Fair Kernel Learning with Dependence Regularizers for Algorithmic Fairness

Bingliang Li
blli18@lzu.edu.cn
Lanzhou University
Lanzhou, Gansu, China

Wanting Su
514330359@qq.com
Hohai University
Nanjing, Jiangsu, China

Ziru Yan
University of California, Irvine
Irvine, California, United State

Leini Chen
The Harvey School
Katonah, New York State, The United State

Abstract

As the application of machine learning grows deeper in many fields, it can significantly influence our society and economy as well as individual life. So fairness, equity, and ethics in machine learning are more concerned than ever. By including a penalty of fairness measurement in the objective function, we can efficiently mitigate the discrimination. In particular, we use Hilbert-Schmidt Independence Criterion (HSIC) to measure dependence between predictors and sensitive variables and use it as the fairness penalty term.

In the beginning we introduce the significance of fairness in machine learning. Then we define HSIC with some simple examples and some computational properties of HSIC.

Then we introduce some methods to fairness constraints. Through comparing with other dependence criteria, HSIC is the best dependence regularizer. We show some extensions of HSIC regularizer.

We then apply the fairness term with Kernel Ridge Regression. This approach illustrates the efficiency of fairness penalty tern on improving model fairness. However, it also shows the trade-off between fairness and predictive error. The experiment results of the model being applied on a real-world dataset of crime prediction show the potential of this approach can improve the fairness of automated decisions based on machine learning.

Finally, we want to deep the application of HSIC and redefine of fairness through casual inference in the fucture.

Keywords: Fairness, Hilbert-Schmidt Independence Criterion, Kernel methods, Regularization

1 Introduction

1.1 Motivation

Machine Learning quickly develops in recent years and has already become one necessary part of many industries. This new technique has a huge influence on multiple aspects of modern society, including education, criminology, economics, etc. A wide range of subjects, including companies, governments, and individuals, are using Machine learning. They feed data into the model, and the model would influence their cognition reversely. However, due to any reason, the model trained from Machine Learning may not be fair. It may include some social bias for many reasons. Under these circumstances, Fairness learning become one important part of the machine learning.

1.2 HSIC and dHSIC in Fairness Learning

The Hilbert-Schmidt Independence Criterion is a widely used technique that can evaluate the dependence between two variables. It is initially used in serval machines learning models such as SVM and GP. Later, due to the wonderful performance of the HSIC, it is widely used in many circumstances, such as feature selection and regularization. Now, some groups of researchers have successfully implied the HSIC as a regulator in Fairness learning and perform well in many conditions. However, the dependence relationship between multiple features cannot be easily evaluated by HSIC. We introduce an improved version of HSIC, dHSIC, as a better technique to perform in this problem. The dHSIC can evaluate the dependence between multiple variables at once, to get the cross dependence. This may be helpful to treat some complicated real-life problems.

1.3 Fairness constraints and Dependence Regularizers

we introduce some fairness constraints and compare different constraints in classification and continual problems. Through comparing with other dependence criteria, HSIC is the best dependence regularizer. We also show some extensions of HSIC regularizer.

1.4 Experimental Result

We illustrate performance in a real-world dataset: *Communities and Crime*. By predicting the per capita violent crime rate subject to race discrimination, we shows the how

2021 CIS Research, April 2021, China

the model improving prediction fairness by applying the above method.

The remainder of the paper is structured as follows. Section 2 describes the background of the problem, introduces notation and related work. In Section 3, we present the fairness constraints and dependence regularizers and some extensions. Experimental results of performance are presented in 4. Conclusions and further work are finalized in Section 5 and Section 6.

2 Background

Fairness is always a controversy in the field of machine learning because even programmers are not sure if their codes include racism, sexism or other discrimination. Machine learning methods are in fact not being fair to all groups of people, so fairness is an issue that people need to be concerned about. There are some earliest methods created to explain the fairness in machine learning. For instance, some people might choose to maintain fairness by decreasing the weight of sensitive features or simply just removing these features. However, according to [11], if we remove the sensitive variables, the model's accuracy will decrease because of informative features. This effect is known in statistics as the omitted variable bias. Another method is to match the prior belief about fairness. Solutions including classifiers which are only used during learning. Training on discrimination-free data is likely to yield more equitable predictions.

