Classifying Anomalies THrough Outer Density Estimation (CATHODE)

Anna Hallin
NHETC, Dept. of Physics and Astronomy
Rutgers University
Piscataway, NJ 08854, USA
anna.hallin@rutgers.edu

Joshua Isaacson
Theoretical Physics Department
Fermi National Accelerator Laboratory
Batavia, IL 60510, USA
isaacson@fnal.gov

Gregor Kasieczka
Institut für Experimentalphysik, Universität Hamburg
Luruper Chaussee 149, 22761 Hamburg, Germany
gregor.kasieczka@uni-hamburg.de

Claudius Krause
NHETC, Dept. of Physics and Astronomy, Rutgers University
Piscataway, NJ 08854, USA
claudius.krause@rutgers.edu

Benjamin Nachman
Physics Division, Lawrence Berkeley National Laboratory
Berkeley, CA 94720, USA
bpnachman@lbl.gov

Tobias Quadfasel
Institut für Experimentalphysik, Universität Hamburg
Luruper Chaussee 149, 22761 Hamburg, Germany
tobias.quadfasel@uni-hamburg.de

Matthias Schlaffer
Département de Physique Nucléaire et Corpusculaire, Université de Genève
Quai Ernest-Ansermet 24, 1205 Genève, Switzerland
matthias.schlaffer@etu.unige.ch

David Shih
NHETC, Dept. of Physics and Astronomy, Rutgers University
Piscataway, NJ 08854, USA
shih@physics.rutgers.edu

Manuel Sommerhalder
Institut für Experimentalphysik, Universität Hamburg
Luruper Chaussee 149, 22761 Hamburg, Germany
manuel.sommerhalder@uni-hamburg.de

Fourth Workshop on Machine Learning and the Physical Sciences (NeurIPS 2021).

Abstract

We propose a new model-agnostic search strategy for hints of new fundamental forces motivated by applications in particle physics. It is based on a novel application of neural density estimation to anomaly detection. Our approach, which we call Classifying Anomalies THrough Outer Density Estimation (CATHODE), assumes potential signal events cluster in phase space in a signal region. However, backgrounds due to known processes are also present in the signal region and too large to directly detect such a signal. By training a conditional density estimator on a collection of additional features outside the signal region, interpolating it into the signal region, and sampling from it, we produce a collection of events that follow the background model. We can then train a classifier to distinguish the data from the events sampled from the background model, thereby approaching the optimal anomaly detector. Using the public LHC Olympics R&D dataset, we demonstrate that CATHODE nearly saturates the best possible performance, and significantly outperforms other approaches in this bump hunt paradigm.

1 Introduction

Motivated by compelling theoretical and experimental evidence, the search for physics beyond the Standard Model (BSM) is a key research goal in particle physics. Given the basically infinite space of BSM theories, it is impossible to perform a dedicated search for every conceivable scenario. The lack of discoveries thus far could be because existing searches do not cover anomalous regions of phase space. Therefore, it is essential to complement the search program with model-agnostic methods.

An important class of such so-called anomaly detection approaches builds on the bump hunt strategy. It assumes that a potential signal is localized in one known feature m (often an invariant mass) and then uses data away from the signal (sideband region or SB) to estimate the background.The exact location of the signal (signal region or SR) is unknown and scanned over m. While broadly sensitive to new physics models with the targeted resonance, bump hunts are not particularly sensitive to any BSM model. Machine learning approaches that enhance the bump hunt use features x other than m to automatically amplify the presence of a potential signal.

Multiple strategies have been proposed for this task. Two key approaches are Classification Without Labels (CWOLA) Hunting [1–3] and Anomaly Detection with Density Estimation (ANODE) [4]. In this paper, we propose a new method combining the best of CWOLA Hunting and ANODE. With *Classifying Anomalies THrough Outer Density Estimation* (CATHODE), we train a density estimator to learn the (usually smooth) background distribution in the SB which we refer to as the "outer" region. Then we interpolate it into the SR, but rather than directly constructing the likelihood ratio as in ANODE, we instead generate *sample events* from the trained, interpolated background density estimator. These sample events should follow $p_{\text{bg}}(x)$ in the SR. Finally, we train a classifier (as in CWOLA Hunting) to distinguish $p_{\text{data}}(x)$ from $p_{\text{bg}}(x)$ in the SR.

