Fair Federated Learning For Deciding Community Improvement Grants

Yuying Duan [1] **Michael Lemmon** [1]

1. Project introduction

This project introduces an approach that utilizes federated learning(McMahan et al., 2023) to decide the equitable allocation of development grants to city communities with the aim of enhancing the quality of life for all city residents. Historical practices sometimes overlook awarding such grants to communities that are disadvantaged with respect to a sensitive social attribute such as racial or ethnic makeup. This project proposes a Machine Learning (ML) method that small cities can use to optimize the communal benefit reaped from these development grant while simultaneously adjusting how "fair" the grant decisions were with respect to sensitive community attributes.

Fairness and bias are important considerations when allocating resources (Deutsch, 1975) (Rawls & Kelly, 2001). Within a city some communities are classified as *socially disadvantaged* with respect to a sensitive community attribute. As an example, a community with a large proportion of elderly residents may be seen as "disadvantaged" with respect to the sensitive attribute of "age". Fairness in this project means that any community resident has an equal opportunity to receive the allocated resources, regardless of whether they belong to an advantaged or disadvantaged group. In statistical terms, this concept can be formalized as *statistical parity* (Mehrabi et al., 2022), which is defined as: the probability that an advantaged individual benefits from allocated resources is *equal* to the probability that a disadvantaged individual benefits from allocated resources. Formally, we can write this as

$$\Pr\{\mathbf{y}=1\,|\,\mathbf{s}=1\}=\Pr\{\mathbf{y}=1\,|\,\mathbf{s}=0\} \tag{1}$$

where $\mathbf{y}=1$ denotes that the individual receives resources and $\mathbf{s}=1/0$ indicates the individual is advantaged/disadvantaged. While the global policy is to decide the allocations that maximize the *communal benefit*, we also want to ensure that our allocations are "fair" with respect to the sensitive attribute.

Our proposed method develops a fair policy using a federated learning framework, wherein each local community trains a local generative model that generates samples with the same distribution as that of the community. These generative models are then sent to a global decision maker who determines how to allocate community improvement grants in a manner that maximizes the communal (global) benefit of these grants. The benefits of developing a global policy in a federated manner (Konečnỳ et al., 2016) (Mothukuri et al., 2021) are 1) it reduces the communication costs needed to support distributed decision making since the local generative models are much smaller than the local raw datasets and 2) it preserves differential privacy because raw data from residents is held privately within the community.

This project's results demonstrate that by selecting a fairness parameter, we can make decisions that trade off the communal benefit of all grants (measured by the aggregate mismatch between community need and allocated grant money) against the global fairness of those decisions as measured by the notion of statistical parity described above. The city, therefore, creates a policy that accommodates its specific needs regarding "fairness" and "communal good" by tuning the fairness parameter. This indicates that our proposed approach is able to work with cities to establish the best policy that addresses the unique needs of their citizens.

2. Proposed method

As we describe in section 1, the use of direct financial grants to individual communities provides a practical approach to enhancing the quality of life for residents with low incomes. These grants can provide direct assistance to residents or be used for community improvement projects that address the unique needs of that community's residents. The federated learning approach (as shown in Figure 1) has each community train a local model for its residents' needs and then transmits those models to the global decision maker (city planner). The city planner then uses these models to generate "fake" residents whose attributes are used to decide how to distribute a fixed budget of grant dollars to individual communities in a manner that maximizes some chosen measure of the global "communal" good.

The proposed work can be divided into the following three tasks:

[1]Department of Electrical Engineering, University of Notre Dame, IN, USA. Correspondence to: Yuying Duan <yduan2@nd.edu>.

