Automating Inference of Binary Microlensing Events with Neural Density Estimation

@AstroKeming

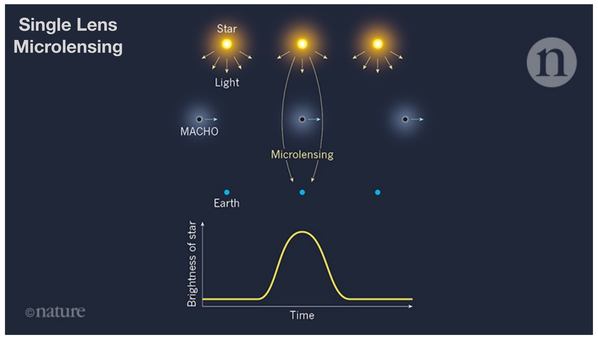
arXiv: 2010.04156

Keming Zhang[1], Joshua S. Bloom[1], B. Scott Gaudi[2], Francois Lanusse[3], Casey Lam[1], Jessica Lu[1]

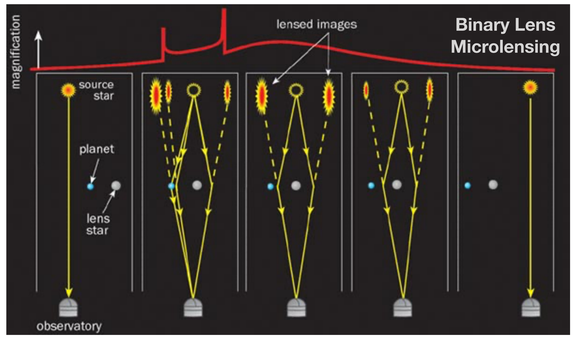
[1]University of California at Berkeley, [2]The Ohio State University,
[3]AIM, CEA, CNRS, Universit'e Paris-Saclay, Universit'e Paris Diderot Sorbonne Paris Cit'e

Berkeley UNIVERSITY OF CALIFORNIA

Microlensing 101



c. nature.com

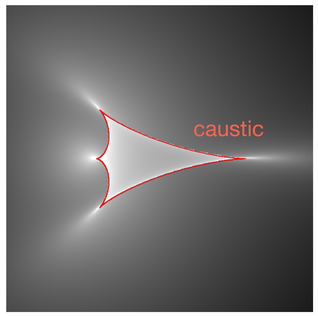


c. physicsworld.com

The gravitational field of stars can act like a magnifying glass: when the apparent trajectory of a foreground *lens* star passes close to 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 (***top left figure***). Binary microlensing events occur when the *lens* is a system of two stars: either a binary star system, or a star-planet configuration (***top right figure***). Such events provide a unique opportunity for exoplanet discovery as the planet-to-star mass ratio may be inferred from the light curve without having to detect light from the star-planet lens itself.

While **single-lens microlensing events are described by a simple analytic expression, 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 (***lower right figure***). This translates to a pathological parameter space for which the likelihood surface suffers from a multitude of local minima that are both narrow and deep; this significantly hampers attempts of direct sampling-based inference without knowledge of the approximate solution.

Here, we present an **automated inference framework based on neural density estimation**, where the fundamental task is to learn distributions from samples with neural networks.



c. J. Yee

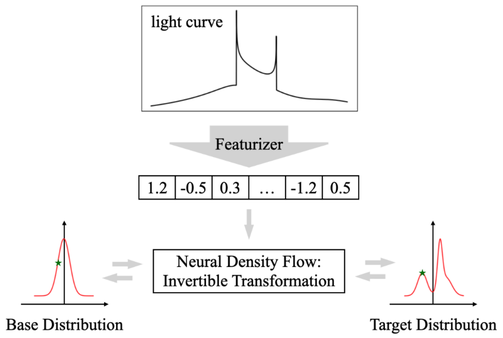
Inference with Neural Density Estimation

We'd like to **learn an approximate posterior which minimizes the KL divergence between the true posterior. To minimize the KL divergence is to maximize likelihood**:

$$\begin{aligned}\phi &= \text{argmin}(D_{\text{KL}}(p(\theta|\mathbf{x}))||\hat{p}_\phi(\theta|\mathbf{x})))\\ &= \text{argmin}(\mathbb{E}_{\theta\sim p(\theta),x\sim p(x|\theta)}[\log(p(\theta|\mathbf{x})) - \log(\hat{p}_\phi(\theta|\mathbf{x}))])\\ &= \text{argmax}(\mathbb{E}_{\theta\sim p(\theta),x\sim p(x|\theta)}[\hat{p}_\phi(\theta|\mathbf{x})])\end{aligned}$$

Note that we're using "maximum likelihood" instead of "maximum posterior" because the microlensing parameters θ are regarded as "data" to be modeled.

We use a **20-block Masked Autoregressive Flow (MAF) for p^(θ|x)**, and a **ResNet-GRU network to extract features** 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 block of the MAF (which is a "MADE") 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 and a 2-layer GRU. 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.



