greedily and solving for all samples simultaneously

After the two estimators for kernel matrices have been constructed, we use them in some experiments to assess their performance in the next section.

## 5 Experiments

**Algorithm 1** SampleAntitheticConsistentFullRankings

**Input:** top$-k$ partial ranking $i_1 \succ i_2 \succ \cdots \succ i_k$, degree $n$
**Returns:** two full rankings $\sigma_1, \sigma_2$ consistent with the given partial ranking
Set $\sigma_1(l) = \sigma_2(l) = i_l$ for $l = 1, \ldots, k$
Obtain a random ordering $j_1, \ldots, j_{n-k}$ of the remaining items $\{1, \ldots, n\} \setminus \{i_1, \ldots, i_k\}$
Let $b_1 < \cdots < b_{n-k}$ be the ordering of $\{1, \ldots, n\} \setminus \{i_1, \ldots, i_k\}$
Set $\sigma_1(b_l) = j_l$ for $l = 1, \ldots, n-k$
Set $\sigma_2(b_l) = j_{n-k-l+1}$ for $l = 1, \ldots, n-k$
Return $\sigma_1, \sigma_2$

In this section, we use the Monte Carlo and antithetic kernel estimators for a variety of machine learning unsupervised and supervised learning tasks: a nonparametric hypothesis test, an agglomerative clustering algorithm and a Gaussian process classifier.

Definition 6 states the antithetic permutation construction with respect to a given permutation for Kendall's distance. In order to consider partial rankings data, we should respect the observed preferences when obtaining the antithetic variate. Algorithm 1 describes how to sample an antithetic permutation and simultaneously respect the constraints imposed by the observed partial ranking. Namely, the antithetic permutation has the observed preferences fixed in the same locations as the original permutation and only reverses the unobserved

**Table 1** Tree purities for the sushi data set using a subsample of 100 users with the full Gram matrix $K$, a censored data set of $topk = 4$ partial rankings for the vanilla Monte Carlo estimator $\widehat{K}$ and the antithetic Monte Carlo estimator $\widehat{K}^a$, with $n_{mc} = 20$ Monte Carlo samples

| | Kendall | Mallows | Semiexp Hamming | Semiexp Cayley | Semiexp Spearman |
|---|---|---|---|---|---|
| $K$ Average | 0.83 | 0.75 | 0.81 | 0.72 | 0.81 |
| $\widehat{K}$ Average | 0.78 (0.052) | 0.79 (0.058) | 0.79 (0.063) | 0.82 (0.040) | 0.78 (0.062) |
| $\widehat{K}^a$ Average | NA | 0.77 (0.050) | NA | NA | NA |

Tree cut at $k = 10$ clusters. The median distance criterion was used to select the inverse of the lengthscale for the semimetric exponential kernels

locations. This corresponds to maximising the Kendall distance between the permutation pair whilst respecting the constraints and ensures that both permutations have the right marginals as stated in Lemmas 1 and 2.

### 5.1 Data sets

**Synthetic data set** The synthetic data set for the nonparametric hypothesis test experiment, where the null hypothesis is $H_0 : P = Q$ and the alternative is $H_1 : P \neq Q$, is the following: the data set from the $P$ distribution is a mixture of Mallows distributions (Diaconis 1988) with the Kendall and Hamming distances. The central permutations are given by the identity permutation and the reverse of the identity, respectively, with lengthscale equal to one. The data set from the $Q$ distribution is a sample from the uniform distribution over $S_n$, where $n = 6$.

**Sushi data set** This data set contains rankings about sushi preferences given by 5000 users (Kamishima et al. 2009). The users ranked 10 types of sushi, and the labels correspond to the user's region This data set is used for the Gaussian and ten for the agglomerative clustering task.

### 5.2 Agglomerative clustering

In this experiment, we used both the full and a censored version of the sushi data set from Sect. 5.1. We used various distances for permutations to compute the estimators for the semimetric matrix between pairs of partial rankings subsets. In order to compute our estimators, we censored the data set by storing the $topk = 4$ partial rankings per user. The Monte Carlo and antithetic kernel estimators were used to obtain negative-type semimetric matrices using the relationship from Equation (2) in the following way:

$$\widehat{D(R, R')^2} = \widehat{K}(R, R) + \widehat{K}(R', R') - 2\widehat{K}(R, R').$$

These matrices were then used as an input to the average linkage agglomerative clustering algorithm (Duda and Hart 1973). The tree purity measure is reported, and it provides way to asses the tree produced by the agglomerative clustering algorithm. It can be computed in the following way: when a dendrogram and all correct labels are given, pick uniformly at random two leaves which have the same label c and find the smallest subtree containing the two leaves. The dendrogram purity is the expected value of $\frac{\#\text{leaves with label c in subtree}}{\#\text{leaves in the subtree}}$ per class. If all leaves in the class are contained in a pure subtree, the dendrogram purity is one. Hence, values close to one correspond to high-quality trees.

In Table 1, the true and estimated purities using the full rankings and the partial rankings data sets are reported. We assumed that the true labels are given by the user's region, and there are ten different possible regions. The true purity corresponds to an agglomerative clustering algorithm using the Gram matrix obtained from the full rankings. We can compute the Gram matrix for the full rankings because we have access to all of the users' rankings over the ten different types of sushi. The antithetic Monte Carlo estimator outperforms the vanilla Monte Carlo estimator in terms of average purity since it is closer to the true purity. It also has a lower standard deviation when estimating the marginalised Mallows kernel.

