Automating Inference of Binary Microlensing Events with Neural Density Estimation

Keming Zhang*
Department of Astronomy
University of California at Berkeley
Berkeley, CA 94720
kemingz@berkeley.edu

Joshua S. Bloom
Department of Astronomy
University of California at Berkeley
Berkeley, CA 94720
joshbloom@berkeley.edu

B. Scott Gaudi
Department of Astronomy
The Ohio State University
Columbus, OH 43210
gaudi.1@osu.edu

François Lanusse
AIM, CEA, CNRS
Université Paris-Saclay
Université Paris Diderot
Sorbonne Paris Cité
F-91191 Gif-sur-Yvette, France
francois.lanusse@cea.fr

Casey Lam
Department of Astronomy
University of California at Berkeley
Berkeley, CA 94720
casey_lam@berkeley.edu

Jessica Lu
Department of Astronomy
University of California at Berkeley
Berkeley, CA 94720
jlu.astro@berkeley.edu

Abstract

Automated inference of binary microlensing events with traditional sampling-based algorithms such as MCMC has been hampered by the slowness of the physical forward model and the pathological likelihood surface. Current analysis of such events requires both expert knowledge and large-scale grid searches to locate the approximate solution as a prerequisite to MCMC posterior sampling. As the next generation, space-based [1] microlensing survey with the Roman Space Observatory [2] is expected to yield thousands of binary microlensing events [3], a new scalable and automated approach is desired. Here, we present an automated inference method based on neural density estimation (NDE). We show that the NDE[2] trained on simulated Roman data not only produces fast, accurate, and precise posteriors but also captures expected posterior degeneracies. A hybrid NDE-MCMC framework can further be applied to produce the exact posterior.

1 Introduction

As mass bends light, when the apparent trajectory of a foreground *lens* system passes close to that of a more distant *source* star, the gravitational field of the *lens* will perturb the light rays from the source resulting in a time-variable magnification. Such are gravitational microlensing events, and the time-series of brightness ("light curves") for those events are recorded by astronomical imaging surveys each night (see [4] for review). Binary microlensing events occur when the *lens* is a system

*Correspondence to: kemingz@berkeley.edu

[2]Neural density *estimator* in this context.

Third Workshop on Machine Learning and the Physical Sciences (NeurIPS 2020), Vancouver, Canada.

of two masses — either a binary star system, or a star-planet configuration. Observation of these events provide a unique opportunity for the discovery of planets as the star-to-planet mass ratio may be inferred from the light curve without having to detect light from the star-planet *lens* itself. The next-generation of microlensing survey with the Roman Space Telescope [2] (hereafter Roman) is estimated to discover thousands of binary microlensing events[3], many with planet-mass companions, over the duration of the 5-year mission span, orders of magnitude more than the dozens of events previously discovered.

While the light-curve of any single-lens event is described by a simple analytic expression ("Paczyński light-curve,") binary microlensing events require numerical forward models that are computationally expensive. In addition, binary microlensing light-curves exhibit extraordinary phenomenological diversity, owning to the different geometrical configurations for which magnification could take place. This translates to a pathological parameter space for which the likelihood surface, in the context of sampling-based Bayesian inference, suffers from a multitude of local minima which are both narrow and deep; this significantly hampers attempts to fully automate inference with sampling-based methods. As a result, binary microlensing events have thus far been analyzed on a case-by-case basis which requires both expert knowledge and expensive grid searches over some parameters, consuming weeks of CPU-hours and human effort. Therefore, there is great challenge to analyze the thousands of binary microlensing events expected to be discovered by Roman.

In this paper, we present an automated inference framework based on neural density estimation (NDE), where the fundamental task is to estimate a posterior function from pre-computed samples from the physical forward model. There has been much progress for NDE, including autoregressive models [5, 6] and flow-based models [7, 8]. While machine learning (ML) has been previously utilized to *discover* and *classify* microlensing events [9, 10, 11], our work represent the first instance for direct parameter inference. We show that the trained NDE generates accurate and precise posteriors effectively in real time, which could be further refined into the exact posterior with MCMC sampling.

