Representer Theorems for Metric and Preference Learning: Geometric Insights and Algorithms

Peyman Morteza
University of Wisconsin-Madison
peyman@cs.wisc.edu

Abstract

We develop a mathematical framework to address a broad class of metric and preference learning problems within a Hilbert space. We obtain a novel representer theorem for the simultaneous task of metric and preference learning. Our key observation is that the representer theorem for this task can be derived by regularizing the problem with respect to the norm inherent in the task structure. For the general task of metric learning, our framework leads to a simple and self-contained representer theorem and offers new geometric insights into the derivation of representer theorems for this task. In the case of Reproducing Kernel Hilbert Spaces (RKHSs), we illustrate how our representer theorem can be used to express the solution of the learning problems in terms of finite kernel terms similar to classical representer theorems. Lastly, our representer theorem leads to a novel nonlinear algorithm for metric and preference learning. We compare our algorithm against challenging baseline methods on real-world rank inference benchmarks, where it achieves competitive performance. Notably, our approach significantly outperforms vanilla ideal point methods and surpasses strong baselines across multiple datasets. Code available at: https://github.com/PeymanMorteza/Metric-Preference-Learning-RKHS

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

1 INTRODUCTION

In machine learning, when given a set of objects or samples with only partial information, a key challenge is to infer their relationships and establish meaningful comparisons across the set. Many real-world AI applications require this capability to make informed decisions despite incomplete or limited observations. For example, large language models (LLMs) rely on ranking or scoring mechanisms to better align their generated responses with human feedback (Ouyang et al., 2022; Rafailov et al., 2023). In computer vision, ranking is crucial in retrieval and similarity-based tasks (Cakir et al., 2019). Similarly, recommender systems must prioritize items according to user preferences (Hsieh et al., 2017).

A common approach to tackling ranking and preference learning problems involves learning a metric to quantify distances between embedded samples or identifying a reference point for proximity-based ranking (Fürnkranz and Hüllermeier, 2010; Kulis et al., 2013; Bellet et al., 2013; Jamieson and Nowak, 2011a). These models are widely applied across domains and extensively studied. However, their standard versions often struggle to capture nonlinear aspects of the problem. While theoretically and empirically well-explored, the systematic study of their kernelized counterparts has many open directions, despite the powerful framework kernel methods provide for modeling complex, nonlinear relationships.

In this paper, we address this gap by introducing a novel mathematical framework designed to investigate a broad spectrum of metric and preference learning problems within Hilbert spaces. Our focus revolves around two common tasks in this domain: the simultaneous task of metric and preference learning from pairwise comparisons (Jamieson and Nowak, 2011a; Massimino and Davenport, 2021; Xu and Davenport, 2020; Canal et al., 2022), and the task of metric learning from triplet comparisons (Ye

et al., 2019; Jain et al., 2016; Mason et al., 2017). In the simultaneous task, given a set of embedded samples $S := \{x_1, ..., x_m\} \subset \mathbb{R}^d$ the learner is searching for an *ideal point* $u \in \mathbb{R}^d$ and a metric that aligns with the given partial binary responses of the form "u prefers x_i over x_j". We show that our framework leads to the first representer theorem for this task. The primary technical challenge in obtaining a representer theorem is that when the problem is lifted to a Hilbert space (potentially infinite dimensional), the ideal point, u, is unknown to the learner, and it may not lie on the subspace spanned by embedded samples. We define the space of *generalized Mahalanobis inner products* on a Hilbert space and demonstrate that the representer theorem can be naturally derived when formulated with respect to the norm induced by the inner product coming from this space. In the task of metric learning from triplet comparison, the learner aims to learn a metric that aligns with binary responses of the form "x_k is closer to x_i than x_j". Here, our framework yields a simple and self-contained representer theorem, offering fresh geometric insights into the derivation of such theorems. As an application, in the case of Reproducing Kernel Hilbert Spaces (RKHSs), we illustrate how our representer theorems can be used to express the solution of the original infinite-dimensional problem by solving a finite-dimensional counterpart. This leads to the development of new nonlinear algorithms specifically designed for these problems. We apply our algorithm to rank inference benchmarks and demonstrate that it is highly competitive, outperforming many strong baseline methods. The contribution of our work can be summarized as follows:

(1) We develop a novel mathematical framework to study a broad spectrum of metric and preference learning problems within Hilbert space. We define the notion of generalized Mahalanobis inner products on a Hilbert space, and precisely characterize their restriction to finite-dimensional subspaces as demonstrated in Theorem 6.

(2) We show that our framework yields the first and novel representer theorem, Theorem 8, for the simultaneous task of metric and preference learning and also leads to a simple and self-contained representer theorem, Theorem 9, for the task of metric learning.

(3) In the case of RKHSs, we show that our representer theorems can be used to express solutions to learning problems using a finite number of kernel terms (Proposition 10). Furthermore, we demonstrate that this formulation leads to the development of new algorithms (Algorithm 1 and Algorithm 2) for nonlinear metric and preference learning.

(4) We present empirical evaluations of our algorithm on both synthetic and real datasets, demonstrating its competitive performance.

We close the introduction by providing an outline of this work. In Section 2, we revisit the simultaneous task of metric and preference learning from pairwise comparisons and the task of metric learning from triplet comparison in finite-dimensional Euclidean spaces. In Section 3, we define the space of generalized Mahalanobis inner products and precisely characterize how they behave when restricted to finite-dimensional subspaces. In Section 4, we utilize our framework and derive our representer theorems for the simultaneous task, and for the triplet task. In Section 5, we investigate the case of RKHS and demonstrate how our representer theorems can be used to convert infinite-dimensional learning problems to finite-dimensional counterparts in Euclidean spaces and present a new kernelized algorithm derived from our representer theorem for metric and preference learning. In Section 9, we evaluate this algorithm on both synthetic and real-world benchmarks to analyze the effect of regularization and compare its performance with other methods. Finally, we conclude our work in Section 8 after reviewing related works in Section 7.

2 METRIC AND PREFERENCE LEARNING PROBLEM IN EUCLIDEAN SPACES

In this section, we revisit the formulation of metric and preference learning in Euclidean spaces. Throughout this section, we work with input space $\mathbb{R}^d$ and $\mathbb{S}^d_+$ the space of $d \times d$ positive definite matrices and $\text{sign}(\cdot)$ denotes the sign function. Next, recall the definition of Mahalanobis distance,

Definition 1 ((Mahalanobis, 1936)). *Given $M \in \mathbb{S}^d_+$, the Mahalanobis distance correspondes to M, denoted by d_M, is defined by,*

$$d^2_M(x, y) = (x - y)^T M(x - y),$$

for $x, y \in \mathbb{R}^d$.

Remark 1. *It's a standard fact that the distance $d_M(\cdot, \cdot)$ is induced by a norm, which in turn is induced by an inner product. The Mahalanobis distance is frequently employed in existing algorithms for metric and preference learning (Kulis et al., 2013; Bellet et al., 2013).*

Next, we revisit the formulation of two primary tasks

in finite-dimensional input space, which we intend to explore in general Hilbert spaces in subsequent sections.

Simultaneous Metric and Preference Learning from Pairwise Comparisons In the task of simultaneous metric and preference learning from paired comparisons(Jamieson and Nowak, 2011a; Xu and Davenport, 2020; Canal et al., 2022), beginning with a set of embedded samples, $S := \{x_1, ..., x_m\} \subset \mathbb{R}^d$, the learner aims to acquire a background preference metric d_M associated with $M \in \mathbb{S}^d_+$ and an ideal point $u \in \mathbb{R}^d$, given the following data,(z^i, y^i), for $1 \leq i \leq n$, where $n \leq \binom{m}{2}$ and $z^i = (z^i_1, z^i_2) \in S \times S$ is a pair of samples from S and $y^i \in \{-1, +1\}$ represents if u prefers z^i_1 over z^i_2 or not in the following sense, $y^i = \text{sign}(d_M(z^i_1, u) - d_M(z^i_2, u))$. Given the provided data, our objective is to learn the ideal point $u \in \mathbb{R}^d$ and the preference metric d_M associated with $M \in \mathbb{S}^d_+$. Next, the learning problem can be formalized by considering the associated Empirical Risk Minimization (ERM) as follows,

$$\min_{A \in \mathbb{S}^d_+, u \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(d^2_A(z^i_1, u) - d^2_A(z^i_2, u), y^i), \quad \text{(PF)}$$

where ℓ is a general loss function. Note that, in the above formulation, we considered a general form of loss that for $1 \leq i \leq n$, depends on $d^2_A(z^i_1, u) - d^2_A(z^i_2, u)$, the preference measure of u between z^i_1 and z^i_2, and response value y^i which indicates the preference sign based on the ground truth metric d^2_M. Our framework works for general loss as explained above. We next elaborate on two special cases of interest for theoretical analysis and practical applications.

- $\ell(d^2_A(z^i_1, u) - d^2_A(z^i_2, u), y^i) := \ell_{0/1}(\text{sign}(d^2_A(z^i_1, u) - d^2_A(z^i_2, u)), y^i)$,
- $\ell(d^2_A(z^i_1, u) - d^2_A(z^i_2, u), y^i) := \ell_{conv}(\big(d^2_A(z^i_1, u) - d^2_A(z^i_2, u)\big) \cdot y^i)$.

In the first item, $\ell_{0/1}(\cdot)$ denotes the $0/1$ loss, which is beneficial for theoretical analysis, such as examining sample complexity. In the second item, ℓ_{conv} represents any convex Lipschitz loss (e.g., Hinge loss), making it practical for applications as one can study its convex relaxation, as demonstrated in works such as (Xu and Davenport, 2020; Canal et al., 2022).

Remark 2. *While Problem PF has been extensively studied, there is currently no kernelization framework for the simultaneous task. The main challenge in developing such a framework is the existence of the ideal point u. In fact, when the problem is lifted to a Hilbert space (potentially infinite-dimensional), there is no guarantee that the ideal point u lies on the subspace spanned by embedded samples, as is the case in the classical setting for other learning problems. We show how to address this by equipping the ambient space with a natural inner product.*

Learning Metrics from Triplet Comparisons In the task of metric learning from triplet comparisons(Ye et al., 2019; Jain et al., 2016; Mason et al., 2017), beginning with a set of embedded samples, $S := \{x_1, ..., x_m\} \subset \mathbb{R}^d$, the learner aims to acquire a background preference metric d_M associated with $M \in \mathbb{S}^d_+$, given the following data, (z^i, y^i), for $1 \leq i \leq n$, where $n \leq \binom{m}{3}$ and $z^i = (z^i_1, z^i_2, z^i_3) \in S \times S \times S$ is a triplet sample and $y^i \in \{+1, -1\}$ represents if z^i_1 is similar to z^i_2 or z^i_3 in the following sense,$y^i = \text{sign}(d_M(z^i_1, z^i_2) - d_M(z^i_1, z^i_3))$. Given the provided data, our objective is to learn the metric d_M associated with $M \in \mathbb{S}^d_+$. Next, the learning problem can be formalized by considering the associated Empirical Risk Minimization (ERM) as follows,

$$\min_{A \in \mathbb{S}^d_+} \frac{1}{n} \sum_{i=1}^{n} \ell(d^2_A(z^i_1, z^i_2) - d^2_A(z^i_1, z^i_3), y^i). \quad \text{(TF)}$$

where ℓ is a general loss function. Similar to the simultaneous task, PF, we can consider the following two special cases of interest for the loss.

