

# Fair Kernel Regression via Fair Feature Embedding in Kernel Space

Austin Okray  
Department of Computer Science  
University of Wyoming  
Laramie, Wyoming  
Email: aokray@uwyo.edu

Hui Hu  
Department of Computer Science  
University of Wyoming  
Laramie, Wyoming  
Email: hhu1@uwyo.edu

Chao Lan  
Department of Computer Science  
University of Wyoming  
Laramie, Wyoming  
Email: clan@uwyo.edu

**Abstract**—In recent years, there have been significant efforts on mitigating unethical demographic biases in machine learning methods. However, very little work is done for kernel methods. In this paper, we propose a novel fair kernel regression method via fair feature embedding (FKR-F<sup>2</sup>E) in kernel space. Motivated by prior works feature processing for fair learning and feature selection for kernel methods, we propose to learn fair feature embeddings in kernel space, where the demographic discrepancy of feature distributions is minimized. Through experiments on three public real-world data sets, we show the proposed FKR-F<sup>2</sup>E achieves significantly lower prediction disparity compared with the state-of-the-art fair kernel regression method and several other baseline methods.

## I. INTRODUCTION

In recent years, we've witnessed a tremendous growth of machine learning applications in real-world problems that have immediate impacts on peoples' lives. However, standardly learned models can have unethical predictive biases against minority peoples; e.g., in recidivism prediction, a commercialized model has significant bias against innocent black defendants [1]; other biases are found in hiring [2], facial verification [3], violence risk assessment in prison [4], etc.

How to learn fair models has become a significant research topic [5], and many methods have been proposed [6]–[12]. They typically sacrifice certain prediction accuracy for improving prediction fairness, bound to the accuracy-fairness tradeoff.

A promising direction is fair kernel learning [13], [14]. By constructing sufficiently complex hypothesis spaces, they are more likely to learn a model that can achieve an efficient accuracy-fairness trade-off. However, this direction is sparsely explored so far. A notable work is fair kernel regression [13], which penalizes a model's predictive bias in kernel space.

In this paper, we propose a novel fair kernel regression method that learns fair feature embeddings (FKR-F<sup>2</sup>E) in the kernel space. It is motivated by the work of Feldman et al [7], which shows that in a properly transformed data space where different demographic groups have similar feature distributions, a standardly learned prediction model will be naturally fair. We thus seek for such a fair transformation in the kernel space.

A major challenge is that kernel space is often implicit, making it hard to find explicit fair transformations therein. To tackle the problem, we borrow ideas from Cao et al [15], which learns feature embeddings in the kernel space for

feature selection. Specifically, we propose to learn fair feature embeddings in the kernel space, such that different demographic groups have similar embedded feature distributions. We propose to measure similarity using mean discrepancy [16].

Through experiments on three real-world data sets, we show the proposed FKR-F<sup>2</sup>E achieves significantly lower prediction bias than the existing fair kernel regression method as well as several non-kernel fair learning methods, without sacrificing a significant amount of prediction accuracy.

The rest of the paper is organized as follows: in Section II, we revisit related works; in Section III, we present the proposed method; in Section IV, experimental results are presented and discussed; our conclusion is in Section V.

## A. Notations and Assumptions

To facilitate discussions in related work, we introduce some notations here. We describe an instance using a triple  $(x, s, y)$ , where  $x$  is a feature vector,  $s$  is a protected demographic (e.g. gender, race) and  $y$  is label. Assume  $s$  is contained in  $x$ .

Similar to prior studies, we assume  $s$  is binary. Let there be  $n$  instances in the training set, among which  $n_u$  belong to the unprotected group ( $s = 0$ ) and  $n_p$  belong to the protected group ( $s = 1$ ). Without loss of generality, we assume the instances are ordered such that the first  $n_u$  ones  $x_1, \dots, x_{n_u}$  are unprotected and the rest  $x_{n_u+1}, \dots, x_n$  are protected.