2 Method

We train a conditional NDE $\hat{p}_\phi(\theta|\mathbf{x})$ to approximate the true posterior $p(\theta|\mathbf{x})$, where θ denotes physical parameters and $\mathbf{x}$ denotes the summary vector of the light curve $x \in \mathbb{R}^{\mathrm{N}}$ with N measurements. The objective is to minimize the Kullback–Leibler (KL) divergence between the two distributions:

$$\phi = \mathrm{argmin}(D_{\mathrm{KL}}(p(\theta|\mathbf{x}))||\hat{p}_\phi(\theta|\mathbf{x}))) \tag{1}$$

$$= \mathrm{argmin}(\mathbb{E}_{\theta\sim p(\theta), x\sim p(x|\theta)}[\log(p(\theta|\mathbf{x})) - \log(\hat{p}_\phi(\theta|\mathbf{x}))]) \tag{2}$$

$$= \mathrm{argmax}(\mathbb{E}_{\theta\sim p(\theta), x\sim p(x|\theta)}[\hat{p}_\phi(\theta|\mathbf{x})]) \tag{3}$$

The NDE is therefore trained with "maximum likelihood"[4] on a training set with physical parameters drawn from the prior ($p(\theta)$) and light-curves drawn from the likelihood, which is the Gaussian measurement noise model[5] on top of the noise-free microlensing light curve $f(\theta)$ (in units of detector count):

$$p(x|\theta) = \mathcal{N}(f(\theta), \sqrt{f(\theta)}). \tag{4}$$

The noise-free light-curve, in turn, is determined by the baseline *source* flux (F_{source}), the magnification time-series produced by the microlensing physical forward model ($A(\theta)$), and the constant *blend* flux, which is the flux from the *lens* star and any other star that are unresolved from the source star: $f(\theta) = A(\theta) \cdot F_{\mathrm{source}} + F_{\mathrm{blend}}$. While this methodology fits perfectly into a class of problem called likelihood-free inference (LFI), we have decided not to adapt this terminology to avoid implications that may be confusing. LFI has been developed primarily to tackle intractable or inaccessible likelihoods in the case of stochastic physical forward models [12]. In our case, the physical forward model is deterministic and the noise realization is a simple Gaussian.

We use a 20-block Masked Autoregressive Flow (MAF) [7] for $\hat{p}(\theta|\mathbf{x})$, and a ResNet-GRU network to extract features ($\mathbf{x}$) from the light curve. In short, conditioned on light-curve features, the MAF transforms a base distribution into the target distribution of the parameter posterior. Each

[3] `https://roman.gsfc.nasa.gov/exoplanets_microlensing.html`

[4] In the context of machine learning literature, the microlensing parameters θ are regarded as "data" rather than latent variables and thus the use of terminology "maximum likelihood" instead of "maximum posterior."

[5] The Poisson "photon-counting" noise is essentially Gaussian because $N_{\mathrm{photon}} \gg 1$

block of the MAF (which is a "MADE" [5]) adapts a fixed ordering of the dimensions and applies affine transformations iteratively for each dimension, subject to the autoregressive condition. We adopt random orderings for each of the 20 block to maximize network expressibility. As binary microlensing often exhibit degenerate, multi-modal solutions, we use a mixture of eight Gaussians for each dimension of the base distribution. The ResNet-GRU network is comprised of a 18-layer 1D ResNet [13] and a 2-layer GRU [14]. Each layer of the ResNet consists of two convolutions and a residual connection. A `MaxPool` layer is applied in between every two ResNet layers, where the sequence length is reduced by half and the feature dimension doubled. The output feature map is then fed to the GRU network where the output feature vector is used as the conditional input to the MAF.

