How do the Score Distributions of Subpopulations Influence Fairness Notions?

Carmen Mazijn[1], Jan Danckaert[1], Vincent Ginis[1,2]

[1]DataLab, Vrije Universiteit Brussel, Belgium [2] School of Engineering and Applied Sciences, Harvard University, USA

VRIJE UNIVERSITEIT BRUSSEL

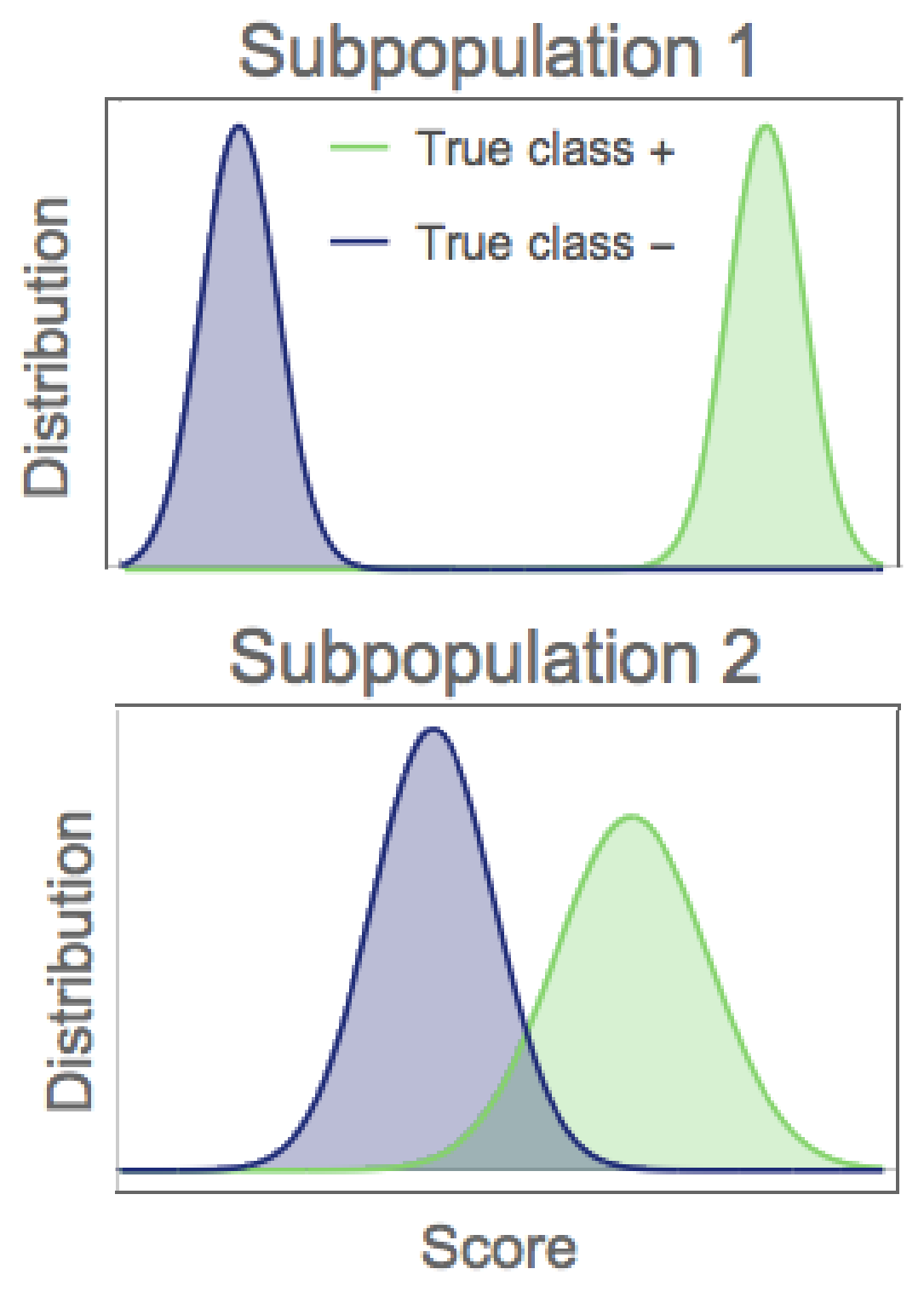
Context

To ensure decision-making algorithms fulfil their role ethically, we study algorithmic fairness. We examine the influence of the score distribution parameters on the Profit-Fairness trade-offs and the Fairness-Fairness trade-offs. We consider a binary classifier and a binary protected feature.

Score distributions

The outputted score determines whether a particular individual will get a reward or not. This decision depends not only on their individual score but also on the score of the others and the type of fairness notion chosen.

Depending on the values of the thresholds, people get the benefit or not. Different thresholds for different subpopulations are possible.



Considered Fairness notions

Profit:

$(TP-FP)/P_{max}$

Equal Accuracy:

$(TP_1 + TN_1)/N_1 - (TP_2 + TN_2)/N_2$

Demographic Parity:

$(TP_1 + FP_1)/N_1 - (TP_2 + FP_2)/N_2$

Equal Opportunity:

$TP_1/(TP_1 + FN_1) - TP_2/(TP_2 + FN_2)$

True Class

Predicted Class	Positive	Negative
Positive	TP	FP
Negative	FN	TN

Conclusions

The score distributions of the positive and negative class of the subpopulations significantly influence the trade-offs.

