α-deep Probabilistic Inference (α-DPI): Efficient Uncertainty Quantification from Exoplanet Astrometry to Black Hole Feature Extraction

He Sun[1] , Katherine L. Bouman[1], Paul Tiede[2,3,4,5,6] , Jason J. Wang[1] , Sarah Blunt[1] , and Dimitri Mawet[1]

[1] California Institute of Technology, 1200 East California Boulevard, Pasadena, CA 91125, USA; hesun@caltech.edu
[2] Perimeter Institute for Theoretical Physics, 31 Caroline Street North, Waterloo, ON, N2L 2Y5, Canada
[3] Department of Physics and Astronomy, University of Waterloo, 200 University Avenue West, Waterloo, ON, N2L 3G1, Canada
[4] Waterloo Centre for Astrophysics, University of Waterloo, Waterloo, ON, N2L 3G1 Canada
[5] Black Hole Initiative at Harvard University, 20 Garden Street, Cambridge, MA 02138, USA
[6] Center for Astrophysics | Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA

Abstract

Inference is crucial in modern astronomical research, where hidden astrophysical features and patterns are often estimated from indirect and noisy measurements. Inferring the posterior of hidden features, conditioned on the observed measurements, is essential for understanding the uncertainty of results and downstream scientific interpretations. Traditional approaches for posterior estimation include sampling-based methods and variational inference (VI). However, sampling-based methods are typically slow for high-dimensional inverse problems, while VI often lacks estimation accuracy. In this paper, we propose α-deep probabilistic inference, a deep learning framework that first learns an approximate posterior using α-divergence VI paired with a generative neural network, and then produces more accurate posterior samples through importance reweighting of the network samples. It inherits strengths from both sampling and VI methods: it is fast, accurate, and more scalable to high-dimensional problems than conventional sampling-based approaches. We apply our approach to two high-impact astronomical inference problems using real data: exoplanet astrometry and black hole feature extraction.

Inferring hidden features from indirect, sparse, and noisy observational data is a fundamental challenge in modern astronomical research. When tackling these inverse problems, it is important to consider uncertainty in the inferred solution via a posterior rather than simply a point estimate (e.g., maximum a posteriori (MAP)). A full posterior guarantees that all possible scientific interpretations have been considered. However, due to the large amount of observational data, the high dimensionality of features to infer, and the potential multimodality of posterior distributions, recovering a full posterior distribution is often computationally challenging. In this paper we propose an efficient and flexible approach for full posterior estimation in inverse problems, and demonstrate the approach on two applications in astronomical inference.

Traditionally, sampling-based approaches, such as importance sampling and Markov Chain Monte Carlo (MCMC) methods, are widely used to solve inference problems in computational science. Importance sampling first generates random samples based on a proposal distribution and then reweights the samples according to their likelihood to approximate the true posterior; MCMC gradually refines the samples using a Markov chain transition distribution to create samples from a posterior distribution. Although these methods have achieved great success in many Bayesian inference problems, they typically suffer from the curse of dimensionality in high-dimensional estimation. Without a good proposal distribution in importance sampling or a good random

initialization in MCMC, sampling approaches are often prohibitively slow to converge (Owen 2013).

To overcome the above limitation, optimization-based inference approaches, such as variational inference (VI), have drawn increased attention. Instead of directly sampling the parameter space, VI methods introduce a parametric density function (e.g., a Gaussian mixture model), and solve the inference problem by minimizing a loss (e.g., Kullback–Leibler (KL) divergence) between the variational density function and the target posterior distribution. VI methods are drastically more efficient than sampling-based approaches since their optimization is gradient based. However, limited by the modeling capacity of the variational density function and the choice of divergence loss, traditional VI methods often estimate overly simplified distributions and sometimes lead to degenerated posterior estimation (e.g., mode collapse; Zhang et al. 2018).

In this paper, we propose a new deep learning Bayesian inference approach, α-deep probabilistic inference (α-DPI), for fast and accurate posterior estimation. Our proposed approach consists of two primary steps: (1) we first use α-divergence VI to optimize a normalizing flow generative model to approximate a posterior; (2) we then use importance sampling to recover a more accurate distribution. By utilizing current state-of-the-art deep neural network architectures, α-DPI inherits advantages from both sampling and optimization methods. This proposed new deep learning approach not only significantly improves the computational efficiency and accuracy of high-dimensional posterior estimation, but also naturally leads to a metric that can be used for model selection when the underlying physical model is unknown. We apply our method to two difficult high-impact astronomical inference problems:

exoplanet astrometry and black hole feature extraction. In both cases, α-DPI achieves improved performance in terms of estimation efficiency and accuracy when compared with traditional methods.

1. Results

1.1. Overview of the Method

α-DPI is a two-step algorithm that combines α-divergence VI, a normalizing flow neural network, and importance sampling for uncertainty quantification with Bayesian inference. Using only the observed measurements y, the goal of α-DPI is to generate samples of the hidden state x that approximate those from the target posterior $p(x|y)$. For a more detailed description of the method please see Section 3.

First, α-DPI learns the weights $θ$ of a normalizing flow neural network $g_θ(\cdot)$ that generates samples from a proposal distribution $q_θ(x)$. This generative network is optimized using an α-divergence VI loss:

$$\theta^\star = \arg\min_\theta D_\alpha[q_\theta(x)\|p(x|y)] \approx \arg\min_\theta \frac{1}{N}\sum_{n=1}^{N}$$
$$\times [\exp[(1-\alpha)(\log p(y|x_n) + \log p(x_n) - \log q_\theta(x_n))]], \tag{1}$$

where $D_\alpha[A\|B]$ is the Renyi's α-divergence of A from B (Rényi 1961; Van Erven & Harremos 2014), $x_n = g_\theta(z_n)$ for $z_n \in \mathbb{R}^{|x|} \sim \mathcal{N}(0, 1)$ such that $x_n \sim q_\theta(x)$, and $p(y|x_n)$ and $p(x_n)$ should be differentiable functions to facilitate gradient-based optimization. In this α-divergence formulation, α is a selected value between 0 and 1. Optimizing with an α-divergence loss encourages the learned distribution $q_\theta(x)$ to be similar to $p(x|y)$. Refer to Section 3.1.1 for a more detailed description of α-divergence and its relation to KL divergence. The resulting generative network will therefore map samples z_n from an independent and identically distributed Gaussian distribution to those from the learned proposal distribution $q_\theta(x)$, where $q_\theta(x)$ approximates the target posterior $p(x|y)$.