- $\ell(d^2_A(z^i_1, z^i_2) - d^2_A(z^i_2, z^i_3), y^i) := \ell_{0/1}(\text{sign}(d^2_A(z^i_1, z^i_2) - d^2_A(z^i_2, z^i_3)), y^i)$,
- $\ell(d^2_A(z^i_1, z^i_2) - d^2_A(z^i_2, z^i_3), y^i) := \ell_{conv}(\big(d^2_A(z^i_1, z^i_2) - d^2_A(z^i_2, z^i_3)\big) \cdot y^i)$.

In the first item, $\ell_{0/1}(\cdot)$ denotes the $0/1$ loss, which is beneficial for theoretical analysis, such as examining sample complexity. In the second item, ℓ_{conv} represents any convex Lipschitz loss (e.g., Hinge loss).

Remark 3. *Note that in the triplet setting, the ideal point is absent. Prior works have explored kernelization frameworks for Mahalanobis metric learning (Chatpatanasiri et al., 2010; Kulis et al., 2013; Jain et al., 2012). The framework developed in this paper yields a straightforward and self-contained representer theorem for the triplet setting. Importantly, it does not rely on the Kernel PCA trick and offer new geometric insights.*

3 SPACE OF GENERALIZED MAHALANOBIS INNER PRODUCTS

In this section, we develop a general framework that can be utilized to investigate a wide range of metric and preference learning problems, including PF and TF, within a general Hilbert space. We establish a couple of technical results that we later leverage to derive our representer theorems. Specifically, we introduce the space of generalized Mahalanobis inner products (see Definition 2) and precisely characterize their restriction to finite-dimensional subspaces, as stated in Theorem 6. Throughout this section, we assume $\mathcal{H}$ is a Hilbert space equipped with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and associated norm $\|\cdot\|_{\mathcal{H}}$ (We review some standard facts and definitions regarding linear operators on Hilbert spaces in the Appendix). To begin, we require a class of Mahalanobis distances in a general Hilbert space. We define the following,

Definition 2 (Space of generalized Mahalanobis inner products)**.** *Space of generalized Mahalanobis inner product on a Hilbert space $\mathcal{H}$ is defined to be the following set,*

$$\mathcal{F}_{\mathcal{H}} := \left\{ A : \mathcal{H} \to \mathcal{H} \mid \begin{array}{l} A \text{ is bounded,} \\ \text{strictly positive, self-adjoint} \end{array} \right\}.$$

Remark 4. *We can view $\mathcal{F}_{\mathcal{H}}$ as the set that parametrizes the generalized Mahalanobis inner product, as we explain next. For $A \in \mathcal{F}_{\mathcal{H}}$, we can consider the associated inner product defined by, $\langle x, y \rangle_A := \langle Ax, y \rangle_{\mathcal{H}}$. It can be seen that for $A \in \mathcal{F}_{\mathcal{H}}$, $\langle \cdot, \cdot \rangle_A$ defines an inner product on $\mathcal{H}$ (See Proposition 3 below). Moreover, when $\mathcal{H}$ is finite-dimensional and an orthonormal basis is chosen with respect to $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, A can be represented by a positive definite matrix M and $\|\cdot\|_A$ coincides with the Mahalanobis norm corresponding to M (See Lemma 4 below). In this sense, elements in $\mathcal{F}_{\mathcal{H}}$ can be viewed as an infinite-dimensional analogue of the standard Mahalanobis norms (Mahalanobis, 1936) in Euclidean spaces. In other words, we can regard elements of $\mathcal{F}_{\mathcal{H}}$ as a generalized version of symmetric, positive-definite matrices in finite-dimensional space.*

Remark 5. *Previous studies (Chatpatanasiri et al., 2010; Kulis et al., 2013; Jain et al., 2012) have explored the learning of a general operator L without additional assumptions and have focused on the distance of the image (e.g., $d_L(x, y) := \|Lx - Ly\|_{\mathcal{H}}$) as the metric corresponding to L. This approach, while capable of kernelizing (TF), fails to work for the kernelization of (PF). Utilizing $\mathcal{F}_{\mathcal{H}}$ offers an important advantage, as its elements are uniquely associated with inner products on $\mathcal{H}$. Leveraging these inner products enables us to develop a novel regularizer that plays a crucial role in formulating our representer theorems, as we will elaborate in Section 4.*

The next proposition shows that associated with any element $\mathcal{F}_{\mathcal{H}}$ comes with a natural inner product.

Proposition 3. *Let $A \in \mathcal{F}_{\mathcal{H}}$, then,*

$$\langle x, y \rangle_A := \langle Ax, y \rangle_{\mathcal{H}},$$

defines an inner product on $\mathcal{H}$ that is equivalent to $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. Conversely, let $g(\cdot, \cdot)$ be an inner product on $\mathcal{H}$ equivalent to $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ then there exist a unique $A \in \mathcal{F}_{\mathcal{H}}$ such that $g = \langle \cdot, \cdot \rangle_A$.

Remark 6. *The proposition 3 establishes a one-to-one correspondence between elements of $\mathcal{F}_{\mathcal{H}}$ and their associated inner products. Consequently, rather than dealing directly with operators, one can operate with these inner products. This eliminates the necessity of referencing a specific basis for deriving represented theorems. Such an approach is particularly useful in general Hilbert spaces (e.g., RKHSs), where a canonical basis is not provided, in contrast to $\mathbb{R}^n$.*

The subsequent lemma demonstrates that when $\mathcal{H}$ is finite-dimensional with a predetermined orthonormal basis, the elements of $\mathcal{F}_{\mathcal{H}}$ align with the standard Mahalanobis norm, corresponding to the representation of the operator with respect to the basis.

Lemma 4. *Let $\mathcal{H}$ be a n-dimensional Hilbert space and $\{e_1, ..., e_n\}$ be an orthonormal basis for $\mathcal{H}$. Let M denote the representation of A with respect to this basis. Then the inner product associated with the Mahalanobis norm coming from M coincides with $\langle \cdot, \cdot \rangle_A$.*

Remark 7. *The lemma implies that for any orthonormal basis, the Mahalanobis inner product of the corresponding matrix aligns with $\langle \cdot, \cdot \rangle_A$. The lemma can be used to perform computations independent of any basis, thereby simplifying the computational process.*

Next, considering a finite-dimensional subspace $V \subset \mathcal{H}$, V can be viewed as a Hilbert space inheriting its inner product from $\mathcal{H}$. We can define $\mathcal{F}_V$ in the same manner as outlined in Definition 2. The following definition delineates which elements in $\mathcal{F}_V$ can be paired with elements in $\mathcal{F}_{\mathcal{H}}$.

Definition 5. *Let $\mathcal{H}$ be a Hilbert space and $V \subset \mathcal{H}$ be a finite-dimensional subspace. Let $A \in \mathcal{F}_{\mathcal{H}}$ and $B \in \mathcal{F}_V$. We say A is Mahalanobis extension of B (or B is a Mahalanobis restrict of A) if,*

$$\langle \cdot, \cdot \rangle_A \big|_V = \langle \cdot, \cdot \rangle_B.$$

We close this section by stating the following theorem which completely characterizes when $A \in \mathcal{F}_{\mathcal{H}}$ on ambient space is the Mahalanobis extension of $B \in \mathcal{F}_V$. This will be used in the next section to obtain our representer theorems.

Theorem 6. *$A \in \mathcal{F}_{\mathcal{H}}$ is Mahalanobis extension of $B \in \mathcal{F}_V$ iff,*

$$B = PA|_V,$$

where,

$$P : \mathcal{H} \to V \subset \mathcal{H},$$

is projection operator with respect to $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. Moreover, each $B \in \mathcal{F}_V$ has at least one Mahalanobis extension.

4 REPRESENTER THEOREMS FOR METRIC AND PREFERENCE LEARNING

In this section, we leverage our framework to delve into metric and preference learning problems in Hilbert space. Specifically, we establish a new representer theorem for PF and present a simple and self-contained representer theorem for TF. We begin by formulating PF and TF in a general Hilbert space $\mathcal{H}$, using the framework developed in Section 3.

Metric and Preference Learning Problem in Hilbert Spaces Here we reformulate PF and TF using the framework developed in Section 3 in general Hilbert spaces. In the task of simultaneous metric and preference learning from paired comparisons, beginning with a set of embedded samples, $S := \{x_1, ..., x_m\} \subset \mathcal{H}$, we aim to learn a background preference metric associated to $A \in \mathcal{F}_{\mathcal{H}}$ and an ideal point $u \in \mathcal{H}$, given the following data,(z^i, y^i), for $1 \leq i \leq n$, where $n \leq \binom{m}{2}$ and $z^i = (z_1^i, z_2^i) \in S \times S$ is a pair of samples from S and $y^i \in \{-1, +1\}$ is defined by, $y^i = \text{sign}(\|z_1^i - u\|_A^2 - \|z_2^i - u\|_A^2)$. Given the provided data, the learning problem can be formalized by considering the associated Empirical Risk Minimization (ERM) as follows,

$$\min_{A \in \mathcal{F}_{\mathcal{H}}, u \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(\|z_1^i - u\|_A^2 - \|z_2^i - u\|_A^2, y^i), \tag{PI}$$

In the task of metric learning from triplet comparisons, we aim to acquire a background preference metric $\|\cdot\|_A$ associated with $A \in \mathcal{F}_{\mathcal{H}}$, given the following data, (z^i, y^i), for $1 \leq i \leq n$, where $n \leq \binom{m}{3}$ and $z^i = (z_1^i, z_2^i, z_3^i) \in S \times S \times S$ is a triplet sample and $y^i \in \{+1, -1\}$ is defined by,$y^i = \text{sign}(\|z_1^i - z_2^i\|_A^2 - \|z_1^i - z_3^i\|_A^2)$. The learning problem can be formalized by considering the associated Empirical Risk Minimization (ERM) as follows,

$$\min_{A \in \mathcal{F}_{\mathcal{H}}} \frac{1}{n} \sum_{i=1}^{n} \ell(\|(z_1^i - z_2^i\|_A^2 - \|z_1^i - z_3^i\|_A^2, y^i). \tag{TI}$$

where ℓ is a general loss function. We refer to Section 2 for detailed discussion on the loss function and other details for these models.

A Representer for Simultaneous Metric and Preference Learning Note that the search space, $\mathcal{F}_{\mathcal{H}} \times \mathcal{H}$, are infinite dimensional when $\mathcal{H}$ has infinite dimension. We would like to convert the infinite-dimensional problem represented by PI into a finite-dimensional equivalent. We demonstrate that a solution to the problem PI can be obtained by solving a finite-dimensional equivalent. Additionally, we show that regularizing the problem with the appropriate norm associated with elements in $\mathcal{F}_{\mathcal{H}}$ allows the entire process to be viewed as a Representer Theorem. First, set $V := \text{span}\{x_1, ..., x_m\}$ and consider the following finite dimensional problem,

$$\min_{A \in \mathcal{F}_V, u \in V} \frac{1}{n} \sum_{i=1}^{n} \ell(\|z_1^i - u\|_A^2 - \|z_2^i - u\|_A^2, y^i), \tag{PFH}$$

Note that V is finite dimensional Hilbert space and it inherits an inner product from $\mathcal{H}$ which we denote it by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ again. $\mathcal{F}_V$ is defined in a same manner as $\mathcal{F}_{\mathcal{H}}$ but the key point here is that $\mathcal{F}_V$ is now finite dimensional space. We would like to know how the solutions of PFH and PI are related. The following observation relates the solutions of PFH to the solutions of PI,

Proposition 7. *Let $A^* \in \mathcal{F}_{\mathcal{H}}, u^* \in \mathcal{H}$ be a solution to PI. Equip $\mathcal{H}$ with $\langle \cdot, \cdot \rangle_{A^*}$ and let $u^* = u^{\perp} + u^T$ where $u^{\perp} \in V^{\perp}$ and $u^T \in V$. Note that the orthogonal decomposition is with respect to $\langle \cdot, \cdot \rangle_{A^*}$ and*

not $\langle\cdot,\cdot\rangle_{\mathcal{H}}$. Then (A^*, u^T) is also a solution to PI. Moreover, let $B^* \in \mathcal{F}_V$ be the Mahalanobis restrict (See Definition 5) of A^*, then (B^*, u^T) is a solution to PFH with same optimal value. Conversely, let (B^*, u) be a solution to PFH, and let $A^* \in \mathcal{F}_{\mathcal{H}}$ be any Mahalanobis extension of B^* (See Definition 5), then (A^*, u) is also a solution to PI with same optimal value.