The symmetry of a model is more important than its expressivity.

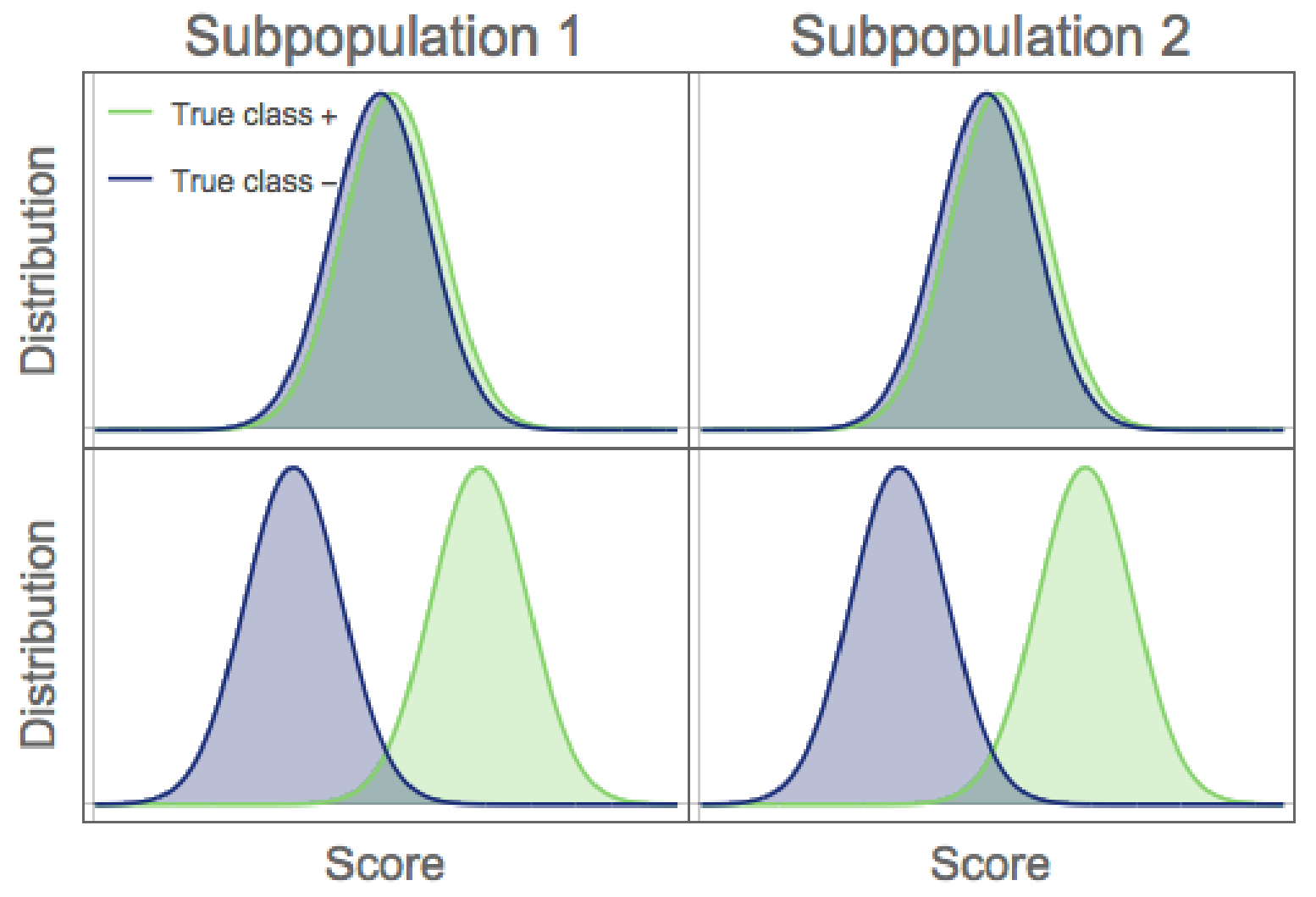
Improving the expressivity for both subpopulations is always good.