Lucy Family Institute May 31, 2023

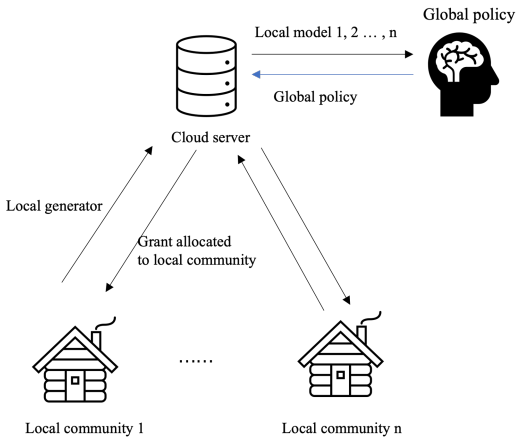


Figure 1. The Federated Learning framework

2.1. Training model for residential income level

This project uses the UCI Adult dataset(Dua & Graff, 2017) to develop a model that predicts an individual's income level based on their profile attributes such as age, work class, and education level. The model is a deep sequential neural network, trained using standard backpropagation techniques. The model inputs are tensors whose elements are 8 categorical one-hot encoded attributes (e.g., education level, occupation, marital status, race) and 6 numerical attributes like age and final weight. The model's output is a value within the range of [0, 1] that represents the likelihood that the individual has an income level greater than a fixed threshold.

This project used the UCI Adult dataset to generate local datasets for simulating the attributes of smaller communities. In particular, we partitioned the UCI dataset into 8 datasets where each dataset represented a different community with varying racial makeup and income profiles.

2.2. Federated Learning that generates community's data distribution

In this section, we employ Generative Adversarial Networks (GAN)(Goodfellow et al., 2014) to generate the race and income distribution for each community. The resulting generators will be sent to the global decision maker to determine a global policy for grant allocation. By using a GAN, we can generate random samples from the same distribution that generated each local community's dataset. This approach addresses the problems associated with ensuring the statistical sampling of communities is independent and identically distributed. So the samples generated by the GAN represent a "generic" resident whose attributes have the same probability distribution of the GAN's training data. This ensures that a particular "real" resident's information cannot be inferred from the generative distribution, thereby protecting the "real" individual's privacy.

2.3. Fair global policy for deciding neighborhood improvement grant

In this part, the global decision maker receives the generators from all communities and makes decisions regarding the allocation of grants to different communities based on the generative race and income distribution of each community. The global policy wants to determine the allocation for each community that meets the aggregate financial needs of the community's low-income individuals. City planners, however, also , want to ensure fairness with respect to a sensitive community attribute that causes a community be classified as *disadvantaged*. In this project, the sensitive attribute is race.

Let us first see how to model a city's decision making process. We let $n = [n_0, n_1, \cdots, n_7]$ denote the need vector of communities whose element i is the grant required by community i to enhance the quality of life for its residents. F denotes the total grant dollars that the government has to allocate. A global policy $p = [p_0, p_1, \cdots, p_7]$ is a vector whose element i represents the percentage of total grant dollars allocated to community i, thus the funding that community i receives from the city's government is $p_i F$. We define the *baseline policy*, p^*, as the policy that solves the following optimization problem:

$$\min_p L(p) = \frac{1}{8}\sum_{i=0}^{7}(p_i F - n_i)^2 \tag{2}$$

where the $L(p)$ is the mean square error between between community's financial need and the grant dollars allocated to that community by the government. Our plan is to *retune* the baseline policy, with respect to a fairness parameter, λ to obtain a policy p_λ whose decisions toward disadvantaged communities are measurably "more fair" than the baseline policy p^*

This project views "fairness" though the lens of distributive justice (Deutsch, 1975), (Rawls & Kelly, 2001). Distributive justice is a social science concept referring to the perceived fairness of a decision that allocates resources to a group of people with differing attributes. That allocation decision is seen as just or fair if any individual has an equal opportunity of receiving the allocated resources regardless of whether that individual belongs to an advantaged or disadvantaged community. In this project, "opportunity of receiving the allocated resources" corresponds to the probability that an individual can benefit from the grant allocated to his or her community. We use race as the sensitive attribute and a community is classified as socially disadvantaged if their majority is non-white. Otherwise, it is a socially advantaged community. This project, therefore, interprets "fairness" as

"distributive justice"

Distributive justice can be expressed as a statistical condition. In particular, we say an allocation policy is fair if it satisfies the notion of statistical parity(Mehrabi et al., 2022) given in equation (1). In this equation $\mathbf{y} = 1$ means the individual benefits from the grant, $\mathbf{s} = 1$ means the individual is socially advantaged (white), $\mathbf{s} = 0$ means the individual is socially disadvantaged (non-white).

