Time Delay Cosmography with a Neural Ratio Estimator

Ève Campeau-Poirier [1 2 3] **Laurence Perreault-Levasseur** [1 2 3 4 5] **Adam Coogan** [1 2 3] **Yashar Hezaveh** [1 2 3 4 5]

Abstract

We explore the use of a Neural Ratio Estimator (NRE) to determine the Hubble constant (H_0) in the context of time delay cosmography. Assuming a Singular Isothermal Ellipsoid (SIE) mass profile for the deflector, we simulate time delay measurements, image position measurements, and modeled lensing parameters. We train the NRE to output the posterior distribution of H_0 given the time delay measurements, the relative Fermat potentials (calculated from the modeled parameters and the measured image positions), the deflector redshift, and the source redshift. We compare the accuracy and precision of the NRE with traditional explicit likelihood methods in the limit where the latter is tractable and reliable, using Gaussian noise to emulate measurement uncertainties in the input parameters. The NRE posteriors track the ones from the conventional method and, while they show a slight tendency to overestimate uncertainties, they can be combined in a population inference without bias.

1. Introduction

Over the past decades, the inflationary ΛCDM model has had striking success in explaining cosmic microwave background (CMB) observations and the detailed evolution of the Universe. The current expansion rate of the Universe, known as the Hubble constant (H_0), is essential for many studies, including understanding the nature of dark energy, neutrino physics, and testing general relativity. In the past decade, the measured values of H_0 from different probes have diverged: the latest CMB and Type Ia supernovae data now disagree at more than 4σ (Riess et al., 2022).

[1]Department of Physics, Université de Montréal, Montréal, Canada [2]Ciela, Montréal, Canada [3]Mila, Montréal, Canada [4]Flatiron Institue, New York, USA [5]Perimeter Institute for Theoretical Physics, Waterloo, Ontario, Canada. Correspondence to: Ève Campeau-Poirier <eve.campeau-poirier@umontreal.ca>.

ICML 2023 Workshop on Machine Learning for Astrophysics, Honolulu, Hawaii, USA. PMLR 202, 2023.

Time delay cosmography can provide an independent measurement of H_0 with different systematics from existing methods. This can be done using the time delays between the multiple images of a strongly lensed variable light source. Previous measurements have achieved a precision between 2% and 8% (Birrer & Treu, 2021) using this method. Meanwhile, 1% precision is required to solve the Hubble tension (Weinberg et al., 2013; Treu et al., 2022). This could be achieved with data available in the next decade with a new generation of survey telescopes. The Rubin Observatory, in particular, is expected to detect thousands of strongly lensed quasars (Oguri & Marshall, 2010).

However, current analysis methods have limitations in terms of complexity and scalability. They rely on likelihood-based approaches, such as Markov Chain Monte Carlo (MCMC) and nested sampling, which require explicit likelihoods and are not amortized. They also require sampling joint posterior distributions of nuisance parameters while only the H_0 marginal is of interest. Hence, they scale poorly as nuisance parameters are included to ensure unbiasied inference.

The simulation-based inference (SBI) framework allows handling complex, high-dimensional data and models that are difficult or intractable to analyze using traditional likelihood-based methods by only relying on the availability of a realistic simulation pipeline. Neural Ratio Estimators (NREs; Cranmer et al. 2015), a specific class of SBI methods, leverage the power of machine learning to allow amortization of the inference process as well as implicit marginalization over large sets of nuisance parameters, providing an efficient way to estimate low-dimensional variables.

We demonstrate the application of an NRE to time delay cosmography by predicting the H_0 posterior distribution given Fermat potentials calculated from modeled lens parameters and image positions, the time delay measurements, and the deflector and source redshifts. We use a Set Transformer architecture (Lee et al., 2019), which allows for the amortization over lensing systems with two or four lensed images by the same model.

While previous works have explored how machine learning can be used for the measurement of H_0 with time-delay cosmography, contributions (e.g. Hezaveh et al. 2017; Levasseur et al. 2017; Morningstar et al. 2019; Pearson et al. 2019; Wagner-Carena et al. 2021; Schuldt et al. 2021; Park

et al. 2021) have been limited to using neural networks (NN) to estimate the lens parameters posterior. The approach presented here is therefore complementary, since it bridges the remaining gap to fully amortize the inference of H_0 from strong lensing data.