Using the public R&D dataset [5] from the LHC Olympics (LHCO) [6], we will show that CATHODE achieves a level of performance (as measured by the significance improvement characteristic) that significantly surpasses the state-of-the-art for various amounts of signal. We also compare CATHODE to a fully supervised classifier (i.e. trained on labeled signal and background events) and an "idealized anomaly detector" (trained on data vs. perfectly simulated background).

2 The dataset

The main dataset used in this study follows the same prescription as in [4], corresponding to an LHC-like dijet resonance search. The background consists of 1M Standard Model (SM)-like events, whereas the signal consists of 1k events following a resonant signal with a resonance mass of 3.5 TeV. Figure 1 shows the distribution of the conditional feature m_{JJ} and the four auxiliary features, used to enhance the bump hunt: m_{J_1}, Δm_J, $\tau_{21}^{J_1}$ and $\tau_{21}^{J_2}$. The figure also highlights the SR, defined as 3.3 TeV $< m_{JJ} <$ 3.7 TeV. Note that this specific choice corresponds to the idealized scenario where the SR is centered around the actual signal resonance, whereas a realistic analysis would involve scanning over different choices.

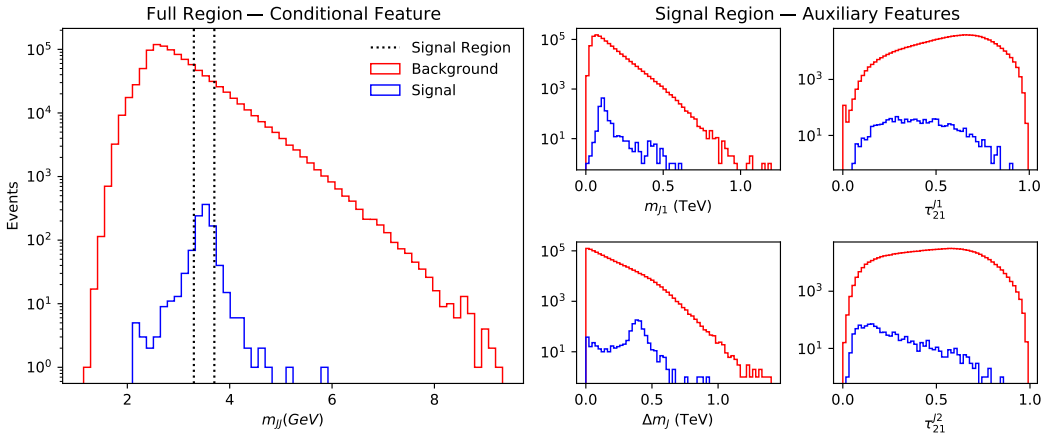

Figure 1: Left: distribution of the conditional feature m_{JJ} separated into signal and background contributions. Right: distributions of each of the four auxiliary features within the signal region.

The SR of the sample defined above, also referred to as the mock data, has a signal-to-background ratio of $S/B = 6 \times 10^{-3}$ and a significance of $S/\sqrt{B} = 2.2$. We also consider lower S/B values.

The data-driven anomaly detection methods will be compared to simulation-based approaches and thus an additional SR background simulation of 272k events following the same distribution as the mock data background are used for training the idealized anomaly detector and the fully supervised benchmark. Furthermore, an evaluation sample is set aside, consisting of 340k background and 20k signal events. These are only used in the final performance comparison.

3 The CATHODE method

3.1 Conditional density estimation and sampling

The first step of the CATHODE method is to train a conditional density estimator on the outer data. Assuming the signal is mostly contained in the SR, the density estimator will learn $p_{\text{data}}(x|m \notin \text{SR}) \approx p_{\text{bg}}(x|m \notin \text{SR})$, where $m = m_{JJ}$ and x are the auxiliary features. A Masked Autoregressive Flow (MAF) with affine transformations [7], whose architecture, training hyperparameters, and feature preprocessing are analogous to [4] is used as density estimator.

