Re-calibrating Photometric Redshift Probability Distributions Using Feature-space Regression

Biprateep Dey, Jeffrey A. Newman, Brett H. Andrews
Dept. of Physics and Astronomy and PITT-PACC
University of Pittsburgh, Pittsburgh, PA 15260, USA
{biprateep, janewman, andrewsb}@pitt.edu

Rafael Izbicki
Dept. of Statistics
Federal University of São Carlos (UFSCar)
São Carlos, Brazil

Ann B. Lee, David Zhao
Dept. of Statistics & Data Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Markus Michael Rau
Dept. of Physics & McWilliams Center for Cosmology
Carnegie Mellon University, Pittsburgh, PA 15213, USA
&
High Energy Physics Division
Argonne National Laboratory, Lemont, IL 60439, USA

Alex I. Malz
Ruhr-University Bochum, Astronomical Institute
German Centre for Cosmological Lensing
Universitätsstr. 150, 44801 Bochum, Germany

Abstract

Many astrophysical analyses depend on estimates of redshifts (a proxy for distance) determined from photometric (i.e., imaging) data alone. Inaccurate estimates of photometric redshift uncertainties can result in large systematic errors. However, probability distribution outputs from many photometric redshift methods do not follow the frequentist definition of a Probability Density Function (PDF) for redshift — i.e., the fraction of times the true redshift falls between two limits z_1 and z_2 should be equal to the integral of the PDF between these limits. Previous works have used the global distribution of Probability Integral Transform (PIT) values to re-calibrate PDFs, but offsetting inaccuracies in different regions of feature space can conspire to limit the efficacy of the method. We leverage a recently developed regression technique that characterizes the local PIT distribution at any location in feature space to perform a local re-calibration of photometric redshift PDFs. Though we focus on an example from astrophysics, our method can produce PDFs which are calibrated at all locations in feature space for any use case.

1 Introduction

Galaxy distance, as measured by redshift, is essential for estimating intrinsic luminosity and 3D location in space, which is crucial information for many astrophysical studies. High-precision redshifts require resource-intensive observations and will only be feasible for a few percent of galaxies

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

in upcoming photometric surveys. Thus, photometric redshifts (photo-z's)—redshifts estimated from imaging alone—will be necessary. Furthermore, accurate photo-z's are critical for some science cases (e.g., weak lensing cosmology), but PDFs from both main methods of photo-z estimation (galaxy spectral template-based and machine learning-based) fail to satisfy the frequentist definition of a PDF for redshift [Dahlen et al., 2013, Kodra, 2019, Schmidt et al., 2020]. The fraction of times the true redshift falls between two limits z_1 and z_2 should equal the integral of a properly-defined PDF between these limits, for any arbitrary subset of the test data.

Current metrics used to measure the quality of calibration, like the distribution of the values of the cumulative distribution function (CDF) evaluated at the true redshift of the object (the Probability Integral Transform or PIT; see Eq. 1) can favor pathological but un-informative PDFs [Schmidt et al., 2020]. Moreover, overall uniformity of PIT values is possible even if particular subsets of the same test data are poorly-calibrated [Zhao et al., 2021]. If the PDFs are well-calibrated, then the distribution of the PIT values of a test sample will be uniform between 0 and 1 or their corresponding CDF will follow the identity line for any arbitrary subset of the test data. The same can be visualized with a P-P plot that shows the empirically calculated CDF versus their theoretical expected values. Ideally, the P-P plot should closely follow the identity line, but it often does not.

Several previous works have studied PDF re-calibration (e.g., Niculescu-Mizil and Caruana 2005, Rau et al. 2015, Kuleshov et al. 2018), though none can ensure that PDFs are well-calibrated at every point in feature space. Bordoloi et al. [2010] described a method to re-calibrate PDFs using a single correction factor based on the overall distribution of PIT values, which ensures a uniform global distribution of PIT values, but this single correction factor is applied to all PDFs and does not account for local variations. Importantly, these local inconsistencies in feature space can be detected using tests like the ones proposed in Jitkrittum et al. [2020] and Zhao et al. [2021], which we will leverage in our method.