### 5.3 Nonparametric hypothesis test with MMD

Let $P$ and $Q$ be probability distributions over $S_n$, the null hypothesis is $H_0 : P = Q$ versus $H_1 : P \neq Q$ using samples $\sigma_1, \ldots, \sigma_n \overset{\text{i.i.d.}}{\sim} P$ and $\sigma'_1, \ldots, \sigma'_m \overset{\text{i.i.d.}}{\sim} Q$. We can estimate a pseudometric between $P$ and $Q$ and reject $H_0$ if the observed value of the statistic is large. The following is an unbiased estimator of the $MMD^2$ (Gretton et al. 2012)

$$\begin{aligned}\widehat{MMD^2}(P, Q) = &\frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} K(\sigma_i, \sigma_j) \\ &+ \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} K(\sigma'_i, \sigma'_j) \\ &- \frac{2}{nm} \sum_{i=1}^{m} \sum_{j \neq i}^{n} K(\sigma_i, \sigma'_j). \end{aligned} \tag{18}$$

This statistic depends on the chosen kernel as can be seen in Eq. (18). If the kernel is characteristic (Sriperumbudur et al. 2011), then the $MMD^2$ is a proper metric over probability distributions. Analogously, we can compute an MMD squared estimator for partial rankings sets, such that

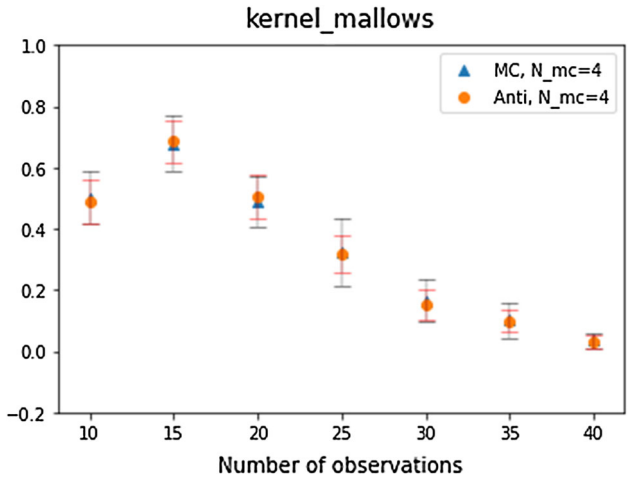


**Fig. 5** Mean $p$ values ($y$-axis) versus number of datapoints in synthetic data set ($x$-axis)

$R_1, \ldots, R_n \overset{\text{i.i.d.}}{\sim} P$ and $R'_1, \ldots, R'_m \overset{\text{i.i.d.}}{\sim} Q$, in the following way

$$\begin{aligned}\widehat{MMD^2}(P, Q) = &\frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} \hat{K}(R_i, R_j) \\ &+ \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \hat{K}(R'_i, R'_j) \\ &- \frac{2}{nm} \sum_{i=1}^{m} \sum_{j \neq i}^{n} \hat{K}(R_i, R'_j). \end{aligned} \tag{19}$$

We used the synthetic data sets for $P$ and $Q$ described in Sect. 5.1 to asses the performance of the Monte Carlo and antithetic kernel estimators in a nonparametric hypothesis test. The data sets consist of rankings over $n = 10$ objects, and we censored them to obtain top$-k$ partial rankings with $k = 3$. We then computed the MMD squared statistic for the samples using the samples from the two populations. Since the non-asymptotic distribution of the statistic from Eq. (19) is not known, we performed a permutation test (Alba Fernández et al. 2007) in order to estimate consistently the null distribution and compute the $p$ value. We did this repeatedly as we varied the number of observations for a fixed number of Monte Carlo samples to see the effect of the sample size in the $p$ value computations. Specifically, Fig. 5 and Table 2 show how the $p$ value computed with the antithetic kernel estimator has lower variance as we vary the number of observations in our data set. Both $p$ values converge to zero since the samples from both populations come from different distributions. In Table 2, we report the standard deviations of the estimated $p$ values. The $p$ value obtained with the antithetic kernel estimator has lower variance across all sample sizes.

### 5.4 Gaussian process classifier

In this experiment, two different kernels were used to compute the estimators for the Gram matrix between different pairs of partial rankings subsets. The matrix was then pro-

**Table 2** Standard deviations for $p$ values computed with the Monte Carlo and antithetic estimators

| # obs | 10 | 15 | 20 | 25 | 30 | 35 | 40 |
|---|---|---|---|---|---|---|---|
| Monte Carlo | 0.0853 | 0.0910 | 0.0830 | 0.1109 | 0.0677 | 0.0596 | 0.0236 |
| Antithetic | 0.0706 | 0.0663 | 0.0712 | 0.0594 | 0.0502 | 0.0363 | 0.0222 |

**Table 3** Averaged over 10 runs with 4 Monte Carlo samples per run, $n = 10, topk = 6$

| | Test accuracy | Train ave-loglik | Test ave-loglik |
|---|---|---|---|
| Mallows | | | |
| Full model | 0.9 | −0.2070 | −0.5457 |
| MC | 0.74 | −0.2486(0.005) | −0.563(0.020) |
| Antithetic | **0.75** | −0.262(0.001) | −0.573(0.002) |
| Gaussian | | | |
| Full model | 0.75 | −0.2215 | −0.7014 |
| MC | 0.72 | −0.2890(0.0245) | −0.5737(0.043) |
| Antithetic | NA | NA | NA |
| Kendall | | | |
| Full model | 0.7 | $-0.311(3.01\times10^{-6})$ | $-0.597(3.5\times10^{-6})$ |
| MC | 0.66 | −0.3575(0.008) | −0.7063(0.052) |
| Antithetic | NA | NA | NA |

The model in bold has the highest test accuracy when the antithetic kernel estimator was used, with respect to all models where only partial rankings data was used to compute the kernel Monte Carlo estimators of the Gram matrix

vided as the input to a Gaussian process classifier (Neal 1998). The Python library GPy (2012) was extended with custom kernel classes for partial rankings which compute both the Monte Carlo and antithetic kernel estimators for partial rankings subsets. Previously, it was only possible to do pointwise evaluations of kernels, but our implementation allows to compute the kernels over pairs of partial ranking subsets by storing the sets in a tensor first.

