Debiasing Recidivism Risk Scores using GANs

Sofía Dudas, Alex Fuster, Matthew Radzihovsky
Department of Computer Science, Stanford University

Abstract

High imprisonment rates in the US have led courts to start using quantitative risk assessment software when making decisions about sentencing and releasing inmates. However, with particular fairness metrics, research has shown a bias against black inmates. To counteract this racial bias, we present a generative adversarial network (GAN) that predicts recidivism scores without imbuing the personal bias and prejudice that inevitably are present with a human decision maker, or even current risk assessment software.

Data and Feature Selection

We used the ProPublica Recidivism dataset, a labeled dataset of ~12k criminals from Broward County. Each data entry consists of demographic information, criminal record data, and COMPAS score (a measurement for predicted risk of recidivism). We construed this as a binary classification problem of recidivism risk, aiming to correctly predict "Low" or "High" risk for each individual.

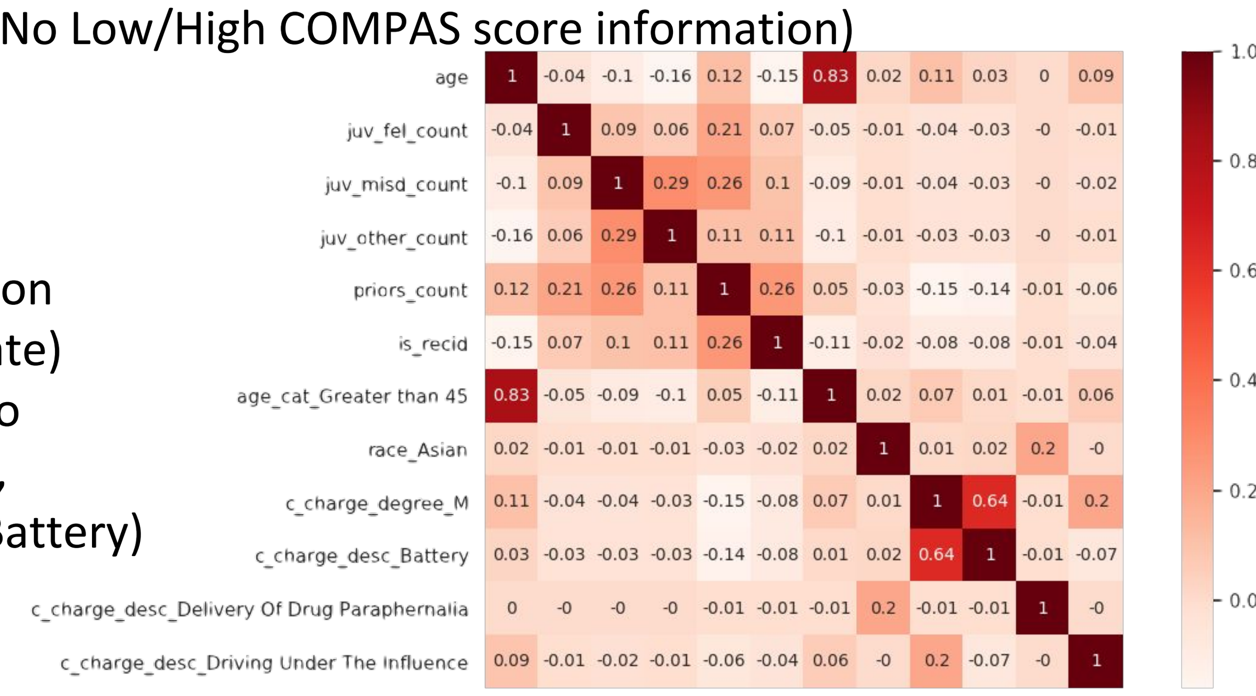
Pre-Processing

- Transformation of categorical variables to multiple indicator variables
- Unlabeled datapoints (No Low/High COMPAS score information)

Filtering

…**of attributes if**

- irrelevant to classification (date of birth, arrest date)
- redundant, according to correlation matrix (age, c_charge description_Battery)