Second, α-DPI performs importance sampling to reweight samples from the learned proposal distribution. In particular, each sample in the set $\{x_j\}$, where $x_j = g_\theta(z_j) \sim q_\theta(x)$, is weighted by

$$w(x_j) = \frac{p(y|x_j)p(x_j)}{q_\theta(x_j)}, \tag{2}$$

so that $p(x_j|y) \propto w(x_j)q_\theta(x_j)$. By simply resampling from the set $\{x_j\}$ according to $\{w(x_j)\}$, α-DPI produces a new set of samples $\{x'_j\}$ that better captures $p(x|y)$.

A critical building block of α-DPI is the normalizing flow generative neural network. This class of invertible networks is widely used in computer vision and machine learning for density function approximation. Normalizing flow generative networks are able to capture complex correlations between parameters in the target posterior. Therefore, unlike simple distribution models typically used for VI (e.g., Gaussian), these networks lead to better proposal distributions that in turn lead to more efficient importance sampling. Additionally, unlike variational autoencoders (VAEs) that rely on efficient optimization via the Gaussian reparameterization trick, normalizing flows are bijective and therefore are not restricted to capturing unimodal distributions. However, note that since a traditional

normalizing flow architecture will result in a continuous generative distribution, target distributions with disconnected modes cannot be captured by a normalizing flow without a connecting "bridge."

α-DPI significantly outperforms our previous work for image reconstruction uncertainty quantification: deep probabilistic imaging (DPI Sun & Bouman 2021). In Sun & Bouman (2021) a single-step KL-divergence VI was used to approximate the target posterior distribution; this formulation would sometimes result in mode collapse. In contrast, we find that formulating the inference using α-divergence, which trades off exploration and exploitation through the α parameter, is far less susceptible to mode collapse. α-divergence VI has the disadvantage of generally producing a distribution $q_\theta(x)$ that samples x's outside of the target posterior. However, by simply following α-divergence VI with importance sampling we can more accurately capture complex multimodal posterior distributions than KL-divergence VI. In this paper, we denote the method presented in Sun & Bouman (2021) as KL-DPI.

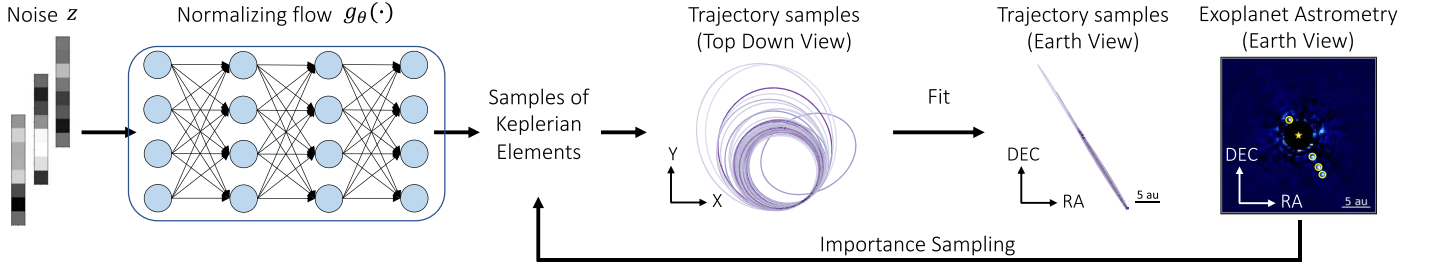### 1.2. Exoplanet Astrometry and Orbital Fitting with Gemini Planet Imager

Detection and characterization of exoplanets using direct imaging is one of the most exciting frontiers in astronomy in the era of 30 m class optical/infrared telescopes. By capturing a series of snapshot images of an exoplanet, we can understand the properties of its planetary system (Yu & Tremaine 2001; Scharf & Menou 2009; Bate et al. 2010; Maire et al. 2019). Fitting an exoplanet's astrometry data allows us to understand the planet's formation and evolution, and sometimes even detect unseen planets (Lacour et al. 2021). A planet's orbit is parameterized by eight Keplerian elements (Blunt et al. 2017, 2020): semimajor axis (a), eccentricity (e), inclination angle (i), argument of periastron of the secondary's orbit (ω), longitude of ascending node (Ω), epoch of periastron passage (τ, defined as a fraction of the orbital period past a reference epoch), parallax (π), and total mass (M_T), as illustrated in Figure 1. In an orbit fitting problem, we estimate the posterior of these parameters based on the exoplanet astrometry data relative to the primary star (e.g., R.A. and decl.) from telescope snapshot images.

We applied α-DPI to a data set from the Gemini Planet Imager (GPI) to infer the orbital parameters of a planet, β Pictoris b (β Pic b) (Wang et al. 2016). The prior distributions on β Pic b's orbital parameters, as mentioned in Wang et al. (2016), are listed in Figure 1. As shown by the exoplanet astrometry data (right image) in Figure 2, β Pic b has an almost edge-on orbit when observed from the Earth. Due to the limited observational data, the posterior of β Pic b's Keplerian elements is multimodal. Figure 3 shows the posterior samples from MCMC, KL-DPI, and α-DPI before and after importance sampling (unless otherwise noted, α-DPI includes the step of importance sampling). The posterior samples are visualized using corner plots, which present the marginal distribution of each Keplerian element and the joint distribution of each pair of Keplerian elements. As can be seen, KL-DPI neglects disconnected minor modes in its approximated posterior; this occurs because the KL loss strongly discourages including any mass from low posterior probability regions in $q_\theta(x)$, which is needed to connect disconnected modes in a posterior when using a normalizing flow model. For this reason, KL-DPI results in a distribution $q_\theta(x)$ that underestimates the
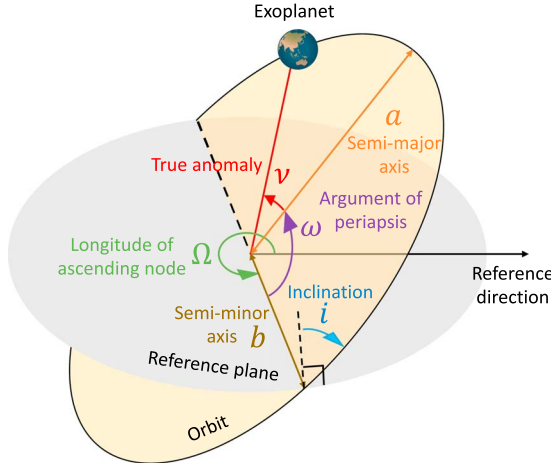
Parameter	Unit	Prior distribution
semimajor axis (a)	astronomical unit (au)	log uniform: $\log U(10, 10^4)$
eccentricity (e)	–	uniform: $U(10^{-8}, 0.99)$
inclination angle (i)	degree $(\circ)$	uniform: $U(0, 180)$
argument of periastron (ω)	degree $(\circ)$	uniform: $U(0, 360)$
longitude of ascending node (Ω)	degree $(\circ)$	uniform: $U(0, 360)$
epoch of periastron (τ)	–	uniform: $U(0, 1)$
parallax (π)	milliarcsecond (mas)	Gaussian: $N(56.95, 0.26)$
total mass (M_T)	solar mass $(M_\odot)$	Gaussian: $N(1.22, 0.08)$