Section 2 introduces the methodology. Section 3 describes the simulations. Section 4 presents the NN architecture and training procedure. Results are presented in section 5.

2. Time-delay cosmography

Gravitational lensing occurs when images from a distant source get distorted by the presence of matter bending spacetime along the line of sight. In strong gravitational lensing, there is formation of multiple images of background sources due to this effect. The lensing equation,

$$\boldsymbol{\beta} = \boldsymbol{\theta} - \boldsymbol{\alpha}\left(\boldsymbol{\theta}\right) , \tag{1}$$

summarizes this phenomenon by retracing the source plane angular position $\boldsymbol{\beta}$ of a ray observed at the image plane angular position $\boldsymbol{\theta}$ after a mass deflector has deviated it by an angle $\boldsymbol{\alpha}$. The lensing potential ψ of the massive object determines the angular deflection $\boldsymbol{\alpha}$ and the convergence κ according to

$$\boldsymbol{\alpha}\left(\boldsymbol{\theta}\right) = \boldsymbol{\nabla}\psi\left(\boldsymbol{\theta}\right) ; \quad \nabla^2\psi\left(\boldsymbol{\theta}\right) = 2\kappa\left(\boldsymbol{\theta}\right) . \tag{2}$$

Gravitational lensing affects the light rays travel time from their source to the observer in two ways : by changing their path length and through the lensing potential itself. The presence of a mass deflector in the light's trajectory lengthens its travel time by an amount proportional to the Fermat potential ϕ, which is fully determined by the mass distribution in the lens and is given by

$$\phi\left(\boldsymbol{\theta}, \boldsymbol{\beta}\right) \equiv \frac{\left(\boldsymbol{\theta} - \boldsymbol{\beta}\right)^2}{2} - \psi\left(\boldsymbol{\theta}\right) . \tag{3}$$

To infer H_0 with time delay cosmography, one observes a multiply-imaged time-varying background source. Each path giving rise to each image is affected by a different Fermat potential, resulting in a different light travel time. This allows the evaluation of the relative travel times between paths Δt, which are called time delays. They are calculated between pairs of images. They are related to H_0 by

$$\Delta t \equiv \frac{D_{\Delta t}}{c} \Delta\phi , \tag{4}$$

where c is the speed of light, $\Delta\phi$ is the difference of Fermat potential at the position of the two distinct images, and $D_{\Delta t}$ is the time delay distance, given by

$$D_{\Delta t} \equiv \left(1 + z_d\right) \frac{D_d D_s}{D_{ds}} . \tag{5}$$

Here, z_d is the deflector redshift, D_d is the diameter angular distance between the observer and the deflector, D_s is the diameter angular distance between the observer and the source, D_{ds} is the diameter angular distance between the deflector and the source. These distances are where the H_0 dependence is contained.

In this framework, the posterior distribution of H_0 generally takes the form

$$P(H_0|\boldsymbol{\Delta t}, \boldsymbol{d}) \propto \int \mathrm{d}\boldsymbol{\zeta}\, P(\boldsymbol{\Delta t}|H_0, \boldsymbol{\zeta}, \boldsymbol{M}) P(\boldsymbol{\zeta}|\boldsymbol{d}, \boldsymbol{M}) P(H_0) \tag{6}$$

where $\boldsymbol{d}$ represents the lensing observation, $\boldsymbol{\zeta}$ is a set of parameters describing the lensing system, and $\boldsymbol{M}$ includes all observational effects (e.g. instrumental noise, point spread function, image covariance matrix, deflector's light, and dust). In this context, the lensing parameters and the observational effects are nuisance parameters that must be integrated out to obtain the marginal distribution of H_0. The main proposal of this work is to replace the traditional Monte Carlo methods to numerically approximate the H_0 posterior.

3. Simulations

In this work, we consider the case where the deflected light is emitted by a variable point source, such as an Active Galactic Nucleus (AGN) or a supernova. We do not consider any light profile for its host galaxy because in the following we assume that the modeling of the lensed image was performed in a previous analysis stage (e.g. with a BNN as in Park et al. 2021). We assume that the source is being distorted by a deflector following as Singular Isothermal Ellipsoid (SIE; Kormann et al. 1994), plus external shear. This model is described by 7 parameters: Einstein radius θ_E, $x-$ and $y-$components of the position (x_d, y_d), axis ratio f and its orientation ϕ_d, and modulus $\gamma_{\rm ext}$ and orientation $\phi_{\rm ext}$ of the external shear. Details about the range of uniform prior used for these parameters, the cosmology, and the variable source are included in Table 1.