For kernel methods, let  $\phi(\cdot)$  be the feature mapping function, and  $f$  be a prediction model mapping from  $\phi(x)$  to  $y$ .

## II. RELATED WORK

### A. Fair Kernel Regression

Perez-Suay et al [13] propose a fair kernel regression method, which directly extends the linear fair learning method [9] to kernel space. Specifically, it minimizes prediction loss while additionally penalizing the correlation between model prediction and demographic feature in the kernel space as:

$$\min_f \sum_{i=1}^n [f(\phi(x_i)) - y_i]^2 + \mu \sum_{i=1}^n (\tilde{f}(x_i) \cdot \tilde{s}_i) + \lambda \Omega(f), \quad (1)$$

where the second term measures predictive bias as the correlation between model prediction and the demographic feature, and  $\tilde{f}(x)$  and  $\tilde{s}$  are centered variables; the last term measures model complexity;  $\mu$  and  $\lambda$  are hyper-parameters. Based on the

Representer Theorem that  $f$  is a linear combination of  $\phi(x_i)$ 's, task (1) admits an analytic solution for the linear coefficients.

Perez et al's method adopts the regularization approach in fair learning, which penalizes predictive bias during learning (e.g., [8], [9]). In this paper, we adopt another popular approach which first constructs a fair feature space and then builds a standard model in it (e.g., [6], [7], [12]). In experiments, we show our method can achieve higher prediction fairness.

### B. Fair Feature Learning and Mean Discrepancy

An effective approach to learn fair models is to first construct a fair feature space and then learn a standard model in it (e.g., [6], [7]). A fair feature space is one where feature distributions of different demographic groups are similar, e.g., different groups have similar CDF's of the new features [7], or the statistical dependence between the new features and the demographic feature is low [6].

In this paper, we develop a new fair kernel regression based on the idea of fair feature learning. Unlike previous studies, we measure feature similarity using mean discrepancy [16]. MD measures distance between distributions and is widely used in machine learning [17]–[19]. Let  $x_1, \dots, x_n$  and  $z_1, \dots, z_m$  be two sets generated from distributions  $P_x$  and  $P_z$  respectively. MD estimates the distance between  $P_x$  and  $P_z$  as

$$MD(P_x, P_z) = \left\| \frac{1}{n} \sum_{i=1}^n \phi(x_i) - \frac{1}{m} \sum_{j=1}^m \phi(z_j) \right\|^2. \quad (2)$$

A technical challenge is that previous fair feature learning approaches assume the feature space is explicit and then modify it to obtain a fairer space. In kernel methods, however, the feature space of  $\phi(x)$  is implicit. To tackle this issue, we propose to construct an explicit fair feature space for  $\phi(x)$ , by learning fair feature embedding functions in the kernel space. This approach is motivated by the literature of feature selection in kernel methods (e.g., [15], [20]).

### C. Feature Selection in Kernel Space

Feature selection is a common practice for improving the robustness and interpretability of machine learning models [21]. However, its practice in kernel methods is not easy, since there is not an explicit feature representation in kernel space. Only a few approaches are proposed, e.g. [15], [20], [22].

Our study is motivated by Cao et al [15]. They propose to learn explicit feature representation in the kernel space, by learning feature embedding function  $\eta$ . They show the optimal function is a linear combination of training instances, and learn such functions by standard methods such as KPCA [23]. After that, instance  $\phi(x)$  is mapped onto  $\eta$  to obtain an explicit feature representation on which feature selection is performed.

Motivated by Cao et al's approach, we propose to learn feature embedding functions that are fair, e.g., different demographic groups have similar distributions in the embedded space. As explained in the previous subsection, similarity is measured by mean discrepancy.