Figure 1. (Left) Illustration of five of the eight Keplerian elements estimated in this work. The additional three elements not illustrated are the total mass of the exoplanet (M_T), the parallax (π), and the eccentricity (e). Eccentricity is defined by semimajor axis (a) and semiminor axis (b) as $e = \sqrt{1 - \frac{b^2}{a^2}}$. (Right) Definitions of Keplerian elements and their prior distributions used in exoplanet β Pic b astrometry.

Figure 2. α-DPI applied to exoplanet astrometry. In this example, we infer the orbital parameters of an exoplanet, β Pic b, based its snapshot images captured by the Gemini Planet Imager (GPI). α-DPI first optimizes the normalizing flow weights, θ, to generate diverse trajectory samples that match the observational data. Then it conducts importance sampling to reweight each trajectory sample to produce an accurate posterior. In the plots, the trajectories with darker colors have higher posterior likelihoods. In the composite exoplanet astrometry image on the far right the locations of the exoplanet at different times have been circled for clarity.

uncertainty of the planet's orbit. In contrast, the normalizing flow learned in α-DPI captures all the posterior modes of the astrometry data. Although α-DPI includes several samples with low data likelihood that connect different posterior modes, it still provides a good proposal distribution that can be used by importance sampling to efficiently generate more accurate posterior samples. The final α-DPI approximate posterior distribution achieves comparable accuracy to the MCMC sampler, but only takes 1.5 hr of computation on a single GTX1080 Ti GPU (time of importance sampling step is negligible compared with normalizing flow training). In comparison, the MCMC method requires more than 24 hr to recover the posterior (Blunt et al. 2020).

We also quantitatively investigate the influence of the α parameter values on both α-DPI's posterior estimation precision and computational time. The precision is evaluated using the maximum mean discrepancy (MMD; Gretton et al. 2012) between the "ground-truth" MCMC posterior, $q_{mc}(\cdot)$, and the optimized normalizing flow posterior, $q_{dpi}(\cdot)$,
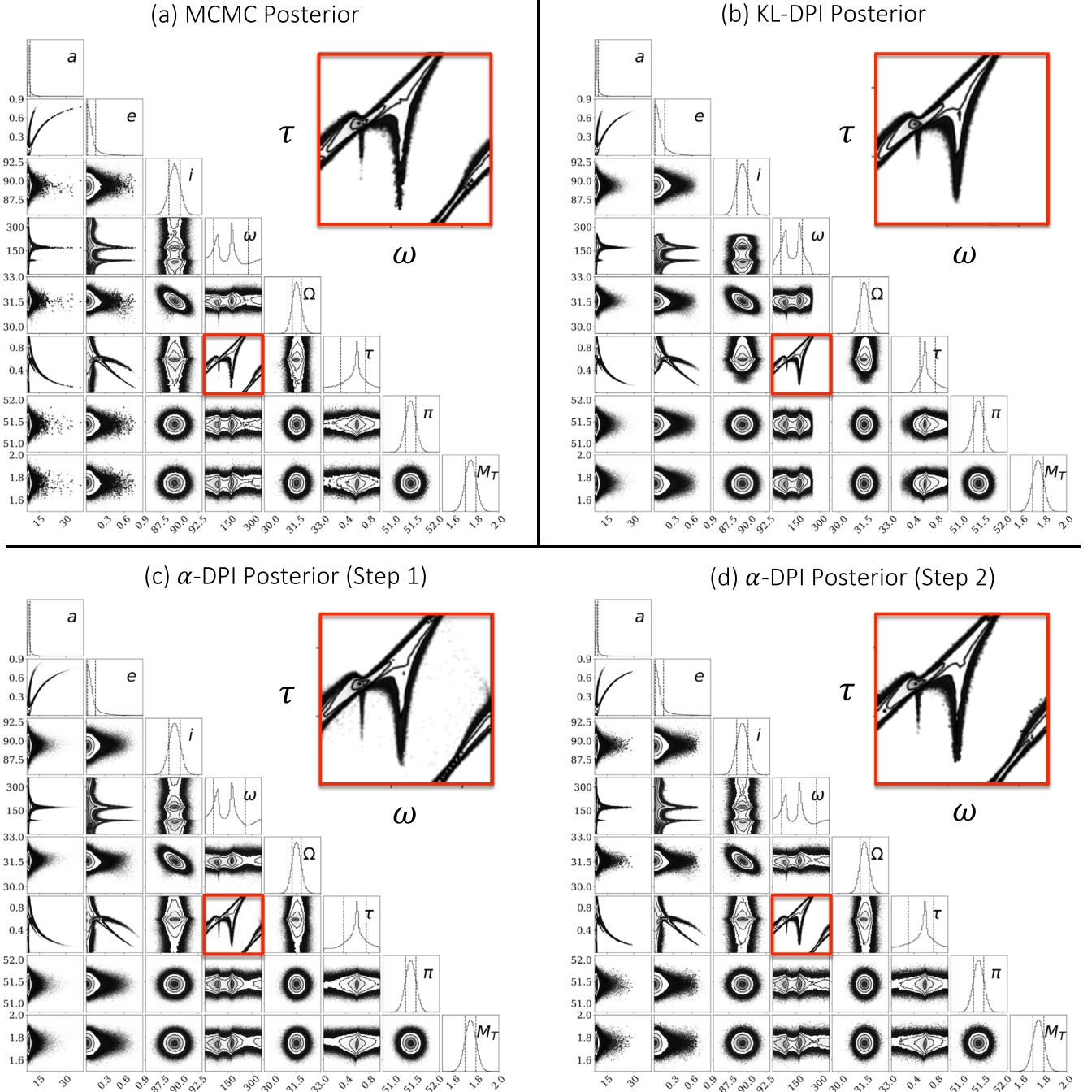
$$MMD^2[q_{mc}(\cdot), q_{dpi}(\cdot)] = \mathbb{E}_{x,x'\sim q_{mc}(\cdot)}[K(x, x')]$$
$$+ \mathbb{E}_{y,y'\sim q_{dpi}(\cdot)}[K(y, y')] - 2\mathbb{E}_{x\sim q_{mc}(\cdot), y\sim q_{dpi}(\cdot)}[K(x, y)], \quad (3)$$

