Automating Inference of Binary Microlensing Events with Neural Density Estimation

Keming Zhang1, Joshua S. Bloom2, B. Scott Gaudi3, Francois Lanusse4, Casey Lam1, Jessica Lu1
1University of California at Berkeley, 2The Ohio State University, 3AIM, CEA, CNRS, Université Paris-Saclay, Université Paris Diderot Sorbonne Paris Cité

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 (θ), time of primary-lens-source closest approach (t0), Einstein ring crossing timescale (tE), impact parameter (u), and finally the finite source size (θ) which is a higher-order effect. We simulate 2L1S events based on the following analytic priors shown to the right:

Noise: We assume an ideal Gaussian measurement noise where the standard deviation of each measurement is the square root of flux measurement error 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 10^8 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 = 10^4) and 5-95 percentile (2\σ) are shown to the right:

\[ t_0 \sim \text{TruncLogNorm}(\text{mean} = 1, \text{scale} = 100, \mu = 10^{-15}, \sigma = 10^{-6}) \]
\[ s_0 \sim \text{Uniform}(0, 2); \quad r \sim \text{LogUniform}(0.2, 2); \quad \sigma \sim \text{Uniform}(0^0.5, 1) \]
\[ \theta \sim \text{Uniform}(0^0.05, \pi^0.05) \]

Note that we’re using “maximum likelihood” instead of “maximum posterior” because the microlensing parameters if are regarded as “data” to be modeled.

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:

\[ \phi = \arg\max_{\theta} \left[ \log p(\theta) + \text{KL}(p(\theta) \parallel q(\theta)) \right] \]

Note that we’re using “maximum likelihood” instead of “maximum posterior” because the microlensing parameters if are regarded as “data” to be modeled.

We use a 20-block Masked Autoregressive Flow (MAF) for p(θ), 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 “MADIE”) adapts a fixed ordering of the dimensions and applies affine transformations iteratively for each dimension, subject to the autoregressive condition. We adopt random orders for each of the 20 blocks 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 16-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.

Inference

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, owing 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 might be inferred from a model that has a 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.

Microlensing 101

(a) NDE posterior for a central-caustic crossing event. 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 = 10^4) and 5-95 percentile (2\σ) are shown to the right:

\[ t_0 \sim \text{TruncLogNorm}(\text{mean} = 1, \text{scale} = 100, \mu = 10^{-15}, \sigma = 10^{-6}) \]
\[ s_0 \sim \text{Uniform}(0, 2); \quad r \sim \text{LogUniform}(0.2, 2); \quad \sigma \sim \text{Uniform}(0^0.5, 1) \]
\[ \theta \sim \text{Uniform}(0^0.05, \pi^0.05) \]

Noise: We assume an ideal Gaussian measurement noise where the standard deviation of each measurement is the square root of flux measurement error 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 10^8 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 = 10^4) and 5-95 percentile (2\σ) are shown to the right:

\[ t_0 \sim \text{TruncLogNorm}(\text{mean} = 1, \text{scale} = 100, \mu = 10^{-15}, \sigma = 10^{-6}) \]
\[ s_0 \sim \text{Uniform}(0, 2); \quad r \sim \text{LogUniform}(0.2, 2); \quad \sigma \sim \text{Uniform}(0^0.5, 1) \]
\[ \theta \sim \text{Uniform}(0^0.05, \pi^0.05) \]

Noise: We assume an ideal Gaussian measurement noise where the standard deviation of each measurement is the square root of flux measurement error 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 10^8 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 = 10^4) and 5-95 percentile (2\σ) are shown to the right:

\[ t_0 \sim \text{TruncLogNorm}(\text{mean} = 1, \text{scale} = 100, \mu = 10^{-15}, \sigma = 10^{-6}) \]
\[ s_0 \sim \text{Uniform}(0, 2); \quad r \sim \text{LogUniform}(0.2, 2); \quad \sigma \sim \text{Uniform}(0^0.5, 1) \]
\[ \theta \sim \text{Uniform}(0^0.05, \pi^0.05) \]