In this work, we develop a local PDF re-calibration procedure that uses an estimate of the local distribution of PIT values (from Zhao et al. [2021]) to calculate a correction factor at any location in feature space. As a proof-of-concept, we train a model to predict photo-z PDFs using galaxy magnitudes and colors, which are measures of the amount of light detected in broad wavelength regions. We use the FlexZBoost [Izbicki and Lee, 2017, Dalmasso et al., 2020] algorithm, which was demonstrated as the best performing photo-z prediction algorithm among the ones compared by Schmidt et al. [2020], though any machine learning algorithm producing PDFs will suffice. FlexZBoost uses gradient boosted decision trees (specifically XGBoost; Chen and Guestrin 2016) to predict photo-z PDFs by minimizing the Conditional Density Estimate (CDE) loss. We use the TEDDY data sets [Beck et al., 2017] to train and test our methods. The TEDDY data set is divided into four subsets which provide us with a test bed for photo-z algorithms to train and test on various distributions of the feature space. We train FlexZBoost on a random 70% subset of the TEDDY-A data set and use the remaining 30% as a calibration/validation set. We take the initial photo-z PDFs produced by FlexZBoost as our starting point, which we re-calibrate using 30% of TEDDY-A as our calibration set and test our methods on TEDDY-B and C data sets. TEDDY-B has the same distribution of features as TEDDY-A, whereas TEDDY-C has a slightly different distribution but spans the same input space.

2 Re-calibration Procedure

Let $\widehat{p}(z|\mathbf{x})$ be the initial estimate of the true PDF $p(z|\mathbf{x})$ of the target variable z (redshift) given the input features $\mathbf{x}$ (galaxy colors and magnitudes). The random variable corresponding to z is denoted by Z. We define the local Probability Integral Transform (PIT) corresponding to this initial estimate as:

$$\widehat{\mathrm{PIT}}(z, \mathbf{x}) = \int_0^z \widehat{p}(z'|\mathbf{x})dz' = \widehat{F}(z|\mathbf{x}) \tag{1}$$

where $\widehat{F}$ is the cumulative distribution function associated with $\widehat{p}$. Using a labeled calibration set (30% of TEDDY-A) and a suitable regression method (XGBoost; Chen and Guestrin 2016 in our case), we estimate the CDF of PIT values as a function of $\mathbf{x}$ following the method described in Zhao et al. [2021]. We perform a separate regression for each value of α on a pre-chosen grid $\mathbb{G} \subset [0, 1]$ of coverage levels to get the CDF of PIT values ($r_\alpha^{\widehat{p}}(\mathbf{x})$):

$$r_\alpha^{\widehat{p}}(\mathbf{x}) \coloneqq \mathbb{P}\left(\widehat{\mathrm{PIT}}(Z, \mathbf{x}) \leq \alpha|\mathbf{x}\right) = \mathbb{P}\left(Z \leq \widehat{F}^{-1}(\alpha|\mathbf{x})|\mathbf{x}\right) \tag{2}$$

If our initial PDFs are locally calibrated, then the relation $r_\alpha^{\widehat{p}} = \alpha$ should hold for any $\mathbf{x}$, i.e., a plot of $r_\alpha^{\widehat{p}}$ vs. α (also called Amortized Local P-P plots or ALP plots) should closely follow the identity line. Most photo-z estimators do not produce locally calibrated PDFs and this relation often does not often hold (see e.g., Fig. 2).