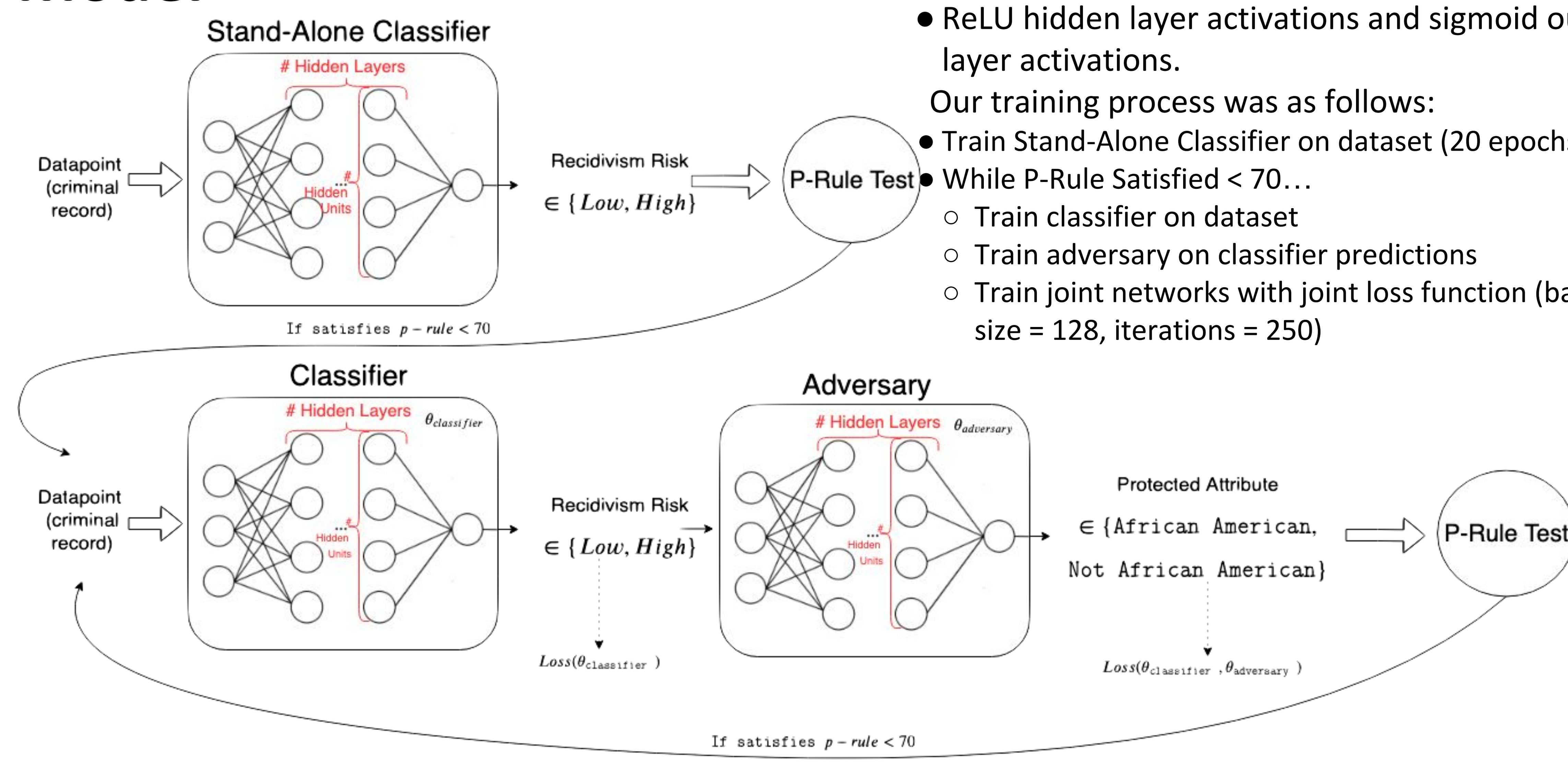


Correlation matrix for all features with correlation > 0.2 with another feature

Recursive Feature Elimination (RFE)

- Recursive elimination of attributes and building a model on remaining attributes
- We found the optimal number of features to be 372.
- In the resulting feature ranking…
  - The most important attributes were: age category, race indicator variables, charge degree
  - The least important attributes were: indicator variables for the charge descriptions.