The mock data in the SB region is split into a training set consisting of 500k events, and a validation set consisting of the remaining SB events. The validation set is reserved for model selection. The losses on these two sets are tracked throughout training for each of the 100 epochs. The ten epochs (model states) with the lowest validation loss are selected for the event sampling.

The interpolation is automatically handled by the MAF. While it was trained on events with $m \notin \text{SR}$ to learn a function $p(x|m)$, it can be queried for any value of m, including $m \in \text{SR}$. A sample of N events are generated from each of the 10 chosen model states, using a kernel density estimate (KDE) fit to the distribution of m_{JJ} in the SR. The KDE was implemented using the Scikit-learn library [8] with a gaussian kernel and a bandwidth of 0.01. The events are then combined and shuffled into a set of $10N$ sample points. This ensembling yields a more representative sample than a single model.

3.2 Classifier

The third step of the CATHODE method is to train a classifier to distinguish the generated background-like events from the SR data. For all the methods under investigation we use the same classifier architecture. This consists of 3 hidden layers with 64 nodes each and a binary cross-entropy loss.

The binary classifier, implemented with PyTorch [9], is trained for 100 epochs, using the Adam [10] optimizer with a learning rate of 10^{-3}. When the classes (SR data or background-like events) are

imbalanced, they are reweighted in the loss computation accordingly, such that they contribute equally.

For this step, we divide the mock data in the SR in half, reserving 60k events for training the classifier and the remaining 60k events for validation (model selection). In a real-life application one would want to perform k-fold cross validation so as to not throw away half of the events. However, as this is a proof of concept we do not employ this here. We sample in total 400k events from the MAF generative model, which are distributed equally into the training and validation set for the classifier.

During training, the loss is recorded on the validation set. The model states of the 10 epochs with the lowest validation losses are used to construct an ensemble prediction, in which the individual predictions of each data point are averaged.

3.3 Anomaly detection

Finally, the trained classifier is applied to data in the SR. The ultimate goal of an anomaly detector is to learn the likelihood ratio between the data and background. In the presence of an anomaly, we have

$$p_{\text{data}}(x) = f_{\text{bg}} p_{\text{bg}}(x) + f_{\text{sig}} p_{\text{sig}}(x)\,, \tag{1}$$

with $f_{\text{sig}} = 1 - f_{\text{bg}} \ll f_{\text{bg}}$ the signal (anomaly) fraction. Although this signal fraction is unknown (along with the form of $p_{\text{sig}}(x)$), the likelihood ratio $p_{\text{data}}(x)/p_{\text{bg}}(x)$ is nevertheless monotonic with the signal-to-background likelihood ratio. Therefore, if the CATHODE method works, the events that are tagged by the classifier as "data-like" should be signal enriched, regardless of the signal.

4 Results

We demonstrate the efficacy of the CATHODE method on the LHCO R&D dataset. and quantify our results using the significance improvement characteristic (SIC), which is $\epsilon_S/\sqrt{\epsilon_B}$ vs. ϵ_S, where ϵ_S and ϵ_B are the signal and background efficiencies of a cut on the classifier score. Note that these efficiencies can only be calculated using the underlying truth labels that we have access to in this dataset. Therefore the SIC curve is being used to demonstrate that the method could find the signal if it was present in the data. In an actual search, where truth labels are not available, one would have to combine the CATHODE method with a background estimation procedure (e.g. sideband interpolation as in the bump hunt) and compute a p-value under the background-only hypothesis.

Figure 2 (left) shows the SIC curves of the different anomaly detection methods trained on our baseline dataset. The comparison also includes the "idealized" anomaly detector, which sets an upper limit on what performance one can expect from an anomaly detection method, and the fully supervised benchmark, corresponding to a dedicated search. The curves show the median value and 68% confidence bands of 10 independent trainings, where all steps of each method (e.g. both density estimator and classifier for CATHODE) have been reinitialized in each run.