where $K(\cdot, \cdot)$ is a kernel function between two samples. A smaller MMD indicates that two distributions are more similar to each other. A radial basis function kernel (Schölkopf 1997)

is used in our MMD calculation. The MMD of normalizing flow samples before and after importance sampling are both reported. As demonstrated in Figure 4 (left), the posterior estimation performs well for α between 0.3 and 0.9, but results in poor performance when α is either large or small. A large α leads to the mode collapse problem and additional importance sampling does not help since a posterior mode has already been neglected; a small α makes the stochastic loss in Equation (1) hard to evaluate with a limited batch size, resulting in a normalizing flow model that fails to converge to an informative proposal posterior. The computational time of α-DPI is defined as the training time for the normalizing flow model to reach a sufficient precision (MMD < 0.1) and stay at this precision for more than 30 iterations. We find the time spent on the importance sampling step (Step 2) is negligible compared with the neural network training, even in the case of a low α. As shown in Figure 4 (right), the computational time does not change significantly given a fixed normalizing flow architecture and a fixed batch size. The computational time using a smaller α is a little longer but varies more among different runs. This again indicates the evaluation of stochastic loss in Equation (1) is difficult with a small α.

Given an appropriate α, the results of α-DPI agree well with the posteriors from Wang et al. (2016), indicating an accurate estimation of the possible orbital configurations for β Pic b,

Figure 3. Corner plots of posterior samples from (a) MCMC, (b) KL-DPI (equivalent to α-DPI with $\alpha = 1.0$, see Section 3.1.1) and α-DPI with $\alpha = 0.5$ before (c; step 1) and after (d; step 2) importance sampling. The zoomed-in figures present the marginal joint distribution of argument of periastron (ω) and epoch of periastron (τ); the presence of disconnected modes is due to the periodicity of both parameters. KL-DPI collapses to a distribution that fails to capture some of the modes in the posterior. In contrast, α-DPI captures all disconnected modes. Although (c) includes a few "bad" samples that should have low probability, after importance sampling the resulting posterior in (d) is nearly identical to the MCMC identified posterior. The contours in the corner plots represent the levels of the posterior likelihoods; points lying outside the posterior contours are the samples whose likelihoods are smaller than a threshold but not negligible.

including the fact it will not transit the star. However, due to its much faster run time, α-DPI lowers the barrier for obtaining orbital posteriors for a large number of systems in the future that have yet to be analyzed fully. For example, Ferrer-Chávez et al. (2021) was able to explore biases in orbit fitting for short orbital arcs (spanning $\sim 1\%$ of a planet's orbit), but chose not to explore longer orbital arcs because current orbit fitting algorithms were prohibitively slow. Thus, α-DPI holds significant promise for handling this inference problem both efficiently and accurately.

Implementation Details. In order to solve for $g_\theta(\cdot)$ using gradient-based optimization, the forward model that maps Keplerian elements to astrometry data is implemented in a differentiable manner. In particular, Kepler's equation is approximated via a Gauss–Newton solver with 10 fixed gradient descent optimization steps. We compare α-DPI ($\alpha = 0.5$ in the divergence loss) to a normalizing flow based KL-divergence VI method (denoted as KL-DPI in this paper) and a parallel-tempered MCMC (PTMCMC) sampler from
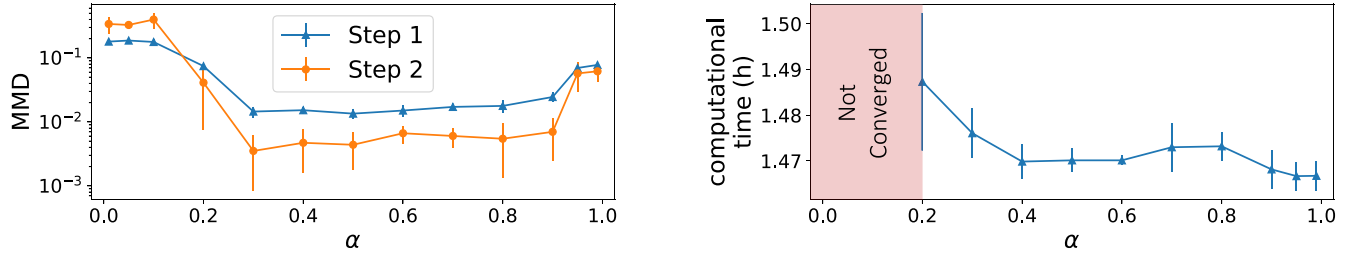
Figure 4. Posterior estimation precision (left) and computational time (right) for α-DPI with different α values ([0.01, 1]). The error bars in both figures are generated based on five independent runs for each α value. The posterior estimation precision is evaluated using the maximum mean discrepancy (MMD) between the converged MCMC samples and α-DPI samples, where a smaller MMD indicates higher precision. Steps 1 and 2 represent the samples before and after importance sampling, respectively. The computational time is defined as the time when the MMD of normalizing flow samples (step 1) drops below 0.1 for 30 consecutive iterations. The normalizing flow models optimized with $0 < \alpha < 0.2$ never achieve sufficient precision (MMD < 0.1) in the allotted time ($\sim$3.5 hr), and thus are marked as "not converged" even if they have converged to a distribution with a larger MMD.
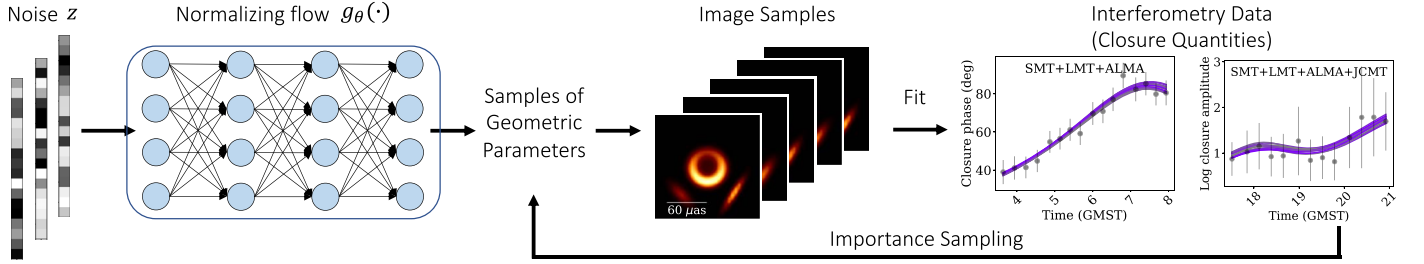
Figure 5. α-DPI applied to VLBI black hole feature extraction. In this example, we infer the geometric parameters (e.g., diameter, position angle, central emission) of a super massive black hole based on the data captured by the Event Horizon Telescope (EHT). We conduct experiments using both simulated interferometric data and real data of M87* captured in 2017. Due to atmospheric turbulence and instrument calibration errors, we fit robust data products referred to as closure quantities. The right two figures show the measurement data (gray dots with error bars) and the ideal corresponding measurements from posterior samples (purple curves; darker curves indicating more likely samples) of two closure quantities used to define the data likelihood. Here the capitalized letters within each plot represent the telescopes used for capturing the corresponding measurements.