Finally, Proposition 7, suggests that if one considers the regularized problem with right norm then we can view it as a representer theorem.

Theorem 8 (Representer Theorem for Simultaneous Metric and Preference Learning)**.** *Let $\lambda > 0$ and consider the following infinite dimensional regularized problem,*

$$\min_{A\in\mathcal{F}_{\mathcal{H}}, u\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \ell(\left\|z_1^i - u\right\|_A^2 - \left\|z_2^i - u\right\|_A^2, y^i) + \lambda \|u\|_A^2, \tag{R}$$

and it's finite dimensional equivalent as follows,

$$\min_{A\in\mathcal{F}_{V}, u\in V} \frac{1}{n}\sum_{i=1}^{n} \ell(\left\|z_1^i - u\right\|_A^2 - \left\|z_2^i - u\right\|_A^2, y^i) + \lambda \|u\|_A^2, \tag{F}$$

Let $A^* \in \mathcal{F}_H$ be any Mahalanobis extension of $B^* \in \mathcal{F}_V$. Then (A^*, u) is a solution to R iff (B^*, u) is a solution to F with same optimal value. In particular this forces $u \in V$.

Remark 8. *Theorem 8 states that by solving a finite dimensional counterpart F, we can find the optimal value of R. If $\mathcal{H}$ is an RKHS associated with a kernel function k, then we can demonstrate that the solutions of F can be expressed in terms of kernel terms (as shown in Proposition 10), which resemble classical representer theorems.*

Remark 9. *Note that Theorem 8 has a regularization term $\|u\|_A^2$ stemming directly from equipping $\mathcal{H}$ with the generalized Mahalanobis inner product discussed in Section 3. In many classical representer theorems, the regularization term arises directly from the original norm $\|\cdot\|_{\mathcal{H}}^2$. However, this classical approach fails to yield a representer theorem for simultaneous tasks. We illustrate this with an example as illustrated in Figure 4.1.*

An Illustrative Example The choice of the inner product on $\mathcal{H}$ play a crucial role in the validity of Theorem 8. Consider the case $\mathcal{H} = \mathbb{R}^2$ with $S = \{x_1, x_2\}$ for $x_1 = e_1 = (1, 0)^T$ and $x_2 = -e_1 = (-1, 0)^T$. In this case, $\mathcal{H} = \mathbb{R}^2, V = \text{span}\{e_1\}$. Next, assume that $u = (0, u_0)^T$ lies on the y-axis. This is illustrated in Figure 4.1 . Let $A \in \mathcal{F}_{\mathcal{H}}$ has the following representation, $A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$, and denotes the metric on $\mathcal{H}$ that we would like to learn. We can think of the distance difference between $\|x_1 - u\|_A^2 - \|x_2 - u\|_A^2$ as preference loss as its sign represents whether u prefers product x_1 or x_2. We would ideally like to find an equivalent problem on V. Following the approach in classical representer theorems, one uses the orthogonal projection of u using $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ n the y axis (i.e. Euclidean projection). One can easily show that the loss depends on the location of u on the projection line and this naive projection does not work. Instead, we suggest projecting along the orthogonal line that is induced by the inner product obtained from $A \in \mathcal{F}_{\mathcal{H}}$ which may not necessarily be aligned with $\langle\cdot,\cdot\rangle_{\mathcal{H}}$. One can easily check that for $w = (1, -1)^T$,$\langle e_1, w\rangle_A = 0$, which means that the orthogonal projection with respect to geometry that is induced by $\langle\cdot,\cdot\rangle_A$ is in fact in the direction of w. This simple insight serves as the foundation for our representer theorem, 8, as demonstrated in Figure 4.1.

A Representer Theorem for Metric Learning from Triplet Comparison Next, we demonstrate how the framework we have established leads to a straightforward and self-contained representer theorem for the triplet learning task.

Theorem 9 (Representer Theorem for the Triplet Task)**.** *Let $A^* \in \mathcal{F}_{\mathcal{H}}$. Then A^* is a solution to*

$$\min_{A\in\mathcal{F}_{\mathcal{H}}} \frac{1}{n}\sum_{i=1}^{n} \ell(\left\|z_1^i - z_2^i\right\|_A - \left\|z_1^i - z_3^i\right\|_A, y^i), \tag{RT}$$

iff $B = PA|_V$, is a solution to,

$$\min_{B\in\mathcal{F}_{V}} \frac{1}{n}\sum_{i=1}^{n} \ell(\left\|(z_1^i - z_2^i\right\|_B - \left\|z_1^i - z_3^i\right\|_B, y^i). \tag{FT}$$

where,

$$P : \mathcal{H} \to V \subset \mathcal{H},$$

is projection operator with respect to $\langle\cdot,\cdot\rangle_{\mathcal{H}}$. And both solutions have the same optimal value.

Proof. The proof immediately follows from Theorem 6. Intuitively speaking, the solution of RT depends solely on the distance induced by A on the subspace V (given there is no ideal point in this task), which is precisely characterized by Theorem 6. □

Remark 10. *Note that the ideal point u is absent in the triplet setting, and a representer theorem for*

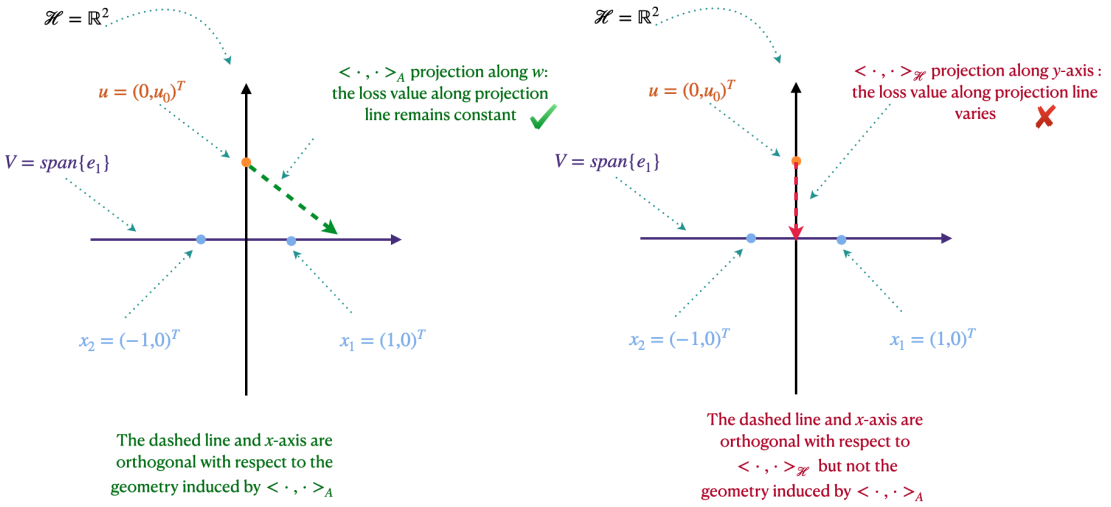

Figure 4.1: Illustration of the variation of preference loss with respect to the underlying geometry. When projecting along the $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, the loss value changes along the projection line. However, when projecting along the line induced by $\langle \cdot, \cdot \rangle_A$, the loss remains constant.

metric learning from triplet comparisons can also be derived from the framework presented in (Chatpatanasiri et al., 2010). Our approach leads to a simple, intuitive, and self-contained representer theorem for TF and has the advantage of not relying on a specific basis for V (e.g., the KPCA trick as in (Chatpatanasiri et al., 2010)). The learner can use their favorite orthonormal basis to transform FT into finite-dimensional Euclidean spaces. We illustrate one possible choice for this process in the next section.

5 KERNELIZED ALGORITHMS FOR METRIC AND PREFERENCE LEARNING

In Section 4, we established representer theorems for metric and preference learning in a general Hilbert space. Now, we turn our attention to the case of reproducing kernel Hilbert spaces (RKHS). Specifically, we aim to apply Theorem 8 and Theorem 9 in the context of RKHS so that we formulate problems in Euclidean spaces for practical application. To this end, we consider the setting where $\mathcal{X} = \mathbb{R}^d$ and k is a kernel on $\mathcal{X}$. Let $\mathcal{H}_k$ denote the RKHS associated with k. Next, we assume that $S' = \{s_1, ..., s_m\} \subset \mathcal{X}$ is a set of embedded samples in $\mathcal{X}$. Next, using k, for $1 \leq i \leq m$, we can associate the following in $\mathcal{H}_k$, $x_i := k(s_i, \cdot)$, and therefore we can transform the problem into $\mathcal{H}_k$ as in Section 4. Now Theorem 8 and Theorem 9 apply to convert infinite dimensional optimization to finite-dimensional equivalent. The following proposition finds an equivalent optimization in finite-dimensional Euclidean space.

Proposition 10. *Assume $S = \{k(\cdot, s_1), ..., k(\cdot, s_m)\}$, forms a linearly independent set. Then:*

- *Solutions of F and FT can be represented in terms of kernel terms as follows,*

$$A = \sum_{i,j=1}^{m} a_{ij} k(\cdot, s_i) \otimes k(\cdot, s_j),$$
$$u = \sum_{i=1}^{m} b_i k(\cdot, s_i),$$

  where in above we identified the space of linear maps on V with $V \otimes V$ using $\langle \cdot, \cdot \rangle_{\mathcal{H}}$.

- *Moreover, by choosing an orthonormal basis for V, there is a one-to-one correspondence between solutions of F with the solutions of the following problem in $\mathbb{R}^m$.*

$$\min_{A \in \mathbb{S}_+^m, u \in \mathbb{R}^m} \sum_{i=1}^{n} \ell\Big(\Big\| \alpha_{z_1^i} - u \Big\|_A^2 - \Big\| \alpha_{z_2^i} - u \Big\|_A^2, y^i \Big) + \lambda \|u\|_A^2 \qquad \text{(EF)}$$

  and one to one correspondence between the solutions of FT with the solutions of the following

Algorithm 1 Kernelized Ideal Point Algorithm: Training

Input:
  Sample set $S = \{s_1, ..., s_m\} \subset \mathbb{R}^d$
  Dataset $\mathcal{D} = \{(z^i, y^i) \mid z^i = (z_1^i, z_2^i) \in S \times S, 1 \leq i \leq n\}$
  Kernel function $k(\cdot, \cdot)$
  Regularization parameter $\lambda > 0$
Solve EF from Proposition 10 using numerical optimization to obtain:
  $A^* \in \mathbb{S}_m^+$ (learned metric)
  $u^* \in \mathbb{R}^m$ (preference/ideal vector)
Output: Trained Ideal Point model:
  Learned parameters: $A^* \in \mathbb{S}_m^+$, $u^* \in \mathbb{R}^m$
  Use these parameters to predict preferences for test sample pairs.