2.1 Introduction to HSIC

The approaches to maintain fairness all apply the Hilbert Schemit framework and HSIC as an indicator. Hilbert-Schmidt Independence Criterion, what we call HSIC, is a kernel based independence measurement method. The general principle of this kind of method is to define cross covariance operators in reproducing Hilbert space. Then from these operators, the appropriate statistics are derived to determine independence. HSIC uses Hilbert Schmidt cross covariance operator, and the independence criterion is obtained by empirical estimation of the operator norm. Here we define C_{xy} as a Hilbert Schmidt operator, and the HSIC is defined as the norm of the Hilbert Schmidt operator of C_{xy}, i.e

$$\mathrm{HSIC}\left(p_{xy}, \mathcal{F}, \mathcal{G}\right) := \left\|C_{xy}\right\|_{\mathrm{HS}}^2 \tag{1}$$

Based on an observed dataset z, the empirical estimation of HSIC can be obtained. The larger the empirical estimate of HSIC, the stronger the correlation between the separable metric spaces. According to [4]if and only if x, y are independent, $\left\|C_{xy}\right\|_S = 0$. Plus, $\left\|C_{xy}\right\|_S = 0$ only exist when $\left\|C_{xy}\right\|_{HS} = 0$ so we can infer thatif and only if x, y are independent, $HSIC\left\|C_{xy}\right\|_{HS} = 0$.

2.2 Computational Properties

As for the computational properties of HSIC, it costs much less time than other kernel methods in computing. Not only that the convergence speed of HSIC is very fast, the cost of computation is also a reason for us to apply HSIC as an independent criterion because we can make our computing process much more efficient by applying it. According to [1]Bach and Jordan, if we utilize incomplete Cholesky decomposition and the kernel has a fast decaying spectrum, we can approximate HSIC accurately and quickly. As for Cholesky decomposition, when it is applicable, its efficiency of solving systems of linear equations is much higher than previous methods. In this case, our working efficiency can be improved a lot.

The paper is organized as follows. We begin our discussion in section 4 where we explain different ways to deal with fairness constraints and introduce the formalism of dependence regularization. Plus, the choice of HSIC or dHSIC for convenient dependence regularizers is also explained in the paper. Then we introduce some possible extensions of regularizers and large scale approximations. The demonstration of experiments and data is in section 4. Section 5 concludes and finalizes the paper.

2.3 Introduction to dHSIC

Based on the theory of HSIC, some group(Pfister1, 2016) has done some expansion. Define vector $X = (X_1, X_2, ..., X_d)$. The variable $X_1, X_2, ..., X_d$ are mutally independent if and only if

$$P_{X1} \times P_{X2} .... \times P_{Xd} = P_{(X_1,...,X_d)} \tag{2}$$

And then we can define dHSIC as following

$$\mathrm{dHSIC}\left(P_{X_1,...X_d}\right) := \left\|\prod(P_{X1} \times P_{X2} .... \times P_{Xd}) - \prod(P_(X_1, ..., X_d)\right\|_{\mathrm{H}}^2 \tag{3}$$

And this is called d-variable Hilbert-Schmidt independence criterion.

Compare to HSIC, the dHSIC can handle multiple variables simultaneously and calculate the cross variable independence relationship. This characteristic allows it to find the independence relationship in a very complicated situation, in the condition when the feature number is more than 1000, which the original HSIC cannot handle. Furthermore, the dHSIC is able to express HSIC as a special case, which allows the researcher to make number d as a hyperparameter in the researching process. That means we can consider dHSIC as a perfect improvement of HSIC.

2.4 HSIC Lasso

HSIC is not only been used in the regularization steps but also in many other steps, like feature selection. Lasso is a very old and commonly used method for feature selection. It can evaluate the importance of the feature by evaluating the linear dependence between the feature and the output variables.

To solve the limitation of the Lasso, some researchers(Yamda, 2014) combine HSIC and Lasso together to create a better version. The HSIC Lasso is defined as

$$min\frac{1}{2}\left\|L-\sum(a_k,K^k)\right\|^2_{Frob}+\lambda||a||_1 \tag{4}$$

Compare to the original Lasso and normal HSIC feature selection method, HSIC Lasso has better performance on multiple circumstances. It is much cheaper than regular HSIC feature selection and much better than Lasso in complicated situations.