We used the sushi data set from Sect. 5.1 with the labels binarised in East Japan or West Japan regions. We selected a random subset of the observations of size 100 and used 80%, for the training set and 20% for the test set. In the Mallows kernel case, we used the median distance heuristic (Takeuchi et al. 2006; Schölkopf and Smola 2002) with the Kendall distance to compute the bandwidth parameter and a scale parameter of 9.5. We performed a grid search over different values of the scale parameter and picked the one that had the largest classification accuracy for the test set.

In Table 3, the results of running the Gaussian process classifier are reported using the marginalised Mallows kernel, the marginalised Gaussian kernel and the marginalised Kendall kernel as well as the corresponding estimators. Since the Mallows kernel is based on the Kendall distance, it is a kernel specifically tailored for permutations and it is the best in terms of predictive performance. In contrast, the Gaussian kernel is a kernel that is suitable for Euclidean spaces and it does not take into account the data type, and it still exhibits good predictive performance. The Kendall kernel does take into account the data type; however, it performs the worst. The full model corresponds to using the Gram matrix computed with the full rankings, and MC and antithetic refer to the Gram matrix obtained with the Monte Carlo and antithetic kernel estimators. We observe that the test and train loglikelihoods obtained with the antithetic kernel estimator have lower variance as expected.

## 6 Conclusion

We addressed the problem of extending kernels to partial rankings by introducing a novel Monte Carlo kernel estimator and explored variance-reduction strategies via an antithetic variates construction. Our schemes lead to a computationally tractable alternative to previous approaches for partial rankings data. The Monte Carlo scheme can be used to obtain an estimator of the marginalised kernel with any of the kernels reviewed herein. The antithetic construction provides an improved version of the kernel estimator for the marginalised Mallows kernel. Our contribution is noteworthy because the computation of most of the marginalised kernels grows super-exponentially with respect to the number of elements in the collection; hence, it quickly becomes intractable for relatively small values of the number of ranked items $n$. An exception is the fast approach for computing the convolution kernel proposed by Jiao and Vert (2015), which is only valid for Kendall kernel. Mania et al. (2016) have showed that the Kendall kernel is not characteristic using non-commutative Fourier analysis to show that it has a degenerate spectrum. For this reason, using other kernels for permutations might be desirable depending on the task at hand.

One possible direction for future work includes the use of explicit feature representations for traditional random features schemes to further reduce the computational cost of the Gram matrix. Another possible application is to use our method with pairwise preference data where users are not necessarily consistent about their preferences. In this type of data, we could still extract a partial ranking from a given user, then sample from the space of the corresponding full rankings consistent with this observed partial ranking and obtain our Monte Carlo kernel estimator. This would benefit from our framework because having a partial ranking is in general more informative that having pairwise comparisons or star ratings.

Another natural direction for future work is to develop variance-reduction sampling techniques for a wider variety of kernels over permutations, and to the extent the theoretical analysis of these constructions to discrete graphs more generally.

**Acknowledgements** We thank the anonymous reviewers for their valuable comments which have improved the quality of our manuscript and Ryan Adams, for insightful discussions. Maria Lomeli and Zoubin Ghahramani acknowledge support from the Alan Turing Institute (EPSRC Grant EP/N510129/1), EPSRC Grant EP/N014162/1, and donations from Google and Microsoft Research. Arthur Gretton thanks the Gatsby Charitable Foundation for financial support. Mark Rowland acknowledges support by EPSRC Grant EP/L016516/1 for the Cambridge Centre for Analysis.



## A Reproducing kernel Hilbert spaces

A reproducing kernel Hilbert space (RKHS) (Berlinet and Thomas-Agnan 2004) over a set $\mathcal{X}$ is a Hilbert space $\mathcal{H}$ consisting of functions on $\mathcal{X}$ such that for each $x \in \mathcal{X}$ there is a function $k_x \in \mathcal{H}$ with the property

$$\langle f, k_x \rangle_{\mathcal{H}} = f(x), \quad \forall f \in \mathcal{H}. \tag{20}$$

The function $k_x(\cdot) = k(x, \cdot)$ is called the *reproducing kernel* of $\mathcal{H}$ (Aronszajn 1950). The space $\mathcal{H}$ is endowed with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and a norm can be defined based

on it such that $\|f\|_{\mathcal{H}} := \sqrt{\langle \cdot, \cdot \rangle_{\mathcal{H}}}$. In order to be a Hilbert space, it needs to contain all limits of Cauchy sequences, i.e. it has to be complete. In the case of the symmetric group of degree $n$, $\mathcal{X} = S_n$, the space is finite-dimensional which guarantees that it is complete. Finally, any symmetric and positive-definite function $k_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ uniquely determines an RKHS. Alternatively, a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *kernel* if there exists a Hilbert space $\mathcal{H}$ and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, y \in \mathcal{X}, k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$. The function $\phi$ is usually referred to as the *feature representation* of $x$. Even though the RKHS induced by the kernel is unique, there can be more than one feature representations that define the same kernel.