(above figure) A ResNet-GRU featurizer turns a raw light curve into a low dimensional vector, which serves as the conditional input to the Masked Autoregressive Flow. To the bottom, the base distribuion is a mixture of 8 gaussians and the target distribution is the posterior.

Training Set

We simulate a dataset of **1 million binary-lens-single-source (2L1S) magnification sequences** with the microlensing code MulensModel within the context of the Roman Space Telescope Microlensing Survey; each sequence contains 144 days at a cadence of 0.01 day, corresponding to the planned Roman cadence of 15 minutes. 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.

Prior: Ignoring orbital motion of both the observer and the binary lens, binary microlensing (2L1S) events are described by seven parameters: **two determine the shape of the caustic**: binary separation (s) and mass ratio (q); **four determine the trajectory**: angle of approach (α), time of primary-lens-source closest approach (t0), Einstein ring crossing timescale (tE), impact parameter (u0), and finally the finite source size (ρ) which is a higher-order effect. We simulate 2L1S events based on the following **analytic priors** shown to the right:

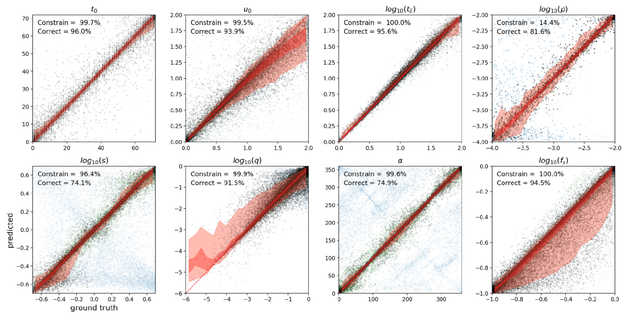
$$t_E \sim \text{TruncLogNorm}(\min = 1, \max = 100, \mu = 10^{1.15}, \sigma = 10^{0.45})$$
$$u_0 \sim \text{Uniform}(0, 2); \quad s \sim \text{LogUniform}(0.2, 5); \quad q \sim \text{Uniform}(10^{-6}, 1)$$
$$\alpha \sim \text{Uniform}(0, 360); \quad \rho \sim \text{LogUniform}(10^{-4}, 10^{-2})$$

Noise: We assume an ideal **Gaussian measurement noise** where the standard deviation of each measurement is the square root of flux measurement in raw detector counts. To simulate a wide range of baseline stellar apparent brightness, the signal-to-noise ratio of the baseline, unmagnified flux is uniformly sampled between 23 to 200 during training.

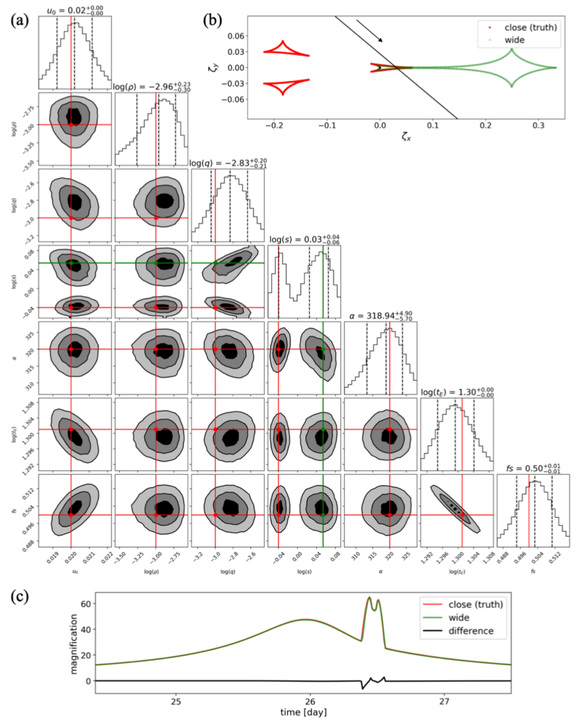
Results

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** on one CPU core. ***The lower right figure*** shows the NDE posterior for an example event which exhibits a classic "close-wide" degeneracy. 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 structure and magnification curve are shown in green.

The NDE posterior uncertainty turns out to be larger than that of the exact posterior computed using e.g. MCMC. We make the following observation. 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. 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 ~thousand steps.



(***above figure***) Predicted vs. ground truth 2L1S parameters for 57,430 test-set 2L1S events. Single-mode NDE posteriors are shown in black dots. For multi-model NDE posteriors, we color-code the solution as follows: those for which the global mode is closest to the ground truth are plotted in black; for cases for which a secondary mode is closest to the true value, that correct, secondary mode is plotted in orange whereas the incorrect global mode is plotted in blue. Red shadows indicates 32-68th percentile (1σ) and 5-95 percentile (2σ) regions. Red-dashed lines shows the diagonal. In the upper left of each subplot, "constrain" refers to the percentage of events whose NDE posterior poses sufficient constraint — the peak posterior probability much be at least twice the prior probability. "Correct" refers to percentage of constrained events whose true parameter lies closest to the global mode.



(above figure) (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. (b) Caustic structure for both close and wide solutions. Arrow indicate direction of source trajectory. (c) Close-up view of magnification curve for both "close" and "wide" solutions, which are hardly distinguishable.