Algorithm 2 Kernelized Ideal Point Algorithm: Testing

Input:
  Test sample pair $z_{\text{test}} = (z_{\text{test},1}, z_{\text{test},2}) \in S \times S$
  Trained model parameters: $A^* \in \mathbb{S}_m^+, u^* \in \mathbb{R}^m$
Compute the α representation using Proposition 10:
  $\alpha_{z_{\text{test},1}}, \alpha_{z_{\text{test},2}} \in \mathbb{R}^m$
Compute the preference score using Proposition 10:
  $s = \|\alpha_{z_{\text{test},1}} - u^*\|_{A^*}^2 - \|\alpha_{z_{\text{test},2}} - u^*\|_{A^*}^2$
Output: Preference prediction based on $\text{sign}(s)$

problem in $\mathbb{R}^m$,

$$\min_{A \in \mathbb{S}_+^m} \sum_{i=1}^{n} \ell(\left\|\alpha_{z_1^i} - \alpha_{z_2^i}\right\|_A^2 - \left\|\alpha_{z_1^i} - \alpha_{z_3^i}\right\|_A^2, y^i), \tag{EFT}$$

where $\mathbb{S}_+^m$ *denotes space of symmetric positive definite matrices on* $\mathbb{R}^m$ *and for* $u \in \mathbb{R}^m$ *and* $A \in \mathbb{S}_+^m$,

$$\|u\|_A^2 = u^T A u,$$

and for $x = k(\cdot, s) \in \mathcal{H}$, $\alpha_x = (\alpha_1, ..., \alpha_m)^T \in \mathbb{R}^m$ *is the representation of* x *with respect to the orthonormal basis obtained by Gram-Schmidt process on elements of* S *and each* α_i *is defined by,*

$$\alpha_i = \frac{1}{\sqrt{D_{i-1} D_i}} \begin{vmatrix} k(s_1, s_1) & \dots & \dots & k(s_1, s_i) \\ \vdots & \ddots & \ddots & \vdots \\ k(s_{i-1}, s_1) & \dots & \dots & k(s_{i-1}, s_i) \\ k(s, s_1) & \dots & \dots & k(s, s_i) \end{vmatrix}$$

and $D_0 = 1$ *and for* $1 \leq i \leq m$, D_i *is defined as follows,*

$$D_i = \begin{vmatrix} k(s_1, s_1) & \dots & \dots & k(s_1, s_i) \\ \vdots & \ddots & \ddots & \vdots \\ k(s_{i-1}, s_1) & \dots & \dots & k(s_{i-1}, s_i) \\ k(s_i, s_1) & \dots & \dots & k(s_i, s_i) \end{vmatrix}.$$

Proof. The first item follows from the fact that elements of S span V. The proof of the second item is a direct application of Lemma 18 relies on the determinant form of the Gram-Schmidt process and Leibniz determinant formula which we recall in the appendix. □

Remark 11. *In Proposition 10, we used the determinant form of the Gram-Schmidt process. Similarly, the computation of* α_x *in Proposition 10 can be efficiently performed using its iterative form. The assumption of linear independence for* S *in Proposition 10 is mainly for clarity. If* S *is not linearly independent, a spanning subset can be selected using the iterative Gram-Schmidt process.*

5.1 Kernelized Algorithms for Metric and Preference Learning

In this subsection, we demonstrate how Proposition 10 can be leveraged to develop a practical algorithm for the kernelized version of ideal point models. Note that both (EF) and (EFT) define optimization problems in Euclidean spaces, which can be solved using numerical optimization algorithms. Thus, given a dataset obtained from either pairwise or triplet comparisons, as described in Section 2, along with a kernel function k, Proposition 10 enables us to solve (F) and (EFT) to learn the metric and/or the ideal point. Once a new sample x arrives, Proposition 10 can be applied to compute its corresponding representation α_x, which is then used to evaluate the metric or preference function under the nonlinear transformation induced by the kernel. In this sense, Proposition 10 also serves as the foundation for a nonlinear algorithm applicable to both (F) and (EFT). We formalize this in Algorithm 1 and Algorithm 2 for the ideal point setting. Similar algorithms can be derived for the triplet setting.

Remark 12. *The sample set* S *in Algorithm 2 varies based on the problem. If a larger set of embedded items is available but only a smaller set of pairwise comparisons is provided, preferences can be predicted using the subspace spanned by this set, as*

Algorithm	Chameleon	FlatLizard
BT (Bradley and Terry, 1952)	0.83±0.03	0.86±0.06
BT-LR (Chu and Ghahramani, 2005)	0.71±0.03	0.84±0.04
BT-GP (Chu and Ghahramani, 2005)	0.75±0.04	0.80±0.05
RC (Negahban et al., 2012)	0.61±0.06	0.66±0.05
RRC (Jain et al., 2020)	0.61±0.03	0.66±0.01
SVD (Cucuringu et al., 2016)	0.72±0.08	0.69±0.05
SVDC (Chau et al., 2022)	0.65±0.06	0.81±0.05
SVDK (Chau et al., 2022)	0.76±0.06	0.68±0.05
Serial (Fogel et al., 2016)	0.79±0.04	0.70±0.05
C-Serial (Chau et al., 2022)	0.80±0.03	0.88±0.01
CC (Chau et al., 2022)	0.66±0.10	0.78±0.08
KCC (Chau et al., 2022)	0.71±0.06	0.78±0.03
Vanilla Ideal Point (Xu and Davenport, 2020; Canal et al., 2022)	0.64±0.06	0.70±0.06
Kernelized Ideal Point (Ours)	**0.83±0.08**	**0.78±0.05**

Table 1: Performance (accuracy) comparison of various methods for rank inference. The reported results for methods other than ideal point models are courtesy of (Chau et al., 2022). The results for the Vanilla Ideal Point Method, inspired by (Xu and Davenport, 2020; Canal et al., 2022), are obtained by numerically solving PF.

described in Proposition 10. In an online setting, where raw unlabeled samples are unavailable in advance, the corresponding α vector in Proposition 10 can be computed by projecting the collected samples onto the training subspace in $\mathcal{H}$.

6 EXPERIMENTS

In this section, we report experimental results to evaluate the performance of Algorithm 1 and Algorithm 2. We use two data sets. **Flatlizard Competition:** This dataset (Whiting et al., 2009) records contests among male flat lizards. **Cape Dwarf Chameleons Contest:** This dataset (Stuart-Fox et al., 2006) documents contests among male chameleons with associated physical measurements.

Experimental Setting and Results We split the dataset into a 70/30 train-test ratio and use it for rank inference. We compare accuracy of Algorithm 1 and Algorithm 2, using RBF kernel with other spectral and probabilistic ranking methods based on pairwise comparisons (see (Chau et al., 2022; Chu and Ghahramani, 2005)). Additionally, we compare the results with the vanilla ideal point method, a variant of which has been studied in prior works (see, for instance, (Canal et al., 2022; Xu and Davenport, 2020; Massimino and Davenport, 2021)). The results, averaged over 10 runs, are reported in Table 1.

7 RELATED WORKS

Representer Theorems Representer theorems are critical in machine learning by converting infinite dimensional learning problems to finite dimensional counterparts. First introduced in approximation theory (Kimeldorf and Wahba, 1971; Wahba, 1990) and later been extended to cover a wide range of learning problems (Schölkopf et al., 2001).

Ranking Algorithms Ranking algorithms from pairwise comparisons can be categorized into two main classes. In the probabilistic setting, generative models are used to estimate the preference probability. For related works in this area, see (Chu and Ghahramani, 2005; Chau et al., 2022) and the references therein. The second class consists of spectral algorithms, which leverage the spectral properties of matrices encoding pairwise comparisons. For a comprehensive overview, see (Vigna, 2016).

Ideal Point Models Prior research has focused on learning efficiency (Jamieson and Nowak, 2011a,b), recovery accuracy in randomized models (Massimino and Davenport, 2021), and extensions to unknown Mahalanobis metrics (Xu and Davenport, 2020) and multiple ideal points (Canal et al., 2022). However, these works are limited to finite-dimensional Euclidean space. Our work is the first to introduce a nonlinear ideal point model and establish a representer theorem for its kernelized version in Reproducing Kernel Hilbert Space (RKHS).

Metric Learning Metric learning primarily follows two popular approaches: pairwise comparisons based on similarity/dissimilarity ((Kwok and Tsang, 2003)) and triplet-based comparisons using relative distances ((Schultz and Joachims, 2003)). For a comprehensive review of these applications, see (Kulis et al., 2013; Bellet et al., 2013; Suárez et al., 2021; Ghojogh et al., 2022) and references therein.

8 CONCLUSION

We developed a mathematical framework for metric and preference learning, leading to a novel representer theorem for their simultaneous learning. Additionally, we derived a self-contained representer theorem for metric learning. In RKHSs, our results enable the transformation of infinite-dimensional problems into finite-dimensional ones, facilitating new nonlinear algorithms. We evaluated our algorithm against other rank inference methods and demonstrated that it achieves competitive performance.

References

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68 (3):337–404, 1950.

A. Bellet, A. Habrard, and M. Sebban. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013.

R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

F. Cakir, K. He, X. Xia, B. Kulis, and S. Sclaroff. Deep metric learning to rank. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1861–1870, 2019.

G. Canal, B. Mason, R. K. Vinayak, and R. Nowak. One for all: Simultaneous metric and preference learning over multiple users. In *NeurIPS*, 2022.

R. Chatpatanasiri, T. Korsrilabutr, P. Tangchanachaianan, and B. Kijsirikul. A new kernelization framework for mahalanobis distance learning algorithms. *Neurocomputing*, 73 (10-12):1570–1579, 2010.

S. L. Chau, M. Cucuringu, and D. Sejdinovic. Spectral ranking with covariates. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 70–86. Springer, 2022.

W. Chu and Z. Ghahramani. Preference learning with gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, pages 137–144, 2005.

M. Cucuringu, I. Koutis, S. Chawla, G. Miller, and R. Peng. Simple and scalable constrained clustering: a generalized spectral method. In *Artificial Intelligence and Statistics*, pages 445–454. PMLR, 2016.

F. Fogel, A. d'Aspremont, and M. Vojnovic. Spectral ranking using seriation. *Journal of Machine Learning Research*, 17(88):1–45, 2016.

G. B. Folland. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.

J. Fürnkranz and E. Hüllermeier. Preference learning and ranking by pairwise comparison. In *Preference learning*, pages 65–82. Springer, 2010.

F. R. Gantmacher. The theory of matrices. *Co., New York*, 2, 1959.

B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley. Spectral, probabilistic, and deep metric learning: Tutorial and survey. *arXiv preprint arXiv:2201.09267*, 2022.

C.-K. Hsieh, L. Yang, Y. Cui, T.-Y. Lin, S. Belongie, and D. Estrin. Collaborative metric learning. In *Proceedings of the 26th international conference on world wide web*, pages 193–201, 2017.

L. Jain, K. G. Jamieson, and R. Nowak. Finite sample prediction and recovery bounds for ordinal embedding. *Advances in neural information processing systems*, 29, 2016.