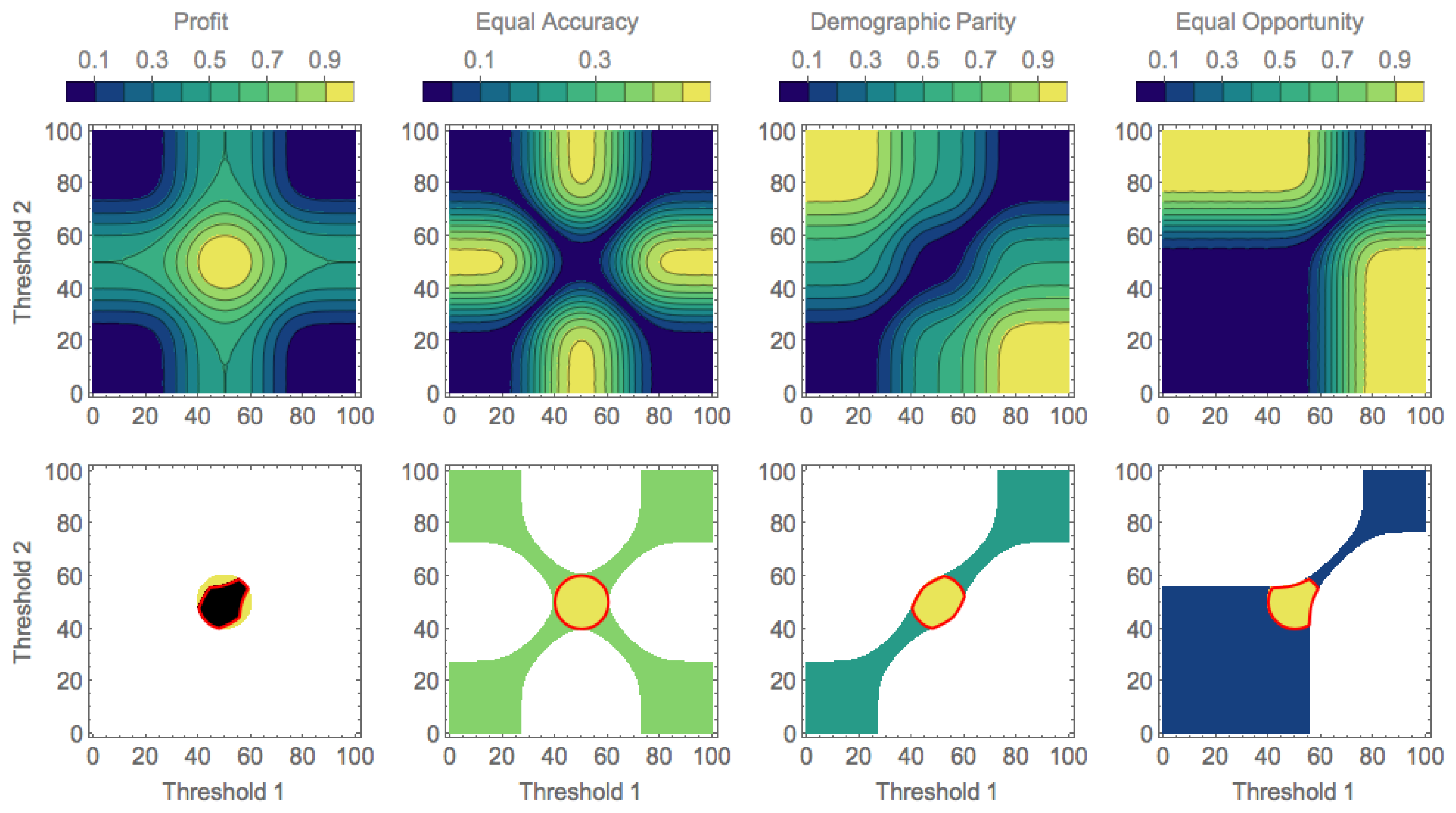
Increasing the difference between the means of the classes is more impactful than declining the variance of the classes.