## III. FAIR KERNEL REGRESSION VIA LEARNING FAIR FEATURE EMBEDDINGS IN KERNEL SPACE (FKR-F<sup>2</sup>E)

In this section, we present the proposed fair kernel regression via learning fair feature embeddings in kernel space (FKR-F<sup>2</sup>E). Recall an individual is  $(x, s, y)$ , where  $x$  is feature vector,  $s$  is binary demographic feature and  $y$  is label. There are  $n$  training instances, where  $x_1, \dots, x_{n_u}$  are from the unprotected group and  $x_{n_u+1}, \dots, x_n$  are from the protected group.

Our proposed method works in two steps: (i) learn fair feature embeddings in kernel space; (ii) build a standard regression model based on the embedded features.

### Step 1. Learn Fair Feature Embeddings in Kernel Space

Our goal is to learn an explicit and fair feature representation for  $\phi(x)$ . To that end, we propose to learn a fair feature embedding function  $\eta$ , such that in the embedded space, the mean discrepancy between the protected group and unprotected group is minimized:

$$\min_{\eta} \left\| \frac{1}{n_u} \sum_{i=1}^{n_u} \langle \phi(x_i), \eta \rangle - \frac{1}{n_p} \sum_{i=n_u+1}^n \langle \phi(x_i), \eta \rangle \right\|^2. \quad (3)$$

Problem (3) cannot be directly solved since there is no explicit representation of  $\phi(x)$ . Motivated by Cao et al [15], we assume the optimal  $\eta$  is a linear combination of training instances:

$$\eta = \sum_{i=1}^n \alpha_i \phi(x_i). \quad (4)$$

To avoid overfitting, we further assume  $\eta$  has a unit norm:

$$\|\eta\|^2 = 1. \quad (5)$$

Solving (3) under constraints (4) and (5), we have that<sup>1</sup>

$$\left( \frac{1}{n_u^2} K_u^T K_u - \frac{2}{n_u n_p} K_u^T \mathbf{1}_u \mathbf{1}_p^T K_p + \frac{1}{n_p^2} K_p^T K_p \right) \alpha = \lambda K \alpha, \quad (6)$$

where  $\alpha = [\alpha_1, \dots, \alpha_n]^T$  is the vector of unknown parameters and  $\lambda$  is an eigenvalue;  $K$  is a standard  $n$ -by- $n$  Gram matrix of all instances;  $K_u$  is  $n_u$ -by- $n$  and  $K_p$  is  $n_p$ -by- $n$  satisfying

$$K = \begin{bmatrix} K_u \\ K_p \end{bmatrix} \quad (7)$$

Formula (6) is a generalized eigenproblem, and  $\alpha$  is the least generalized eigenvector. After  $\alpha$  is solved, we obtain the first explicit and fair feature of  $\phi(x)$  in the kernel space as

$$\langle \phi(x), \eta \rangle = \sum_{i=1}^n \alpha_i k(x, x_i). \quad (8)$$

The above analysis gives the first fair feature embedding function  $\eta$  in the kernel space. Now we present how to obtain the second  $\eta'$ , and the rest can be derived in similar fashions.

The second optimal embedding  $\eta'$  is obtained in a similar fashion as  $\eta$ , with an additional constraint that it should be orthogonal to the previously obtained embeddings:

$$\min_{\eta'} \left\| \frac{1}{n_u} \sum_{i=1}^{n_u} \langle \phi(x_i), \eta' \rangle - \frac{1}{n_p} \sum_{i=n_u+1}^n \langle \phi(x_i), \eta' \rangle \right\|^2 \quad (9)$$

$$s.t. \quad \eta' = \sum_{i=1}^n \alpha'_i \phi(x_i), \quad \|\eta'\|^2 = 1, \quad \eta^T \eta' = 0.$$

<sup>1</sup>Detailed arguments are in Appendix A.

Solving (9) shows that  $\alpha' = [\alpha'_1, \dots, \alpha'_n]^T$  is the second least generalized eigenvector of the same eigenproblem (6)<sup>2</sup>.

By similar arguments, we can show the linear coefficients of  $k$  optimal fair embeddings  $\eta_1, \dots, \eta_k$  are the least  $k$  generalized eigenvectors of the eigenproblem (6).