Model



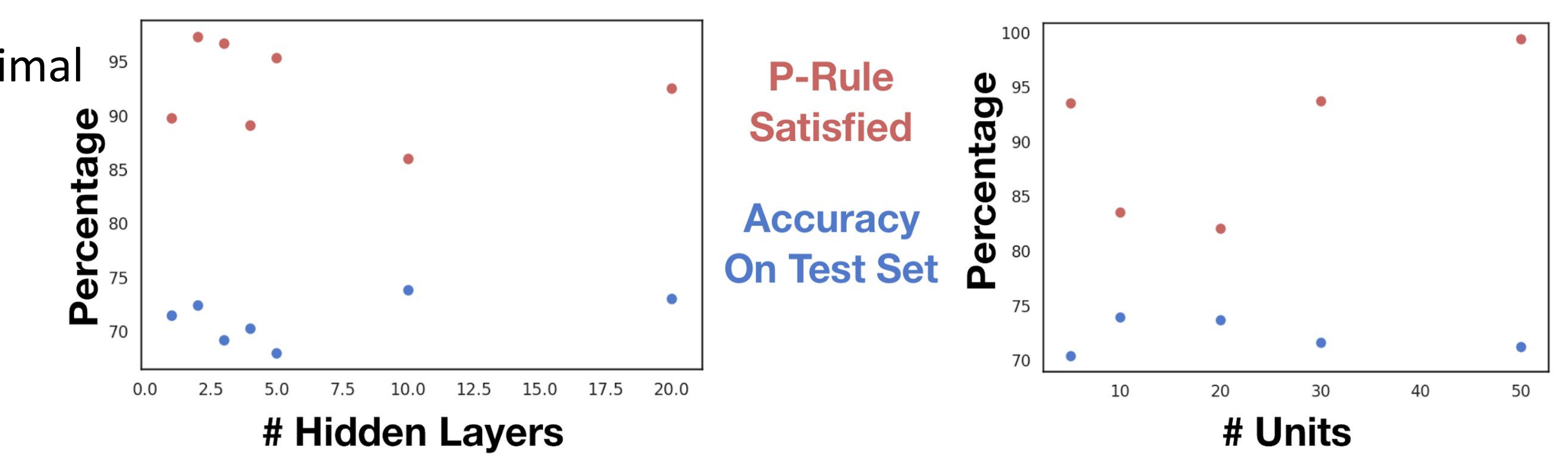
- Densely connected sequential neural networks
- ReLU hidden layer activations and sigmoid output layer activations.

Our training process was as follows:

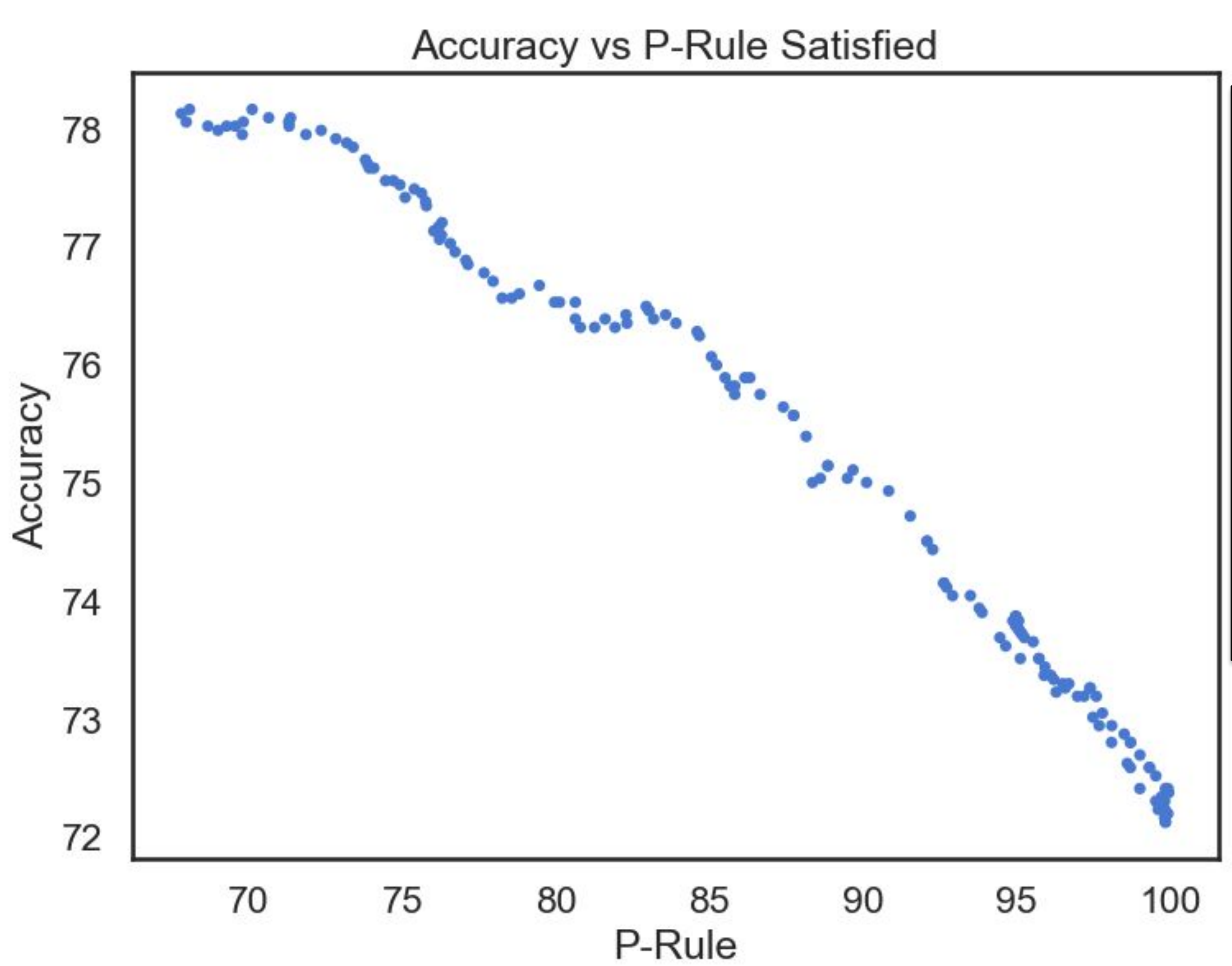
- Train Stand-Alone Classifier on dataset (20 epochs)
- While P-Rule Satisfied < 70…
  - Train classifier on dataset
  - Train adversary on classifier predictions
  - Train joint networks with joint loss function (batch size = 128, iterations = 250)

As seen in our analysis (to the right), optimal number of…

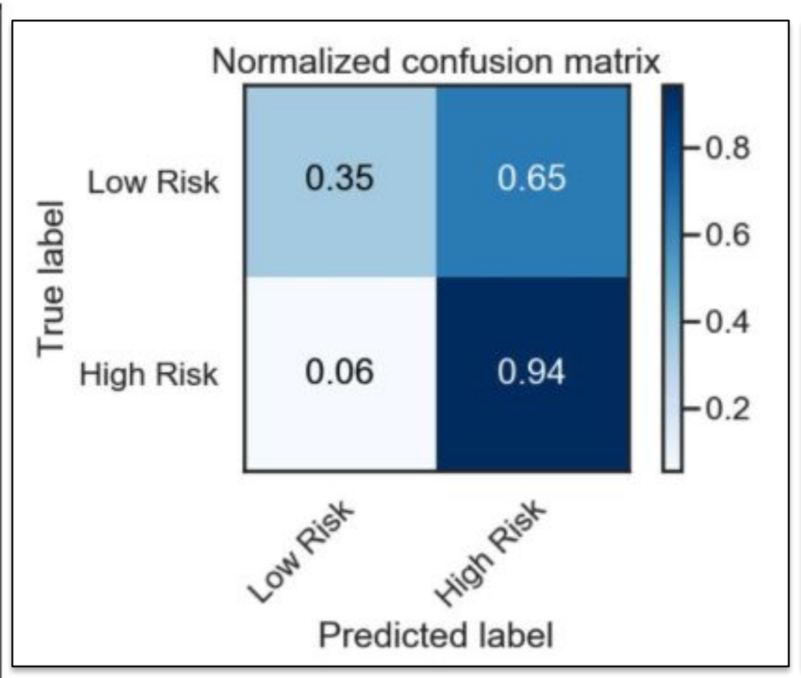
- Hidden Layers = 2
- Units per Hidden Layer = 50