We compute time delay distances according to Equation (5). The H_0 value, the source and the deflector redshifts are drawn from uniform prior distributions detailed in Table 1. We assume a flat ΛCDM cosmology. With the Fermat potential at the image positions and the time delay distance, we calculate the time delays from Equation (4) and relative Fermat potentials from Equation (3), meaning that doubles have one time delay - Fermat potential pair, while quads have three.

For the noise model, the goal is to emulate the results of a standard analysis, which models the system parameters from the lensing observation and measures the time delays from the image light curves. Therefore, we add Gaussian

Table 1. Prior distributions of all the parameters needed to generate Fermat potentials and time delays in our framework

Parameter	Distribution
Cosmology	
Hubble constant (km s^{-1} Mpc^{-1})	$H_0 \sim \mathrm{U}(50, 90)$
Dark energy density	$\Omega_\Lambda = 0.7$
Matter energy density	$\Omega_m = 0.3$
Deflector	
Redshift	$z_d \sim \mathrm{U}(0.04, 0.5)$
Position ($''$)	$x_d, y_d \sim \mathrm{U}(-0.3, 0.3)$
Einstein radius ($''$)	$\theta_E \sim \mathrm{U}(0.5, 2.0)$
Axis ratio	$f \sim \mathrm{U}(0.30, 0.99)$
Orientation (rad)	$\varphi_d \sim \mathrm{U}(-\pi/2, \pi/2)$
External Shear	
Modulus	$\gamma_{\text{ext}} \sim \mathrm{U}(0, 0.2)$
Orientation (rad)	$\varphi_{\text{ext}} \sim \mathrm{U}(-\pi/2, \pi/2)$
Variable point light source	
Redshift	$z_s \sim \mathrm{U}(1, 3)$
Position ($''$)	$x_s, y_s = (0, 0)$

noise to the lensing parameters, the image positions and the source position. As standard deviations, we use each parameter's average error from the BNN in Park et al. (2021). From those noisy estimates, we compute the Fermat potentials. For the time delays, we add Gaussian noise to the ones generated with the true parameters. This replicates the uncertainty yielded by the light curve measurements, as well as the mass-sheet degeneracy (Park et al., 2021). Table 2 summarizes all the standard deviations of the Gaussian noise distributions.

Table 2. Standard deviation of the Gaussian noise distributions used to mimic the uncertainties of lens modeling, time delay measurements, and image position measurements

Observables	Noise standard deviation
Time delays (days)	0.35
Image positions ($''$)	0.001
Deflector	
Position ($''$)	0.005
Einstein radius ($''$)	0.011
Ellipticities	0.039
External shear	
Components	0.02
Active galactic nucleus	
Position ($''$)	0.012

4. Methods

4.1. Neural Ratio Estimation

In this work, we train a Neural Ratio Estimator to learn the posterior distribution of H_0. At its core, a NRE learns the ratio between two distributions of the parameters of interest $\mathbf{\Theta}$ (in our case H_0), and the simulated observations x: the joint distribution $p(\mathbf{x}, \mathbf{\Theta})$, which we can sample using our simulator, and the product of the marginals $p(\mathbf{x})\, p(\mathbf{\Theta})$, which we can sample by pairing randomly simulations and parameters sampled from the prior.

Assigning the class label $y = 1$ to the joint distribution and the class label $y = 0$ to the product of the marginals, the optimal discriminator $\mathbf{d}^*$ that classifies samples from these two distributions converges to the decision function

$$\mathbf{d}^*(\mathbf{x}, \mathbf{\Theta}) = p(y = 1 \mid \mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{\Theta})}{p(\mathbf{x}, \mathbf{\Theta}) + p(\mathbf{x})\, p(\mathbf{\Theta})} \quad (7)$$

The ratio $r(\mathbf{x} \mid \mathbf{\Theta})$ between the distributions can be written as a function of the discriminator :

$$r(\mathbf{x} \mid \mathbf{\Theta}) \equiv \frac{p(\mathbf{x}, \mathbf{\Theta})}{p(\mathbf{x})\, p(\mathbf{\Theta})} = \frac{\mathbf{d}^*(\mathbf{x}, \mathbf{\Theta})}{1 - \mathbf{d}^*(\mathbf{x}, \mathbf{\Theta})} \quad (8)$$

The product between the estimator of r learnt by the NRE, $\hat{r}(\mathbf{x} \mid \mathbf{\Theta})$, and the prior distribution yields a posterior distribution estimator. To conduct an inference with a trained Neural Ratio Estimator, the estimator $\hat{r}(\mathbf{x} \mid \mathbf{\Theta})$ is calculated multiple times for the same observation, but with different parameter values at each computation.