To re-calibrate the original PDF estimates $\widehat{p}(z|\mathbf{x})$ such that the relation $r_\alpha^{\tilde{p}} \approx \alpha$ holds for the new PDFs, $\tilde{p}$, for a new unseen test data set, we define, $\beta \coloneqq r_\alpha^{\widehat{p}}$ for each $\alpha \in \mathbb{G}$ and define a new cumulative distribution function, $\tilde{F}$, such that:

$$\tilde{F}^{-1}\left(\beta|\mathbf{x}\right) = \widehat{F}^{-1}\left(\alpha|\mathbf{x}\right) \tag{3}$$

Then by construction the new PDFs, $\tilde{p}$, will be calibrated since

$$r_\beta^{\tilde{p}}(\mathbf{x}) = \mathbb{P}\left(Z \leq \tilde{F}^{-1}(\beta|\mathbf{x})|\mathbf{x}\right) = r_\alpha^{\widehat{p}} = \beta \tag{4}$$

Now, for $\tilde{z} = \tilde{F}^{-1}\left(\beta|\mathbf{x}\right)$ we will have,

$$\int_0^{\tilde{z}} \tilde{p}(z'|\mathbf{x})dz' = \beta \tag{5}$$

$$\implies \tilde{p}(\tilde{z}|\mathbf{x}) - \tilde{p}(0|\mathbf{x}) = \frac{d\beta}{dz'} = \frac{d\beta}{d\alpha}.\frac{d\alpha}{dz'} \tag{6}$$

Eqs. 3 and 1 imply that $\tilde{p}(\tilde{z}|\mathbf{x}) = \tilde{p}(z|\mathbf{x})$ and Eq. 2 implies that $\frac{d\alpha}{dz'} = \widehat{p}(z|\mathbf{x})$. It is not physical to have any object at redshift 0 so we can assume $\tilde{p}(0|\mathbf{x}) = 0$. This gives us the relation:

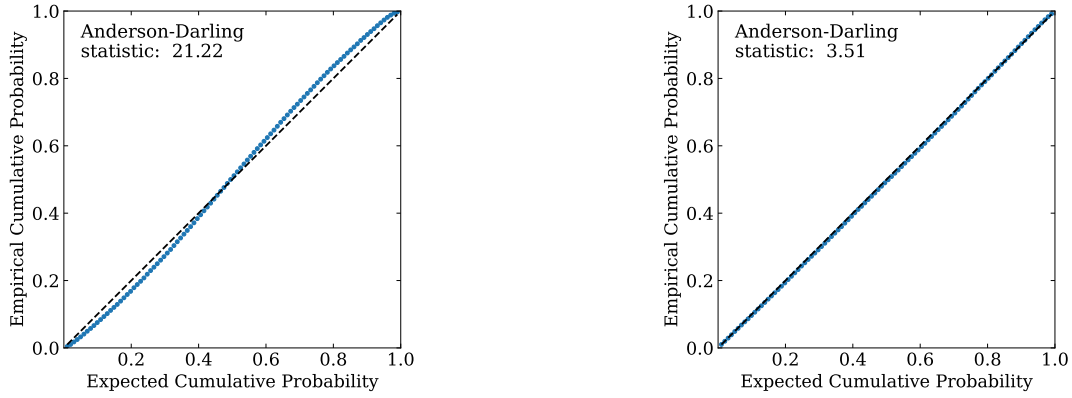
$$\tilde{p}(z|\mathbf{x}) = \widehat{p}(z|\mathbf{x}).\frac{d\beta(\alpha)}{d\alpha} \tag{7}$$

This means that our corrected PDF equals the initial PDF multiplied by a correction factor which is the local PIT distribution evaluated at the coverage corresponding to various redshifts. This relation is very similar to what Bordoloi et al. [2010] uses to re-calibrate photo-z PDFs except now the correction factor is calculated using the local PIT distribution rather than the empirical distribution obtained from the calibration set as a whole.

The local CDF of PIT values ($r_\alpha^{\widehat{p}}$) is often noisy (see Fig. 2). So we implement the re-calibration method defined by Eq. 7, with a smooth version of $\beta(\alpha)$ and its derivative. Since, $\beta(\alpha)$ is the CDF corresponding to the PIT distribution, it should be a monotonic non-decreasing function of α. Therefore, we use a basis of I-spline functions to fit a smooth model using non-negative least square regression [Ramsay, 1988]. I-splines are monotonic functions, a linear combination of which with non-negative coefficients gives us any arbitrary monotonic non-decreasing function. We use splines of order 3 and use a basis of 5 splines. The number of basis splines to use is a hyper-parameter and controls the level of smoothing. The derivative of I-splines can be obtained analytically (M-splines), which gives us a smooth representation of the correction factor when evaluated on a grid of α corresponding to the discrete grid of z on which $\widehat{p}$ has been evaluated.