In general, the equation 1 will not be satisfied, but we can quantify how close the policy, p, is to achieving statistical parity through the risk difference function:

$$\begin{aligned} R[p] \quad &= \quad \Pr\{\mathbf{y} = 1 \text{ under } p \,|\, \mathbf{s} = 1\} \\ &\quad -\Pr\{\mathbf{y} = 1 \text{ under } p \,|\, \mathbf{s} = 0\} \end{aligned} \tag{3}$$

The risk difference function in our case can be estimated by:

$$\hat{R}[p] = \left| \frac{\sum_{i=1}^{n} p_{di} F}{\sum_{i=1}^{n} |D_{di}|} \right| - \left| \frac{\sum_{j=1}^{m} p_{aj} F}{\sum_{j=1}^{m} |D_{aj}|} \right|$$

In this formula, p_{di} represents the percentage of the grant allocated to disadvantaged community i, and F represents the total grant dollars. Thus, $\sum_{i=1}^{n} p_{di} F$ represents the grant dollars allocated to disadvantaged communities. $|D_{di}|$ represents the number of people in disadvantaged community i, and $\sum_{i=1}^{n} |D_{di}|$ represents the total number of people in disadvantaged communities. Likewise, p_{aj} represents the percentage of the grant dollars allocated to advantaged community j and $\sum_{j=1}^{m} p_{aj} F$ represents the total grant dollars allocated to advantaged communities. $|D_{aj}|$ represents the number of people in advantaged community j, and $\sum_{j=1}^{m} |D_{aj}|$ represents the total number of people in advantaged communities.

The first term of the risk function, $\frac{\sum_{i=1}^{n} p_{di} F}{\sum_{i=1}^{n} |D_{di}|}$, divides the total grant allocated to disadvantaged communities by the total population in those communities. Likewise, the second term of the risk function, $\frac{\sum_{i=1}^{m} p_{ai} F}{\sum_{i=1}^{m} |D_{ai}|}$ divides the total grant allocated to advantaged communities by the total population in those communities.

A more fair global policy p can be obtained by solving:

$$\min_{p} L_\lambda(p) = L(p) + \lambda \hat{R}[p] \tag{4}$$

In this equation $\lambda > 0$ is a fairness parameter that we select to optimally tradeoff communal benefit (i.e., the mean square error (MSE) between the money that a community needs and the funding that the government gives it) against the fairness of that policy (measured by the risk difference). In general, increasing the fairness parameter will result in decreasing the risk difference (improving fairness) and increasing the MSE value (reducing communal benefit).

3. Techniques we used in our project

In this section, we present the technique employed in our project. In Section 2.2, our proposed framework involves the transmission of local communities' generative model to the global decision maker. The generative model we used in our project is a Generative Adversarial Network (GAN)

3.1. Generative Model

GANs: GANs (Goodfellow et al., 2014) are used to generate data from the same distribution as the training data. The GAN (as shown by Figure.2) has two components: a generator G: $\mathbb{R}^m \to \mathbb{R}^n$ and a discriminator D: $\mathbb{R}^n \to [0, 1]$, both G and D are multi-layer neural networks. The generator G(z) generates fake samples from a prior distribution P_z on a noise variable z and learns a generative distribution P_G to match the real data distribution P_{data}. The discriminator $D(x)$ is a binary logistic regression classifier whose input is generative data and real data whose output is the probability that x is from real data distribution rather than from generative distribution. The objective function of the GAN

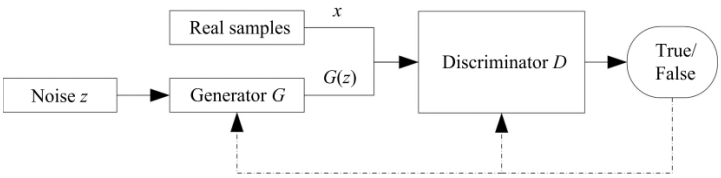


Figure 2. The GAN framework

is:

$$\begin{aligned} \min_G \max_D V(D, G) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \\ \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))] \end{aligned} \tag{5}$$

For a fixed generator G, the optimal discriminator D is :

$$D^*(\boldsymbol{x}) = \frac{p_{\text{data}}(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})} \tag{6}$$

When D is optimal, the objective function is:

$$\begin{aligned} C(G) &= \min_G V(G, D^*) \\ &= \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}}[\log D^*(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}}[\log(1 - D^*(G(\boldsymbol{z})))] \\ &= \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}}[\log D^*(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x} \sim p_g}[\log(1 - D^*(\boldsymbol{x}))] \\ &= -\log(4) + 2 \cdot JSD(p_{\text{data}} \| p_g) \end{aligned}$$

$$\text{where,} \quad JSD(p||q) = \frac{1}{2} \int \left(p \log \frac{2p}{p+q} + q \log \frac{2q}{p+q} \right) d\mu \tag{7}$$

The objective function with an optimal D^* is indeed an estimator of the JS divergence of real data distribution and

current generative distribution. The global minimum of $C(G)$ is achieved when $p_g = p_{data}$. At that point, $C(G) = -\log 4$.