## B Proofs

**Lemma 1** *Let $R \subseteq S_n$ be a top-$k$ partial ranking, let $\sigma \in R$. Then, there exists a unique solution to the problem*

$$\arg\max_{\sigma' \in R} d_\tau(\sigma, \sigma').$$

*Moreover, it can be calculated directly; if the preference partial ranking corresponding to $R$ is given by $a_1 \succ \cdots \succ a_k$, so that the full ranking $\sigma \in R$ satisfies $\sigma(1) = a_1, \ldots, \sigma(k) = a_k$, then the unique distance-maximising permutation $\sigma'$ is given by*

$$\begin{aligned} \sigma'(i) &= a_i && \text{for } i = 1, \ldots, k, \\ \sigma'(k + j) &= \sigma(n + 1 - j) && \text{for } j = 1, \ldots, n - k. \end{aligned}$$

*In this case, we have $d_\tau(\sigma, \sigma') = \binom{n-k}{2}$.*

***Proof*** We use the interpretation of Kendall's tau distance as counting numbers of discordant pairs. From this perspective, it straightforwardly follows that all permutations in $R$ can be at most distance $\binom{n-k}{2}$ from one another, since any pair of items involving one of the top-$k$ items must necessarily be concordant for any two permutations in $R$. It also follows straightforwardly that there is a unique permutation in $R$ for which this distance is realised, given when all pairs of items not in the top-$k$ have the opposite ordering to that under $\sigma$, from which the formula in the statement follows. □

**Lemma 2** *If $R \subseteq S_n$ is a top-$k$ partial ranking, then if $\sigma \sim Unif(R)$, then $A_R(\sigma) \sim Unif(R)$.*

***Proof*** The proof is immediate from Lemma 1, since $A_R$ is bijective on $R$. □

**Lemma 4** *Let $R$ be a top-$k$ ranking $a_1 \succ \cdots \succ a_l \succ [n] \setminus \{a_1, \ldots, a_l\}$, and let $\sigma, \nu \in R$. Then $d_\tau(\sigma, \nu) = \binom{n-l}{2} - d_\tau(A_R(\sigma), \nu)$.*

***Proof*** As for the proof of Lemma 3, we use the "discordant pairs" interpretation of the Kendall distance. Note that if a distinct pair $\{x, y\} \in [n]^{(2)}$ has at least one of $x, y \in \{a_1, \ldots, a_l\}$, then by virtue of the fact that $\sigma, A_R(\sigma), \nu \in R$, any pair of these permutations is concordant for $x, y$. Now observe that any distinct pair $x, y \in [n] \setminus \{a_1, \ldots, a_l\}$ is discordant for $\sigma, \nu$ iff it is concordant for $A_R(\sigma), \nu$, from the construction of $A_R(\sigma)$ described in Lemma 1. The total number of such pairs is $\binom{n-l}{2}$, so we have $d_\tau(\sigma, \nu) + d_\tau(A_R(\sigma), \nu) = \binom{n-l}{2}$, as required. □

**Lemma 5** *Let $R \subseteq S_n$ be a top-$k$ partial ranking, let $\nu \in S_n$ be arbitrary. There is a unique closest element in $R$ to $\nu$. In other words, $\arg\min_{\sigma \in R} d_\tau(\sigma, \nu)$ is a set of size 1.*

***Proof*** We use the interpretation of the Kendall distance as the number of discordant pairs between two permutations. Let $R$ be the top-$k$ partial ranking given by $x_1 \succ \cdots \succ x_k \succ [n] \setminus \{x_1, \ldots, x_k\}$, and let $X = \{x_1, \ldots, x_k\}$. We decompose the Kendall distance between $\sigma \in R$ and $\nu$ as follows:

$$\begin{aligned} d_\tau(\sigma, \nu) = &\sum_{x, y \in X, x \neq y} \mathbb{1}_{x,y \text{ discordant for } \sigma, \nu} \\ &+ \sum_{x \in X, y \notin X} \mathbb{1}_{x,y \text{ discordant for } \sigma, \nu} \\ &+ \sum_{x, y \notin X, x \neq y} \mathbb{1}_{x,y \text{ discordant for } \sigma, \nu}. \end{aligned} \tag{21}$$

As $\sigma$ varies in $R$, only some of these terms vary. In particular, it is only the third term that varies with $\sigma$, and it is minimised at 0 by the permutation $\sigma$ in $R$ which is in accordance with $\nu$ on the set $[n] \setminus X$. □

**Lemma 6** *Let $\sigma \in R$, and $\nu \in S_n$. We have the following decomposition of the distance $d(\sigma, \nu)$*

$$d_\tau(\sigma, \nu) = d_\tau(\sigma, \Pi_R(\nu)) + d_\tau(\Pi_R(\nu), \nu).$$

***Proof*** We compute directly with the discordant pairs definition of the Kendall distance. Again, let $R$ be the partial ranking $x_1 \succ \cdots \succ x_k$, and let $X = \{x_1, \ldots, x_k\}$. We decompose the Kendall distance between $\sigma \in R$ and $\nu$ as before:

$$\begin{aligned} d_\tau(\sigma, \nu) = &\sum_{x, y \in X, x \neq y} \mathbb{1}_{x,y \text{ discordant for } \sigma, \nu} \\ &+ \sum_{x \in X, y \notin X} \mathbb{1}_{x,y \text{ discordant for } \sigma, \nu} \\ &+ \sum_{x, y \notin X, x \neq y} \mathbb{1}_{x,y \text{ discordant for } \sigma, \nu}. \end{aligned} \tag{22}$$

By the construction of $\Pi_R(\nu)$ in the proof of Lemma 5, we have that

$$d(\Pi_R(\nu), \nu) = \sum_{x,y \in X, x \neq y} \mathbb{1}_{x,y \text{ discordant for } \sigma,\nu} + \sum_{x \in X, y \notin X} \mathbb{1}_{x,y \text{ discordant for } \sigma,\nu},$$

i.e. the first two terms of the decomposition in Equation (22). Similarly, we have

$$d(\Pi_R(\nu), \sigma) = \sum_{x,y \notin X, x \neq y} \mathbb{1}_{x,y \text{ discordant for } \sigma,\nu},$$

and so the result follows. □

**Lemma 7** *Let $\sigma \in R$, and let $\nu \in R'$. We have the following relationship between $d_\tau(A_R(\sigma), \nu)$ and $d_\tau(\sigma, \nu)$*

$$d_\tau(A_R(\sigma), \nu) = d_\tau(\sigma, \nu) + \binom{n-k}{2} - 2d_\tau(\sigma, \Pi_R(\nu)). \tag{11}$$