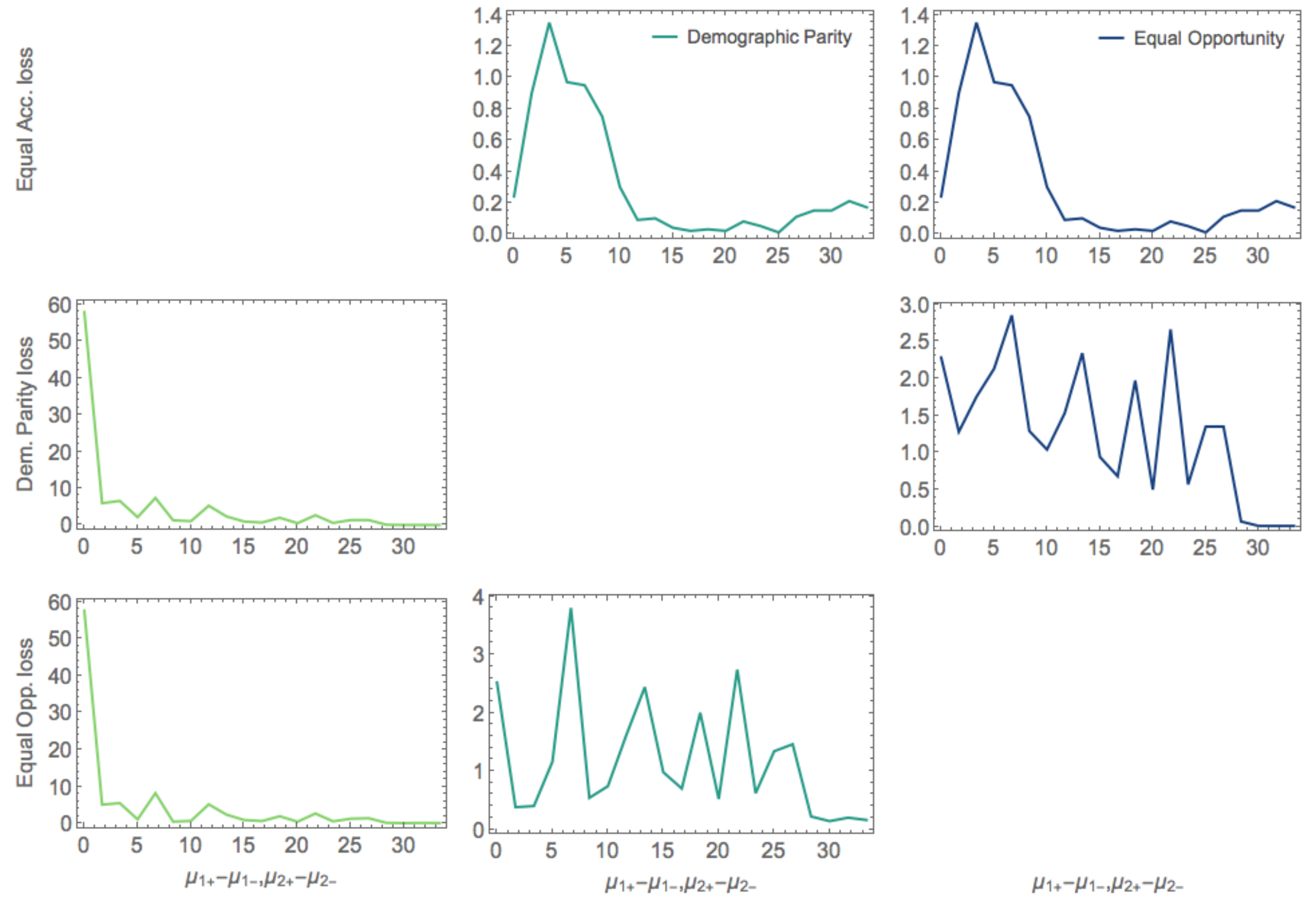
References

Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023 (2018).
Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. CoRR abs/1609.05807 (2016). arXiv:1609.05807 http://arxiv.org/abs/1609.05807
Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2020. On the Applicability of ML Fairness Notions. arXiv preprint arXiv:2006.16745 (2020).