3 Data

Training data is generated within the context of the Roman Space Telescope Cycle-7 design (see [3]). We simulate a dataset of 10^6 binary-lens-single-source (2L1S) magnification sequences with the microlensing code `MulensModel` [15]; each sequence contains 144 days at a cadence of 0.01 day, corresponding to the planned Roman cadence of 15 minutes [3]. These sequences are chosen to have twice the length of the 72-day Roman observation window to facilitate training with a realistic lensing occurrence times in the Roman window (see below). The magnification sequences are converted into light-curves during training on the fly by multiplying with the baseline pre-magnification *source* flux before adding the constant *blend* flux and applying noise. The ratio of the *source* flux and the constant *blend* flux is described by the source flux fraction $f_s = F_{\rm source}/(F_{\rm source} + F_{\rm blend}) < 1$. We approximate the f_s distribution of simulated Roman events in [3] with a broken power-law (see Figure 12 of their paper). For simplicity, we limit ourselves to $0.1 < f_s < 1$ and truncate the tail of relatively uncommon $0.01 < f_s < 0.1$ events. The distribution is written as $\log(f_s) \sim \sqrt{\rm Uniform(0,1)} - 1$.

Prior: Ignoring orbital motion of both the observer and the binary lens, binary microlensing (2L1S) events are described by seven parameters. Three are in common with Paczyński single-lens-single-source (1L1S) events: time of primary-lens-source closest approach (t_0), Einstein ring crossing timescale (t_E), and impact parameter relative to the lens center-of-mass (u_0). Additional four are unique to binary events: binary lens separation (s), mass ratio (q), angle of approach (α), and the finite source size (ρ). We simulate 2L1S events based on the following analytic priors:

$$\begin{gathered} t_E \sim \mathrm{TruncLogNorm}(\min = 1, \max = 100, \mu = 10^{1.15}, \sigma = 10^{0.45}) \\ u_0 \sim \mathrm{Uniform}(0, 2); \quad s \sim \mathrm{LogUniform}(0.2, 5); \quad q \sim \mathrm{Uniform}(10^{-6}, 1) \\ \alpha \sim \mathrm{Uniform}(0, 360); \quad \rho \sim \mathrm{LogUniform}(10^{-4}, 10^{-2}) \end{gathered} \tag{5}$$

During dataset simulation t_0 is fixed at mid-sequence ($t_0 = 72$) and during training, a random 72-day segment is chosen from the 144-day data; this allows the model to adapt to any 2L1S event where t_0 lies somewhere within the 72-day observation window, effectively prescribing a uniform prior on t_0. The truncated normal distribution for t_E is an approximation of a statistical analysis based on OGLE-IV data [16]. The lower limit of $q = 10^{-6}$ corresponds to the mass ratio between Mercury and a low-mass ($M \sim 0.1 M_\odot$) M-dwarf star, highlighting the superb sensitivity of Roman.

Noise: We assume an ideal Gaussian measurement noise (Equation 4) where the standard deviation of each measurement is the square root of flux measurement in raw detector counts. Studies of bulge star population shows that the apparent magnitude largely lies within the range of 20m to 25m ([3]: Figure 5). The Roman/WFIRST Cycle 7 design has the zeropoint magnitude (1 count/s) at 27.615m for the W149 filter. With exposure time at 46.8s, the aforementioned magnitude range corresponds to signal-to-noise (*S/N*) ratios between 230 and 20, which we uniformly sample during training.

Training: As previously discussed, the f_s, t_0, and base *S/N* are randomly selected for each light curve during training; this allows for large number of realizations for any single magnification sequence, which acts as data augmentation. Each light-curve is then individually scaled by their 10th percentile value as preprocessing so that the input data remains agnostic of the baseline, pre-magnification flux and the source fraction. Network optimization is performed with ADAM [17] at an initial learning rate of 0.0005 and batch size 32, which is scheduled to decay by a factor of 0.1 at the 20th and 30th epochs. Training ends at the 40th epoch. 10% data is reserved as a validation set to monitor for over-fitting. Each training epoch takes ~ 1 hour on a NVidia Titan X GPU.