After that, we obtain a  $k$ -dimensional explicit fair feature representation of  $\phi$  in the kernel space, i.e.,

$$\phi_{FS}(x) = [\langle \phi(x), \eta_1 \rangle, \dots, \langle \phi(x), \eta_k \rangle]^T. \quad (10)$$

*Step 2. Learn a Standard Regression Model on  $\phi_{FS}(x)$*

Given an explicit fair feature representation  $\phi_{FS}(x)$ , we learn a standard regression model based on it. Let  $x_1, \dots, x_n$  be  $n$  training instances. One can easily verify that

$$\phi_{FS}(x_i) = [\langle \phi(x_i), \eta_1 \rangle, \dots, \langle \phi(x_i), \eta_k \rangle]^T = (K_{:,i}^T A)^T \quad (11)$$

where  $K_{:,i}$  is the  $i^{th}$  column Gram matrix  $K$ , and  $A$  is an  $n$ -by- $k$  matrix with column  $j$  being the linear coefficient vector of  $\eta_j$  (e.g., the first column is  $\alpha$  and the second column is  $\alpha'$ ).

Then, one can obtain an  $n$ -by- $k$  training sample matrix

$$X_{FS} = \begin{bmatrix} (\phi_{FS}(x_1))^T \\ \vdots \\ (\phi_{FS}(x_n))^T \end{bmatrix} = \begin{bmatrix} K_{:,1}^T A \\ \vdots \\ K_{:,n}^T A \end{bmatrix} = K^T A = KA. \quad (12)$$

Now, we learn a regression model  $\beta \in \mathbb{R}^k$  on  $X_{FS}$  by

$$\min_{\beta} \|X_{FS} \cdot \beta - Y\|^2 + \gamma \|\beta\|^2, \quad (13)$$

where  $\gamma$  is a regularization coefficient.

For any testing instance  $z$ , we first compute its explicit fair feature representation

$$\phi_{FS}(z) = [\langle \phi(z), \eta_1 \rangle, \dots, \langle \phi(z), \eta_k \rangle]^T, \quad (14)$$

and then compute its prediction as

$$\hat{y} = \phi_{FS}(z)^T \beta. \quad (15)$$

For classification tasks, one can simply threshold  $\hat{y}$ .

#### IV. EXPERIMENT

##### A. Data Sets

We experimented on three public data sets, namely, the Credit Default data set<sup>3</sup>, the Community Crime data set<sup>4</sup>, and the COMPAS data set<sup>5</sup>.

The original Credit Default data set contains 30,000 individuals described by 23 attributes. We treated ‘education level’ as the sensitive variable, and binarized it into higher education and lower education as in [12]; ‘default payment’ is treated as the binary label. We removed individuals with missing values and down-sampled the data set from 30,000 to 20,000. Our preprocessed data sets are published at <sup>6</sup>.

<sup>2</sup>Detailed arguments are in Appendix B.

<sup>3</sup><https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>

<sup>4</sup><http://archive.ics.uci.edu/ml/datasets/communities+and+crime>

<sup>5</sup><https://github.com/propublica/compas-analysis>

<sup>6</sup><https://uwymachinelearning.github.io/>

The Communities Crime data set contains 1,993 communities described by 101 informative attributes. We treated the ‘fraction of African-American residents’ as the sensitive feature, and binarized it so that a community is ‘minority’ if the fraction is above 0.5 and ‘majority’ otherwise. Label is the ‘community crime rate’, and we binarized it into high if the rate is above 0.5 and low otherwise.

The COMPAS data set contains 18,317 individuals with 40 features (e.g., name, sex, race). We down-sampled the data set to 16,000 instances and 15 numerical features (e.g. name is removed). Similar to [24], we treated ‘race’ as the sensitive feature and ‘risk of recidivism’ as the binary label.