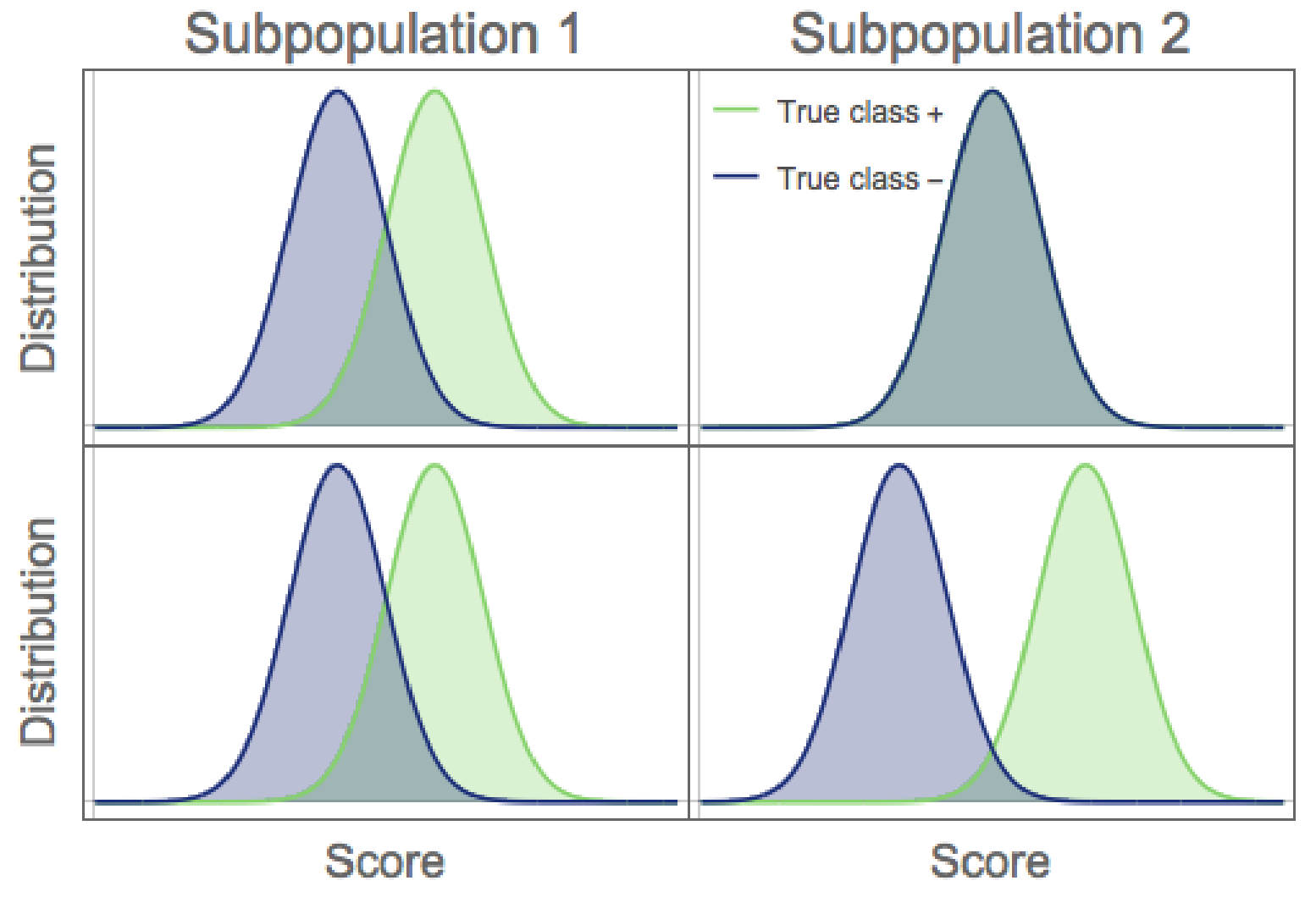
Analysis I

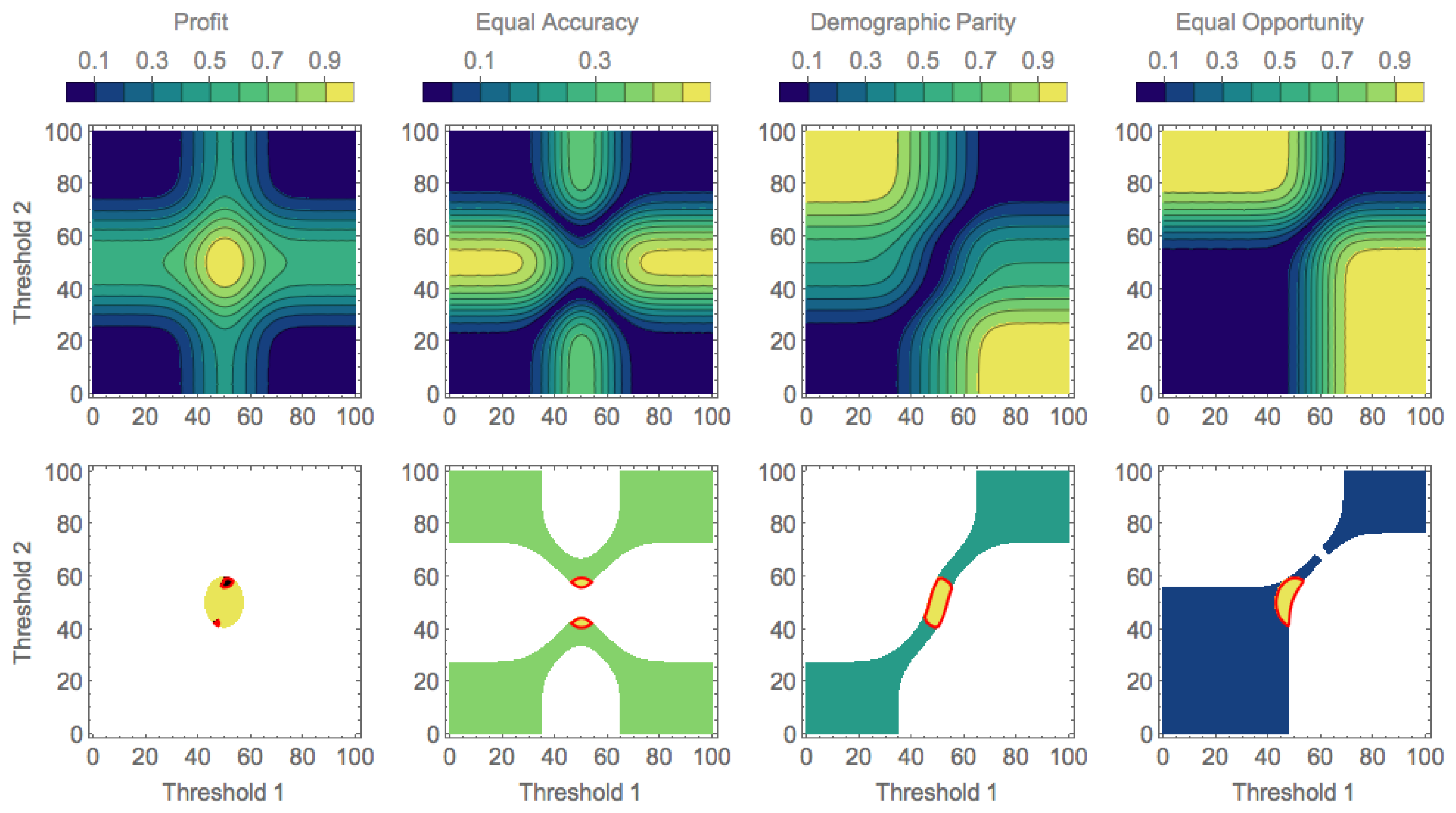


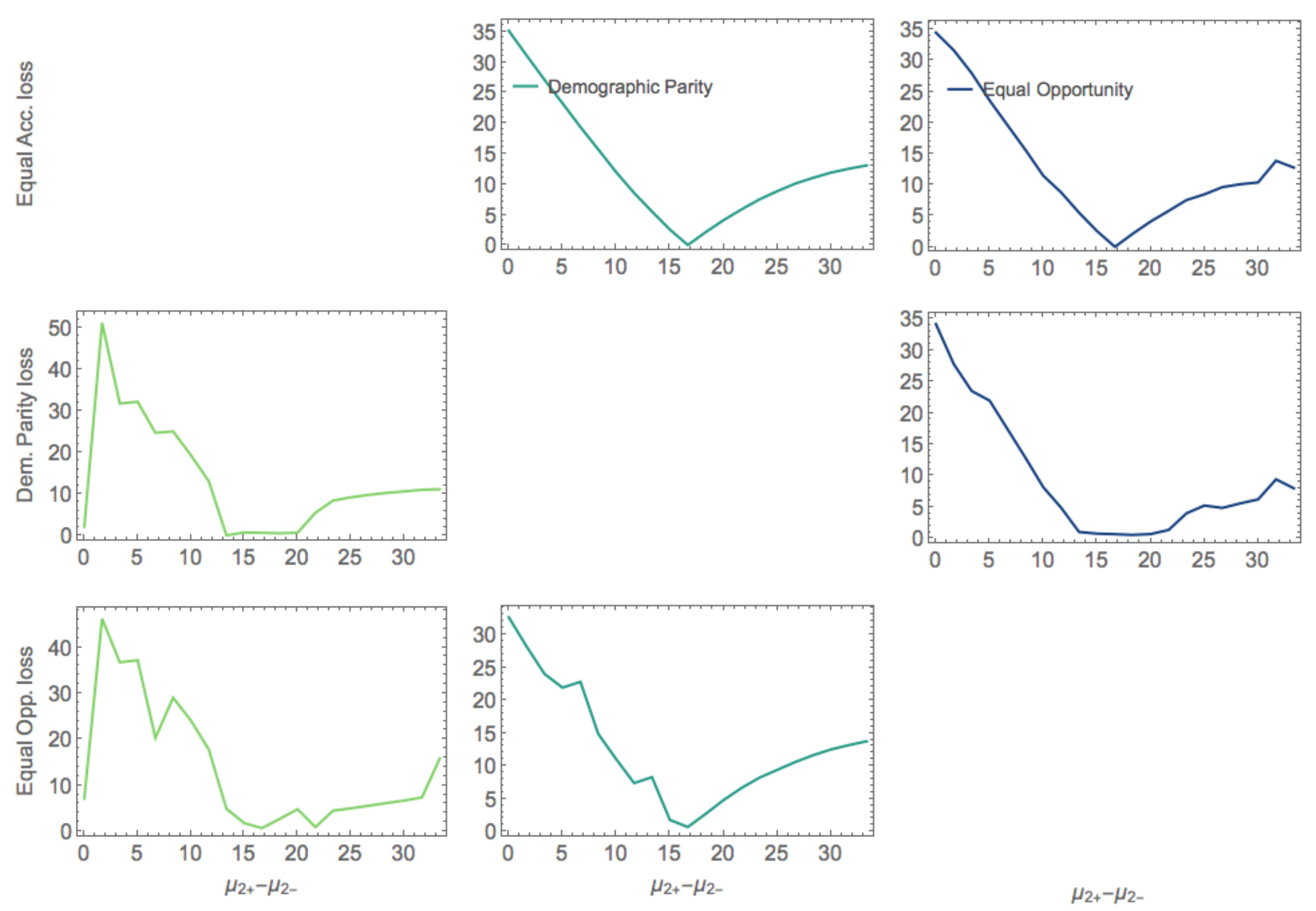