Researchers (Arjovsky & Bottou, 2017) have shown that GANs can be difficult to train when the real distribution's support lies on a submanifold of the data space. It is easy to calculate that $JSD\left(p_{\text{data}} \| p_g\right)$ keeps a fixed value: log2. In this case, the gradients vanish for the generator update and the generator will not update to match the real distribution. This problem can be addressed using the Wasserstein GAN.

Wasserstein GANs The Wasserstein GAN (Arjovsky et al., 2017) (WGAN) is motivated by the GAN's issues with the real distributions that are not continuous. In the WGAN, instead of using the JS divergence, they use the Wasserstein distance (W-distance) to measure the difference between two distributions. The objective function of the WGAN is:

$$\min_G \max_D \mathbb{E}_{x\sim\mathbb{P}_r}\left[D(x)\right] - \mathbb{E}_{z\sim p(z)}\left[D\left(G(z)\right]\right. \\ +\lambda \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathbb{P}_x}\left[\left(\|\nabla_{\boldsymbol{x}} D(\boldsymbol{x})\|_2 - 1\right)^2\right] \tag{8}$$

The objective function of the WGAN with an optimal discriminator is an estimate of the W distance between the generative distribution P_θ and real distribution P_r:

$$W\left(\mathbb{P}_r, \mathbb{P}_\theta\right) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x\sim\mathbb{P}_r}[f(x)] - \mathbb{E}_{x\sim\mathbb{P}_\theta}[f(x)]$$

There are two main reasons why the objective function of WGAN makes sense. Firstly, the W-distance, used in the WGAN, is a continuous and differentiable function with respect to the generator parameters. This allows for effective gradient-based optimization of the generator. Secondly, the WGAN discriminator is designed to use the W-distance, which will not collapse when the real data's distribution lies on a submanifold whose dimension is less than that of the generator's output.

In our project, we are using the WGAN to generate the race and income distribution, which has a joint distribution over both continuous and discrete variables. The real distribution lies on a submanifold immersed in the data space, while the generator's output is always a continuous distribution. By using the WGAN, we can ensure that the training gradients used to update the generator will not vanish, thereby guaranteeing that the distance between the generative distribution and the real distribution will decrease over time.

4. Result and evaluation

This section has three parts: clustering data into 8 different communities, training local WGANs to generate local data distributions and creating a desired global allocation policy.

4.1. Clustering Data

The K-means algorithm is used to partition the dataset into 8 distinct groups, where each group corresponds to a community with different racial makeup and income profiles.

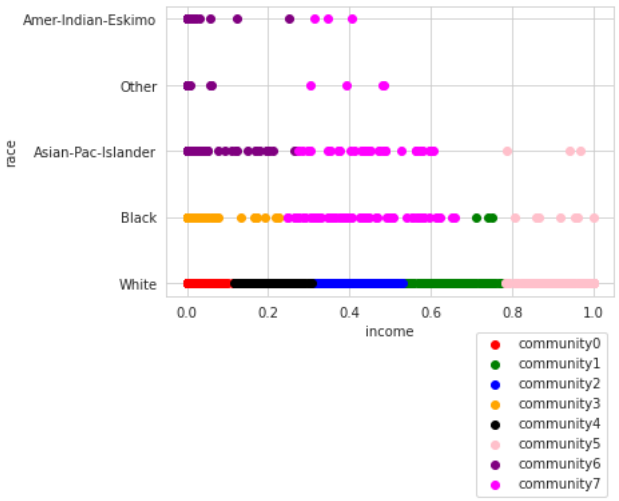


Figure 3. Clustered Data

The data clusters are visualized in Figure 3, with the income level represented on the x-axis ranging from 0 to 1, and the race of the samples depicted on the y-axis. This represents the dataset that was used to solve the fair grant allocation problem.