***Proof*** We begin by observing that, by Lemma 6, we have

$$d(\sigma, \nu) = d(\sigma, \Pi_R(\nu)) + d(\Pi_R(\nu), \nu), \tag{23}$$

and

$$d(A_R(\sigma), \nu) = d(A_R(\sigma), \Pi_R(\nu)) + d(\Pi_R(\nu), \nu). \tag{24}$$

Now, from Lemma 4, we have that $d(A_R(\sigma), \Pi_R(\nu)) = \binom{n-k}{2} - d(\sigma, \Pi_R(\nu))$. Hence, the result follows. □

**Lemma 8** *Let $R, R' \subseteq S_n$ be top-k rankings, in preference notation given by*

$$R : a_1 \succ \cdots \succ a_l \succ [n] \setminus \{a_1, \ldots, a_l\},$$
$$R' : b_1 \succ \cdots \succ b_m \succ [n] \setminus \{b_1, \ldots, b_m\}.$$

*If $\nu \sim Unif(R')$, then $\Pi_R(\nu)$ is a full ranking with distribution $Unif(R'')$, where $R'' \subseteq R$ is the partial ranking given by*

$$R'' : a_1 \succ \cdots \succ a_l \succ b_{i_1} \succ \cdots \succ b_{i_q} \succ [n] \setminus \{a_1, \ldots, a_l, b_1, \ldots, b_m\},$$

*where $\{b_{i_1}, \ldots, b_{i_q}\} = \{b_1, \ldots, b_m\} \setminus \{a_1, \ldots, a_l\}$, and $i_j < i_{j+1}$ for all $j = 1, \ldots, q-1$.*

***Proof*** We first show that $\Pi_R$ maps $R'$ into $R''$. This is straightforward, as given $\nu \in R'$, we first observe that $\Pi_R(\nu) \in R$, and so the full ranking $\Pi_R(\nu)$ is consistent with the partial ranking

$$a_1 \succ \cdots \succ a_l \succ [n] \setminus \{a_1, \ldots, a_l\}.$$

Next, since $\Pi_R(\nu)$ is concordant with $\nu$ for all pairs outside the set $\{a_1, \ldots, a_l\}$, $\Pi_R(\nu)$ must be consistent with the partial ranking

$$b_{i_1} \succ \cdots \succ b_{i_q} \succ [n] \setminus \{a_1, \ldots, a_l, b_1, \ldots, b_m\}.$$

Putting these two facts together shows that the full ranking $\Pi_R(\nu)$ must be consistent with the partial ranking

$$a_1 \succ \cdots \succ a_l \succ b_{i_1} \succ \cdots \succ b_{i_q} \succ [n] \setminus \{a_1, \ldots, a_l, b_1, \ldots, b_m\}.$$

Thus, given $\nu \sim \text{Unif}(R')$, the distribution of $\Pi_R(\nu)$ is supported on $R''$. To show that it is uniform, we now argue that equally many rankings in $R'$ are mapped to each ranking in $R''$. To see this, we observe that the pre-image of a ranking in $R''$ is the set of all rankings in $R'$ which are concordant with it on all pairs in $[n] \setminus \{a_1, \ldots, a_l, b_1, \ldots, b_m\}$. The number of such rankings is independent of the selected ranking in $R''$, and so the statement of the lemma follows. □

**Lemma 9** *Let $R'' \subseteq R \subseteq S_n$ be top-k partial rankings. Then for $\sigma \in R$, we have*

$$A_{R''}(\Pi_{R''}(\sigma)) = \Pi_{R''}(A_R(\sigma)).$$

***Proof*** We begin by introducing preference-style notation for $R$ and $R''$. Let $R$ be the top-$k$ ranking given by $a_1 \succ \cdots \succ a_l \succ [n] \setminus \{a_1, \ldots, a_l\}$, and let $R''$ be the partial ranking given by $a_1 \succ \cdots \succ a_l \succ a_{l+1} \succ \cdots \succ a_m \succ [n] \setminus \{a_1, \ldots, a_m\}$. Let $\sigma \in R$, and let the elements of $[n] \setminus \{a_1, \ldots, a_m\}$ be given by $b_1, \ldots, b_q$, with indices chosen such that $\sigma$ corresponds to the full ranking

$$a_1 \succ \cdots a_m \succ b_1 \succ \cdots \succ b_q.$$

Then, the ranking $A_{R''}(\Pi_{R''}(\sigma))$ is given by

$$a_1 \succ \cdots a_m \succ b_q \succ \cdots \succ b_1,$$

and a straightforward calculation shows that this is also the case for $\Pi_{R''}(A_R(\sigma))$, as required. □

**Lemma 10** *Let $R'' \subseteq R \subseteq S_n$ be top-k partial rankings, given in preference notation by*

$$R : a_1 \succ \cdots \succ a_l \succ [n] \setminus \{a_1, \ldots, a_l\},$$
$$R'' : a_1 \succ \cdots \succ a_l \succ a_{l+1} \succ \cdots a_m \succ [n] \setminus \{a_1, \ldots, a_m\}.$$

*Let $\alpha$ be the number of unranked elements under $R$, and let $\beta$ be the additional number of elements ranked under $R''$ relative to $R$. Then for $\sigma \in R$, we have*

$$d_\tau(\sigma, \Pi_{R''}(\sigma)) = ((n-l)-(m-l))(m-l) + \binom{m-l}{2} - d_\tau(A_R(\sigma), \Pi_{R''}(A_R(\sigma))).$$