3 Fairness Constraints and Dependence Regularizers

There are some different notions of fairness: disparate treatment, disparate impact, and disparate mistreatment. In the classification problem, these notions derive different misclassification measures which represent equality in certain groups - conditional probabilities for the two sensitive feature groups, i.e.,$P(\cdot|z=0)$ and $P(\cdot|z=1)$. In order to satisfy these requirements, the covariance measure of decision boundary unfairness [18] was proposed, which can convert the covariance onstraints for the fairness notion of interest with respect to corresponding misclassification measures to convex-concave constraints. Also, these misclassification measures are designed to make the predictor independent of sensitive variables in any conditional probability.

When considering the notion of statistical parity with continuous labels, HSIC which uses the entire spectrum of the cross-covariance operator while ICA only adopts the largest singular value is indeed a dependence criterion under all circumstances, and regularization allows imposing structural assumptions and inductive biases onto the problem at hand. Hence, adding dependence regularization into the objective function is a good way to control the dependence between the predictor and sensitive variables .

Meanwhile, there are other approaches to deal with fairness constraints, such as Group-Fairness in influence maximization [17] and Casual Bayesian networks [8].

3.1 Dependence Regularizers with HSIC

There are several methods to measure the dependence between two variables, and here is a table comparing them:

Table 1. Comparison of five independence test methods

	Strong association	Robust	Fast	Confidence interval
Pearson	False	False	True	True
Kendall's Tau	False	True	True	True
Distance correlation	True	False	False	True
HSIC	True	False	False	Sort of
Tau*	True	True	False	True

HSIC is a measurement method with strong correlation and confidence interval for testing operation. At the same time, it has no limitation on use scenarios, but it does not have good robustness, easily interfered by outliers. Because of these advantages, HSIC is a favorable independent regular term choice.

3.2 Extensions of the Regularizers

To avoid the sensitivity of dependence measure to kernel parameters, the normalized version [6] uses the normalized cross-covariance operator $V_{sx}:=\sum_{ss}^{-1/2}\sum_{sx}\sum_{xx}^{-1/2}$ to replace $\sum_{sx}$. Hence,when $\varphi(\cdot)$ and $\psi(\cdot)$ are finite dimensional the fair learning is the following optimization problem:

$$\widehat{\beta}:=argmin_\beta\left\{\frac{1}{n}\|y-X\beta\|^2{}_2+\lambda\|\beta\|^2{}_2+\eta\left\|\overbrace{V}_{sx}\beta\right\|^2_2\right\} \tag{5}$$

This leads to a closed-form solution.

Besides, in case of that $\varphi(\cdot)$ and $\psi(\cdot)$ are finite dimensional,we could only partially normalize the cross-covariance operator with respect to hyperparameters from l and formulate the following learning problem:

$$\widehat{\beta}:=argmin_\beta\left\{\frac{1}{n}\|y-X\beta\|^2{}_2+\lambda\|\beta\|^2{}_2\right\}+\eta\left\|\widehat{\sum_{ss}^{-1/2}}\widehat{\sum_{sx}}\beta\right\|^2_2$$

This also leads to a closed-form solution.

In fair kernel ridge [6], considering now using the explicit feature mapping $x_i\rightarrow\phi(x_i)$ we have $f=\phi x$ and thus can recast optimization as:

$$min\left\{\frac{1}{\lambda}V(y,\phi\beta)+\beta^T\beta+\delta\beta^T\phi^THLH\phi\beta\right\} \tag{6}$$

This gives priority to the evaluations :

$$f\sim N\left(0,\phi(I+\delta\phi^THLH\phi)^{-1}\phi^T\right) \tag{7}$$

Then the GP model is given by:

$$f\sim GP\left(0,k(\cdot,\cdot)-k^T{}_X(KHLH+\delta^{-1}I)^{-1}HLHk_X\right) \tag{8}$$

where $k_X=[k(\cdot,x_1),...,k(\cdot,x_n)]^T$ for any training set $\{x_i\}^n{}_{i=1}$.