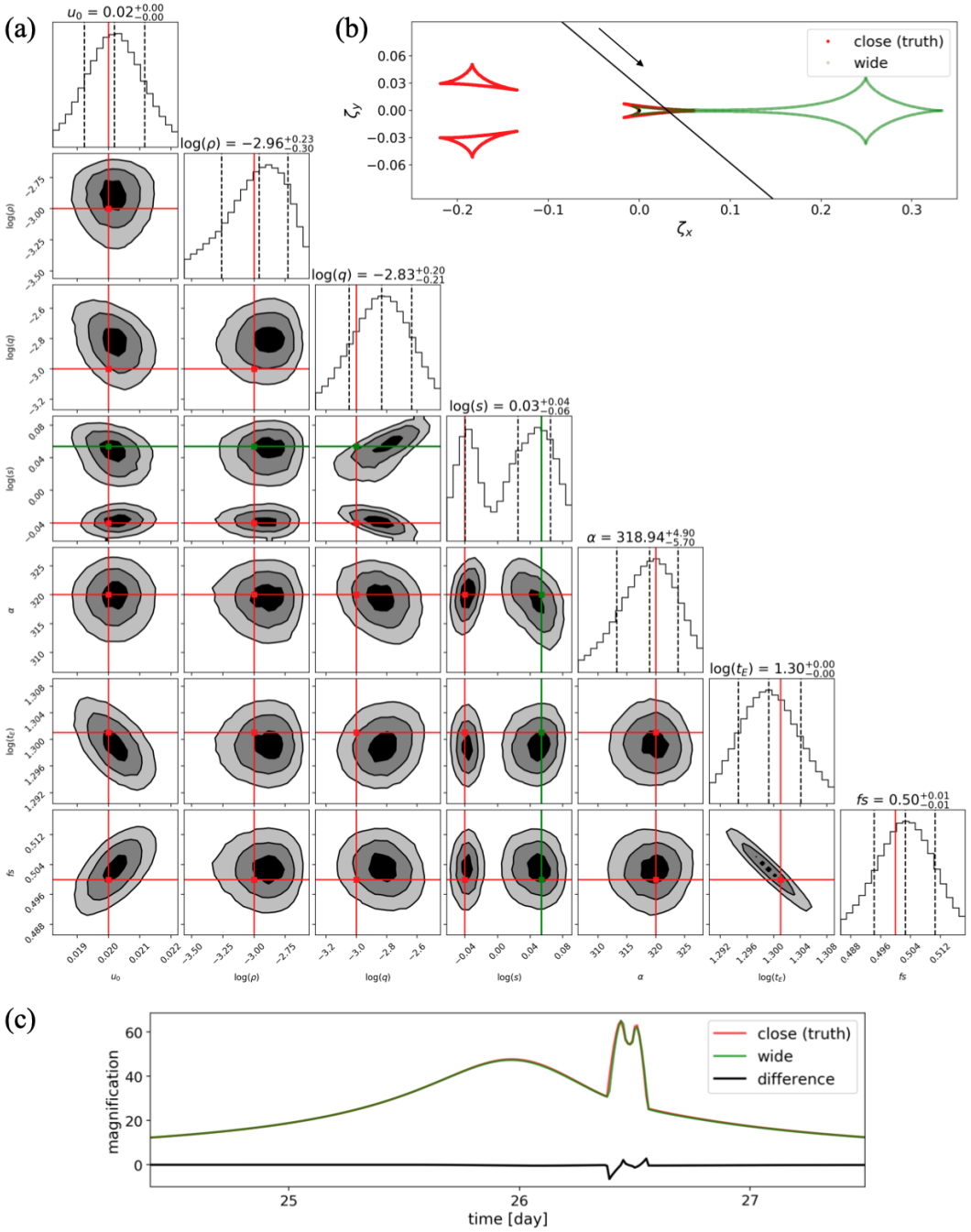

Figure 1: (a) NDE posterior for a central-caustic crossing event. The ground truth "close" solution is marked with red cross-hairs while the degenerate "wide" solution is marked with green lines. Log is base-10. (b) Caustic structure for both close and wide solutions. Arrow indicate direction of source trajectory. (c) Close-up view of the magnification curves for both "close" and "wide" solutions, which are hardly distinguishable. Error-bars would be hardly seen on the scale of the figure if shown. Caustic crossing occurs around 0.5 days after $t_0 = 26$, which is shifted 10 days away from the center of the 72-day observation window for generality.