We see that overall, CATHODE either matches or hugely outperforms the other weakly supervised methods (ANODE and CWOLA Hunting) across the entire range of signal efficiencies. At lower signal efficiencies, CATHODE reaches a maximum SIC of 14 compared to ANODE's 6.5 and CWOLA Hunting's 11.

The previous discussion has been focused on the benchmark scenario with 1000 signal events injected into the background sample (corresponding to $S/B \approx 0.6\%$ and $S/\sqrt{B} \approx 2.2$ in the signal region). The performance of the approaches at lower signal rates is explored in Figure 2 (right). Here, each method is evaluated 10 times for their maximum achieved significance at lower values of the signal/background ratio, each time with a different random separation of signal and background events into training, validation and evaluation sets. The hatched 68% confidence bands thus include the variance due to different realizations of the signal, along with reinitialized neural network trainings.

We see that CATHODE retains the highest significance improvement among the anomaly detection methods down to a signal fraction of about 0.25 %, after which point none of the methods can raise the total significance to at least 3 σ. We also see that across the entire range of relevant S/B values, CATHODE saturates the upper threshold set by the idealized anomaly detector. Therefore, the degradation in the performance of CATHODE as S/B decreases is not due to a failure of the CATHODE method, but is being set by limited statistics of the data in the signal region.

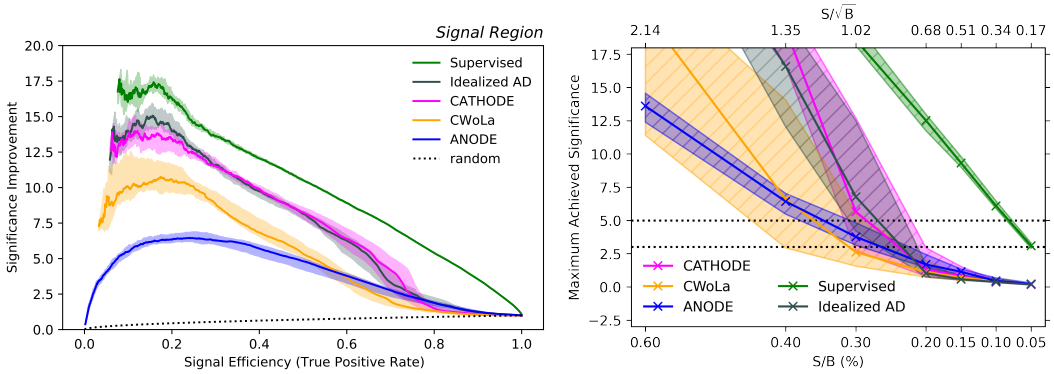

Figure 2: Left: significance improvement of the various anomaly classifiers as a function of the signal efficiency. Solid lines denote the median value of 10 fully independent trainings on the same fixed dataset, whereas the uncertainty bands contain 68% of these runs. Right: Achieved maximum significance, which is computed by multiplying the original significance by the maximum significance improvement, as a function of decreasing signal injections. The hatched 68% confidence bands here also incorporate the uncertainty from 10 different realizations of the signal and background events.

The CATHODE method as also been tested for its stability with respect to artificially introduced correlations, yielding near-optimal stability. More details on this are included in the Appendix A.1.

5 Conclusions

We have proposed a new method to detect anomalous regions in data that operates without specific model assumptions beyond the existence of a bump. This new CATHODE protocol can be seen as a synergy of the CWOLA Hunting and ANODE algorithms. Using the LHCO R&D dataset, we empirically test this procedure. Here, CATHODE achieves superior performance, as measured by the significance improvement characteristic, compared to CWOLA Hunting and ANODE. CATHODE reaches a maximal SIC of 14, compared to 11 for CWOLA Hunting and up to 6.5 for ANODE. Interestingly, CATHODE closely approximates an idealized anomaly detector's performance and—for low signal efficiency—even approaches a fully supervised classifier. The overall ability of anomaly detection algorithms decreases with the amount of signal present, and CATHODE achieves a significance of at least 3 σ down to an S/B ratio in the signal region of about 0.25%.