If we want to ensure that x has as much information as possible during feature extraction and try to make x independent of sensitive variables. We can also use HSIC to maximize the dependence between the projected and the original data,[11]. Here is the projected :

$$\begin{aligned}V^*&=\underset{V}{arg\,max}\left\{\frac{HSIC\left(\widehat{XV},\overbrace{X}\right)}{HSIC(\widehat{XV},\widehat{S})}\right\}\\&=\underset{V}{arg\,max}\left\{\frac{Tr(V^T(\widehat{X}^T\widehat{X}\widehat{X}^T\widehat{X})V)}{Tr(V^T(\widehat{X}^T\widehat{S}S^T\widehat{X})V)}\right\}\end{aligned} \tag{9}$$

4 Experiments

In this section, we compared the performance of HSIC regularized Kernel Ridge Regression (FKR) with ordinary Kernel Ridge Regression (KRR) on *Communities and Crime* dataset[2].

4.1 Dataset Description and Pre-processing

This dataset contains 1999 instances with 127 features. We will use relative features such as the percentage of people under the poverty level, the number of police cars to predict per capita violent crime rate in different communities in the U.S.

Process of data pre-processing:

- Remove non-predictive features such as numeric code for county and community name.
- Remove columns with missing values. At this stage, we obtained the 1993×100 data matrix described in the paper.
- Remove multicollinearity according to the VIF value of each feature.
- Manuel feature selection: there are three sensitive variables left: *racepctblack*, *racePctAsian*, *racePctHisp*. After calculating the Correlation Coefficient r between each of the three variables with per capita violent crime rate, we will simply remove *racePctAsian*, *racePctHisp* to simplify the experiments, as their r values are lower than 0.2, which indicates we can see them as not related to our dependent variable[7].
- Feature standardization.

4.2 Model Setup

We used different combination of parameters to test the performance of the FKR model to see the influence of penalty parameter λ and μ, and to compare the difference between FKR and KRR. We used median heuristic combined with randomly sample drawing to generate θ_k, and use 5-fold cross validation to choose best parameter combination.

Table 2. Model Parameters

Parameter	Description
θ_k	Kernel bandwidth parameter for k
θ_l	Kernel bandwidth parameter for l
λ	L2 regularization penalty parameter
μ	Penalty hyperparameter for unfairness

θ_k: To use median heuristic[3] to set bandwidth parameter for kernel k, for a given set of observations $x_1, \ldots, x_n$, first calculate $\ell = \text{medain}\left(\left\|x_i - x_j\right\|_2\right)$, then the Gaussian RBF/squared exponential kernel is parameterized as:

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{\ell^2}\right)$$

After calculating ℓ, we randomly draw 10 samples around its value. Besides the median heuristic method, we also uniformly draw 15 values between $\left[e^{-15}, 1.0\right]$ as potential values for θ_k. There are other methods to determine θ_k, which we will discuss at 6.1.

θ_l: as parameters from ℓ are free to adjust, we will set $\theta_l = 0.5$ for reproducing results in [6].

λ: we use $e^{-20}, e^{-19}, \ldots, e^{-10}$ as λ list.

μ: we chose 7 different μ in the interval [0, 10]. As high value of penalty hyperparameter for unfairness representing more fair model, we can compare the trade-off between fairness and prediction accuracy.

Besides hyperparameter tuning, we modified the model to use an approximation of Distance Correlation (and by extension HSIC[16]) via a chi-squared distribution[15][9] to reduce model complexity from $O(m^3)$ to $O(n \log n)$, we then compared the performance between the two approaches.

4.3 Experiments Results

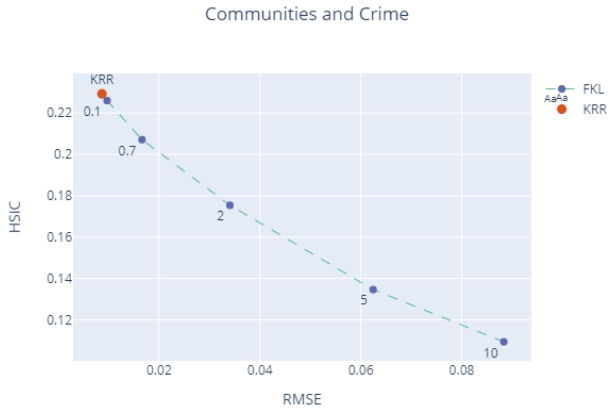

Figure 1. The trade-off curve of unfairness/prediction error, also shows the comparisons between fairness-regularized kernel ridge regression and ordinary KRR.