4.2. Set Transformer Architecture

For the architecture of the discriminator, we use a Set Transformer (Lee et al., 2019) to make use of the fact that different lensing configurations (doubles or quads) can have different number of time delay-relative Fermat potential pairs, and that those pairs are permutation invariant. We also explored Deep Sets (Zaheer et al., 2017), however in our experiments they were outperformed by the Set Transformer, and so we only report on the latter.

The NRE takes as inputs the measured time delays, the modeled relative Fermat potentials, a H_0 value, the source's redshift, and the deflector's redshift. See Appendix A Figure 3 for the specific details of the architecture.

4.3. Training

The training set, the validation set, and the test set contain 1,280,000 examples, 160,000 examples and 26,500 examples, respectively. The dataset is composed of approximately 45% doubles and 55% quads. We train the neural network on batches of 1,000 examples with a binary cross entropy loss as the objective function. At each batch, we draw a new realization of noise for the time delays, the parameters, the

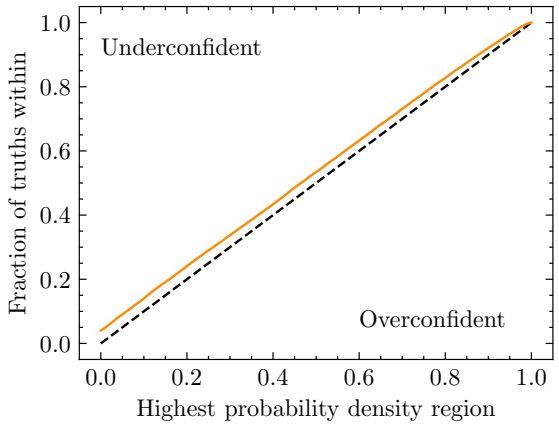

Figure 1. Coverage diagnostic of the NRE. A perfectly consistent distribution would fall on the dashed line. An underconfident distribution would lay on the top-left area, while an overconfident distribution would be in the bottom-right region. The NRE coverage, represented by the orange solid line, indicates a weak underconfident behaviour.

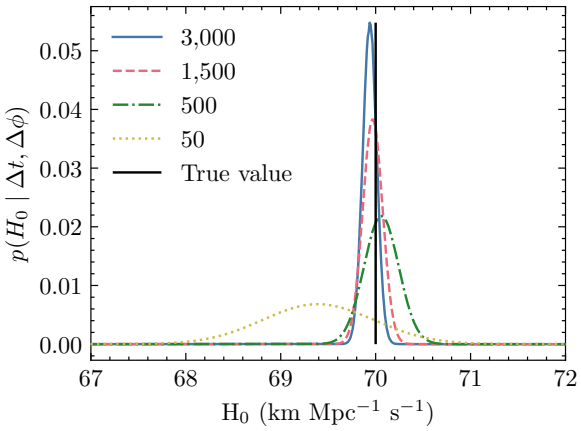

Figure 2. Population inferences of H_0 with the NRE. The blue solid line, the pink dashed line, the green dashed-dotted line, and the yellow dotted line represent populations of 3,000, 1,500, 500 and 50 lensing systems, respectively. The true value $H_0 = 70$ km s^{-1} Mpc^{-1} is indicated by the vertical black solid line. It falls inside the 2σ interval for all populations.

image positions, and the source position. We then compute the Fermat potentials. The training lasts for 5,000 epochs. The learning rate starts at 1×10^{-4}, and decreases by half every 500 epochs, as it was the optimal schedule we found through a hyperparameter search.