ptemcee (Vousden et al. 2016) using 1000 Markov chains with 45,000 iterations in each chain (first 40,000 steps as the burn-in). Since the MCMC sampler uses a large number of chains and iterations, we believe it converges to close to the ground-truth posterior. Therefore, we use the resulting MCMC posterior in order to evaluate the inference accuracy of α-DPI.

1.3. Black Hole Feature Extraction with the Event Horizon Telescope

Very long baseline interferometry (VLBI) has enabled the reconstruction of high-resolution astronomical images using sparse data from a multitelescope synthetic aperture. By joining radio telescopes from across the globe, the Event Horizon Telescope (EHT) collaboration captured the first picture of a black hole, M87* (Event Horizon Telescope Collaboration et al. 2019a). This marked a new era for black hole astronomy, since scientists can now study black holes by directly observing their event-horizon scale structure. However, VLBI data for black hole imaging is typically very sparse and noisy: only seven telescopes at five geographic sites were used for collecting the M87* data that led to the first black hole image —this sparse telescope data was heavily corrupted by atmospheric turbulence and instrumentation calibration errors. In order to deliver reliable scientific interpretations, it is important to carefully characterize the uncertainty in features of the black hole image, including diameter, width, asymmetry, and position angle.
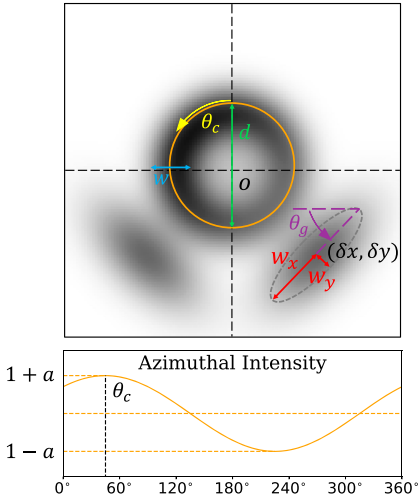
In this section, we apply our proposed α-DPI algorithm to both simulated and real EHT M87* data, and fit geometric models to understand important black hole properties, as shown in Figure 5. We choose to parameterize the black hole image as the summation of a crescent (an asymmetric ring with a central emission disk) and multiple additional elliptical Gaussians, as illustrated in Figure 6. Elliptical Gaussians are introduced to account for extended flux in the black hole image. The number of ellipses to include is unknown a priori when analyzing real observational data. Therefore, we first attempt models with different number of Gaussian ellipses (0–3) in our following experiments, and subsequently select the best geometric model by comparing the evidence lower bounds (ELBO) of different models, denoted as m. In particular, by expanding out the equation of $p(y|m)$ (i.e., the probability of data under a given model) one can see that, if the true posterior has been well approximated (i.e., $D_{KL} \approx 0$), the ELBO function identifies which model best describes the data without overinterpretation:

$$\log p(y|m) = D_{KL}[q_\theta(x) \| p(x|y, m)]$$
$$- \mathbb{E}_{x \sim q_\theta(x)}[-\log p(y|x, m) - \log p(x|m) + \log q_\theta(x)]$$
$$\geqslant -\mathbb{E}_{x \sim q_\theta(x)}[-\log p(y|x, m) - \log p(x|m) + \log q_\theta(x)]$$
$$= \text{ELBO}(m).$$

(4)

Since α-DPI samples often approximate the true posterior well, we assume that D_{KL} is negligible compared with the ELBO and choose the preferred model, m by selecting the model with the largest ELBO value (Penny 2012).

Simulated EHT data is generated using a crescent model with two elliptical Gaussians. Figure 7 (top) presents the ELBO when fitting simulated data using models of varying complexity. As expected, the true model—a crescent with two Gaussian

Parameter	Unit	Prior distribution
crescent diameter (d)	microarcsecond (μas)	uniform: $U(20, 100)$
crescent width (fwhm, w)	microarcsecond (μas)	uniform: $U(1, 40)$
crescent asymmetry (a)	—	uniform: $U(0, 1)$
crescent position angle (θ_c)	degree $(\circ)$	uniform: $U(0, 360)$
crecent flux (V_c)	—	uniform: $U(0, 2)$
flux ratio of central emission disk and crescent (V_d)	—	uniform: $U(0, 1)$
central coordinates of additional Gaussian components $(\delta x_k, \delta y_k)$	microarcsecond (μas)	uniform: $U(-200, 200)$
half major and minor axes of additional Gaussian components $(w_{x,k}, w_{y,k})$	microarcsecond (μas)	uniform: $U(0, 100)$
Gaussian component's position angle $(\theta_{g,k})$	degree $(\circ)$	uniform: $U(0, 90)$
Gaussian component's flux $(V_{g,k})$	—	uniform: $U(0, 2)$

Figure 6. (Left) Illustration of a black hole geometric model. The model consists of two parts: (1) an asymmetric ring with a central emission disk (the ring and disk share the same diameter), and (2) multiple Gaussian ellipses (index by k). The azimuthal intensity of the ring follows a sinusoidal function. The flux of the ring, central disk and Gaussian ellipses are V_c, V_d, and $V_{g,k}$, respectively, which are neglected in the figure. (Right) Definitions of black hole geometric parameters and their prior distributions used for simulated and real M87* interferometry data.

ellipses—has the largest ELBO value. The bottom figure in Figure 7 (bottom; (a)) shows the corner plots of α-DPI posterior samples assuming a two-ellipse crescent model (cross correlation between the crescent and ellipses are neglected for brevity). The ground-truth values (blue lines) successfully lie in the recovered posterior. More interestingly, since the parameters of two Gaussian ellipses are interchangeable, their marginal posteriors are identical to each other; α-DPI successfully identifies this predictable symmetry.