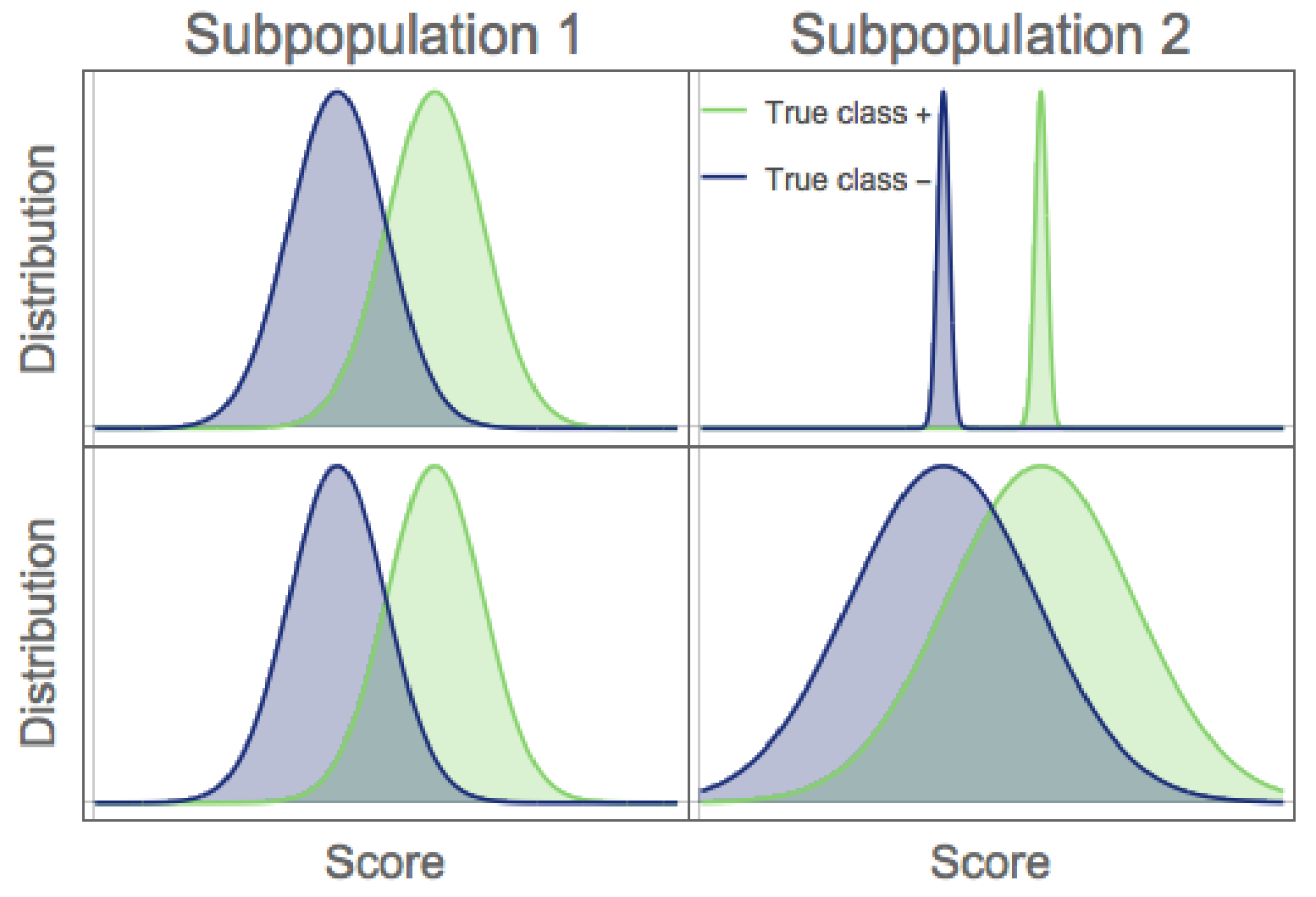
Analysis II

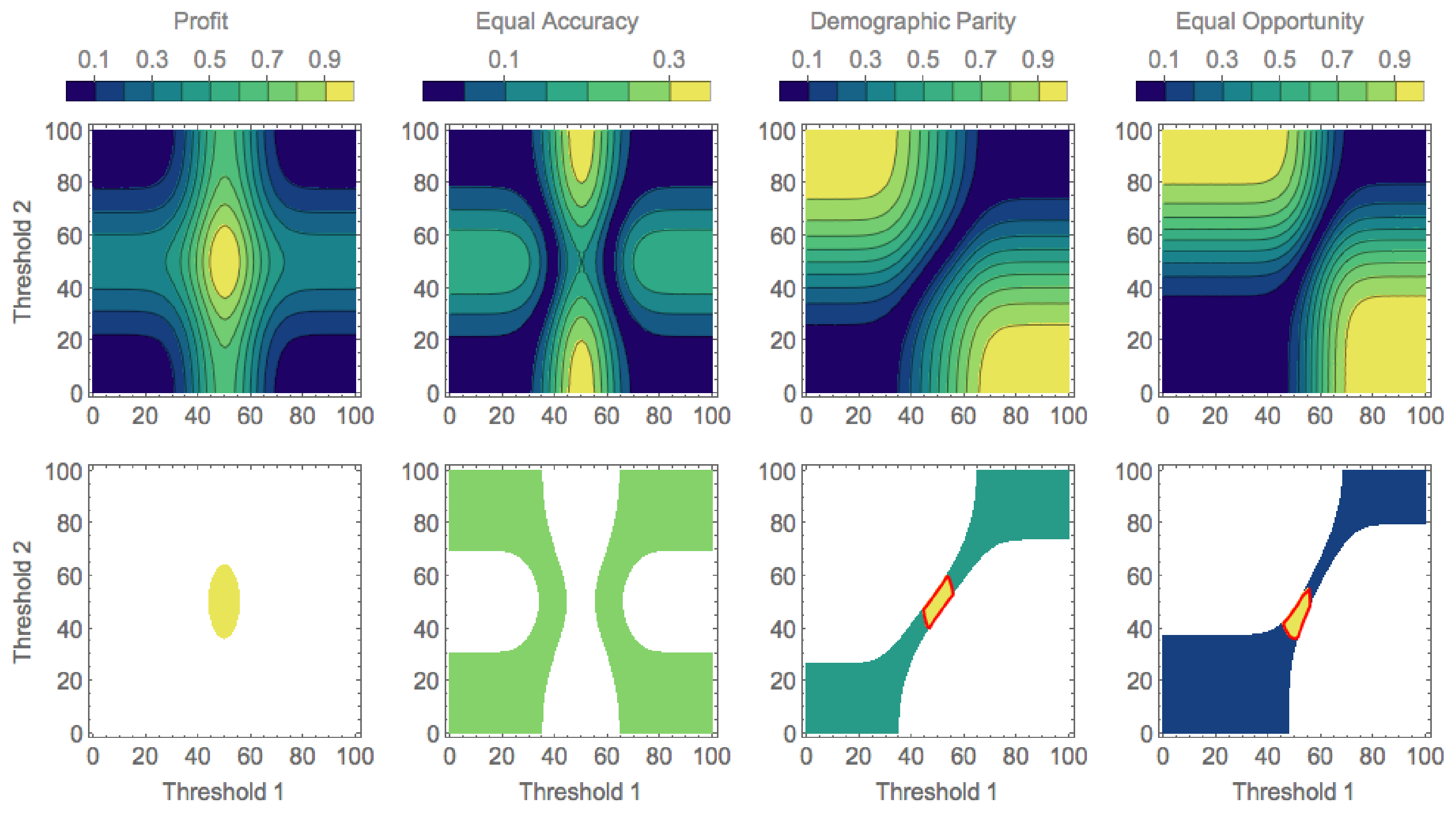


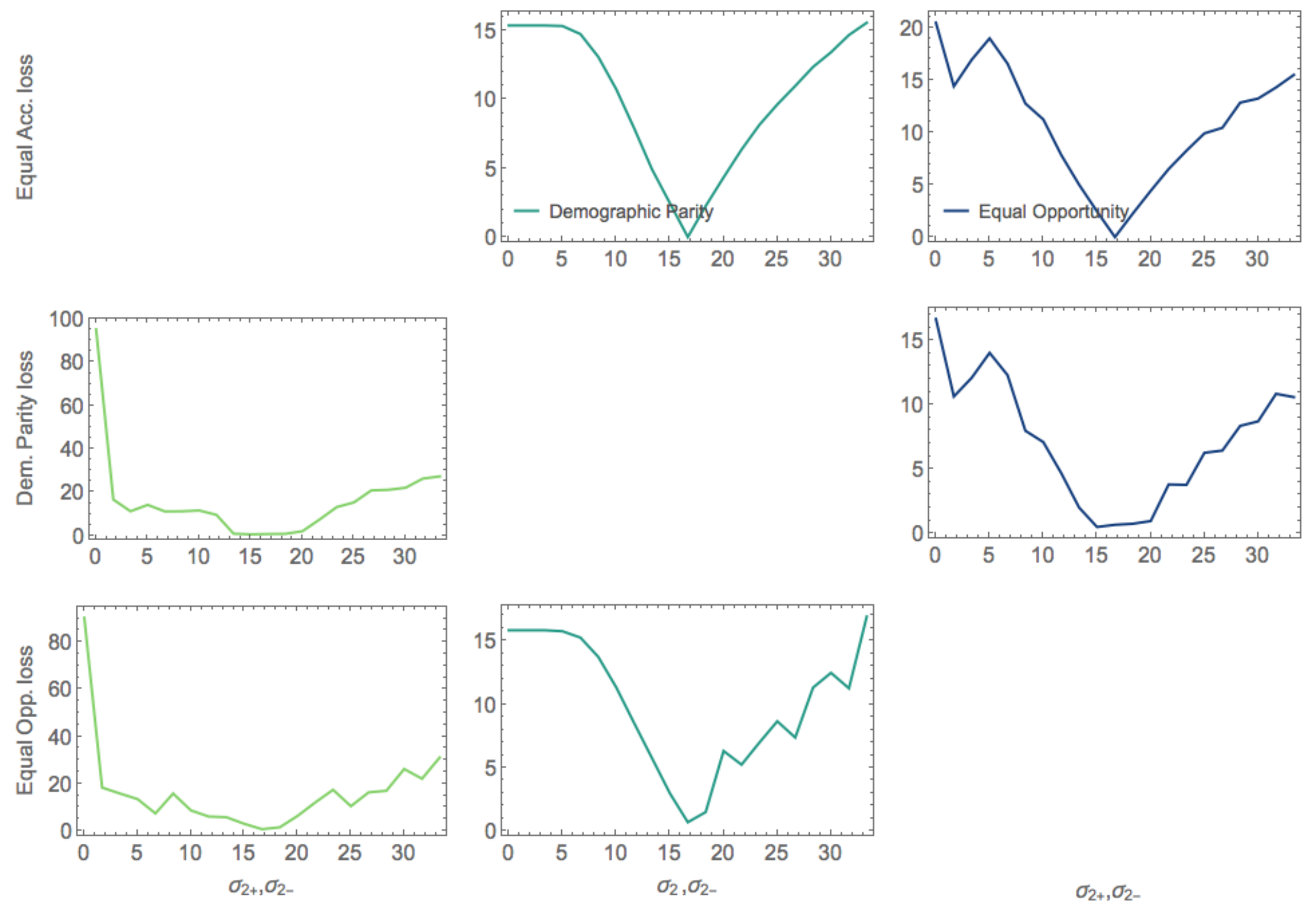




Analysis III


Subpopulation 1
Subpopulation 2
True class +
True class –
Distribution
Score


Profit
Equal Accuracy
Demographic Parity
Equal Opportunity
Threshold 2
Threshold 1


Equal Acc. loss
Dem. Parity loss
Equal Opp. loss
Demographic Parity
Equal Opportunity