5. Results and Discussion

In our framework, the general posterior in Equation (6) takes the specific form

$$P(H_0|\Delta t, \Delta\phi, z_d, z_s) \propto P(\Delta t|H_0, \Delta\phi, z_d, z_s)P(\Delta\phi)P(H_0) \tag{9}$$

$$P(\Delta\phi) \propto \int \mathrm{d}\boldsymbol{\zeta}\, P(\Delta\phi|\boldsymbol{\zeta})P(\boldsymbol{\zeta}) \tag{10}$$

where $P(\Delta t|H_0, \Delta\phi)$ and $P(\boldsymbol{\zeta})$ are normal distributions, $P(\Delta\phi|\boldsymbol{\zeta})$ is a delta function, and $P(H_0)$ is a uniform distribution. We sample this posterior with PolyChord (Handley et al., 2015a;b) and find agreement with the NRE posteriors, as shown in some representative examples in Appendix B. To assess the NRE's accuracy, we perform a coverage test (Hermans et al., 2021; Cole et al., 2022) using the highest posterior density (HPD) interval of the NRE on the noisy examples from the test set. Results are displayed in Figure 1. The NRE shows a slightly underconfident behaviour, which is preferable to overconfidence.

Moreover, the NRE offers a significant improvement in the analysis speed. With PolyChord, the posterior sampling process requires from 20 to 40 minutes on a CPU, and is not amortized. By contrast, once trained, the NRE only requires $\sim$1 second to estimate the posterior of H_0 for a given lens, making the analysis more than 1000 times faster.

We perform a population inference of H_0. We simulate noisy data from multiple lensing systems (doubles and quads), fixing $H_0 = 70$ km s^{-1} Mpc^{-1}. Figure 2 shows the population inferences of 3,000, 1,500, 500 and 50 lensing systems. The NRE appears unbiased because all posteriors enclose the truth in their 2σ interval.

One of the main advantages of a simulation-based approach such as the NRE over traditional maximum-likelihood methods is that it implicitly marginalizes over nuisance parameters (Hermans et al., 2019). This is because, even though the simulator samples all parameters to generate the mock data, the classes and the loss function are independent of the nuisance parameters. While here our simulations remained simple, including further nuisance parameters in the inference is now reduced to simulating them.

Another important advantage of SBI methods is that they do not require any assumption about the form of the posterior. The complexity of the posterior is only limited by the simulations themselves, which can include complex environment, noise, selection effects, etc. In contrast, traditional explicit-likelihood methods require an analytical form for both the prior and the likelihood to compute the posterior distribution. These often imply simplistic priors, and simplifying assumptions about the model's parametrization, which can introduce biases in the inference.

A notable source of bias is the mass sheet degeneracy (Falco et al., 1985). In this paper, we do not consider explicitly the mass sheet degeneracy. However, we chose the noise distributions so that the uncertainty on H_0 could reach 8% frequently, which is the error budget estimated by (Birrer & Treu, 2021) when accounting for the mass sheet degeneracy.

6. Conclusion

In this work, we used an NRE to infer H_0 from the time delays, the relative Fermat potentials, and the source and deflector redshifts of strong lensing systems. This work bridges the gap to completely amortize the inference of H_0 from time delay cosmography, bringing down the inference time by a factor of more than 1000 from at least 20 minutes with POLYCHORD to about 1 second per lens. Moreover, combining measurements from a population of 3,000 lenses suggests that our estimator is unbiased.

We assumed that the parameters describing the deflector could be estimated with a precision similar to that of BNNs published in the literature (Park et al., 2021). To improve this work, more complex simulations incorporating environmental effects, such as the mass sheet degeneracy, as well as more inputs to break it, like velocity dispersion measurenments, could be used to train the NRE.

We expect the NRE to fully leverage the upcoming large datasets of strong lensing observations to reach the 1% precision needed to solve the Hubble tension. Its implicit marginalization over nuisance parameters can take into account as many possible biases as can be simulated, while guaranteeing the accuracy of the inference.

Acknowledgments

This work was in part supported by Schmidt Futures, a philanthropic initiative founded by Eric and Wendy Schmidt as part of the Virtual Institute for Astrophysics (VIA). The work was in part supported by computational resources provided by Calcul Québec and the Digital Research Alliance of Canada. Y.H. and L.P.L. acknowledge support from the Canada Research Chairs Program, the National Sciences and Engineering Council of Canada through grants RGPIN-2020-05073 and 05102, and the Fonds de recherche du Québec through grants 2022-NC-301305 and 300397.

References

Birrer, S. and Treu, T. Tdcosmo - v. strategies for precise and accurate measurements of the hubble constant with strong lensing. *Astronomy & Astrophysics*, 649:A61, 2021. doi: 10.1051/0004-6361/202039179. URL https://doi.org/10.1051/0004-6361/202039179.