3 Results and Discussion

We apply our method to re-calibrate PDFs obtained from FlexZBoost on both the TEDDY-B and TEDDY-C data sets and observe improvements in calibration. In addition to using P-P plots for a visual comparison, we use the Anderson-Darling (AD) statistic which is a weighted mean-squared difference between the theoretical and empirical CDFs of the PIT distributions. A lower value will indicate that the two CDFs are more similar. For the sake of brevity, we present the results from TEDDY-C as it is a more challenging scenario given the difference in the distributions of the calibration and test sets. Fig. 1 shows the P-P plot for the global distribution of PIT values for the entire TEDDY-C test set before and after the re-calibration procedure. The initial P-P plot (Fig. 1a) deviates from the identity line but after re-calibrating using the local distribution of PIT values follows the identity line closely (Fig. 1b), which is also evident from the significant decrease in the value of the AD statistic. Local calibration is a stronger requirement than global calibration so we expect that PDFs that are well calibrated locally will also be well calibrated globally.



(a) Original

(b) Re-calibrated

Figure 1: P-P plot of the global distribution of PIT values for the test sample (TEDDY-C). The blue dots show the empirical CDF of the PIT values calculated as a function of their theoretical expected value. Ideally the empirical CDF and the theoretical CDF should be equal and follow the identity line (black dashed line). Fig. 1a shows the P-P plot for the PDFs predicted by FlexZBoost and Fig. 1b shows the P-P plot for the same data set after local re-calibration. We see that after re-calibration empirical and expected CDFs match closely which is also quantified by a significant decrease in the value of the Anderson-Darling Statistic.

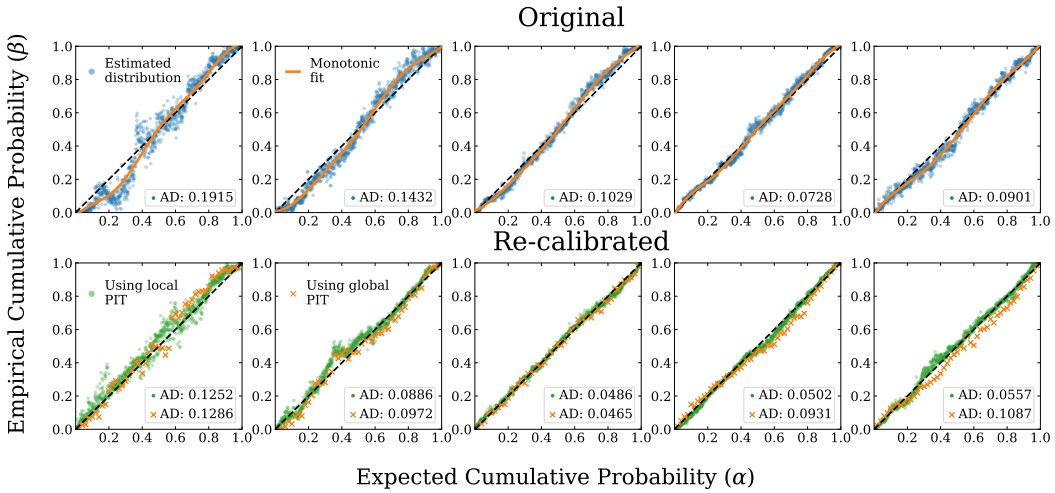


Figure 2: P-P plot of the local distribution of PIT values for 5 randomly selected objects from the test sample (TEDDY-C). The blue dots show the empirical CDF for the PIT distribution estimated as a function of their theoretical expected value using the method described in Zhao et al. [2021]. Ideally the empirical CDF and the theoretical CDF should be equal and follow the identity line (black dashed line). The orange curve in top row shows the (least squares) fit to the blue dots using a basis of 5 I-splines. This fitted model is used to calculate the correction factor from eq. 7. The top row shows the local P-P plot for the original set of PDFs, whereas the bottom row shows the local P-P plot for the globally (orange crosses) and locally (green dots) re-calibrated PDFs for the same set of objects. After re-calibration, the local P-P plots are closer to the identity line and have a lower value of the Anderson-Darling (AD) statistic. The re-calibration done using the local PIT distribution tends to perform better than the re-calibration done using the global distribution of PIT values, as seen from the lower value of the AD statistic for a majority of the cases.