We also investigate the posterior distribution of M87* geometric features using the real EHT data set (Event Horizon Telescope Collaboration et al. 2019b) that produced the first black hole image in 2019.[7] The top left of Figure 7 presents the mean reconstructed images obtained from α-DPI samples under models of increasing complexity, along with the corresponding computational time and ELBO values. Figure 7 also presents the uncertainty of estimated black hole feature parameters, visualized by the corner plots (b) of α-DPI samples using the model with the largest ELBO (a crescent with two ellipses). As seen in the corresponding table on the top right of Figure 7, the recovered posterior of black hole features aligns well with the sampling-based results first presented in Event Horizon Telescope Collaboration et al. (2019c).

In addition to the inference accuracy, we also quantitatively study α-DPI's efficiency by comparing its computational time to a representative nested sampling method (Speagle 2020). In this section, the computational time is defined as the training time for α-DPI to converge to the optimal loss or for nested sampling to reach the stationary state. Figure 7 presents results obtained using models with increasing complexity, ranging from 6 to 1024 dimensions. As shown in the top left of Figure 7, α-DPI scales very well with the dimension of the model, remaining within the same order of magnitude for computational time across all models (refer to the computational time log-plot of Figure 7). As α-DPI is primarily an optimization-based problem, its computational time directly depends on the normalizing flow network architecture (depth,
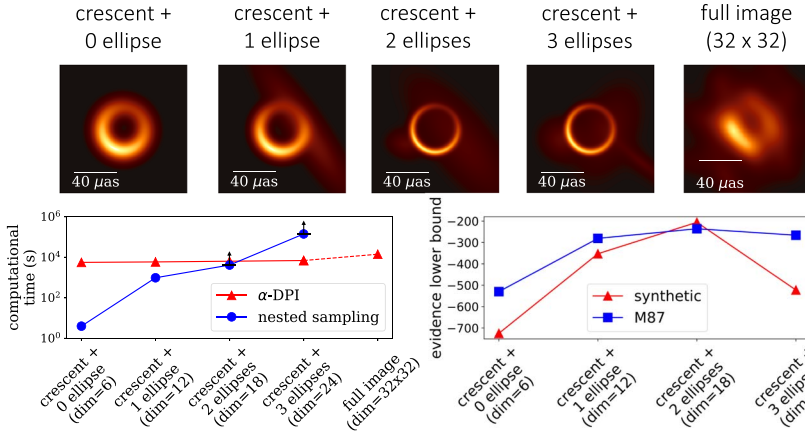
width, activation function, etc.) instead of the model dimension. As introduced in Section 3.3, our experiments only need to use neural network architectures with the same depth (the number of affine coupling layers) but varying widths (the number of neurons in each layer) to handle geometric models with different dimensions. α-DPI's computational time only slightly changes as the neural network width proportionally increases with the model dimension since it can make use of GPU parallel acceleration. As a comparison, nested sampling cannot be easily parallelized, so its computational time increases significantly with the model dimension (Salomone et al. 2018). Note that since α-DPI is much more computationally efficient than the baseline sampling methods, it easily scales up to estimating posteriors of full images with $32 \times 32 = 1024$ parameters (i.e., pixels), which was demonstrated in Sun & Bouman (2021) using KL-DPI. However, such a high-dimensional full image uncertainty quantification task is not easily achievable using current sampling-based methods, and thus for computational reasons previous methods are restricted to characterizing images with far fewer pixels (Broderick et al. 2020). α-DPI's efficiency will become increasingly critical in future observations using the next-generation EHT (ngEHT) where we would like to fit higher-dimensional geometric models or images to understand higher-resolution features of black holes (Raymond et al. 2021).

Implementation Details. The observational data in VLBI black hole observations are the Fourier components of the astronomical signals, which are referred to as "visibilities" in radio astronomy (Thompson et al. 2017). In practical observations, the visibilities are usually contaminated by atmospheric turbulence, so we define the α-DPI target posterior using two robust data products derived from visibilities: closure phase and closure amplitude (Thompson et al. 2017; Chael et al. 2018). Since the exact likelihood of these closure quantities cannot be explicitly defined, we use Gaussian distributions to approximate their likelihood in our implementation. This Gaussian approximation is reasonable with high signal-to-noise ratio (SNR) measurement data. We use $\alpha = 0.9$ in the α-DPI divergence loss for the black hole imaging; this α value is empirically chosen by finding the lowest α where fewer than