***Proof*** Again, we denote $\{b_1, \ldots, b_q\} = [n] \setminus \{a_1, \ldots, a_m\}$, with indices chosen such that $\sigma$ corresponds to the full ranking $a_1 \succ \cdots \succ a_m \succ b_1 \succ \cdots \succ b_q$. From earlier arguments, we have

$$d_\tau(\sigma, \Pi_{R''}(\sigma)) = \sum_{\substack{x \in \{a_{l+1}, \cdots, a_m\} \\ y \in \{a_{l+1}, \ldots, a_m\}}} \mathbb{1}_{(x,y) \text{ discordant for } \sigma, \Pi_{R''}(\sigma)} + \sum_{\substack{x \in \{a_{l+1}, \cdots, a_m\} \\ y \in \{b_1, \ldots, b_q\}}} \mathbb{1}_{(x,y) \text{ discordant for } \sigma, \Pi_{R''}(\sigma)}.$$

Now observe that for $a_i, a_j$ with $l+1 \leq i < j \leq m$, this pair is discordant for the pair of rankings $\sigma, \Pi_{R''}(\sigma)$ iff $a_j \succ a_i$ under $\sigma$ iff $a_i \succ a_j$ w.r.t $A_R(\sigma)$ iff $a_i, a_j$ are concordant for the pair of rankings $A_R(\sigma), \Pi_{R''}(A_R(\sigma))$. Hence, we have

$$\sum_{\substack{x \in \{a_{l+1}, \cdots, a_m\} \\ y \in \{a_{l+1}, \ldots, a_m\}}} \mathbb{1}_{(x,y) \text{ discordant for } \sigma, \Pi_{R''}(\sigma)} + \sum_{\substack{x \in \{a_{l+1}, \cdots, a_m\} \\ y \in \{a_{l+1}, \ldots, a_m\}}} \mathbb{1}_{(x,y) \text{ discordant for } A_R(\sigma), \Pi_{R''}(A_R(\sigma))} = \binom{\beta}{2}.$$

By analogous reasoning, we have

$$\sum_{\substack{x \in \{a_{l+1}, \cdots, a_m\} \\ y \in \{b_1, \ldots, b_q\}}} \mathbb{1}_{(x,y) \text{ discordant for } \sigma, \Pi_{R''}(\sigma)} + \sum_{\substack{x \in \{a_{l+1}, \cdots, a_m\} \\ y \in \{b_1, \ldots, b_q\}}} \mathbb{1}_{(x,y) \text{ discordant for } A_R(\sigma), \Pi_{R''}(A_R(\sigma))} = (\alpha - \beta)\beta.$$

Altogether, these statements yield the result of the lemma. □

## C Expectation of the Kernel Monte Carlo estimator

***Proof*** For distinct $i, j = 1, \ldots, I$, let $\left\{\sigma_n^{(i)}\right\}_{n=1}^{N_i}$ be an independent and identically distributed (i.i.d.) sample from $p(\sigma \mid R_i)$ and $\left\{\sigma_m^{(j)}\right\}_{m=1}^{N_j}$ be an i.i.d. sample from $p(\sigma \mid R_j)$. Then,

$$\mathbb{E}\left(\widehat{K}(R_i, R_j)\right) = \frac{1}{N_i N_j} \sum_{n=1}^{N_i} \sum_{m=1}^{N_j} \mathbb{E}\left(K(\sigma_n^{(i)}, \sigma_m^{(j)})\right) \quad (25)$$

By linearity of expectation, since the samples are identically distributed, the expectation in the summand above reduces to

$$= \sum_{\sigma \in R_i} \sum_{\sigma' \in R_j} K(\sigma, \sigma') p(\sigma \mid R_i) p(\sigma' \mid R_j)$$

as required. Hence, the kernel Monte Carlo estimator is unbiased for the off-diagonal elements of the kernel matrix. In the diagonal case the expectation is biased but consistent since,

$$\begin{aligned}
\mathbb{E}\left(\widehat{K}(R_i, R_i)\right) &= \frac{1}{N_i^2}\left(\sum_{n=1}^{N_i}\sum_{m=1}^{N_i} \mathbb{E}\left(K(\sigma_n^{(i)}, \sigma_m^{(i)})\right)\right) \\
&= \frac{1}{N_i^2}\sum_{n=1}^{N_i}\mathbb{E}\left[\sum_{m \neq n}^{N_i} \mathbb{E}\left(K(\sigma_n^{(i)}, \sigma_m^{(i)} \mid \sigma_n^{(i)})\right)\right. \\
&\quad \left. + \mathbb{E}(K(\sigma_n^{(i)}, \sigma_n^{(i)} \mid \sigma_n^{(i)}))\right] \\
&= \frac{1}{N_i^2}\sum_{n=1}^{N_i}\mathbb{E}\left[(N_i - 1)\mathbb{E}\left(K(\sigma^{(i)}, \sigma'^{(i)} \mid \sigma_n^{(i)})\right)\right. \\
&\quad \left. + \mathbb{E}\left(K(\sigma'^{(i)}, \sigma'^{(i)})\right)\right] \\
&= \frac{(N_i - 1)}{N_i}\mathbb{E}_{\sigma,\sigma'}\left(K(\sigma^{(i)}, \sigma'^{(i)})\right) \\
&\quad + \frac{1}{N_i}\mathbb{E}_{\sigma'}\left(K(\sigma'^{(i)}, \sigma'^{(i)})\right).
\end{aligned}$$

□

## D Variance of Kernel Monte Carlo estimator with i.i.d. samples

***Proof*** The variance of the Kernel Monte Carlo estimator with uniform weights is the following:

$$\begin{aligned}
\mathrm{Var}\left[\widehat{K}(R_i, R_j)\right] &= \frac{1}{N_i^2 N_j^2}\mathrm{Var}\left[\sum_{n=1}^{N_i}\sum_{m=1}^{N_j} K(\sigma_n^{(i)}, \sigma_m^{(j)})\right] \\
&= \frac{1}{N_i N_j^2}\left[\mathrm{Var}\left(\sum_{m=1}^{N_j} K(\sigma_1^{(i)}, \sigma_m^{(j)})\right)\right]
\end{aligned}$$