##### B. Experiment Design

On each data set, we randomly chose 75% of the instances for training and used the rest for testing. We evaluated each method over 50 random trials and reported its average performance and standard deviation.

We compared the proposed FKR-F<sup>2</sup>E with the existing fair kernel regression [13], and several other non-kernel methods. For each compared method, we set its hyper-parameters as described in the original paper.

For FKR-F<sup>2</sup>E, we used polynomial kernel on both Credit and Community Crime data sets and sigmoid kernel on the COMPAS data set. For polynomial kernel, we grid-searched its optimal degree in  $\{3, 4, 5, 6\}$  and optimal additive coefficient in  $[10^{-3}, 10^{-2}, 10^{-1}, 10^0, 10^1]$ . For sigmoid kernel, we grid-searched its optimal  $c$  among 5 values in the logarithmic range of  $[10^{-4}, 10^1]$ , and we used the default  $\gamma$  in Scikit-Learn [25] (i.e., inverse of feature dimension). For the ridge regression regularization coefficient  $\lambda$ , we grid-searched an optimal value among 6 values in the logarithmic range  $[10^{-3}, 10^2]$ .

Finally, an important hyper-parameter is the number of feature embeddings  $k$ . We experimented with 4 values, namely,  $\frac{n}{250}$ ,  $\frac{n}{200}$ ,  $\frac{n}{150}$  and  $\frac{n}{100}$ . In experiment these values yielded good generalization performance on all data sets.

We evaluated model accuracy using the standard classification error (Error), and evaluated model fairness using a popular measure called statistical disparity (SD) [11], defined as:

$$SD(f, S) = |p(f(x) = 1 \mid s = 1) - p(f(x) = 1 \mid s = 0)|. \quad (16)$$

Finally, all experiments were run on the Teton Computing Environment at the University of Wyoming’s Advanced Research Computing Center (<https://doi.org/10.15786/M2FY47>), and our FFE implementation is at <https://github.com/aokray/FFE>.

##### C. Classification Results and Discussions

Our classification results are summarized in Table I.

Our first observation is that FKR-F<sup>2</sup>E consistently achieves lower statistical disparity than the existing fair kernel regression method (and other baselines) across the three data sets. This implies that fair feature embedding is an effective approach for learning fair models in kernel space.

We notice the superior fairness of FKR-F<sup>2</sup>E is not achieved without any cost. In general, it has slightly higher prediction error than the existing fair kernel regression and other baselines.

Method	Credit Default		Communities Crime		COMPAS	
	SD	Error	SD	Error	SD	Error
FKR-F <sup>2</sup> E	<b>.0021</b> ±.0017	.2277±.0050	<b>.0392</b> ±.0267	.1384±.0125	<b>.0025</b> ±.0018	.2307±.0057
FKRR [13]	.0079±.0011	<b>.2001</b> ±.0054	.0968±.0722	.1208 ±.0054	.0041±.0013	<b>.2190</b> ±.0089
FLR [9]	.0779±.0571	.2412±.0469	.0898±.0971	.1166±.0189	.0408±.0162	.2428±.0917
FRR [11]	.0186±.0016	.2914±.0186	.3062±.0452	<b>.1102</b> ±.0128	.0182±.0042	.2276±.0040
FPCA [12]	.1716±.0149	.4025±.0382	.0859±.0479	.1731±.0089	.2806±.0182	.3204±.1032

TABLE I: Classification Performance of Different Methods across Different Data Sets. For polynomial kernel, we set degree as 4 and additive coefficient as 0.1. For sigmoid kernel, we used  $c = 0.01$  and  $\gamma$  as the inverse of feature dimension.  $k$  is set to  $\frac{n}{250}$ .

However, we argue the loss of accuracy is small compared with the increase of fairness. For example, on the Credit Default data set, FKR-F<sup>2</sup>E lowers prediction disparity by at least 75% =  $(0.0079-0.0021)/0.0079$  but only increases prediction error by at most 13% =  $(0.2277-0.2001)/0.2001$ . We thus argue this method has a more efficient accuracy-fairness trade-off.