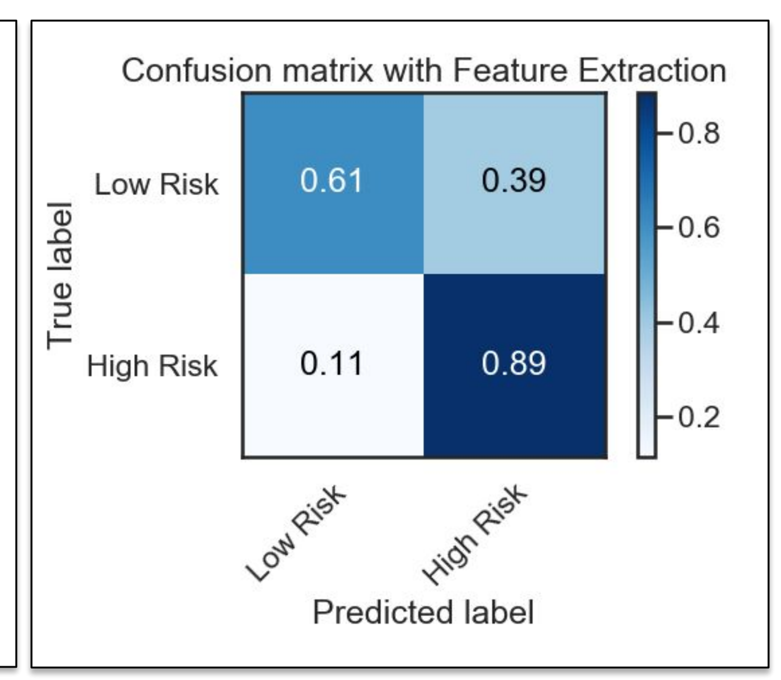


Results and Analysis:



- Plot of Accuracy vs P-Rule Satisfied using feature selection
- **Maximum Accuracy**: 78%, **Maximum P-Rule**: 99%
- Ability to tune fairness at the expense of accuracy





- Confusion Matrices Before and After Feature Extraction.
- False positives down 26%
- Increased precision with just 5% decrease on recall
- Balance letting free potential recidivists (recall) versus falsely predicting individuals as high risk (precision)

Future Direction

- Penalize false positives more than false negatives in the loss function - improve precision of model
- Provide adversary with additional features as input (not protected attributes) to the adversarial network, giving better predictive ability to the adversary and enabling more effective de-biasing
- Experimentation with weighting the classifier and adversary loss differently in loss function for the fitting of the joint model with de-biasing
- Experimentation with process for jointly-fitting the classifier and adversary

References & Acknowledgments:

[1] S. Acharya, "Tackling bias in machine learning,"medium.com, 2018.
[2] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-Adversarial Training of Neural Networks," arXiv e-prints, p. arXiv:1505.07818, May 2016.
[3] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," Proceeding AIES '18 Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, no. 1, pp. 335–340, 2018.
[4] C. Wadsworth, F. Vera, and C. Piech, "Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction," arXiv e-printss, p. arXiv:1807.00199v1, June 2018