If we use the law of total variance, then

$$
\begin{aligned}
&\mathrm{Var}\left(\sum_{m=1}^{N_j} K(\sigma_1^{(i)}, \sigma_m^{(j)})\right) \\
&= \mathrm{Var}\left(\mathbb{E}\left[\sum_{m=1}^{N_j} K(\sigma_1^{(i)}, \sigma_m^{(j)}) \mid \sigma_1^{(i)}\right]\right) \\
&\quad + \mathbb{E}\left(\mathrm{Var}\left[\sum_{m=1}^{N_j} K(\sigma_1^{(i)}, \sigma_m^{(j)}) \mid \sigma_1^{(i)}\right]\right) \\
&= {N_j}^2 \mathrm{Var}\left(\mathbb{E}\left[K(\sigma_1^{(i)}, \sigma_1^{(j)}) | \sigma_1^{(i)}\right]\right) \\
&\quad + N_j \mathbb{E}\left(\mathrm{Var}\left[K(\sigma_1^{(i)}, \sigma_1^{(j)}) | \sigma_1^{(i)}\right]\right) \\
&= N_j^2 \mathrm{Var}\left(\sum_{\sigma' \in R_j} K(\sigma_1^{(j)}, \sigma') p(\sigma' \mid R_j)\right) \\
&\quad + N_j \mathbb{E}\left(\sum_{\sigma' \in R_j} K(\sigma_1^{(i)}, \sigma')^2 p(\sigma' \mid R_j)\right) \\
&\quad - N_j \mathbb{E}\left(\left(\sum_{\sigma' \in R_j} K(\sigma_1^{(i)}, \sigma') p(\sigma' \mid R_j)\right)^2\right) \\
&= N_j^2 \sum_{\sigma \in R_i} p(\sigma' \mid R_i) \left(\sum_{\sigma' \in R_j} p(\sigma' \mid R_j) K(\sigma, \sigma')\right)^2 \\
&\quad - N_j^2 \left(\sum_{\sigma \in R_i} \sum_{\sigma' \in R_j} K(\sigma, \sigma') p(\sigma \mid R_i) p(\sigma' \mid R_j)\right)^2 \\
&\quad + N_j \sum_{\sigma \in R_i} p(\sigma \mid R_i) \left(\sum_{\sigma' \in R_j} p(\sigma' \mid R_j) K(\sigma, \sigma')\right)^2 \\
&\quad - N_j \sum_{\sigma' \in R_j} \sum_{\sigma \in R_i} K(\sigma, \sigma')^2 p(\sigma' \mid R_j) p(\sigma \mid R_i)
\end{aligned}
$$

So the variance for the Monte Carlo kernel estimator is given by

$$
\begin{aligned}
&\mathrm{Var}\left[\widehat{K}(R_i, R_j)\right] \\
&= \frac{1}{N_i}\left[\sum_{\sigma \in R_i} p(\sigma \mid R_i) \left(\sum_{\sigma' \in R_j} p(\sigma' \mid R_j) K(\sigma, \sigma')\right)^2\right. \\
&\quad \left. - \left(\sum_{\sigma \in R_i} \sum_{\sigma' \in R_j} K(\sigma, \sigma') p(\sigma \mid R_i) p(\sigma' \mid R_j)\right)^2\right] \\
&\quad + \frac{1}{N_i N_j}\left[\sum_{\sigma \in R_i} p(\sigma \mid R_i) \left(\sum_{\sigma' \in R_j} p(\sigma' \mid R_j) K(\sigma, \sigma')\right)^2\right. \\
&\quad \left. - \sum_{\sigma \in R_i} \sum_{\sigma' \in R_j} K(\sigma, \sigma')^2 p(\sigma \mid R_i) p(\sigma' \mid R_j)\right].
\end{aligned}
$$

□

**Table 4** Table of partial rankings subset cardinalities for a given number of items and number of observed preferences (by overlapping pairs)

| $n$/#pairs | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 3 | 3 | 1 | – | – |
| 4 | 12 | 4 | 1 | – |
| 5 | 60 | 20 | 5 | 1 |
| 6 | 360 | 120 | 30 | 6 |
| 7 | 2520 | 840 | 210 | 42 |
| 8 | 20,160 | 6720 | 1680 | 336 |
| 9 | 181,440 | 60,480 | 15,120 | 3024 |
| 10 | 1,814,400 | 604,800 | 151,200 | 30,240 |

## E Factorial growth of the space of consistent full rankings for a given partial ranking

See Table 4.

## References


Aronszajn, N.: Theory of reproducing kernels. Trans. Am. Math. Soc. **68**(3), 337–404 (1950)

Berg, C., Christensen, J.P.R., Ressel, P.: Harmonic Analysis on Semigroups. Springer, New York (1984)

Berlinet, A., Thomas-Agnan, C.: Reproducing Hilbert Spaces for Probability and Statistics. Kluwer Academic Publishers, Dordrecht (2004)

Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: ACM workshop on Computational Learning Theory (1992)

Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs. Biometrika **39**, 324–345 (1952)

Busse, L.M., Orbanz, P., Buhmann, J.M.: Cluster analysis of heterogeneous rank data. In: International Conference on Machine Learning (2007)

Caron, F., Teh, Y.W., Murphy, T.B.: Bayesian nonparametric Plackett–Luce models for the analysis of preferences for college degree programmes. Ann Appl Stat **8**, 1145–1181 (2014)

Chen, Y., Fan, J., Ma, C., Wang, K.: Spectral method and regularized MLE are both optimal for top-k ranking (2017). arXiv:1707.09971