4 Results and Discussions

The trained model is able to generate accurate and precise posteriors samples at a rate of 10^5 per second on one GPU, effectively in real-time. This compares to the ~ 1 per second simulation speed of the forward model `MulensModel` on one CPU core. Figure 1a shows the NDE posterior for an example event which exhibits a classic "close-wide" degeneracy. The NDE posterior tightly constrains all seven parameters including the finite source effect (ρ), although the light curve realization is at the nosiest level seen during training ($S/N_{base} = 20$). In addition, the expected covariances, especially among u_0, t_E, and f_s, and between s and q, are as expected. The close-wide degeneracy is exhibited by the bimodal distribution in s-space (close: $s < 1$, wide: $s > 1$). The degenerate, wide solution ($s = 10^{0.055}$; all else equal) as well as its caustic[6] structure and magnification curve are shown in green in Figure 1.

The precision of a posterior sample is determined by two kinds of uncertainty: data uncertainty and model uncertainty. As neural networks in practice are not infinitely expressive, in the limit of the highest-quality data, model uncertainty is expected to dominate over data uncertainty. This is the case for Roman data. Indeed, by increasing the baseline *S/N* from 20 to 200, we do not see significant improvement in the precision of the NDE posterior. Applied to much noisier and more sparsely sampled ground-based data we expect that data uncertainties will dominate over model uncertainties, thus allowing the NDE posterior to converge towards the exact posterior. To obtain the exact posterior from the NDE posterior, a hybrid NDE-MCMC framework is used where samples from the NDE posterior are used to initialize MCMC chains. For the current example, we evaluated the likelihood of 800 NDE posterior samples and used the top 16 to seed 16 MCMC chains, which allowed for MCMC burn-in in $\sim$thousand steps. The performance of the NDE and the NDE-MCMC hybrid framework will be systematically analyzed in future work.

Broader impact

The dissemination of trained NDE models can allow researchers with less access to expensive computational facilities to more easily draw their own inferences on public data. Development of automated inference techniques, such as presented herein, has a learning curve that is much less steep than traditional sample-based inference which requires domain expertise and vast experience. Therefore, such work may reduce the barrier to entry and allow for a wider engagement from outside the microlensing community.

Acknowledgement

K.Z. and J.S.B. are supported by a Gordon and Betty Moore Foundation Data-Driven Discovery grant. J.S.B. is partially sponsored by a faculty research award from Two Sigma. K.Z. thanks the LSSTC Data Science Fellowship Program, which is funded by LSSTC, NSF Cybertraining Grant 1829740, the Brinson Foundation, and the Moore Foundation; his participation in the program has benefited this work. B.S.G. is supported by NASA grant NNG16PJ32C and the Thomas Jefferson Chair for Discovery and Space Exploration. This work is supported by the AWS Cloud Credits for Research program.

References

[1] David P. Bennett and Sun Hong Rhie. Simulation of a Space-based Microlensing Survey for Terrestrial Extrasolar Planets. *The Astrophysical Journal*, 574(2):985, August 2002. ISSN 0004-637X. doi: 10.1086/340977. URL `https://iopscience.iop.org/article/10.1086/340977/meta`. Publisher: IOP Publishing.

[2] D. Spergel, N. Gehrels, C. Baltay, D. Bennett, J. Breckinridge, M. Donahue, A. Dressler, B. S. Gaudi, T. Greene, O. Guyon, C. Hirata, J. Kalirai, N. J. Kasdin, B. Macintosh, W. Moos, S. Perlmutter, M. Postman, B. Rauscher, J. Rhodes, Y. Wang, D. Weinberg, D. Benford, M. Hudson, W.-S. Jeong, Y. Mellier, W. Traub, T. Yamada, P. Capak, J. Colbert, D. Masters, M. Penny, D. Savransky, D. Stern, N. Zimmerman, R. Barry, L. Bartusek, K. Carpenter,

[6] Points of infinite magnification. See: https://microlensing-source.org/concept/caustics-overview/

E. Cheng, D. Content, F. Dekens, R. Demers, K. Grady, C. Jackson, G. Kuan, J. Kruk, M. Melton, B. Nemati, B. Parvin, I. Poberezhskiy, C. Peddie, J. Ruffa, J. K. Wallace, A. Whipple, E. Wollack, and F. Zhao. Wide-Field InfrarRed Survey Telescope-Astrophysics Focused Telescope Assets WFIRST-AFTA 2015 Report. *arXiv:1503.03757 [astro-ph]*, March 2015. URL http://arxiv.org/abs/1503.03757. arXiv: 1503.03757.