For the FKR model, we can see in Figure 1 that the value of HSIC, which is used to measure unfairness, is lower than the ordinary KRR model. However, for all the optimal parameter combinations, KRR has a lower RMSE value than the FKR model. Such trade-off shows that though penalty parameter μ can help improve algorithmic fairness, it still has a negative impact on model accuracy.

The most significant correlations between hyperparameters, fairness, and prediction accuracy are the correlation between μ and RMSE, μ and HSIC. Figure 2 shows that μ has

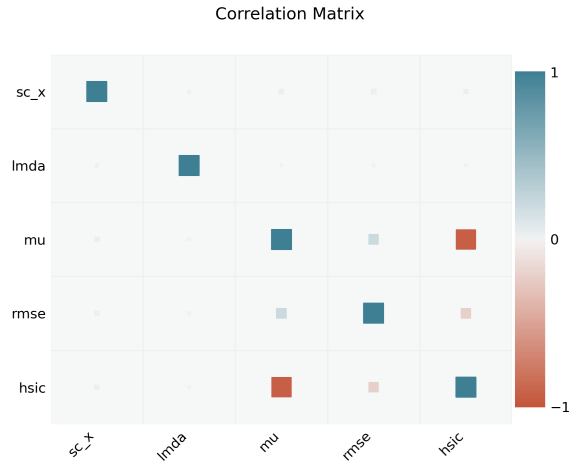

Figure 2. Correlation matrix for different model parameters, RMSE, and HSIC. The size will be bigger if two variables have a strong correlation. Red means positive correlation, blue means negative correlation.

a strong negative correlation with model fairness and a relatively weak positive correlation with RMSE, so by increasing μ, we can significantly improve the fairness performance of the model without losing too much predict accuracy.

Additionally, by using approximation of HSIC, the time of grid search with cross-validation parameter estimation is 477.048s. By calculating the accurate value of HSIC(without using any approximation), the CV grid search cost 22 hours 32 minutes 23 seconds. All experiments are finished on an Alienware Area 51M laptop with Intel® Core™ i7 10700.

5 Conclusions

There are huge methods to constraint fairness. HSIC is suit to be a dependence regularizer for continuous labels compared with other dependence criteria. Also, the extensions of HSIC are various in kernel ridge regression and perform better compared with ICA algorithms.

The introduced methods show promising performance in crime prediction subject to race discrimination, allowing to strike favorable tradeoffs between the predictive performance and its fairness in terms of statistical dependence. And by tuning the fairness penalty parameter in the model, allows us to input sensitive variables to the model while keeping the prediction fair.

6 Further Discussion

6.1 Parameter Setting

In some situations, kernels that parameterized by median heuristic can lead to poor performance[3], especially when the test is performed on high dimension data[14][13], the parameter set by median heuristic may fail.

Bayesian Kernel Embedding (BKE) approach[3]: Inside of using median heuristic, BKE setting the bandwidth of RBF kernels by maximizing the marginal likelihood. For a prior on the kernel mean embedding μ_θ, we can define a likelihood linking via the empirical mean embedding estimator $\widehat{\mu_\theta}$ to link it to the observations $\{x_i\}_{i=1}^n$, which allow us to infer the posterior distribution of the kernel mean embedding. Then by learning a posterior distribution over the hyperparameter space $\mathcal{H}_{k_\theta}$, we can find the optimal hyperparameter θ.

6.2 Model Modification

In our experiments, we modified our model to use approximation of HSIC based on chi-squared distribution. We tried to use some of the approximations from [19], but the package kerpy, which provided such tests are not compatible with our Python environment setting, and unfortunately we don't have enough time to update the code. The approximation of HSIC we used in the model is provided by hyppo[9], a Python package. Code implementation of an article is vital for the further development of an algorithm, so our code is publish at [5].

6.3 Application of HSIC in casual inference

In the causal inference, if we want to pay attention to the influence of x on y, there must be the influence of mediating variables between them. If these mediating variables are independent from each other, it will facilitate our research. Therefore, in the DAG [12] test method, we can use dHSIC to test the joint independence of the variables between the mediating variable sets, namely the error items. As HSIC is easy to cause curse of dimensionality and the calculation cost is high, dHSIC performs better in this aspect.