Cole, A., Miller, B. K., Witte, S. J., Cai, M. X., Grootes, M. W., Nattino, F., and Weniger, C. Fast and credible likelihood-free cosmology with truncated marginal neural ratio estimation. *Journal of Cosmology and Astroparticle Physics*, 2022(09):004, sep 2022. doi: 10.1088/1475-7516/2022/09/004. URL https://dx.doi.org/10.1088/1475-7516/2022/09/004.

Cranmer, K., Pavez, J., and Louppe, G. Approximating Likelihood Ratios with Calibrated Discriminative Classifiers. *arXiv e-prints*, art. arXiv:1506.02169, June 2015.

Falco, E. E., Gorenstein, M. V., and Shapiro, I. I. On model-dependent bounds on H(0) from gravitational images : application to Q 0957+561 A, B. *Astrophysical Journal Letters*, 289:L1–L4, February 1985. doi: 10.1086/184422.

Handley, W. J., Hobson, M. P., and Lasenby, A. N. polychord: nested sampling for cosmology. *Monthly Notices of the Royal Astronomical Society: Letters*, 450 (1):L61–L65, apr 2015a. doi: 10.1093/mnrasl/slv047. URL https://doi.org/10.1093%2Fmnrasl%2Fslv047.

Handley, W. J., Hobson, M. P., and Lasenby, A. N. polychord: next-generation nested sampling. *Monthly Notices of the Royal Astronomical Society*, 453(4):4385–4399, sep 2015b. doi: 10.1093/mnras/stv1911. URL https://doi.org/10.1093%2Fmnras%2Fstv1911.

Hermans, J., Begy, V., and Louppe, G. Likelihood-free mcmc with amortized approximate ratio estimators, 2019. URL https://arxiv.org/abs/1903.04057.

Hermans, J., Delaunoy, A., Rozet, F., Wehenkel, A., Begy, V., and Louppe, G. A trust crisis in simulation-based inference? your posterior approximations can be unfaithful, 2021. URL https://arxiv.org/abs/2110.06581.

Hezaveh, Y. D., Levasseur, L. P., and Marshall, P. J. Fast automated analysis of strong gravitational lenses with convolutional neural networks. *Nature*, 548(7669):555–557, aug 2017. doi: 10.1038/nature23463. URL https://doi.org/10.1038%2Fnature23463.

Kormann, R., Schneider, P., and Bartelmann, M. Isothermal elliptical gravitational lens models. , 284:285–299, April 1994.

Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., and Teh, Y. W. Set transformer: A framework for attention-based permutation-invariant neural networks. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Levasseur, L. P., Hezaveh, Y. D., and Wechsler, R. H. Uncertainties in parameters estimated with neural networks: Application to strong gravitational lensing. *The Astrophysical Journal*, 850(1):L7, nov 2017. doi: 10.3847/2041-8213/aa9704. URL https://doi.org/10.3847%2F2041-8213%2Faa9704.

Morningstar, W. R., Levasseur, L. P., Hezaveh, Y. D., Blandford, R., Marshall, P., Putzky, P., Rueter, T. D., Wechsler, R., and Welling, M. Data-driven reconstruction of gravitationally lensed galaxies using recurrent inference machines. *The Astrophysical Journal*, 883(1):14, sep 2019. doi: 10.3847/1538-4357/ab35d7. URL https://dx.doi.org/10.3847/1538-4357/ab35d7.

Oguri, M. and Marshall, P. J. Gravitationally lensed quasars and supernovae in future wide-field optical imaging surveys. *Monthly Notices of the Royal Astronomical Society*, pp. no–no, apr 2010. doi: 10.1111/j.1365-2966.2010.16639.x. URL https://doi.org/10.1111%2Fj.1365-2966.2010.16639.x.

Park, J. W., Wagner-Carena, S., Birrer, S., Marshall, P. J., Lin, J. Y.-Y., and Roodman, A. Large-scale gravitational lens modeling with bayesian neural networks for accurate and precise inference of the hubble constant. *The Astrophysical Journal*, 910(1):39, mar 2021. doi: 10.3847/1538-4357/abdfc4. URL https://doi.org/10.3847%2F1538-4357%2Fabdfc4.