[3] Matthew T. Penny, B. Scott Gaudi, Eamonn Kerins, Nicholas J. Rattenbury, Shude Mao, Annie C. Robin, and Sebastiano Calchi Novati. Predictions of the *WFIRST* Microlensing Survey. I. Bound Planet Detection Rates. *The Astrophysical Journal Supplement Series*, 241 (1):3, February 2019. ISSN 1538-4365. doi: 10.3847/1538-4365/aafb69. URL https://iopscience.iop.org/article/10.3847/1538-4365/aafb69.

[4] B. Scott Gaudi. Microlensing Surveys for Exoplanets. *Annual Review of Astronomy and Astrophysics*, 50(1):411–453, September 2012. ISSN 0066-4146, 1545-4282. doi: 10.1146/annurev-astro-081811-125518. URL http://www.annualreviews.org/doi/10.1146/annurev-astro-081811-125518.

[5] Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. MADE: Masked Autoencoder for Distribution Estimation. In *International Conference on Machine Learning*, pages 881–889. PMLR, June 2015. URL http://proceedings.mlr.press/v37/germain15.html. ISSN: 1938-7228.

[6] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. *arXiv:1609.03499 [cs]*, September 2016. URL http://arxiv.org/abs/1609.03499. arXiv: 1609.03499.

[7] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked Autoregressive Flow for Density Estimation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2338–2347. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/6828-masked-autoregressive-flow-for-density-estimation.pdf.

[8] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. 2017. URL https://arxiv.org/abs/1605.08803.

[9] Lukasz Wyrzykowski, Alicja E. Rynkiewicz, Jan Skowron, Szymon Kozlowski, Andrzej Udalski, Michal K. Szymański, Marcin Kubiak, Igor Soszyński, Grzegorz Pietrzyński, Radoslaw Poleski, Pawel Pietrukowicz, and Michal Pawlak. OGLE-III MICROLENSING EVENTS AND THE STRUCTURE OF THE GALACTIC BULGE. *The Astrophysical Journal Supplement Series*, 216(1):12, January 2015. ISSN 0067-0049. doi: 10.1088/0067-0049/216/1/12. URL https://doi.org/10.1088%2F0067-0049%2F216%2F1%2F12. Publisher: IOP Publishing.

[10] D. Godines, E. Bachelet, G. Narayan, and R. A. Street. A machine learning classifier for microlensing in wide-field surveys. *Astronomy and Computing*, 28:100298, July 2019. ISSN 2213-1337. doi: 10.1016/j.ascom.2019.100298. URL http://www.sciencedirect.com/science/article/pii/S2213133719300113.

[11] Przemek Mróz. Identifying microlensing events using neural networks. *arXiv:2008.11930 [astro-ph]*, August 2020. URL http://arxiv.org/abs/2008.11930. arXiv: 2008.11930.

[12] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, May 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1912789117. URL https://www.pnas.org/content/early/2020/05/28/1912789117. Publisher: National Academy of Sciences Section: Physical Sciences.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. doi: 10.1109/CVPR.2016.90. ISSN: 1063-6919.

[14] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://www.aclweb.org/anthology/D14-1179.

[15] R. Poleski and J. C. Yee. Modeling microlensing events with MulensModel. *Astronomy and Computing*, 26:35–49, January 2019. ISSN 2213-1337. doi: 10.1016/j.ascom.2018.11.001. URL http://www.sciencedirect.com/science/article/pii/S221313371830026X.

[16] Przemek Mróz, Andrzej Udalski, Jan Skowron, Radosław Poleski, Szymon Kozłowski, Michał K. Szymański, Igor Soszyński, Łukasz Wyrzykowski, Paweł Pietrukowicz, Krzysztof Ulaczyk, Dorota Skowron, and Michał Pawlak. No large population of unbound or wide-orbit Jupiter-mass planets. *Nature*, 548(7666):183–186, August 2017. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature23276. URL http://www.nature.com/articles/nature23276.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *ICLR*, 2015. URL http://arxiv.org/abs/1412.6980. arXiv: 1412.6980.