Fig. 2 shows the local P-P plots for a random subset of galaxies from TEDDY-C. The top row shows the local distribution of PIT values as inferred using a regression method trained on the calibration set and a smooth representation of the same obtained by fitting a monotonic function to the data using a basis of I-spline functions. After the re-calibration procedure was applied (both using global and local PIT distributions) to the entire TEDDY-C data set, we use half of TEDDY-C to train the regression model on the re-calibrated PDFs to predict the CDF of PITs for the other half of the data set. The bottom row shows a comparison of the local P-P plots re-calibrated using both local and global distribution of PIT values. We see that the P-P plots follow the identity line more closely after global re-calibration and perform even better with local re-calibration. This is again evident from the decrease in the value of the AD statistic after re-calibration. We find that the local re-calibration method tends to outperform the global re-calibration in most cases.

This work shows that PDFs can be re-calibrated using local information and produce better uncertainty estimates. Though the method works reasonably well when the distribution of features for the calibration set is slightly different from the test set, we expect performance to worsen if the distribution of features for the test set is drastically different. A systematic study to understand the performance of this method for various distributions of input features will be performed in a future work.

Broader Impact

Machine learning algorithms are being increasingly used in decision making processes, including situations where human lives are at stake like medical diagnostics and autonomous transportation. It is therefore important that the algorithms produce accurate estimates of prediction uncertainty along with their predictions, so that we know how much confidence we can place on their predictions. This works aims to improve the quality of machine learning based predictions by proposing a general purpose method to produce well calibrated uncertainty estimates. The methods developed in this work can help us make better and hopefully unbiased decisions informed by machine learning methods.

Acknowledgments

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. AST-2009251. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

This work was carried out in part at "Quarks to Cosmos with AI", a conference supported by the NSF AI Institute: Physics of the Future, NSF PHY-2020295. This work used the Extreme Science and Engineering Discovery Environment (XSEDE; Towns et al. 2014), which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges-2 system, which is supported by NSF award number ACI-1928147, at the Pittsburgh Supercomputing Center (PSC).

Argonne National Laboratory's work was supported by the U.S. Department of Energy, Office of Science, under contract DE-AC02-06CH11357.

We would like to thank Michael Stanley and Andresa Campos for their help and comments on this work.

References

R. Beck, C. A. Lin, E. E. O. Ishida, F. Gieseke, R. S. de Souza, M. V. Costa-Duarte, M. W. Hattab, and A. Krone-Martins. On the realistic validation of photometric redshifts. MNRAS, 468(4): 4323–4339, July 2017. doi: 10.1093/mnras/stx687.

R. Bordoloi, S. J. Lilly, and A. Amara. Photo-z performance for precision cosmology. MNRAS, 406 (2):881–895, August 2010. doi: 10.1111/j.1365-2966.2010.16765.x.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL `http://doi.acm.org/10.1145/2939672.2939785`.

Tomas Dahlen, Bahram Mobasher, Sandra M. Faber, Henry C. Ferguson, Guillermo Barro, Steven L. Finkelstein, Kristian Finlator, Adriano Fontana, Ruth Gruetzbauch, Seth Johnson, Janine Pforr, Mara Salvato, Tommy Wiklind, Stijn Wuyts, Viviana Acquaviva, Mark E. Dickinson, Yicheng Guo, Jiasheng Huang, Kuang-Han Huang, Jeffrey A. Newman, Eric F. Bell, Christopher J. Conselice, Audrey Galametz, Eric Gawiser, Mauro Giavalisco, Norman A. Grogin, Nimish Hathi, Dale Kocevski, Anton M. Koekemoer, David C. Koo, Kyoung-Soo Lee, Elizabeth J. McGrath, Casey Papovich, Michael Peth, Russell Ryan, Rachel Somerville, Benjamin Weiner, and Grant Wilson. A Critical Assessment of Photometric Redshift Methods: A CANDELS Investigation. ApJ, 775(2): 93, October 2013. doi: 10.1088/0004-637X/775/2/93.