Finally, we see fair kernel methods generally achieve lower statistical disparity than other fair learning methods, suggesting their promisingness for fair machine learning.

#### D. Sensitivity Analysis

In this section, we examined the performance of FKR-F<sup>2</sup>E on the Communities Crime data set under different configurations.

We first examined its performance with different choices of kernel. Results on testing samples averaged over 50 random trials are reported in Figure 1. We see that polynomial kernel achieves the highest prediction fairness, with slightly higher prediction error. Sigmoid kernel is the second best, and linear kernel does not give low disparity. This supports our hypothesis that why fair kernel methods are promising – they construct a complex hypothesis space that is more likely to include models with efficient fairness-accuracy trade-off.

Next, we examined performance with polynomial kernel under different  $k$  (number of feature embeddings). Results are shown in Figure 2. We see that smaller  $k$  generally leads to higher prediction fairness and slightly higher prediction error. The former phenomenon implies that only the least eigenvectors of problem (6) can effectively minimize the mean discrepancy between two groups. The latter is easy to understand – higher feature dimension provides more information for building an accurate prediction model. However, the variation versus  $k$  seems quite limited, suggesting our method has robust classification performance.

#### V. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a novel fair kernel regression method FKR-F<sup>2</sup>E. It first learns a set of fair feature embeddings in the kernel space, and then standardly learns a prediction model in the embedded space. Through experiments across three real-world data sets, we show it achieves significantly lower bias in prediction compared with the state-of-the-art fair kernel regression method as well as several non-kernel fair

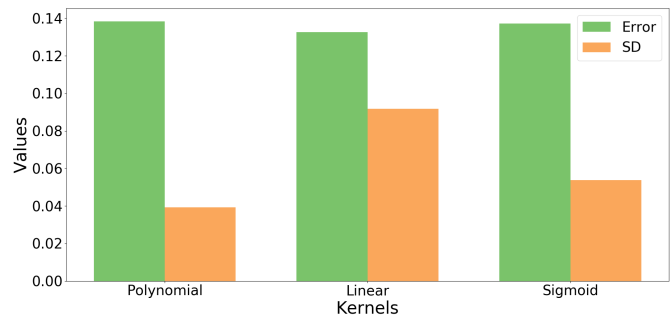


Fig. 1: Sensitivity analysis for varying kernels with their approximately optimal number of features selected.

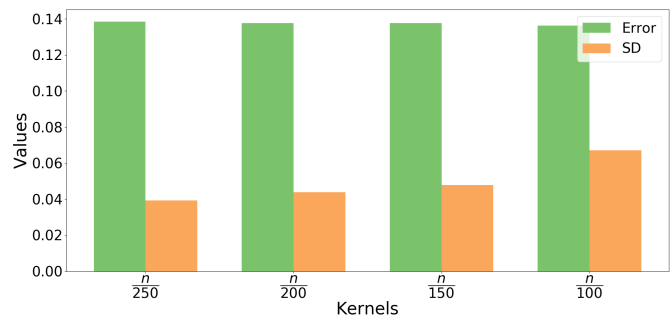


Fig. 2: Performance versus  $k$ .

learning methods, while sacrificing only a small amount of prediction accuracy.

#### REFERENCES

- [1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, “Machine bias: Theres software used across the country to predict future criminals. and its biased against blacks,” *ProPublica*, 2016.
- [2] M. Hoffman, L. B. Kahn, and D. Li, “Discretion in Hiring\*,” *The Quarterly Journal of Economics*, 2017.
- [3] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain, “Face recognition performance: Role of demographic information,” *IEEE Transactions on Information Forensics and Security*, 2012.
- [4] M. D. Cunningham and J. R. Sorensen, “Actuarial models for assessing prison violence risk: Revisions and extensions of the risk assessment scale for prison (rasp),” *Assessment*, 2006.
- [5] Press *et al.*, “Preparing for the future of artificial intelligence,” 2016.
- [6] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, “Learning fair representations,” in *ICML*, 2013.