4.2. Training GAN to generate community level data

In this section, WGANs are trained to generate race and income data of the local communities. In a WGAN, the generator is a multi-output feedforward neural network with two 2 layers, having 128, 64 dimensions respectively. Each output of the WGAN represents one feature (race or income). We apply a sigmoid activation function to numerical feature (income) output and softmax activation function to the categorical feature output (race). These outputs are concatenated together and directly feed into the discriminator. The discriminator is a feedforward neural network with 3 hidden layers, having 64,16,8 dimensions respectively. We train the WGAN for 3000 epochs. In each epoch, we update discriminator using Adam with a learning rate 0.0001 and then update the generator using Adam with a learning rate 0.0001. After finishing the training process, we just take the argmax output and turn into a one-hot encoded output.

We evaluate the WGAN's ability to capture the input data's true distribution in two different ways. The first approach is visual, where we plot the real and generative data distributions for race and income. This is shown in Fig. 6 and from this figure we can see that for all outputs, the generative

and true distributions are qualitatively similar to each other. The second way to evaluate the WGAN's performance is to compute the JS divergence between the real and generative distributions for race and income. Tables 1-8 show the JS-divergence for communities 1-8. In these tables we see JS divergence for race is 0-0.12 and for income it is .05-.172. These appear to be relatively small values, especially when considered along with the distributions shown in Fig. 6. Our experimental results thereby support the assertion that the WGAN was able to learn a meaningful model of each communities racial and income distribution.

feature	race	income
JS divergence	0.000	0.050

Table 1. The JS divergence between generative distribution and real distributions of race and income for community0

feature	race	income
JS divergence	0.120	0.153

Table 2. The JS divergence between generative distribution and real distributions of race and income for community1

feature	race	income
JS divergence	0.000	0.133

Table 3. The JS divergence between generative distribution and real distributions of race and income for community2

feature	race	income
JS divergence	0.000	0.172

Table 4. The JS divergence between generative distribution and real distributions of race and income for community3

feature	race	income
JS divergence	0.014	0.123

Table 5. The JS divergence between generative distribution and real distributions of race and income for community4

feature	race	income
JS divergence	0.061	0.155

Table 6. The JS divergence between generative distribution and real distributions of race and income for community5

feature	race	income
JS divergence	0.085	0.112

Table 7. The JS divergence between generative distribution and real distributions of race and income for community6

feature	race	income
JS divergence	0.036	0.167

Table 8. The JS divergence between generative distribution and real distributions of race and income for community7

4.3. Developing a global policy for deciding neighborhood improvement grants

This section will decide how to allocate grant dollars between the communities. Assuming we have a total budget of F dollars to improve eight communities, our objective is to allocate funds in a way that meets the financial need of low-income residents while being fair with respect to race. To develop the global grant allocation policy, as we described in section 2.3, the objective function we use is equation (4) in section 2.3, which is:

$$\begin{aligned}\min_p L_\lambda(p) &= L(p) + \lambda \hat{R}(p) \\ &= \frac{1}{8}\sum_{i=0}^{7}(p_i F - n_i)^2 \\ &+ \lambda \left| \frac{\sum_{i=1}^{n} p_{di} F}{\sum_{i=1}^{n} |D_{di}|} - \frac{\sum_{j=1}^{m} p_{aj} F}{\sum_{j=1}^{m} |D_{aj}|} \right|\end{aligned} \tag{9}$$

The *financial need* for community i , is denoted as n_i and is given by $\sum_{j=1}^{|D_i|} Relu(x_{pl} - x_{ij})$. The parameter,$x_{pl}$, is a threshold that classifies residents as low-income. In this project we set $x_{pl} = \$30,000$ per year. The variable x_{ij} denotes the j-th resident's income in community i. The individual financial need for resident j in community i is the difference between x_{pl} and x_{ij} when $x_{pl} > x_{ij}$ and is zero otherwise. Thus, community i's aggregate financial need is given by the formula $\sum_{j=1}^{|D_i|} Relu(x_{pl} - x_{ij})$. As describe in section 2.3, the first term in equation (9)'s cost function is the squared mismatch between the allocated granted dollars, $p_i F$ and the financial need, n_i, of low-income residents. The second term in this cost function is the risk difference (i.e., statistical parity) weighted by the *fairness parameter*, λ.