L. Jain, A. Gilbert, and U. Varma. Spectral methods for ranking with scarce data. In *Conference on Uncertainty in Artificial Intelligence*, pages 609–618. PMLR, 2020.

P. Jain, B. Kulis, J. V. Davis, and I. S. Dhillon. Metric and kernel learning using a linear transformation. *The Journal of Machine Learning Research*, 13(1):519–547, 2012.

K. G. Jamieson and R. Nowak. Active ranking using pairwise comparisons. *Advances in neural information processing systems*, 24, 2011a.

K. G. Jamieson and R. D. Nowak. Low-dimensional embedding using adaptively selected ordinal data. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1077–1084. IEEE, 2011b.

G. Kimeldorf and G. Wahba. Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95, 1971.

B. Kulis et al. Metric learning: A survey. *Foundations and Trends® in Machine Learning*, 5(4): 287–364, 2013.

J. T. Kwok and I. W. Tsang. Learning with idealized kernels. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 400–407, 2003.

S. Lang. *Algebra*, volume 211. Springer Science & Business Media, 2012.

P. C. Mahalanobis. On the generalized distance in statistics. National Institute of Science of India, 1936.

B. Mason, L. Jain, and R. Nowak. Learning low-dimensional metrics. *Advances in neural information processing systems*, 30, 2017.

A. K. Massimino and M. A. Davenport. As you like it: Localization via paired comparisons. *Journal of Machine Learning Research*, 22(186):1–39, 2021.

S. Negahban, S. Oh, and D. Shah. Iterative ranking from pair-wise comparisons. *Advances in neural information processing systems*, 25, 2012.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *Computational Learning Theory: 14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT 2001 Amsterdam, The Netherlands, July 16–19, 2001 Proceedings 14*, pages 416–426. Springer, 2001.

M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. *Advances in neural information processing systems*, 16, 2003.

A. J. Smola and B. Schölkopf. *Learning with kernels*, volume 4. Citeseer, 1998.

L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968, 2009.

D. M. Stuart-Fox, D. Firth, A. Moussalli, and M. J. Whiting. Multiple signals in chameleon contests: designing and analysing animal contests as a tournament. *Animal Behaviour*, 71(6):1263–1271, 2006.

J. L. Suárez, S. García, and F. Herrera. A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges. *Neurocomputing*, 425:300–322, 2021.

S. Vigna. Spectral ranking. *Network Science*, 4(4): 433–445, 2016.

G. Wahba. *Spline models for observational data*. SIAM, 1990.

M. J. Whiting, J. K. Webb, and J. S. Keogh. Flat lizard female mimics use sexual deception in visual but not chemical signals. *Proceedings of the Royal Society B: Biological Sciences*, 276(1662): 1585–1591, 2009.

A. Xu and M. Davenport. Simultaneous preference and metric learning from paired comparisons. *Advances in Neural Information Processing Systems*, 33:454–465, 2020.

H.-J. Ye, D.-C. Zhan, and Y. Jiang. Fast generalization rates for distance metric learning: Improved theoretical analysis for smooth strongly convex distance metric learning. *Machine Learning*, 108: 267–295, 2019.

Supplementary Materials
Representer Theorems for Metric and Preference Learning: Geometric Insights and Algorithms

In the supplementary material, we provide additional experimental results and a more detailed description of the experimental setup (see Section 9), present background information, review standard facts (see Section 11), and include detailed proofs (see Section 10) for our theorems from the main text.

9 ADDITIONAL EXPERIMENTS

In this section, we present additional experimental results on synthetic data to better understand the advantages of our method over previous approaches. We conduct numerical experiments using a nonlinear data distribution to demonstrate the effectiveness of our approach. The results are summarized in Table 2. Below, we explain the data distribution setup, training details, different learning settings, and the experimental results.We then report and compare our experimental results with those from prior work.

9.1 Data Distribution Setup

The data distribution consists of two concentric circles with Gaussian noise of variance 0.4. In this configuration, the left portion of the larger circle and the right portion of the smaller circle are labeled as 0, while the remaining areas are labeled as 1. This is illustrated in Figure 9.1, where label 0 is shown in red and label 1 in blue. For our experiments, we assume that the user prefers all points labeled 0 over those labeled 1. The objective is to identify the ideal point u and the positive semi-definite (PSD) matrix A that align with these preferences.

Remark 13. *Why did we select this configuration? The configuration presented above is chosen for its non-linear representation. It is intuitively clear that neither an ideal point nor the Mahalanobis metric in $\mathbb{R}^2$ can fully capture a user's preferences, for data distribution illustrated in Figure 9.1. To effectively address preference learning in this context, it is necessary to transform the data distribution non-linearly using a kernel and then perform preference learning in the transformed space generated by the kernel. This is precisely the focus of our framework discussed in Section 5.*

9.2 Training Details

All reported results in Table 2 are averaged over three runs, with the number in parentheses indicating the standard deviation. All training is performed under the following loss function $\ell(z_1, z_2, y) := \text{hinge}\left(y \cdot \left(\|z_1 - u\|_A^2 - \|z_2 - u\|_A^2\right)\right) + \lambda\|u\|_A^2$ for a data pair (z_1, z_2) and label y. The training loss is optimized using the Adam optimizer with a learning rate of $lr = 0.01$, and the PSD condition is enforced using Cholesky decomposition. The training-to-test data ratio is 0.3. The testing error is reported using the 0/1 loss. The kernel k_C defined by $k_C(x, y) := \langle x, y\rangle + \|x\|^2 \|y\|^2$ is used in the experiments. Please see https://github.com/PeymanMorteza/Metric-Preference-Learning-RKHS for other details.

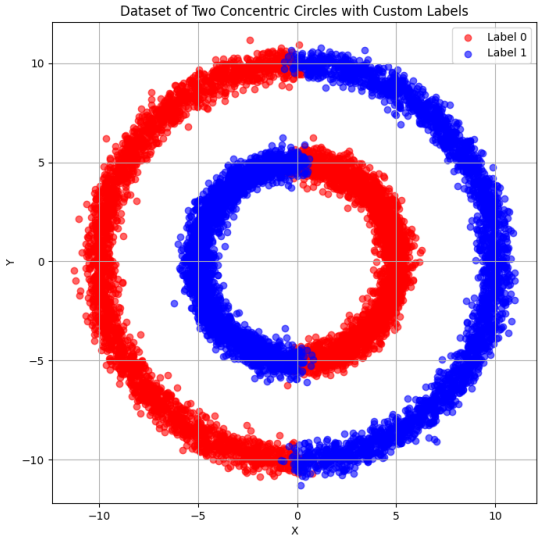

Figure 9.1: The data distribution is supported along two concentric circles with Gaussian noise of variance 0.4. In this setup, the left portion of the larger circle and the right portion of the smaller circle are labeled as 0, while the remaining regions are labeled as 1. This is illustrated in the figure above, with label 0 depicted in red and label 1 in blue. For our experiments, we assume that the user prefers all points labeled 0 over those labeled 1, and the objective is to search for the ideal point u and the PSD matrix A that align with these preferences.

Learning Method	**Average Training Loss (Std Dev)**	**Testing Error(Std Dev)**
No Kernel, No Regularization	1.07(0.04)	62.0% (8.3)
Kernel k_C, No Regularization	0.2 (0.3)	17.0 % (25)
Kernel k_C, With Regularization **(ours)**	0.25 (0.2)	0.6% (0.9)

Table 2: **Comparison of Different Learning Settings:** When regularizing the problem in the kernelized setting using the norm inspired by our Representer Theorem (See Theorem 8 and Proposition 10)) , we achieve a solution with low training error that generalizes well to the test data. See Subsection 9.3 for further explanation.

9.3 Comparison of Different Learning Methods

The results are reported in Table 2. The following experimental setting are considered.

- **1. No Kernel, No Regularization:** This setting corresponds to what is considered in prior work (e.g. (Canal et al., 2022)) in simultaneous task of metric and preference learning. Both the training loss and testing error are high. This is primarily due to the non-linear data representation, which this framework is not equipped to handle effectively.
- **2. Kernel k_C, No Regularization:** The kernel k_C (defined above) is used to transform the training data, and the associated optimization problem, as given by Proposition 10, is solved with no regularization (i.e., $\lambda = 0$). Small training loss with a non-trivial testing error indicates benign overfitting. This can be attributed to the lack of a regularization parameter, which plays a key role in our Representer Theorem (Theorem 8).
- **3. Kernel k_C, With Regularization (Ours):** The kernel k_C (defined below) is used to transform the training data, and the associated optimization problem, as given by Proposition 10, is solved with

non-zero regularization(i.e., $\lambda = 0.0001$). Both the training loss and test loss are low, indicating that the method finds a solution that generalizes well in this data setting, particularly when the problem is regularized with the appropriate norm. (a s explained in Theorem 8 and Proposition 10).

10 PROOFS

In this section, we provide complete proof for the theorems introduced in the previous sections.

Proposition 11 (Proposition 3 from the main body)**.** *Let $A \in \mathcal{F}_{\mathcal{H}}$, then,*

$$\langle x, y\rangle_A := \langle Ax, y\rangle_{\mathcal{H}},$$

defines an inner product on $\mathcal{H}$ that is equivalent to $\langle\cdot,\cdot\rangle_{\mathcal{H}}$. Conversely, let $g(\cdot,\cdot)$ be an inner product on $\mathcal{H}$ equivalent to $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ then there exist a unique $A \in \mathcal{F}_{\mathcal{H}}$ such that $g = \langle\cdot,\cdot\rangle_A$.

Proof. First, we show for $A \in \mathcal{F}_{\mathcal{H}}$, $\langle\cdot,\cdot\rangle_A$ defines an inner product. This part easily follows from the definition. For $x \neq 0$, we have,

$$\langle x, x\rangle_A = \langle Ax, x\rangle_{\mathcal{H}} > 0,$$

where we used the fact that A is a (strictly) positive operator. Next,

$$\langle x, y\rangle_A = \langle Ax, y\rangle_{\mathcal{H}} = \langle x, Ay\rangle_{\mathcal{H}} = \langle Ay, x\rangle_{\mathcal{H}} = \langle y, x\rangle_A,$$

where we used the fact that A is self-adjoint and $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is an inner product. Linearity for each coordinate follows from the fact that A is a linear operator and $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is an inner product. Next, since A is a bounded operator there exist a_1 such that for $x \in \mathcal{H}$,

$$\langle x, Ax\rangle_{\mathcal{H}} \leq a_1\langle x, x\rangle_{\mathcal{H}},$$

on the other hand, since A is strictly positive, there exist a_2 such that,

$$a_2\langle x, x\rangle_{\mathcal{H}} \leq \langle x, Ax\rangle_{\mathcal{H}},$$

above two implies that $\langle\cdot,\cdot\rangle_A$ is equivalent to $\langle\cdot,\cdot\rangle_{\mathcal{H}}$. Therefore, $\langle\cdot,\cdot\rangle_A$ is an inner product on $\mathcal{H}$ that is equivalent to $\langle\cdot,\cdot\rangle_{\mathcal{H}}$. Conversely, let g be an inner product on $\mathcal{H}$ that is equivalent to $\langle\cdot,\cdot\rangle_{\mathcal{H}}$, therefor there exist $c_1, c_2 > 0$, such that for all $y \in \mathcal{H}$,

$$c_1\langle y, y\rangle_{\mathcal{H}} \leq g(y, y) \leq c_2\langle y, y\rangle_{\mathcal{H}}.$$