N. Dalmasso, T. Pospisil, A. B. Lee, R. Izbicki, P. E. Freeman, and A. I. Malz. Conditional density estimation tools in python and R with applications to photometric redshifts and likelihood-free cosmological inference. *Astronomy and Computing*, 30:100362, January 2020. doi: 10.1016/j.ascom.2019.100362.

Rafael Izbicki and Ann B. Lee. Converting High-Dimensional Regression to High-Dimensional Conditional Density Estimation. *arXiv e-prints*, art. arXiv:1704.08095, April 2017.

Wittawat Jitkrittum, Heishiro Kanagawa, and Bernhard Schölkopf. Testing goodness of fit of conditional density models with kernels. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, virtual online, August 3-6, 2020*, volume 124 of *Proceedings of Machine Learning Research*, pages 221–230. AUAI Press, 2020. URL http://proceedings.mlr.press/v124/jitkrittum20a.html.

Dritan Kodra. The galaxy morphology-density relation at high redshift with candels. PhD Thesis, January 2019. URL http://d-scholarship.pitt.edu/35716/.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2801–2809. PMLR, 2018. URL http://proceedings.mlr.press/v80/kuleshov18a.html.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In Luc De Raedt and Stefan Wrobel, editors, *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*, volume 119 of *ACM International Conference Proceeding Series*, pages 625–632. ACM, 2005. doi: 10.1145/1102351.1102430. URL https://doi.org/10.1145/1102351.1102430.

J. O. Ramsay. Monotone regression splines in action. *Statistical Science*, 3(4):425–441, 1988. ISSN 08834237. URL http://www.jstor.org/stable/2245395.

Markus Michael Rau, Stella Seitz, Fabrice Brimioulle, Eibe Frank, Oliver Friedrich, Daniel Gruen, and Ben Hoyle. Accurate photometric redshift probability density estimation - method comparison and application. MNRAS, 452(4):3710–3725, October 2015. doi: 10.1093/mnras/stv1567.

S. J. Schmidt, A. I. Malz, J. Y. H. Soo, I. A. Almosallam, M. Brescia, S. Cavuoti, J. Cohen-Tanugi, A. J. Connolly, J. DeRose, P. E. Freeman, M. L. Graham, K. G. Iyer, M. J. Jarvis, J. B. Kalmbach, E. Kovacs, A. B. Lee, G. Longo, C. B. Morrison, J. A. Newman, E. Nourbakhsh, E. Nuss, T. Pospisil, H. Tranin, R. H. Wechsler, R. Zhou, R. Izbicki, and LSST Dark Energy Science Collaboration. Evaluation of probabilistic photometric redshift estimation approaches for The Rubin Observatory Legacy Survey of Space and Time (LSST). MNRAS, 499(2):1587–1606, December 2020. doi: 10.1093/mnras/staa2799.

John Towns, Timothy Cockerill, Maytal Dahan, Ian Foster, Kelly Gaither, Andrew Grimshaw, Victor Hazlewood, Scott Lathrop, Dave Lifka, Gregory D. Peterson, Ralph Roskies, J. Ray Scott, and Nancy Wilkins-Diehr. Xsede: Accelerating scientific discovery. *Computing in Science Engineering*, 16(5):62–74, 2014. doi: 10.1109/MCSE.2014.80.

David Zhao, Niccolò Dalmasso, Rafael Izbicki, and Ann B Lee. Diagnostics for conditional density models and bayesian inference algorithms. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 125, 2021.

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]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (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? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
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)? [No] The work described is still in the early stages and some additional characterization of the performance of our methods need to be done. The code will be released after that.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (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)? [Yes]
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? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (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]