---

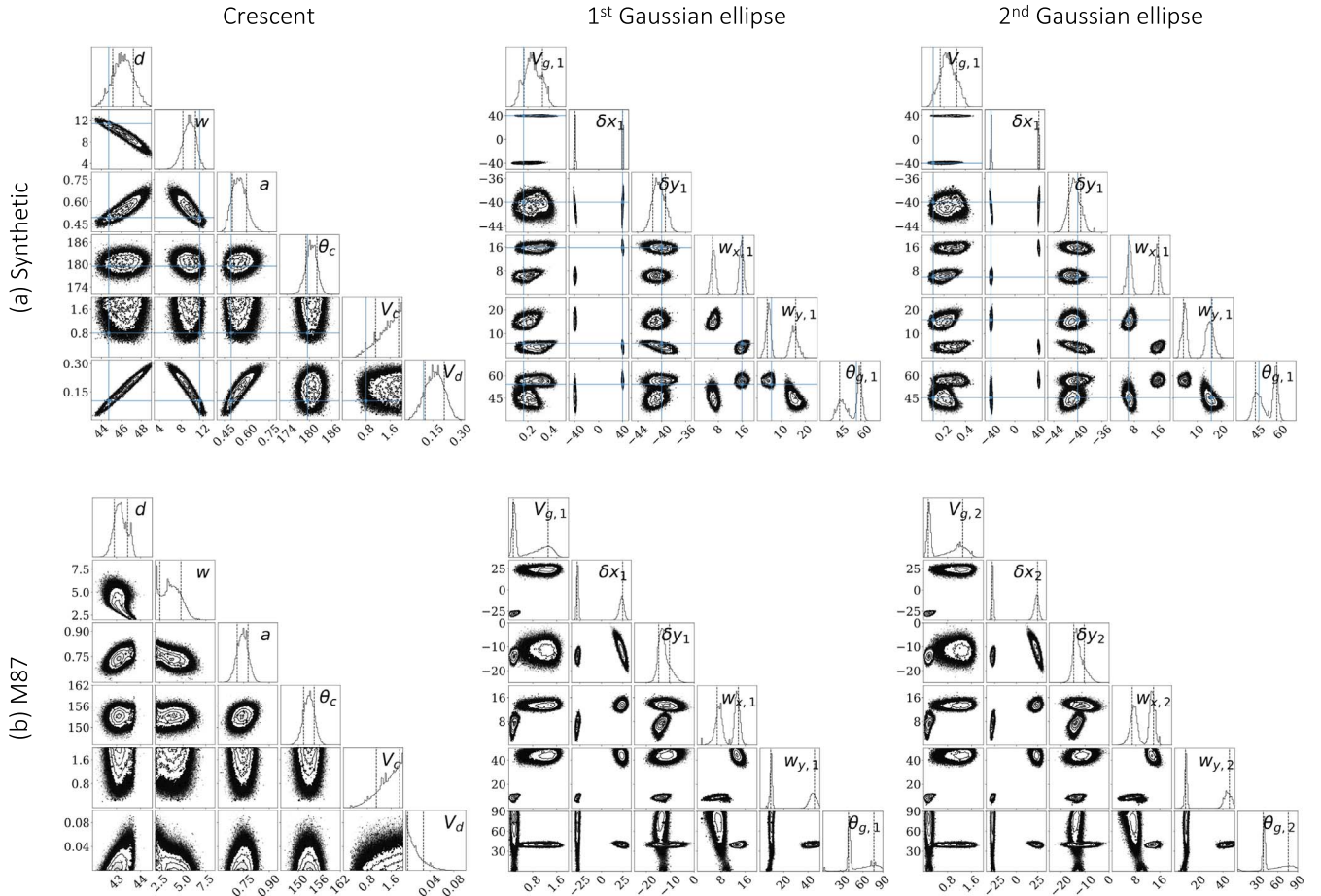[7] This data set is publicly available online at https://eventhorizontelescope.org/for-astronomers/data.

	crescent + 0 ellipse	crescent + 1 ellipse	crescent + 2 ellipses	crescent + 3 ellipses	full image (32 x 32)

Comparison between the α-DPI estimated M87* parameters (2 Gaussian ellipses) and the original M87* feature extraction results (Event Horizon Telescope Collaboration et al. 2019c).

Code	THEMIS	dynesty	α-DPI
d (μas)	$43.5^{+0.14}_{-0.14}$	$43.4^{+0.27}_{-0.26}$	$43.2^{+0.32}_{-0.25}$
$\log_{10}(\hat{s})$	$-0.87^{+0.09}_{-0.20}$	$-1.31^{+0.31}_{-0.52}$	$-1.00^{+0.12}_{-0.17}$
$\log_{10}(V_d)$	$-2.14^{+0.43}_{-0.62}$	$-2.63^{+0.41}_{-0.60}$	$-1.94^{+0.40}_{-0.62}$
a	—	—	$0.74^{+0.031}_{-0.032}$
θ_c ($\circ$)	$153.0^{+2.0}_{-2.4}$	$148.5^{+1.4}_{-1.2}$	$153.1^{+1.5}_{-1.6}$

* $\hat{s} = \frac{2\sigma\sqrt{2\ln 2}}{d}$ is the crescent's sharpness.

Figure 7. (Top) Computational times and ELBO values obtained with models of increasing complexity. The computational time of nested sampling grows more than polynomially as the geometric model dimension increases. In contrast, the computational time of α-DPI stays within the same order of magnitude. Blue arrows on the sampling method's computational time indicate that there is evidence that this method has not fully converged in the allotted time. A crescent with two elliptical Gaussian ellipses is identified as the best model for both synthetic and real M87* data according to the ELBO. This choice correctly identifies the true underlying model used for generated synthetic data. It also agrees with the model choice selected and presented in Event Horizon Telescope Collaboration et al. (2019c) for the real M87* data (top right table; THEMIS and dynesty are two benchmark EHT feature extraction pipelines). (Bottom) Corner plots visualizing posterior samples of black hole geometric parameters (a crescent with two elliptical Gaussian ellipses). (a) and (b) figures are the results of synthetic and real M87* data, respectively. In the synthetic data experiment, the ground-truth values (blue lines) are contained in the recovered posterior. Since the two Gaussian ellipses are interchangeable, in both synthetic and real data experiment α-DPI correctly identifies that the marginal posteriors of two Gaussian ellipses are identical. The contours in the corner plots represent the levels of the posterior likelihoods; points lying outside the posterior contours are the samples whose likelihoods are smaller than a threshold but not negligible.

50% of unique α-DPI samples are rejected by importance resampling. In the 32×32 image posterior estimation, rather than using the prior distribution defined for the geometric model in

Figure 6, we introduce an image prior/regularizer that combines maximum entropy (Skilling & Bryan 1984) and total squared variation (Bouman & Sauer 1993; Kuramochi et al. 2018)

regularizers; since the observational data is sparse, adding an image regularizer is necessary to recover an interpretable posterior that highlights black hole images with desirable low-level image statistics.

2. Discussion

α-DPI provides a general framework for uncertainty quantification in Bayesian inference problems. It is broadly applicable to inference problems in astronomical science: the only requirement is a differentiable forward model, which is usually available (e.g., black hole interferometry) or can be approximated using numerical methods (e.g., exoplanet astrometry). α-DPI merges the strengths of both sampling-based and optimization-based inference methods, resulting in a very efficient and scalable approach that still maintains posterior estimation accuracy. By pairing a normalizing generative model along with an α-divergence VI loss, α-DPI is less frequently plagued by the mode collapse failures than KL-DPI. All the above strengths make α-DPI a promising new method in not only astronomy, but also many other high-dimensional scientific inference problems, such as seismic tomography (Gao et al. 2021), inverse material design (Sanchez-Lengeling & Aspuru-Guzik 2018), and biomedical imaging (Sun & Bouman 2021).

3. Method Background and Details

In this section, we first describe in detail the two steps of α-DPI: (1) α-divergence VI (α-VI) with a normalizing flow neural network (Section 3.1), and (2) importance sampling (Section 3.2). Then we describe the implementation details (Section 3.3), including neural network architectures and optimization choices made for the two astronomical applications studied in this work.