Next, for a fixed $x \in \mathcal{H}$, consider the following functional on $\mathcal{H}$,

$$\begin{aligned}\phi_x &: \mathcal{H} \to \mathbb{R}\\ \phi_x(y) &= g(x, y),\end{aligned}$$

ϕ_x is a linear functional on $\mathcal{H}$ because g is an inner product (e.g. linear on each coordinate). Next, we show that ϕ_x is a bounded functional,

$$|\phi_x(y)| = g(x, y) \leq \sqrt{g(x, x)g(y, y)} \leq C\,\|y\|_{\mathcal{H}},$$

where $C := \sqrt{c_2 \cdot g(x, x)}$ and we used Cauchy-Schwartz inequality and the fact that g is equivalent to $\langle\cdot,\cdot\rangle_{\mathcal{H}}$. Therefore, by Riesz representation theorem, there exist a unique $z_x \in \mathcal{H}$ so that,

$$\phi_x(y) = \langle z_x, y\rangle_{\mathcal{H}},$$

Now define,

$$\begin{aligned}A &: \mathcal{H} \to \mathcal{H},\\ A(x) &= z_x,\end{aligned}$$

It is clear from the above construction that A is a linear operator. Next we show A is strictly positive,

$$\langle x, Ax\rangle_{\mathcal{H}} = \langle x, z_x\rangle_{\mathcal{H}} = \phi_x(x) = g(x,x) \geq c_1\langle x, x\rangle_{\mathcal{H}},$$

To show that A is self-adjoint, for $x, y \in \mathcal{H}$, we have,

$$\langle y, Ax\rangle_{\mathcal{H}} = \langle y, z_x\rangle_{\mathcal{H}} = \phi_x(y) = g(x,y) = g(y,x) = \phi_y(x) = \langle x, z_y\rangle_{\mathcal{H}} = \langle Ay, x\rangle_{\mathcal{H}},$$

where we used the fact that g is symmetric. To show uniqueness, assume there are two operators $A, B \in \mathcal{F}_{\mathcal{H}}$ such that for all $x, y \in \mathcal{H}$,

$$\langle x, Ay\rangle_{\mathcal{H}} = \langle x, By\rangle_{\mathcal{H}}.$$

Above implies that for all $x, y \in \mathcal{H}$,

$$\langle x, (A-B)y\rangle_{\mathcal{H}} \equiv 0 \implies Ay = By \implies A = B,$$

and we are done. □

The subsequent lemma demonstrates that when $\mathcal{H}$ is finite-dimensional with a predetermined orthonormal basis, the elements of $\mathcal{F}_{\mathcal{H}}$ align with the standard Mahalanobis norm, corresponding to the representation of the operator with respect to the basis.

Lemma 12 (Lemma 4 from the main body)**.** *Let $\mathcal{H}$ be a n-dimensional Hilbert space and $\{e_1, ..., e_n\}$ be an orthonormal basis for $\mathcal{H}$. Let M denote the representation of A with respect to this basis. Then the inner product associated with the Mahalanobis norm coming from M coincides with $\langle\cdot,\cdot\rangle_A$.*

Proof. Let $v, w \in \mathcal{H}$. We have,

$$v = \sum_{i=1}^{n} \langle v, e_i\rangle_{\mathcal{H}} e_i,$$
$$w = \sum_{i=1}^{n} \langle w, e_i\rangle_{\mathcal{H}} e_i,$$

therefore, representation of v, w are given by,

$$V = (v_1, ..., v_n)^T \in \mathbb{R}^n,$$
$$W = (w_1, ..., w_n)^T \in \mathbb{R}^n,$$

where for $1 \leq i \leq n$, $v_i = \langle v, e_i\rangle_{\mathcal{H}}$ and $w_i = \langle w, e_i\rangle_{\mathcal{H}}$. Next, for $1 \leq i \leq n$ and $1 \leq j \leq n$, $M_{ij} = \langle Ae_i, e_j\rangle_{\mathcal{H}}$. Next, we have,

$$\begin{aligned}\langle v, w\rangle_A = \langle v, Aw\rangle_{\mathcal{H}} &= \langle \sum_{i=1}^{n} \langle v, e_i\rangle_{\mathcal{H}} e_i, A(\sum_{i=1}^{n} \langle w, e_i\rangle_{\mathcal{H}} e_i)\rangle_{\mathcal{H}} \\ &= \langle \sum_{i=1}^{n} v_i e_i, \sum_{i=1}^{n} w_i A(e_i)\rangle_{\mathcal{H}} \\ &= \sum_{1\leq i,j\leq n} v_i w_j \langle e_i, Ae_j\rangle_{\mathcal{H}} = V^T M W.\end{aligned}$$

□

Definition 13. *Let $\mathcal{H}$ be a Hilbert space and $V \subset \mathcal{H}$ be a finite-dimensional subspace. Let $A \in \mathcal{F}_{\mathcal{H}}$ and $B \in \mathcal{F}_V$. We say A is Mahalanobis extension of B ((or B is a Mahalanobis restrict of A) if,*

$$\langle\cdot,\cdot\rangle_A\big|_V = \langle\cdot,\cdot\rangle_B.$$

The following theorem completely characterizes when $A \in \mathcal{F}_{\mathcal{H}}$ on ambient space is the Mahalanobis extension of $B \in \mathcal{F}_V$.

Theorem 14. *$A \in \mathcal{F}_{\mathcal{H}}$ is Mahalanobis extension of $B \in \mathcal{F}_V$ iff,*

$$B = PA|_V,$$

where,

$$P : \mathcal{H} \to V \subset \mathcal{H},$$

is projection operator with respect to $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. Moreover, each $B \in \mathcal{F}_V$ has at least one Mahalanobis extension.

Proof. Consider the following orthogonal decomposition with respect to $\langle \cdot, \cdot \rangle_{\mathcal{H}}$,

$$\mathcal{H} = V \oplus V^{\perp}.$$

First, consider $A \in \mathcal{F}_{\mathcal{H}}$. Consider the projection operator with respect to $\langle \cdot, \cdot \rangle_{\mathcal{H}}$,

$$P : \mathcal{H} \to V,$$

and set,

$$\begin{aligned} &B : V \to V \\ &B(x) = PA(x). \end{aligned}$$

It is clear from the definition that B is bounded and linear. Next, for $x \in V$, we also have,

$$\langle Bx, x \rangle_{\mathcal{H}} = \langle PAx, x \rangle_{\mathcal{H}} = \langle Ax, x \rangle_{\mathcal{H}} > c \left\| x \right\|_{\mathcal{H}}^2,$$

which shows that B is positive. Next, for $x, y \in V$,

$$\langle Bx, y \rangle_{\mathcal{H}} = \langle PAx, y \rangle_{\mathcal{H}} = \langle Ax, y \rangle_{\mathcal{H}} = \langle x, Ay \rangle_{\mathcal{H}} = \langle x, PAy \rangle_{\mathcal{H}},$$

where we used the fact that $x, y \in V$ and A is self-adjoint. This shows that B is self-adjoint. Next, for $x, y \in V$, we have,

$$\langle x, y \rangle_A = \langle Ax, y \rangle_{\mathcal{H}} = \langle PAx + (Ax)^{\perp}, y \rangle_{\mathcal{H}} = \langle PAx, y \rangle_{\mathcal{H}} = \langle x, y \rangle_B,$$

therefore we showed that A is Mahalanobis extension of $B = PA|_V$. Conversely, assume $A \in \mathcal{F}_{\mathcal{H}}$ is a Mahalanobis extension of B then setting $C = PA|_V$ we know that,

$$\langle \cdot, \cdot \rangle_C = \langle \cdot, \cdot \rangle_B,$$

and it follows from Proposition 3 that $B = C$. Finally, for $B \in \mathcal{F}_V$, define $A := B \oplus \mathrm{Id}$ by,

$$\begin{aligned} &A : \mathcal{H} \to \mathcal{H} \\ &A(h^T + h^{\perp}) := B(h^T) + h^{\perp}, \end{aligned}$$

where $h = h^T + h^{\perp}$ for $h^T \in V$ and $h^{\perp} \in V^{\perp}$. It can be easily checked that A is positive, self-adjoint, and bounded (See Lemma 15). It also follows from the definition of A that,

$$A\big|_V = B,$$

therefore B has at least one Mahalanobis extension and we are done. □

Lemma 15. *For $A \in \mathcal{F}_V$, define $B := A \oplus \mathrm{Id}$ by,*

$$B : \mathcal{H} \to \mathcal{H}$$
$$B(h^T + h^\perp) := A(h^T) + h^\perp.$$

where the following orthogonal decomposition with respect to $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is considered,

$$\mathcal{H} = V \oplus V^\perp,$$

and $h \in \mathcal{H}$ is represented as $h = h^T + h^\perp$. Then B is positive, self-adjoint, and bounded.

Proof. It is clear from the definition that B is a linear operator. To show that B is positive,

$$\langle Bx, x \rangle_{\mathcal{H}} = \langle A(x^T) + x^\perp, x^T + x^\perp \rangle_{\mathcal{H}} = \langle A(x^T), x^T \rangle_{\mathcal{H}} + \langle x^\perp, x^\perp \rangle_{\mathcal{H}} > c \left\| x^T \right\|_{\mathcal{H}}^2 + \left\| x^\perp \right\|_{\mathcal{H}}^2 > \min(1, c) \left\| x \right\|_{\mathcal{H}}^2 > 0,$$

where we used the fact that A is a positive operator (with constant c) and applied the Pythagorean theorem. To show that B is bounded we have,

$$\langle Bx, Bx \rangle_{\mathcal{H}} = \langle A(x^T) + x^\perp, A(x^T) + x^\perp \rangle_{\mathcal{H}} = \langle A(x^T), A(x^T) \rangle_{\mathcal{H}} + \langle x^\perp, x^\perp \rangle_{\mathcal{H}},$$

and boundedness follows from the fact that A is bounded. To show that B is self-adjoint, we have,

$$\begin{aligned} \langle Bx, y \rangle_{\mathcal{H}} = \langle A(x^T) + x^\perp, y \rangle_{\mathcal{H}} &= \langle x^\perp, y \rangle_{\mathcal{H}} + \langle A(x^T), y \rangle_{\mathcal{H}} \\ &= \langle x^\perp, y^\perp \rangle_{\mathcal{H}} + \langle x^T, A(y^T) \rangle_{\mathcal{H}} \\ &= \langle x, y^\perp \rangle_{\mathcal{H}} + \langle x, A(y^T) \rangle_{\mathcal{H}} \\ &= \langle x, By \rangle_{\mathcal{H}}. \end{aligned}$$

Thus, B is positive, self-adjoint, and bounded. □

Proposition 16 (Proposition 7 from the main body)**.** *Let $A^* \in \mathcal{F}_{\mathcal{H}}, u^* \in \mathcal{H}$ be a solution to PI. Equip $\mathcal{H}$ with $\langle \cdot, \cdot \rangle_{A^*}$ and let $u^* = u^\perp + u^T$ where $u^\perp \in V^\perp$ and $u^T \in V$. Note that the orthogonal decomposition is with respect to $\langle \cdot, \cdot \rangle_{A^*}$ and not $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. Then (A^*, u^T) is also a solution to PI. Moreover, let $B^* \in \mathcal{F}_V$ be the Mahalanobis restrict (See Definition 5) of A^*, then (B^*, u^T) is a solution to PFH with same optimal value. Conversely, let (B^*, u) be a solution to PFH, and let $A^* \in \mathcal{F}_{\mathcal{H}}$ be any Mahalanobis extension of B^* ((See Definition 5) then (A^*, u) is also a solution to PI with same optimal value.*

