# Fair Representation Learning with Maximum Mean Discrepancy Distance Constraint

**Alexandru Lopotenco**[1], **Ian Tong Pan**[3], **Jack Zhang**[1], **Guan Xiong Qiao**[2]
[1]University of Pennsylvania, College of Arts and Sciences
[2]University of Pennsylvania, School of Engineering and Applied Sciences
[3]University of Pennsylvania, The Wharton School
lopo@sas.upenn.edu, tompan@wharton.upenn.edu, jzhang0@sas.upenn.edu, alanqiao@seas.upenn.edu

**Abstract**

Unsupervised learning methods such as principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and autoencoding are regularly used in dimensionality reduction within the statistical learning scene. However, despite a pivot toward fairness and explainability in machine learning over the past few years, there have been few rigorous attempts toward a generalized framework of fair and explainable representation learning. Our paper explores the possibility of such a framework that leverages maximum mean discrepancy to remove information derived from a protected class from generated representations. For the optimization, we introduce a binary search component to optimize the Lagrangian coefficients.

## Introduction

One method to satisfy general fairness constraints is to alter the data such that it still explains the structure, yet it is penalized for probabilistic fairness conditions that induce equality of odds and demographic disparity. Such an approach is detailed in (Zemel et al. 2013). We will pursue a similar goal, but instead, we will penalize by the distance between the distribution of subsets of the data that have different protected attributes.

## Maximum Mean Discrepenacy

Maximum Mean Discrepancy (MMD) is a popular distribution distance introduced by (Gretton et al. 2008) which focuses on using means directly to establish the distance. While (Gretton et al. 2008) provides the theoretical definition of MMD, we will use the function's plug-in estimator. Formally, for instances $X = (X_i)_{i\leq m}$ and $Y = (Y_i)_{i\leq n}$ sampled from distributions $p$ and $q$, respectively, an estimator for the MMD is provided by (Gretton et al. 2008) as

$$\text{MMD}(U, V) = [U(X) + V(Y) - UV(X, Y)]^{\frac{1}{2}},$$

where we have that for the Gaussian kernel $k$, we define

$$U(X) = \frac{1}{m^2}\sum_{i,j\leq m} k(X_i, X_j), V(Y) = \frac{1}{n^2}\sum_{i,j\leq n} k(Y_i, Y_j)$$

and $UV(X, Y) = \frac{2}{mn}\sum_{i\leq m,j\leq n} k(u_i, v_j)$.



For our setup, suppose that we are working with a data set $X$, for which $A$ is a binary, protected feature. We aim to transform $x_i \to y_i$, where $y_i$ are the transformed vectors that embed $x_i$'s. Let $A(x)$ be the value at the attribute $A$ of datapoint $x$. Then, say that $N^{(p)} = \{i : A(x_i) = p\}$ for $p \in \{0, 1\}$ and for a possible transformation of $X$, say $Y$ let $Y^{(p)} = \{y_i : i \in N^{(p)}\}$. To achieve fairness between the two populations of classes of $A$, we want to minimize the difference between the underlying distributions of the samples $Y^{(0)}$ and $Y^{(1)}$.

## Constrained Optimization

Define $f(Y; X)$ as the loss characteristic to $f$ induced by the representation $Y$ of the original dataset. For instance, for PCA that would be the reconstruction error, and for T-SNE that would be KL divergence.

Now we impose a bound on the MMD, say $\delta$ and then resolve the constrained optimization problem (1)

$$\begin{aligned} &\text{argmin}_{Y\in\mathbb{R}^p} && f(Y; X) \\ &\text{s.t.} && \text{MMD}(Y^{(0)}, Y^{(1)})^2 \leq \delta^2 \end{aligned}$$

We solve this problem using the Lagrangian method and to minimize the gradient of the following objective

$$f(Y; X) + \beta\text{MMD}(Y^{(0)}, Y^{(1)})^2.$$

This is a non-convex problem, hence we propose Algorithm 1 which performs a binary search to find a decent value of $\beta$ and runs gradient descent for each instance of $\beta_k$.

## Alternative Approach

We assume that $f$ is a differentiable function. If this does not hold, or if $f$ is hard to work with, then an alternative approach would be to substitute $f(Y; X)$ with $\text{MMD}(\hat{Y}, Y)$ in (1), where $\hat{Y} = \text{argmin}_Y f(Y; X)$.

## Experiments

We empirically demonstrate how our Algorithm 1 differs from the baseline when applied to T-SNE, first introduced by (van der Maaten and Hinton 2008), keeping $f(Y; X)$ to be KL divergence. In order to evaluate the performance of our

Algorithm 1: Fair optimization for unsupervised dimensionality reduction algorithm with loss function f