Pearson, J., Li, N., and Dye, S. The use of convolutional neural networks for modelling large optically-selected strong galaxy-lens samples. *Monthly Notices of the Royal Astronomical Society*, 488(1):991–1004, 06 2019. ISSN 0035-8711. doi: 10.1093/mnras/stz1750. URL https://doi.org/10.1093/mnras/stz1750.

Riess, A. G., Yuan, W., Macri, L. M., Scolnic, D., Brout, D., Casertano, S., Jones, D. O., Murakami, Y., Anand, G. S., Breuval, L., Brink, T. G., Filippenko, A. V., Hoffmann, S., Jha, S. W., D'arcy Kenworthy, W., Mackenty, J., Stahl, B. E., and Zheng, W. A Comprehensive Measurement of the Local Value of the Hubble Constant with 1 km s $^{-1}$ Mpc $^{-1}$ Uncertainty from the Hubble Space Telescope and the SH0ES Team. *The Astrophysical Journal Letters*, 934(1):L7, July 2022. ISSN 2041-8205, 2041-8213. doi: 10.3847/2041-8213/ac5c5b. URL https://iopscience.iop.org/article/10.3847/2041-8213/ac5c5b.

Schuldt, S., Suyu, S. H., Meinhardt, T., Leal-Taixé, L., Cañameras, R., Taubenberger, S., and Halkola, A. Holismokes - iv. efficient mass modeling of strong lenses through deep learning. *Astronomy & Astrophysics*, 646:A126, 2021. doi: 10.1051/0004-6361/202039574. URL https://doi.org/10.1051/0004-6361/202039574.

Treu, T., Suyu, S. H., and Marshall, P. J. Strong lensing time-delay cosmography in the 2020s. *The Astronomy and Astrophysics Review*, 30(1):8, November 2022. ISSN 1432-0754. doi: 10.1007/s00159-022-00145-y. URL https://doi.org/10.1007/s00159-022-00145-y.

Wagner-Carena, S., Park, J. W., Birrer, S., Marshall, P. J., Roodman, A., Wechsler, R. H., and Collaboration), L. D. E. S. Hierarchical inference with bayesian neural networks: An application to strong gravitational lensing. *The Astrophysical Journal*, 909(2):187, mar 2021. doi: 10.3847/1538-4357/abdf59. URL https://dx.doi.org/10.3847/1538-4357/abdf59.

Weinberg, D. H., Mortonson, M. J., Eisenstein, D. J., Hirata, C., Riess, A. G., and Rozo, E. Observational probes of cosmic acceleration. *Physics Reports*, 530(2):87–255, 2013. ISSN 0370-1573. doi: https://doi.org/10.1016/j.physrep.2013.05.001. URL https://www.sciencedirect.com/science/article/pii/S0370157313001592. Observational Probes of Cosmic Acceleration.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R., and Smola, A. Deep sets, 2017. URL https://arxiv.org/abs/1703.06114.

A. Neural network variable dimensions

Figure 3 and Table 3 illustrate our Set Transformer architecture. The first self-attention block computes multi-head attention between the time delay - relative Fermat potential pairs belonging to the same lensing system. The second self-attention block repeats the operation with the output of the first one. After, the features are aggregated by computing multi-head attention between a learnable seed vector and them. At each step, we use 6 attention heads of dimension 64. The H_0 value, z_d and z_s are concatenated to the result, which is then fed sequentially to 3 linear layers, each of 768 neurons. There is a ELU activation functions before and after the second layer. The whole neural network counts 2,224,514 parameters.

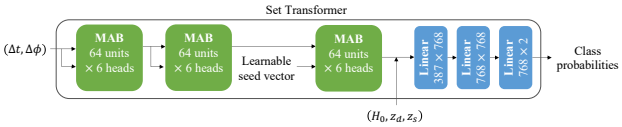

Figure 3. Architecture of the discriminator. The green squares represent the multihead attention blocks (MAB), and the blue rectangles, the linear layers. ELU activation functions are used after the first and the second linear layers. At inference, a softmax is added after the last layer.

At inference time, we apply a softmax function the final output to retrieve the class probabilities. We then insert the probability of the class with label $y = 1$ in Equation (8) to estimate the distribution ratio. The latter is equivalent to the posterior density at the input H_0 because the prior is uniform.

Table 3. Input sizes for each operation in the deep set ratio estimator.