- [7] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *KDD*, 2015.
- [8] T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang, "Controlling attribute effect in linear regression," *ICDM*, 2013.
- [9] Kamishima, A. Akaho, and Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *ECML-PKDD*, 2012.
- [10] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Innovations in Theoretical Computer Science Conference*. ACM, 2012.
- [11] D. McNamara, C. S. Ong, and R. C. Williamson, "Provably fair representations," *CoRR*, vol. abs/1710.04394, 2017.
- [12] S. Samadi, U. Tantipongpipat, J. Morgenstern, M. Singh, and S. Vempala, "The price of fair pca: One extra dimension," in *NIPS*, 2018.
- [13] A. Pérez-Suay, V. Laparra, G. Mateo-García, J. Muñoz-Marí, L. Gómez-Chova, and G. Camps-Valls, "Fair kernel learning," in *Joint European Conf. Machine Learning and Knowledge Discovery in Databases*, 2017.
- [14] M. Olfat and A. Aswani, "Convex formulations for fair principal component analysis," *CoRR*, vol. abs/1802.03765, 2019.
- [15] B. Cao, D. Shen, J.-T. Sun, Q. Yang, and Z. Chen, "Feature selection in a kernel space," in *ICML*, 2007.
- [16] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *JMLR*, 2012.
- [17] J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, and A. J. Smola, "Correcting sample selection bias by unlabeled data," in *NIPS*, 2007.
- [18] S. J. Pan, J. T. Kwok, Q. Yang *et al.*, "Transfer learning via dimensionality reduction," in *AAAI*, 2008.
- [19] A. Gretton, A. J. Smola, J. Huang, M. Schmittfull, K. M. Borgwardt, and B. Schölkopf, "Covariate shift by kernel mean matching," 2009.
- [20] Y. Grandvalet and S. Canu, "Adaptive scaling for feature selection in svms," in *NIPS*, 2002.
- [21] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," in *Data Classification: Algorithms and Applications*, 2014.
- [22] L. Yang, S. Lv, and J. Wang, "Model-free variable selection in reproducing kernel hilbert space," *JMLR*, 2016.
- [23] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *International Conf. Artificial Neural Networks*, 1997.
- [24] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, 2017.
- [25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *JMLR*, 2011.

## VI. APPENDIX

### A. Derivation of Eigen-Problem (6)

We will show how to derive (6) by solving (3) under constraints (4) and (5). Recall that  $\eta = \sum_{i=1}^n \alpha_i \phi(x_i)$  where  $\alpha_i$ 's are unknown parameters. Rewrite the objective in (3) as

$$\begin{aligned}
 J(\eta) &= \left\| \frac{1}{n_u} \sum_{i=1}^{n_u} \langle \phi(x_i), \eta \rangle - \frac{1}{n_p} \sum_{i=n_u+1}^n \langle \phi(x_i), \eta \rangle \right\|^2 \\
 &= \frac{1}{n_u^2} \left( \sum_{i=1}^{n_u} \langle \phi(x_i), \eta \rangle \right)^2 + \frac{1}{n_p^2} \left( \sum_{i=n_u+1}^n \langle \phi(x_i), \eta \rangle \right)^2 \\
 &\quad - \frac{2}{n_u n_p} \left( \sum_{i=1}^{n_u} \langle \phi(x_i), \eta \rangle \right) \left( \sum_{i=n_u+1}^n \langle \phi(x_i), \eta \rangle \right) \\
 &= \frac{\alpha^T K_u^T K_u \alpha}{n_u^2} + \frac{\alpha^T K_p^T K_p \alpha}{n_p^2} - \frac{2\alpha^T K_u^T \mathbf{1}_u \mathbf{1}_p^T K_p \alpha}{n_u n_p} \\
 &= \alpha^T M \alpha = J(\alpha),
 \end{aligned} \tag{17}$$