**Input:** Dataset $X$ of $\mathbb{R}^d$, vanilla embeddings $\hat{Y}$, $\delta$, $\rho, \beta_{\max}$
**Initialize:** $Y_0 = \hat{Y}, l \in \mathbb{N}, \beta_{\min} = 0, \epsilon_k, k = 1, \dots l, M$
**for** $k = 1$ **to** $l$: **do**
  $\beta_k \leftarrow \frac{1}{2}(\beta_{\max} + \beta_{\min})$
  Solve
  $Y_k = \text{argmin}_{Y \in \mathbb{R}} f(Y; X) + \beta_k \text{MMD}(Y^{(0)}, Y^{(1)})$ via gradient descent by initializing $Y_k = Y_{k-1}$ and until $\|\nabla_Y f(Y; X) + \beta_k \nabla_Y \text{MMD}(Y^{(0)}, Y^{(1)})\| \leq \epsilon_k$
  **if** $\text{MMD}(Y_k^{(0)}, Y_k^{(1)}) \geq \delta$ : **then**
    $\beta_{\min} \leftarrow \beta_k$
  **end if**
  **if** $\text{MMD}(Y_k, \hat{Y}) \geq \rho$ : **then**
    $\beta_{\max} \leftarrow \beta_k$
  **end if**
  **if** $\text{MMD}(Y_k^{(0)}, Y_k^{(1)}) \leq \delta$ and $\text{MMD}(Y_k, \hat{Y}) \leq \rho$ : **then**
    Return $Y_k$ and end
  **end if**
**end for**
**Output:** If did not end, $Y_l$ the final value

algorithm in a real-world scenario, we applied it to a subset of 1000 samples from the Census Income dataset, publicly available from the UCI Machine Learning Repository, which had categorical features one-hot encoded (Barry and Ronny 1996). We executed our algorithm (implemented in PyTorch) based on the original dataset, T-SNE embeddings ($\hat{Y}$) over a range of deltas, using gender as the protected attribute. We show some of the results, notably at roughly 1/2 and 1/50 of the baseline MMD in Figure 2 and Figure 3, respectively. These embeddings were visualized using Matplotlib, with points color-coded to indicate both gender (blue/red) and income level in green and orange (above or below $50,000/year, respectively).

## Results

We notice in Figure 1 that T-SNE without fairness constraints clearly separates women from men, even if this is unintended.

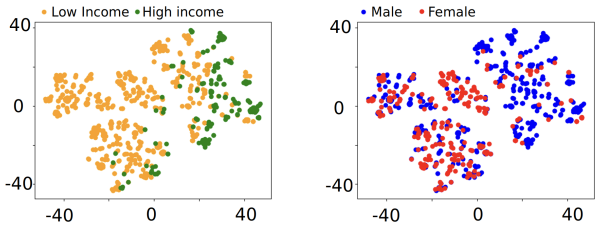


Figure 1: T-SNE with no fairness constraint. Left: $< 50K$ and $>= 50K$ Income Points, Right: Male Female Points

After running our algorithm by setting $\delta$ equal to roughly $\frac{1}{2}$ of T-SNE's MMD and $\rho = 0.1$ we see in Figure 2 that the separation between men and women is less apparent, while the income classes are still clustered. We also tried setting $\delta$ to $\frac{1}{50}$ of the vanilla T-SNE MMD, which is shown in Figure 3. It is obvious that males and females are scattered, indicating strong fairness. Surprisingly, the income classes remain relatively separated, hence our algorithm has satisfactorily preserved the dataset structure despite a fairness constraint.

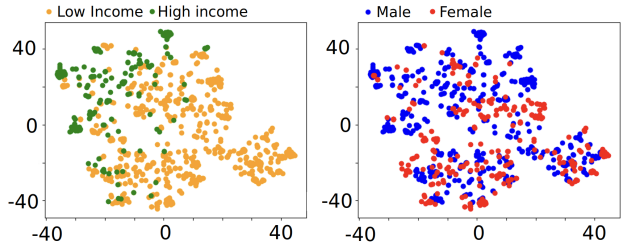


Figure 2: Algorithm 1 with $\delta$ set to $\frac{1}{2}$ of the original MMD. Left: $< 50K$ and $>= 50K$ Income Points, Right: Male Female Points

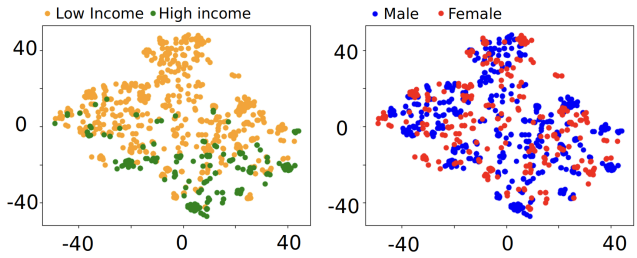


Figure 3: Algorithm 1 with $\delta$ set to $\frac{1}{50}$th of the original MMD. Left: $< 50K$ and $>= 50K$ Points, Right: Male Female Points

## Conclusion

We have successfully introduced a framework for fair representation learning and dimensionality reduction via constrained optimization which can optimize against any fairness threshold. The power of our algorithm is its applicability to numerous models such as autoencoders and PCA, and we encourage further experimentation in these directions.

## References

Barry, B.; and Ronny, K. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Scholkopf, B.; and Smola, A. J. 2008. A Kernel Method for the Two-Sample Problem. *CoRR*, abs/0805.2368.

van der Maaten, L.; and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86): 2579–2605.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning Fair Representations. In Dasgupta, S.; and McAllester, D., eds., *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, 325–333. Atlanta, Georgia, USA: PMLR.