Proof. First note that by Proposition 3, $\langle \cdot, \cdot \rangle_{A^*}$ is equivalent to $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ therefore $\mathcal{H}$ equipped with $\langle \cdot, \cdot \rangle_{A^*}$ is a Hilbert space. Next, decompose u^* into $u^T \in V$ and $u^\perp \in V^\perp$ where the orthogonal decomposition in with respect to $\langle \cdot, \cdot \rangle_{A^*}$ and write,

$$u^* = u^T + u^\perp,$$

For $1 \leq i \leq n$, we have,

$$\begin{aligned} &\ell(\left\| z_1^i - u^* \right\|_{A^*}^2 - \left\| z_2^i - u^* \right\|_{A^*}^2, y^i) \\ =&\ell(\left\| z_1^i - (u^T + u^\perp) \right\|_{A^*}^2 - \left\| z_2^i - (u^T + u^\perp) \right\|_{A^*}^2, y^i) \\ =&\ell(\left\| (z_1^i - u^T) - u^\perp \right\|_{A^*}^2 - \left\| (z_2^i - u^T) - u^\perp \right\|_{A^*}^2), y^i) \\ =&\ell(\left\| (z_1^i - u^T) \right\|_{A^*}^2 + \left\| u^\perp \right\|_{A^*}^2 - \left\| (z_2^i - u^T) \right\|_{A^*}^2 - \left\| u^\perp \right\|_{A^*}^2, y^i) \\ =&\ell(\left\| (z_1^i - u^T) \right\|_{A^*}^2 - \left\| (z_2^i - u^T) \right\|_{A^*}^2, y^i). \end{aligned}$$

Above implies that,

$$\frac{1}{n} \sum_{i=1}^n \ell(\left\| z_1^i - u^* \right\|_{A^*}^2 - \left\| z_2^i - u^* \right\|_{A^*}^2, y^i) = \frac{1}{n} \sum_{i=1}^n \ell(\left\| z_1^i - u^T \right\|_{A^*}^2 - \left\| z_2^i - u^T \right\|_{A^*}^2, y^i).$$

Therefore if (A^*, u^*) solves PI then (A^*, u^T) also solves PI. Now by Proposition 3, it is clear that (B^*, u^T) solves PFH. The converse follows similarly. Let (B^*, v), $v \in V$ be a solution to PFH. Let $A^* \in \mathcal{F}_{\mathcal{H}}$ be any Mahalanobis extension. We claim that (A^*, v) also solves *PI*. If not, there exist (A_1^*, u_1) with smaller loss. Then arguing as above we know that the loss for (A_1^*, u_1) would be same as the loss for (A_1^*, u_1^T) which would be same as the loss for (B_1^*, u_1^T). Therefore (B_1^*, u_1^T) has smaller loss that (B^*, v) which is a contradiction. □

Theorem 17 (Representer Theorem for Simultaneous Metric and Preference Learning)**.** *Let $\lambda > 0$ and consider the following infinite dimensional regularized problem,*

$$\min_{A \in \mathcal{F}_{\mathcal{H}}, u \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(\left\|z_1^i - u\right\|_A^2 - \left\|z_2^i - u\right\|_A^2, y^i) + \lambda \left\|u\right\|_A^2, \tag{R}$$

and it's finite dimensional equivalent as follows,

$$\min_{A \in \mathcal{F}_V, u \in V} \frac{1}{n} \sum_{i=1}^{n} \ell(\left\|z_1^i - u\right\|_A^2 - \left\|z_2^i - u\right\|_A^2, y^i) + \lambda \left\|u\right\|_A^2, \tag{F}$$

Let $A^* \in \mathcal{F}_H$ be any Mahalanobis extension of $B^* \in \mathcal{F}_V$. Then (A^*, u) is a solution to R iff (B^*, u) is a solution to F with same optimal value. In particular this forces $u \in V$.

Proof. Let (A^*, u) be a solution to R. By Proposition 7, we get a smaller loss for the regularized term when we project on V (with respect to $\langle \cdot, \cdot \rangle_{A^*}$) while keeping the value of the other term unchanged, therefore $u \in V$. Next, we claim that (B^*, u) is a solution to F. If not, there exist a solution (B_1, v_1) with smaller loss. Let A_1 be any Mahalanobis extension of B_1 obtained from 6. It follows that (A_1, v_1) has a same loss value for R as (B_1, v_1) for F. Therefore, (A_1, v_1) has smaller loss than (A^*, u) which is a contradiction. The converse follows with a similar argument and we are done. □

Lemma 18. *Consider the same setting as in Theorem 20 and let $x \in \mathcal{H}$ and write $x = x^T + x^\perp$ where the decomposition is with respect to $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. Then x^T can be represented by $\alpha_x = (\alpha_1, ..., \alpha_m)^T \in \mathbb{R}^m$ as follows,*

$$x^T = \sum_{i=1}^{m} \langle x, e_i \rangle_{\mathcal{H}} e_i = \sum_{i=1}^{m} \alpha_i e_i,$$

and α_i is defined by,

$$\alpha_i = \frac{1}{\sqrt{D_{i-1} D_i}} \begin{vmatrix} \langle x_1, x_1 \rangle_{\mathcal{H}} & \dots & \dots & \langle x_1, x_i \rangle_{\mathcal{H}} \\ \vdots & \ddots & \ddots & \vdots \\ \langle x_{i-1}, x_1 \rangle_{\mathcal{H}} & \dots & \dots & \langle x_{i-1}, x_i \rangle_{\mathcal{H}} \\ \langle x, x_1 \rangle_{\mathcal{H}} & \dots & \dots & \langle x, x_i \rangle_{\mathcal{H}} \end{vmatrix}.$$

Proof. This simply follows from *Leibniz formula for determinants* applying to e_i obtaining from Theorem 20. □

11 BACKGROUND

Here we review some math background for completeness. We refer to standard references for a more detailed discussion.

11.1 Leibniz Formula

Theorem 19. *Let A be a $n \times n$ square matrix. Then we have the following for its determinant,*

$$\det(A) = \sum_{\tau \in S_n} \operatorname{sign}(\tau) \prod_{i=1}^{n} a_{i\tau(i)}$$

, where a_{ij} denotes the i and j entries and S_n denotes the symmetric group on n letters and for $\tau \in S_n$, $\text{sign}(\tau)$ *denotes the sign of the permutation.*

Remark 14. *When computing the formal determinant, as exemplified in Theorem 20, we calculate a formal determinant where the last row consists of vectors and the other rows consist of scalars. In this scenario, the determinant can be interpreted using the Leibniz formula mentioned earlier, where for $\tau \in S_n$,* $\text{sign}(\tau) \prod_{i=1}^{n} a_{i\tau(i)}$ *represents the product of $n-1$ scalars and one vector. Consequently, the resultant outcome can be viewed as a vector.*

11.2 Tensor Product and Musical Isomorphism

Here, we review some properties of the tensor product and the dual of a vector space, as utilized in the statement of Proposition 10. These properties are well-documented in standard references; see, for example, (Lang, 2012). Let V and W be two vector spaces. Their tensor product $V \otimes W$ is a vector space consisting of all formal sums:

$$V \otimes W = \left\{ \sum_{i=1}^{n} v_i \otimes w_i \mid v_i \in V, w_i \in W, n \in \mathbb{N} \right\},$$

such that linearity is preserved in each coordinate. Specifically, for $a \in \mathbb{R}$:

$$a \cdot \sum_{i=1}^{n} v_i \otimes w_i = \sum_{i=1}^{n} (a \cdot v_i) \otimes w_i = \sum_{i=1}^{n} v_i \otimes (a \cdot w_i).$$

When V is n-dimensional with a basis $\{v_1, \ldots, v_n\}$, and W is m-dimensional with a basis $\{w_1, \ldots, w_m\}$, it can be shown that $V \otimes W$ is nm-dimensional, with a basis given by $\{v_i \otimes w_j \mid 1 \leq i \leq n, 1 \leq j \leq m\}$. Next, let V be a finite-dimensional vector space with an inner product $\langle \cdot, \cdot \rangle_g$. The dual space V^* consists of all linear functionals on V. The space of linear maps, such as $T : V \to V$, can be represented as $V \otimes V^*$. Moreover, V can be naturally identified with V^* via the so-called *musical isomorphism*, defined as:

$$\flat : v \in V \mapsto \langle v, \cdot \rangle_g \in V^*,$$

$$\sharp : \langle v, \cdot \rangle_g \in V^* \mapsto v \in V.$$

Therefore, the space of linear maps on V, identified with $V \otimes V^*$, can further be identified with $V \otimes V$ using the musical isomorphism. This identification is utilized in the statement of Proposition 10.

11.3 Gram-Schmidt Process

In this section, we recall the determinant form of the Gram-Schmidt process. We refer to (Gantmacher, 1959) for detailed discussion. Let $\mathcal{H}$, equipped with $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ be an inner product space.

Theorem 20 (Gram-Schmidt Determinant Formula)**.** *Let $\{x_1, ..., x_m\} \in \mathcal{H}$ be linearly independent and set,*

$$V = \text{span}\{x_1, ..., x_m\} \subset \mathcal{H}.$$

Then $\{e_i\}_{i=1}^{m}$ forms an orthonormal basis for V where for $1 \leq i \leq m$, e_i is defined by,

$$e_i = \frac{1}{\sqrt{D_{i-1} D_i}} \begin{vmatrix} \langle x_1, x_1 \rangle_{\mathcal{H}} & \ldots & \ldots & \langle x_1, x_i \rangle_{\mathcal{H}} \\ \vdots & \ddots & \ddots & \vdots \\ \langle x_{i-1}, x_1 \rangle_{\mathcal{H}} & \ldots & \ldots & \langle x_{i-1}, x_i \rangle_{\mathcal{H}} \\ x_1 & \ldots & \ldots & x_i \end{vmatrix},$$

where for each $1 \leq i \leq m$, we compute a formal $i \times i$ determinant and $D_0 = 1$ and for $1 \leq i \leq m$, D_i is defined as follows,

$$D_i = \begin{vmatrix} \langle x_1, x_1 \rangle_{\mathcal{H}} & \dots & \dots & \langle x_1, x_i \rangle_{\mathcal{H}} \\ \vdots & \ddots & \ddots & \vdots \\ \langle x_{i-1}, x_1 \rangle_{\mathcal{H}} & \dots & \dots & \langle x_{i-1}, x_i \rangle_{\mathcal{H}} \\ \langle x_i, x_1 \rangle_{\mathcal{H}} & \dots & \dots & \langle x_i, x_i \rangle_{\mathcal{H}} \end{vmatrix}.$$

Remark 15. *When computing the formal determinant, as exemplified in Theorem 20, we calculate a formal determinant where the last row consists of vectors and the other rows consist of scalars. In this scenario, the determinant can be interpreted using the Leibniz formula mentioned earlier, where for $\tau \in S_n$, $\operatorname{sign}(\tau) \prod_{i=1}^{n} a_{i\tau(i)}$ represents the product of $n-1$ scalars and one vector. Consequently, the resultant outcome can be viewed as a vector.*

11.4 Linear Operators on Hilbert space

Here, we review some standard facts about bounded operators on Hilbert spaces. For a more detailed discussion, see (Folland, 1999). Throughout this subsection, we assume $\mathcal{H}$ is a (real) Hilbert space equipped with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and associated norm $\|\cdot\|_{\mathcal{H}}$.