6.4 Redefine fairness by casual inference

When considering what fairness is, we want to combine causal inference with fairness. For example, meeting the requirements of misclassification measures does not necessarily guarantee a certain degree of fairness. Take ovarall misclassification rate [18] as an example:

$$P\left(\widehat{y} \neq y | z = 0\right) = P\left(\widehat{y} \neq y | z = 1\right)$$

If the sensitive variable is independent of y, the sensitive variable can form a chain with y through x to have an effect on y. So the inverse probability weighting method in causal inference or direct intervention mediator variables [10] can be used to calculate the controlled direct effect and remove the false correlation arrow so that x is dependent of z to measure the relationship between x and y.

This is a not-so-general insight, and we hope that in the future, through further learning, we can further define fairness through causal inference.

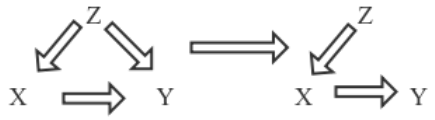

Figure 3. The causal diagram after z and y are independent

References

[1] Francis R Bach and Michael I Jordan. 2002. Kernel independent component analysis. *Journal of machine learning research* 3, Jul (2002), 1–48.

[2] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[3] Seth Flaxman, Dino Sejdinovic, John P Cunningham, and Sarah Filippi. 2016. Bayesian learning of kernel embeddings. *arXiv preprint arXiv:1603.02160* (2016).

[4] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. 2005. Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory*. Springer, 63–77.

[5] Bingliang Li. 2021. HSIC-regularized-Kernel-Ridge-Regression. https://github.com/BingliangLI/HSIC-regularized-Kernel-Ridge-Regression.

[6] Zhu Li, Adrian Perez-Suay, Gustau Camps-Valls, and Dino Sejdinovic. 2019. Kernel dependence regularizers and gaussian processes with applications to algorithmic fairness. *arXiv preprint arXiv:1911.04322* (2019).

[7] David S Moore, William I Notz, and William Notz. 2006. *Statistics: Concepts and controversies*. Macmillan.

[8] Luca Oneto and Silvia Chiappa. 2020. Fairness in machine learning. In *Recent Trends in Learning From Data*. Springer, 155–196.

[9] Sambit Panda, Satish Palaniappan, Junhao Xiong, Eric W Bridgeford, Ronak Mehta, Cencheng Shen, and Joshua T Vogelstein. 2019. hyppo: A Comprehensive Multivariate Hypothesis Testing Python Package. *arXiv preprint arXiv:1907.02088* (2019).

[10] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

[11] Adrián Pérez-Suay, Valero Laparra, Gonzalo Mateo-García, Jordi Muñoz-Marí, Luis Gómez-Chova, and Gustau Camps-Valls. 2017. Fair kernel learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 339–355.

[12] Niklas Pfister, Peter Bühlmann, Bernhard Schölkopf, and Jonas Peters. 2016. Kernel-based tests for joint independence. *arXiv preprint arXiv:1603.00285* (2016).

[13] Aaditya Ramdas, Sashank Jakkam Reddi, Barnabás Póczos, Aarti Singh, and Larry Wasserman. 2015. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.

[14] Sashank Reddi, Aaditya Ramdas, Barnabás Póczos, Aarti Singh, and Larry Wasserman. 2015. On the high dimensional power of a linear-time two sample test under mean-shift alternatives. In *Artificial Intelligence and Statistics*. PMLR, 772–780.

[15] Cencheng Shen and Joshua T Vogelstein. 2019. The Chi-Square Test of Distance Correlation. *arXiv preprint arXiv:1912.12150* (2019).

[16] Cencheng Shen and Joshua T Vogelstein. 2020. The exact equivalence of distance and kernel methods in hypothesis testing. *AStA Advances in Statistical Analysis* (2020), 1–19.

[17] Alan Tsang, Bryan Wilder, Eric Rice, Milind Tambe, and Yair Zick. 2019. Group-fairness in influence maximization. *arXiv preprint arXiv:1903.00967* (2019).

[18] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. 2019. Fairness Constraints: A Flexible Approach for Fair Classification. *J. Mach. Learn. Res.* 20, 75 (2019), 1–42.

[19] Qinyi Zhang, Sarah Filippi, Arthur Gretton, and Dino Sejdinovic. 2018. Large-scale kernel methods for independence testing. *Statistics and Computing* 28, 1 (2018), 113–130.