This gain likely comes from the two main innovations in the construction: using one density estimator instead of two (as in ANODE) simplifies the learning task, and only comparing events inside (either real or transported to) the signal region removes the problem of correlations for the classification task.

Given the widespread interest in deploying anomaly detection methods, the increased significance improvement and robustness of CATHODE compared to previous approaches should directly translate into more sensitive searches and might also enable applications outside of particle physics.

Acknowledgments

The work of AH, CK and DS was supported by DOE grant DOE-SC0010008. The work of BN was supported by the Department of Energy, Office of Science under contract number DE-AC02-05CH11231. GK, TQ, and MSo acknowledge the support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2121 "Quantum Universe" – 390833306. This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics. The work of MSc was supported by the Alexander von Humboldt Foundation. This research was supported in part by the National Science Foundation under Grant No. NSF PHY-1748958. JI, CK, and MSc thank Christina Gao for her contributions in the early phase of this project.

A Appendix

A.1 Performance in the presence of correlations

In a realistic application of anomaly detection, the signal and its properties are unknown. Therefore, one needs to be able to choose the set of auxiliary variables x as arbitrarily as possible, in order to gain generic discrimination power through them. However, some anomaly detection algorithms (e.g. CWoLa Hunting) are known to break down once there are significant correlations between x and m_{JJ}, thus limiting the choice of candidates for x.

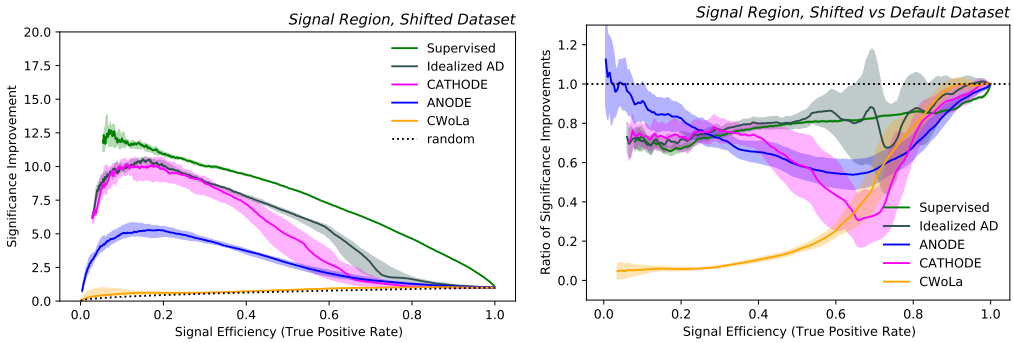

Figure 3: Left: Significance improvement of the various anomaly classifiers as a function of the signal efficiency on the shifted dataset. The solid lines are deduced from a median value of 10 fully independent trainings on the same training, validation and evaluation set. The uncertainty bands are defined the same way as in Figure 2. Right: The ratio between the significance improvement with and without the shift on the data applied.

As in [4], we test this effect by introducing an artificial correlation between x and m_{JJ} via shifting the features m_{J_1} and Δm in each event according to

$$\begin{aligned} m_{J_1} &\to m_{J_1} + 0.1 m_{JJ} \\ \Delta m &\to \Delta m + 0.1 m_{JJ} \end{aligned} \tag{2}$$

The CATHODE method is applied to the shifted dataset in the otherwise same setup as described in Section 3. The same benchmark methods as in Figure 2 are tested on this shifted data analogously and compared in Figure 3.

We see that to varying degrees, each of the different anomaly detection methods (as well as the supervised classifier) suffer from a performance loss due to the shift. In more detail:

1. Most notably, the CWoLa Hunting performance breaks down. This is completely expected, because the classifier can trivially deduce from the difference in m_{JJ} distribution whether a data point comes from the signal region or sideband, rather than learning the desired likelihood ratio.
2. Interestingly, the performances of the idealized anomaly detector and the supervised classifier also degrade due to the shift in x, with the degradation largest at lower signal efficiencies. We surmise that this is due to the fact that the classifiers are trained on x alone and not m_{JJ}; adding m_{JJ} to x then is effectively like smearing x by another independent random variable. This in turn makes the signal less localized relative to background, which would degrade the performance of even an optimal classifier—especially at lower signal efficiencies where the classifier is benefitting most from the localization of the signal relative to the background.
3. The ANODE method involves density estimation alone and not the classifier, which means that it does not have the same sensitivities to correlations that CWoLa Hunting does. However, we see from the ratio plot in Figure 3 that there is a drop in the performance of ANODE due to the shifted features, primarily at higher signal efficiencies. We attribute this to a combination of a more smeared out and difficult-to-find signal (as in the previous case), as well as worse density estimation in the presence of correlated or noisy features.

4. Finally, we come to the CATHODE method. Since CATHODE involves both density estimation and classification, we can think of it as a hybrid of ANODE and the idealized anomaly detector. Indeed, from Figure 3 (right), it is striking how CATHODE's performance degradation appears to be a "sum" of that of ANODE at higher signal efficiencies and the idealized anomaly detector at lower signal efficiencies. From Figure 3 (left), we see that at lower signal efficiencies, CATHODE is still comparable to the idealized anomaly detector and supervised classifier. Therefore, whatever is degrading the performances of the latter two is also affecting CATHODE in a similar way. Meanwhile, at higher signal efficiencies, CATHODE is noticeably worse than the idealized anomaly detector and seems to be tracking ANODE instead. Here we may be seeing the additional effect of poorer density estimation as for ANODE.

A.2 Code and data

The code for this paper can be found at https://github.com/HEPML-AnomalyDetection/CATHODE. The LHC Olympics R&D dataset can be found at https://zenodo.org/record/4287846.

References

[1] E. M. Metodiev, B. Nachman, and J. Thaler, "Classification without labels: Learning from mixed samples in high energy physics," *JHEP*, vol. 10, p. 174, 2017. DOI: 10.1007/JHEP10(2017)174. arXiv: 1708.02949 [hep-ph].

[2] J. H. Collins, K. Howe, and B. Nachman, "Anomaly Detection for Resonant New Physics with Machine Learning," *Phys. Rev. Lett.*, vol. 121, no. 24, p. 241 803, 2018. DOI: 10.1103/PhysRevLett.121.241803. arXiv: 1805.02664 [hep-ph].

[3] ——, "Extending the search for new resonances with machine learning," *Phys. Rev.*, vol. D99, no. 1, p. 014 038, 2019. DOI: 10.1103/PhysRevD.99.014038. arXiv: 1902.02634 [hep-ph].

[4] B. Nachman and D. Shih, "Anomaly Detection with Density Estimation," *Phys. Rev. D*, vol. 101, p. 075 042, 2020. DOI: 10.1103/PhysRevD.101.075042. arXiv: 2001.04990 [hep-ph].

[5] G. Kasieczka, B. Nachman, and D. Shih, *R&D Dataset for LHC Olympics 2020 Anomaly Detection Challenge*, version v4, Zenodo, Apr. 2019. DOI: 10.5281/zenodo.4536377. [Online]. Available: https://doi.org/10.5281/zenodo.4536377.

[6] G. Kasieczka *et al.*, "The LHC Olympics 2020: A Community Challenge for Anomaly Detection in High Energy Physics," Jan. 2021. arXiv: 2101.08320 [hep-ph].

[7] G. Papamakarios, T. Pavlakou, and I. Murray, "Masked Autoregressive Flow for Density Estimation," *arXiv e-prints*, arXiv:1705.07057, arXiv:1705.07057, May 2017. arXiv: 1705.07057 [stat.ML].

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[9] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[10] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2014. arXiv: 1412.6980 [cs.LG].

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We highlight which aspects are idealized and how they would be approached in a realistic analysis.
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] In the appendix.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] These choices are either listed in the paper or can be found in the referenced sources.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We draw the error bars differently depending on whether the underlying data has been reshuffled. In any case, the model trainings have been reinitialized 10 times.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No] It can be found in the reference.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] A link to the code to reproduce the results of this paper is provided in the appendix.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] Our usage is in agreement with their license.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]