Definition 21. *A linear operator,*

$$A : \mathcal{H} \to \mathcal{H},$$

is called bounded if one of the following equivalent conditions holds,

- *A is continuous at $0 \in \mathcal{H}$,*
- *A is continuous,*
- *There exist $c > 0$ such that $\|Ax\|_{\mathcal{H}} \leq c$ for all $x \in \mathcal{H}$ with $\|x\|_{\mathcal{H}} \leq 1$,*
- *There exist $c > 0$ such that $\|Ax\|_{\mathcal{H}} \leq c \|x\|_{\mathcal{H}}$ for all $x \in \mathcal{H}$,*

Remark 16. *When $\mathcal{H}$ is n-dimensional (e.g., $\mathbb{R}^n$ equipped with the standard inner product), the linear map A can be represented by an $n \times n$ matrix M which depends on the basis chosen for $\mathcal{H}$. Therefore we can think of bounded linear operators as natural generalizations of matrices.*

Theorem 22 (Riesz Representation Theorem)**.** *Let $\phi : \mathcal{H} \to \mathbb{R}$ be a bounded functional. Then there exist a unique $z_\phi \in \mathcal{H}$, denoted by Riesz representation of ϕ, such that for all $x \in \mathcal{H}$,*

$$\phi(x) = \langle x, z_\phi \rangle_{\mathcal{H}}.$$

Definition 23. *Let*

$$A : \mathcal{H} \to \mathcal{H},$$

be a bounded linear operator. The adjoint of A, denoted by A^* is bounded linear operator on $\mathcal{H}$ that for any $x, y \in \mathcal{H}$ satisfies,

$$\langle Ax, y \rangle = \langle x, A^* y \rangle.$$

An operator A is called self-adjoint if $A = A^*$

Remark 17. *The existence and uniqueness of the adjoint follow from the Riesz representation theorem. In the case where $\mathcal{H}$ is n-dimensional (e.g., $\mathbb{R}^n$ equipped with the standard inner product), and the linear map A is represented by an $n \times n$ matrix M, then the adjoint of A corresponds to its transpose, denoted as M^T. Therefore, self-adjoint operators on Hilbert spaces can be considered as a generalization of symmetric matrices.*

Definition 24. *A bounded linear operator on $\mathcal{H}$ is called strictly positive (or bounded from below) if there exist $c > 0$ such that for any $x \in \mathcal{H}$,*

$$\langle x, Ax \rangle_{\mathcal{H}} > c \cdot \|x\|^2_{\mathcal{H}}.$$

Remark 18. *In the case where $\mathcal{H}$ is n-dimensional (e.g., $\mathbb{R}^n$ equipped with the standard inner product), and the linear map A is represented by an $n \times n$ matrix M, the positivity of A translates to M being a positive definite matrix. Consequently, positive operators on Hilbert spaces can be regarded as a generalization of positive definite matrices.*

We next recall the notion norm equivalence of vector spaces,

Definition 25. *Let g_1, g_2 be two inner product on a Hilbert space $\mathcal{H}$. We say g_1 and g_2 are equivalent if there exist constant c_1 and c_2 such that for all $x \in \mathcal{H}$,*

$$c_1 g_1(x, x) \leq g_2(x, x) \leq c_2 g_1(x, x).$$

We next recall the definition of space of generalized Mahalanobis inner product from main body,

Definition 26 (Space of generalized Mahalanobis inner products)**.** *Space of generalized Mahalanobis inner product on a Hilbert space $\mathcal{H}$ is defined the following set,*

$$\mathcal{F}_{\mathcal{H}} := \{A : \mathcal{H} \to \mathcal{H} | A \text{ is bounded, strictly positive, and self-adjoint}\}.$$

In light of the above discussion, we can regard elements of $\mathcal{F}_{\mathcal{H}}$ as a generalized version of symmetric, positive-definite matrices in finite-dimensional space.

11.5 Kernels and RKHS

In this subsection, we review some standard concepts related to kernels and reproducing kernel Hilbert spaces that is used in the main body of the paper. These can be found in standard references such as (Smola and Schölkopf, 1998; Song et al., 2009). Throughout this discussion, let $\mathcal{X}$ denote the feature space.

Definition 27. *A (real) kernel on $\mathcal{X}$ is a mapping*

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R},$$

such that, it is symmetric, i.e. for any $x, y \in \mathcal{X}$,

$$k(x, y) = k(y, x),$$

and for any $\{x_1, ..., x_n\} \subset \mathcal{X}$ the corresponding $n \times n$ Gram matrix K defined by,

$$K_{ij} = k(x_i, x_j),$$

is positive definite.

Given a kernel function k is given one can construct a special Hilbert space associated to it called universal RKHS. First, consider the following vector space,

$$\mathcal{H} := \{f | f : \mathcal{X} \to \mathbb{R}\},$$

and consider the feature map,

$$\Phi : \mathcal{X} \to \mathcal{H}$$
$$\Phi(x) := k(x, .).$$

Next, let $\mathcal{H}_k$ denotes the vector space generated by $\Phi(\mathcal{X})$. More formally,

$$\mathcal{H}_k := \{\sum_{i=1}^{m} a_i k(., x_i) | a_i, x_i \in \mathbb{R}, m \in \mathbb{N}\} \subset \mathcal{H}$$

One can define the inner product on $\mathcal{H}_k$ first by,

$$\langle k(x,.), k(y,.)\rangle := k(x,y),$$

and extend above to $\mathcal{H}_k$ by linearity. Notice that above implies that for $f \in \mathcal{H}_k$,

$$f(x) = \langle f, k(x,.)\rangle., \qquad (*)$$

which is referred to as reproducing property. The completion of $\mathcal{H}_k$ (also denoted by $\mathcal{H}_k$) is termed as the "reproducing kernel Hilbert space" associated with the kernel k. It can be demonstrated that there exists a unique RKHS linked to a kernel function, whose elements reside in $\mathcal{H}$, as stated in the following theorem.

Theorem 28 ((Aronszajn, 1950)). *Let k be a symmetric positive definite kernel. Then there exist a unique Hilber space of functions for which k is reproducing kernel.*

References

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3): 337–404, 1950.

A. Bellet, A. Habrard, and M. Sebban. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013.

R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

F. Cakir, K. He, X. Xia, B. Kulis, and S. Sclaroff. Deep metric learning to rank. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1861–1870, 2019.

G. Canal, B. Mason, R. K. Vinayak, and R. Nowak. One for all: Simultaneous metric and preference learning over multiple users. In *NeurIPS*, 2022.

R. Chatpatanasiri, T. Korsrilabutr, P. Tangchanachaianan, and B. Kijsirikul. A new kernelization framework for mahalanobis distance learning algorithms. *Neurocomputing*, 73(10-12):1570–1579, 2010.

S. L. Chau, M. Cucuringu, and D. Sejdinovic. Spectral ranking with covariates. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 70–86. Springer, 2022.

W. Chu and Z. Ghahramani. Preference learning with gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*, pages 137–144, 2005.

M. Cucuringu, I. Koutis, S. Chawla, G. Miller, and R. Peng. Simple and scalable constrained clustering: a generalized spectral method. In *Artificial Intelligence and Statistics*, pages 445–454. PMLR, 2016.

F. Fogel, A. d'Aspremont, and M. Vojnovic. Spectral ranking using seriation. *Journal of Machine Learning Research*, 17(88):1–45, 2016.

G. B. Folland. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.

J. Fürnkranz and E. Hüllermeier. Preference learning and ranking by pairwise comparison. In *Preference learning*, pages 65–82. Springer, 2010.

F. R. Gantmacher. The theory of matrices. *Co., New York*, 2, 1959.

B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley. Spectral, probabilistic, and deep metric learning: Tutorial and survey. *arXiv preprint arXiv:2201.09267*, 2022.

C.-K. Hsieh, L. Yang, Y. Cui, T.-Y. Lin, S. Belongie, and D. Estrin. Collaborative metric learning. In *Proceedings of the 26th international conference on world wide web*, pages 193–201, 2017.

L. Jain, K. G. Jamieson, and R. Nowak. Finite sample prediction and recovery bounds for ordinal embedding. *Advances in neural information processing systems*, 29, 2016.

L. Jain, A. Gilbert, and U. Varma. Spectral methods for ranking with scarce data. In *Conference on Uncertainty in Artificial Intelligence*, pages 609–618. PMLR, 2020.

P. Jain, B. Kulis, J. V. Davis, and I. S. Dhillon. Metric and kernel learning using a linear transformation. *The Journal of Machine Learning Research*, 13(1):519–547, 2012.

K. G. Jamieson and R. Nowak. Active ranking using pairwise comparisons. *Advances in neural information processing systems*, 24, 2011a.

K. G. Jamieson and R. D. Nowak. Low-dimensional embedding using adaptively selected ordinal data. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1077–1084. IEEE, 2011b.

G. Kimeldorf and G. Wahba. Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95, 1971.

B. Kulis et al. Metric learning: A survey. *Foundations and Trends® in Machine Learning*, 5(4):287–364, 2013.

J. T. Kwok and I. W. Tsang. Learning with idealized kernels. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 400–407, 2003.

S. Lang. *Algebra*, volume 211. Springer Science & Business Media, 2012.

P. C. Mahalanobis. On the generalized distance in statistics. National Institute of Science of India, 1936.

B. Mason, L. Jain, and R. Nowak. Learning low-dimensional metrics. *Advances in neural information processing systems*, 30, 2017.

A. K. Massimino and M. A. Davenport. As you like it: Localization via paired comparisons. *Journal of Machine Learning Research*, 22(186):1–39, 2021.

S. Negahban, S. Oh, and D. Shah. Iterative ranking from pair-wise comparisons. *Advances in neural information processing systems*, 25, 2012.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36: 53728–53741, 2023.

B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *Computational Learning Theory: 14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT 2001 Amsterdam, The Netherlands, July 16–19, 2001 Proceedings 14*, pages 416–426. Springer, 2001.

M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. *Advances in neural information processing systems*, 16, 2003.

A. J. Smola and B. Schölkopf. *Learning with kernels*, volume 4. Citeseer, 1998.

L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968, 2009.

D. M. Stuart-Fox, D. Firth, A. Moussalli, and M. J. Whiting. Multiple signals in chameleon contests: designing and analysing animal contests as a tournament. *Animal Behaviour*, 71(6):1263–1271, 2006.

J. L. Suárez, S. García, and F. Herrera. A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges. *Neurocomputing*, 425:300–322, 2021.

S. Vigna. Spectral ranking. *Network Science*, 4(4):433–445, 2016.

G. Wahba. *Spline models for observational data*. SIAM, 1990.

M. J. Whiting, J. K. Webb, and J. S. Keogh. Flat lizard female mimics use sexual deception in visual but not chemical signals. *Proceedings of the Royal Society B: Biological Sciences*, 276(1662):1585–1591, 2009.

A. Xu and M. Davenport. Simultaneous preference and metric learning from paired comparisons. *Advances in Neural Information Processing Systems*, 33:454–465, 2020.

H.-J. Ye, D.-C. Zhan, and Y. Jiang. Fast generalization rates for distance metric learning: Improved theoretical analysis for smooth strongly convex distance metric learning. *Machine Learning*, 108:267–295, 2019.