We minimize the weighted cost function in equation (9) for various values of the fairness parameter. The outcome is shown in Figure 4, where it can be observed that increasing the fairness parameter decreases the risk difference, thereby indicating a more fair or "just" allocation of grant dollars. It is important to note, however, that this improvement in fairness is achieved at the expense of increasing the total squared mismatch $\sum_i (p_i F - n_i)^2$ between community need and allocated grant dollars.

In particular, the baseline policy, p^*, is obtained when we ignore statistical parity (i.e. $\lambda = 0$) as shown by point A in Figure 4. For the baseline policy, we see the risk difference

is about 0.35, indicating that the probability of an individual from an advantaged community benefiting from the grant is 35% higher than that of an individual from a disadvantaged community. By increasing the fairness parameter λ, we will obtain a fairer policy. One such solution is shown by point B in Figure 4. In this case, the risk difference for point B decreased to 0.15, implying a smaller disparity between the probability of advantaged individuals benefiting from the grant more than disadvantaged individuals. Policy B is therefore considered to be fairer than Policy A, as it reduces the bias towards the disadvantaged community. But this gain in fairness increased the mismatched need from 5 in policy A to 8.

Figure 4 suggests a way to quantitatively tradeoff fairness against the unmet need. In particular, decision makers would probably select a target risk difference level and see how much unmet need would be left. If it is desired to reduce that unmet need, one could increase the total grant dollar budget, F. As shown in Fig. 5, increasing F simply shifts the fairness/mismatch curve to the left. How much that total budget, F, should be increased can be readily determined from Figure 5, thereby providing a tool for budget planning as well.

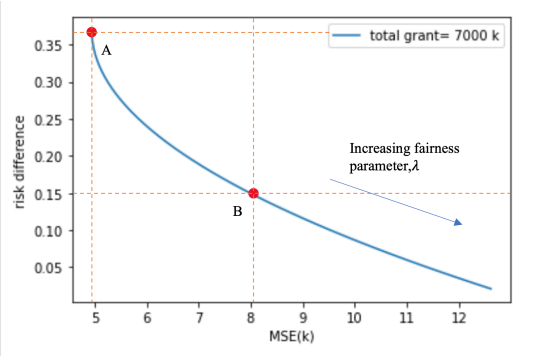


Figure 4. Policy's mean square error vs. fairness

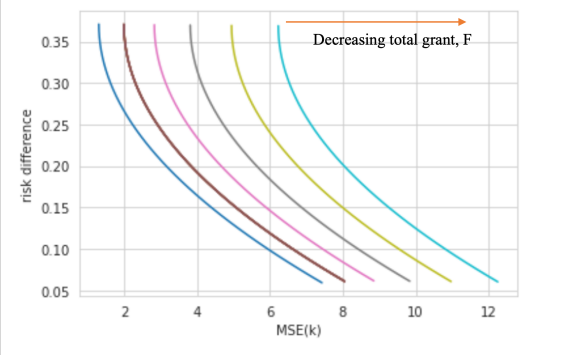


Figure 5. Policy's mean square error vs. fairness as total grant decreasing

5. Conclusion

In this project, we use the UCI Adult dataset as an example to demonstrate that the proposed fair federated method can develop a policy that is fairer than the baseline. By selecting the fairness parameter, one can trade off the the MSE in unmet community need against fairness of an allocation policy to meet those needs. Future work will develop connections with community leaders to better understand how this quantitative approach to achieving distributive justice can made into a practical decision-making tool for civic leaders.

References

Arjovsky, M. and Bottou, L. Towards principled methods for training generative adversarial networks, 2017.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein gan, 2017.

Deutsch, M. Equity, equality, and need: What determines which value will be used as the basis of distributive justice? *Journal of Social Issues*, 31:137–149, 1975.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks, 2014.