Operation	**Input sizes**
First multihead attention block	example set size $\times$ 2
Second multihead attention block	example set size $\times$ 384
Third multihead attention block	features : example set size $\times$ 384
	learnable seed vector : example set size $\times$ 1 $\times$ 384
Concatenating H_0 and the redshifts	384
First linear layer	387
Second linear layer	768
Third linear layer	768
Ratio estimation (see Equation (8))	2

B. Examples of individual posteriors

In Figure 4, we compare the NRE results on 6 representative test examples with those of nested sampling performed with the package POLYCHORD (Handley et al., 2015a;b). Each plot is associated to a different lensing system and a different H_0 value. The nested sampling and the NRE posteriors are respectively indicated by the blue dashed line and the red solid line. The NRE shows a good agreement with the nested sampling posteriors. Moreover, each NRE posterior is a factor of about 1000 faster to obtain, taking only $\sim$1 sec on an NVidia V100 GPU, whereas sampling with POLYCHORD requires a minimum of 20 minutes per posterior.

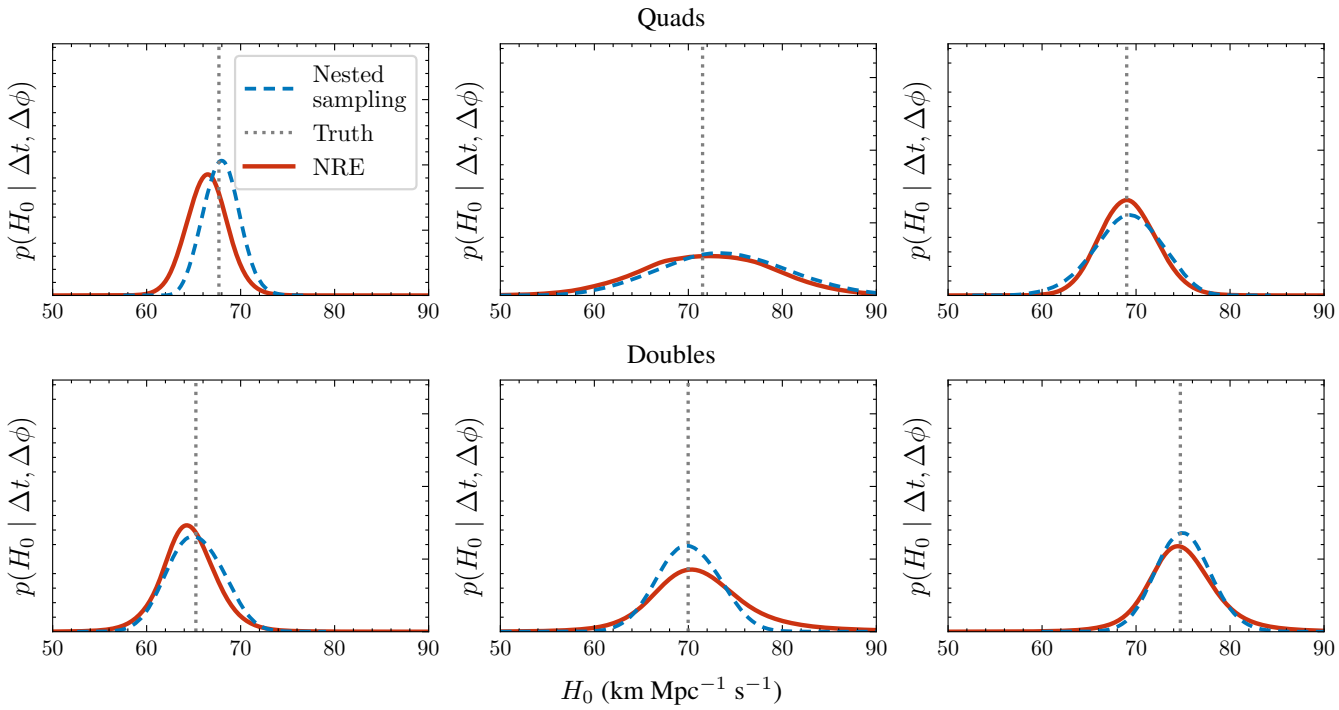

Figure 4. Six H_0 inferences on examples from the test set. The first row show results for three different quads, and the last row, for three different doubles. The true H_0 value is also different on each plot. The blue dashed line indicates the true posterior distribution (computed with nested sampling and the true likelihood), the grey dotted line represents the true value, and the red solid line is the NRE posterior distribution. The difference between the two distributions is noticeable, but they still agree well with each other.