Chen, Y., Welling, M., Smola, A.: Super-Samples from Kernel Herding. In: UAI (2010)

Chierichetti, F., Dasgupta, A., Haddadan, S., Kumar, R., Lattanzi, S.: Mallows models for top-k lists. In: Advances in Neural Information Processing Systems (2018)
Cortes, C., Vapnik, V.N.: Support-vector networks. J. Mach. Learn. **20**, 273–297 (1995)
Deza, M.M., Deza, E.: Encyclopedia of Distances. Springer, Berlin, Heidelberg (2009)
Diaconis, P.: Group Representations in Probability and Statistics. Institute of Mathematical Statistics lecture notes (1988)
Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. Wiley, New York (1973)
Dudley, R.: Real Analysis and Probability. Cambridge University Press, Cambridge (2002)
Fernández, V.A., Gamero, M.D.J., García, J.M.: A test for the two-sample problem based on empirical characteristic functions. Comput. Stat. Data Anal. **52**, 3730–3748 (2007)
Fink, A.M., Jodeit, M.: On Chebyshev's other inequality. Inequal. Stat. Probab. **5**, 115–120 (1984)
Fligner, M.A., Verducci, J.S.: Distance based ranking models. J R Stat Soc Ser B **48**, 359–369 (1986)
Fukumizu, K., Sriperumbudur, B., Gretton, A., Schölkopf, B.: Characteristic kernels on groups and semigroups. In: Neural Information processing systems (2009)
GPy. since 2012. GPy: A Gaussian process framework in python. http://github.com/SheffieldML/GPy
Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A Kernel two sample test. J. Mach. Learn. Res. **13**, 723–773 (2012)
Hammersley, J.M., Morton, K.W.: A new Monte Carlo technique: antithetic variates. Math. Proc. Camb. Philos. Soc. **52**, 449–475 (1956)
Haussler, D.: Convolution Kernels on Discrete Structures (1999)
Irurozki, E., Calvo, B., Lozano, J.A.: PerMallows: an R package for Mallows and Generalized Mallows models. J. Stat. Softw. (2016a)
Irurozki, E., Calvo, B., Lozano, J.A.: Sampling and learning Mallows and Generalized Mallows models under the Cayley distance. Methodol. Comput. Appl. Probab. **20**, 1–35 (2016b)
James, G.D.: The Representation Theory of the Symmetric Groups. Springer, Berlin (1978)
Jiao, Y., Vert, J.P.: The Kendall and Mallows Kernels for permutations. In: International conference for Machine learning (2015)
Kamishima, T., Akaho, S.: Efficient clustering for orders. In: Zighed, D.A., Tsumoto, S., Ras, Z.W., Hacid, H. (eds.) Mining Complex Data, pp. 261–279. Springer, Berlin (2009)
Knuth, D.: The Art of Computer Programming, vol. 3. Addison-Wesley, Boston (1998)
Kondor, R., Barbosa, M.: Ranking with kernels in Fourier space. In: Conference on Learning Theory (2010)
Kondor, R., Howard, A., Jebara, T.: Multi-object tracking with representations of the symmetric group. In: AISTATS (2007)
Lebanon, G., Mao, Y.: Non-parametric modeling of partially ranked data. J. Mach. Learn. Res. **9**, 2401–2429 (2008)
Luce, R.D.: Individual Choice Behavior: A Theoretical Analysis. Wiley, New York (1959)
Mallows, C.L.: Non-null ranking models. Biometrika **44**, 114–130 (1957)
Mania, H., Ramdas, A., Wainwright, M.J., Jordan, M.I., Recht, B.: On kernel methods for covariates that are rankings (2016)
Mukherjee, S.: Estimation in exponential families. Ann. Stat. **44**, 853–875 (2016)
Neal, R.M.: 1998. Regression and Classification Using Gaussian Process Priors. Bayesian Statistics 6
Plackett, L.R.: The analysis of permutations. J. R. Stat. Soc. Ser. C (1974)
Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press, Cambridge (2006)
Ross, S.M.: Simulation Fourth Edition, Statistical Modeling and Decision Science. Academic, Cambridge (2006)
Schölkopf, B., Smola, A.: Learning with Kernels: Support Vector Machines, Regularisation, Optimisation and Beyond. MIT press, Cambridge (2002)
Schölkopf, B., Smola, A.J., Müller, K.R.: Kernel principal component analysis. In: Schölkopf, B., Burges, C.J.C., Smola, A.J. (eds.) Advances in Kernel Methods. MIT Press, Cambridge (1999)
Sejdinovic, D., Sriperumbudur, B., Gretton, A., Fukumizu, K.: Equivalence of distance-based and RKHS-based statistics in hypothesis testing. Ann. Stat. **41**, 2263–2291 (2013)
Serfling, R.J.: Approximation Theorems of Mathematical Statistics. Wiley, New York (1980)
Shawer-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge (2004)
Sriperumbudur, B.K., Fukumizu, K., Lanckriet, G.R.G.: Universality, characteristic Kernels and RKHS embedding of measures. J. Mach. Learn. Res. **12**, 2389–2410 (2011)
Stanley, R.P.: Enumerative Combinatorics, vol. 1. Cambridge University Press, Cambridge (2000)
Takeuchi, I., Le, Q.V., Sears, T.D., Smola, A.: Nonparametric quantile estimation. J. Mach. Learn. Res. **7**, 1231–1264 (2006)
Tsuda, K., Kin, T., Asai, K.: Marginalised kernels for biological sequences. Bioinformatics **18**, S268–S275 (2002)
Vitelli, V., Sørensen, Ø., Crispino, M., Frigessi, A., Arjas, E.: Probabilistic preference learning with the mallows rank model. J. Mach. Learn. Res. **18**, 158-1 (2017)
Wendland, H.: Scattered Data approximation. Cambridge University Press, Cambridge (2005)