where  $M$  is a symmetric matrix defined as

$$M = \frac{1}{n_u^2} K_u^T K_u - \frac{2}{n_u n_p} K_u^T \mathbf{1}_u \mathbf{1}_p^T K_p + \frac{1}{n_p^2} K_p^T K_p, \tag{18}$$

matrix  $K_u \in \mathbb{R}^{n_u \times n}$  has  $k(x_i, x_j)$  at row  $i$  column  $j$  and matrix  $K_p \in \mathbb{R}^{n_p \times n}$  has  $k(x_{n_u+i}, x_j)$  at row  $i$  column  $j$ ;  $\mathbf{1}_u \in \mathbb{R}^{n_u}$  and  $\mathbf{1}_p \in \mathbb{R}^{n_p}$  are vectors of ones, and  $\alpha = [\alpha_1, \dots, \alpha_n]$ .

Next, it is easy to verify that constraint (4) satisfies

$$\eta^T \eta = \alpha^T K \alpha = 1, \tag{19}$$

where  $K \in \mathbb{R}^{n \times n}$  is the standard Gram matrix.

Thus we need to solve

$$\min_{\alpha} J(\alpha) \quad \text{s.t.} \quad \alpha^T K \alpha = 1. \tag{20}$$

The Lagrange function is

$$L(\alpha, \lambda) = J(\alpha) + \lambda(\alpha^T K \alpha - 1). \tag{21}$$

Setting  $\frac{\partial L(\alpha, \lambda)}{\partial \alpha} = 0$  and solving for  $\alpha$  gives (6).

### B. Derivation of the Solution to (9)

Here, we show why solution to (9) is also a solution to the eigen-problem (6). Let  $\alpha$  be the coefficient vector for the first feature embedding  $\eta$  (known), and  $\alpha'$  be the coefficient vector of the second embedding  $\eta'$  (unknown). The new constraint when learning  $\eta'$  can be written as

$$\eta^T \eta' = \left( \sum_{i=1}^n \alpha_i \phi(x_i) \right) \left( \sum_{i=1}^n \alpha'_i \phi(x_i) \right) = \alpha^T K \alpha' = 0. \tag{22}$$

Thus we need to solve

$$\min_{\alpha'} J(\alpha') \quad \text{s.t.} \quad (\alpha')^T K \alpha' = 1, \tag{23}$$

and

$$(\alpha')^T K \alpha = 0. \tag{24}$$

The Lagrange function is

$$L(\alpha', \lambda_1, \lambda_2) = J(\alpha') + \lambda_1((\alpha')^T K \alpha' - 1) + \lambda_2 \alpha'^T K \alpha. \tag{25}$$

Setting  $\frac{\partial L(\alpha', \lambda_1, \lambda_2)}{\partial \alpha'} = 0$  and left-multiplying both sides by  $\alpha^T$ ,

$$\alpha^T M \alpha' - 2\lambda_1 \alpha^T K \alpha' - \lambda_2 \alpha^T K \alpha = 0. \tag{26}$$

Since  $\alpha^T K \alpha' = 0$  and  $\alpha'^T K \alpha' = 1$ , we have

$$(\alpha')^T M \alpha = \lambda_2. \tag{27}$$

Further, from (6) we know  $\alpha$  is a generalized eigenvector of  $M$  satisfying  $M \alpha = \lambda K \alpha$ . Thus (27) becomes

$$(\alpha')^T M \alpha = (\alpha')^T \lambda K \alpha = \lambda (\alpha')^T K \alpha = 0 = \lambda_2, \tag{28}$$

where the second equality is due to the new constraint (22).

Comparing (21) and (25), with  $\lambda_2 = 0$ , we see  $\alpha$  and  $\alpha'$  have the same Lagrange function. Thus it is easy to show they are solution to the same generalized eigenproblem (6).