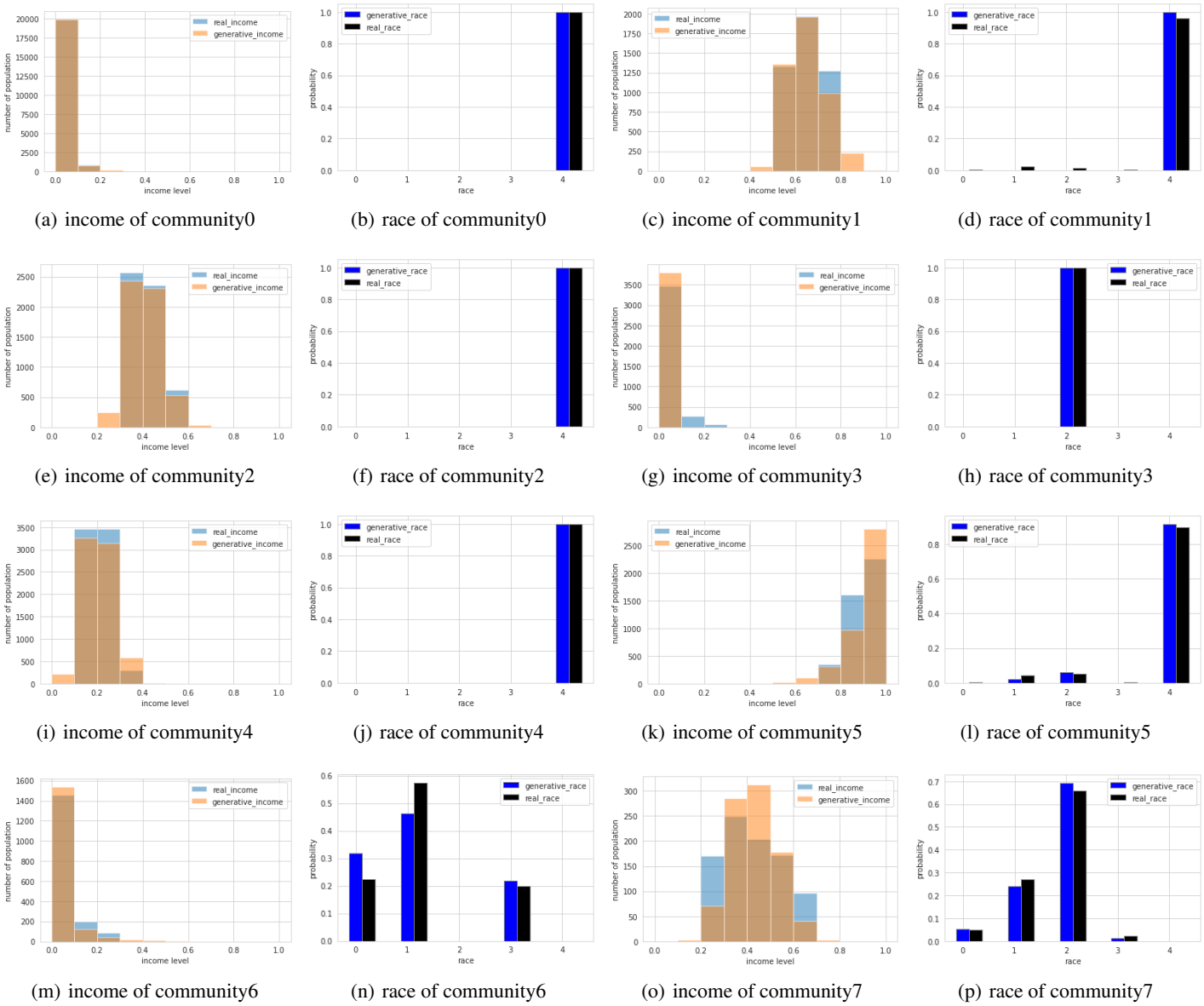
Konečnỳ, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data, 2023.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning, 2022.

Mothukuri, V., Parizi, R. M., Pouriyeh, S., Huang, Y., Dehghantanha, A., and Srivastava, G. A survey on security and privacy of federated learning. *Future Generation Computer Systems*, 115:619–640, 2021.

Rawls, J. and Kelly, E. I. Justice as fairness: A restatement. 2001.



(a) income of community0 (b) race of community0 (c) income of community1 (d) race of community1

(e) income of community2 (f) race of community2 (g) income of community3 (h) race of community3

(i) income of community4 (j) race of community4 (k) income of community5 (l) race of community5

(m) income of community6 (n) race of community6 (o) income of community7 (p) race of community7

Figure 6. The real and generative distributions of race and income for 8 communities