3.1. α-divergence Variational Inference (α-VI) with Normalizing Flow

VI is an approach that solves an optimization problem to estimate a posterior distribution. In VI, a family of density functions parameterized by θ, $q_\theta(x)$, is defined. Then, optimization algorithms are used to find the parameters, $\theta^\star$, that best match the variational density function to the target posterior distribution, $p(x|y)$. The optimization problem of VI is

$$\theta^\star = \arg\min_\theta D[q_\theta(x)\|p(x|y)], \qquad (5)$$

where $D[\cdot\|\cdot]$ is a divergence function that measures the similarity between two distributions. Two major factors that influence the performance of VI are the objective divergence function and the modeling capacity of the variational density function.

3.1.1. Renyi's α-divergence

In traditional VI, KL divergence (Kullback & Leibler 1951) is typically used as the objective loss function:

$$\theta^\star = \arg\min_\theta D_{KL}[q_\theta(x)\|p(x|y)]$$
$$= \arg\min_\theta \mathbb{E}_{x\sim q_\theta(x)}[-\log p(x|y) + \log q_\theta(x)]$$
$$= \arg\min_\theta \mathbb{E}_{x\sim q_\theta(x)}[-\log p(y|x) - \log p(x) + \log q_\theta(x)].$$
$$(6)$$

This KL loss is equivalent to jointly optimizing the MAP loss (first two terms) along with the entropy of the variational distribution (last term). Although KL-VI theoretically should produce an optimized density function that well matches the true posterior, in practice KL-VI usually leads to zero-forcing effects. In other words, the variational model tries to avoid including samples from low-probability regions, often resulting in underestimating the posterior and thus the uncertainty of inferred parameters. Zero-forcing effects are especially detrimental when estimating posteriors with disconnected modes.

In α-DPI, we alternatively define our VI objective function using the Renyi's α-divergence (Li & Turner 2016):

$$\theta^\star = \arg\min_\theta D_\alpha[q_\theta(x)\|p(x|y)]$$
$$= \arg\min_\theta \frac{1}{\alpha-1} \log \mathbb{E}_{x\sim q_\theta(x)}\left[\left(\frac{p(x|y)}{q_\theta(x)}\right)^{1-\alpha}\right]$$
$$= \arg\min_\theta \frac{1}{\alpha-1} \log \mathbb{E}_{x\sim q_\theta(x)}$$
$$\times \{\exp[(1-\alpha)(\log p(y|x) + \log p(x) - \log q_\theta(x))]\}.$$
$$(7)$$

Renyi's α-divergence is a more general class of similarity metrics between two distributions $q_\theta(x)$ and $p(x|y)$—when $\alpha \to 1$ α-divergence converges to a KL divergence, $D_{KL}[q_\theta(x)\|p(x|y)]$, making α-VI the same as KL-VI, which only exploits samples from $q_\theta(x)$ to compute the similarity between distributions; when $\alpha \to 0$ α-divergence converges to $-\log \int_{q_\theta(x)>0} p(x)dx$, making α-VI the same as a maximum likelihood estimation (MLE) of θ (Amari et al. 2001), which needs to explore the full probability space of $p(x)$. The former case is very efficient in computation but sometimes lacking in posterior estimation accuracy, while the latter one produces a very accurate posterior but is computationally slow. By tuning the α value in $D_\alpha[q_\theta(x)\|p(x|y)]$, we balance the exploitation and the exploration in our posterior estimation, so α-DPI can be both efficient and accurate. In contrast with KL-VI, α-VI ($0 < \alpha < 1$) can tolerate a few samples with low likelihood if it leads to a distribution that better captures multiple modes in the true posterior. Please refer to the Appendix for more detailed derivations.

3.1.2. Normalizing Flow

In order to solve the stochastic optimization problem in Equation (7) efficiently, the variational density function should be efficient in both sampling ($x\sim q_\theta(x)$) and evaluating the likelihood of a sample (i.e., computing $\log q_\theta(x)$). Traditional VI typically uses a simple family of density functions, such as those from the exponential family, to facilitate efficient sampling and evaluation. However, these simple variational models sacrifice the ability to capture complex multimodal distributions. To improve inference accuracy, in α-DPI we propose to use a more flexible neural network based density function: a normalizing flow network. Normalizing flows are a class of deep generative models that are widely used in computer vision and machine learning for complex density function approximation. They parameterize a probability density $q_\theta(\cdot)$ in an implicit manner by transforming a simple

base distribution $\pi(\cdot)$ using an invertible neural network,

$$x = g_\theta(z), \quad z = g_\theta^{-1}(x), \tag{8}$$

where $z \sim \pi(z)$ is an arbitrary sample from the base distribution (e.g., independent and identically distributed Gaussian), and $x \sim q_\theta(x)$ is a sample from the approximated distribution $q_\theta(x)$. According to the "change of variables theorem,"

$$q_\theta(x) = \pi(z) \left| \det \frac{dg_\theta(z)}{dz} \right|^{-1} = \pi(g_\theta^{-1}(x)) \left| \det \frac{dg_\theta^{-1}(x)}{dx} \right|, \tag{9}$$

where $\det \frac{dg_\theta^{-1}(x)}{dx}$ is the determinant of the neural network function's Jacobian matrix. Therefore, the probability of a random sample from $q_\theta(x)$ is easy to evaluate when the Jacobian matrix is computationally tractable, as is the case with commonly used normalizing flow forms including NICE (Dinh et al. 2014), Real-NVP (Dinh et al. 2016), and Glow (Kingma & Dhariwal 2018). As described in Section 3.3, in this work we choose to use a Real-NVP generative network to parameterize $q_\theta(x)$.

3.2. Importance Sampling

Importance sampling is a Monte Carlo technique for estimating posterior distributions. It first generates random samples based on a proposal distribution $q(x)$, and then reweights the samples based on their posterior probabilities $p(x|y)$, where the weight of each sample x_j is $w_j = \frac{p(x_j \mid y)}{q(x_j)} \propto \frac{p(y \mid x_j)p(x_j)}{q(x_j)}$. In α-DPI, importance sampling is applied after α-VI to further improve the accuracy of the normalizing flow approximated posterior.

A normalizing flow, by definition, is a continuous bijective function. As a result, to capture multiple disconnected modes in a posterior distribution, a normalizing flow model must include low-probability samples that connect the modes. Therefore, to approximate the posterior more accurately, these samples should be removed from the approximated distribution. Since the learned normalizing flow distribution $q_\theta(x)$ is already close to the true posterior after α-VI, $w_j = \frac{p(x_j \mid y)}{q(x_j)} \approx 1$ and for most samples $x_j \sim q_\theta(x)$. In this case, applying the importance reweighting to samples from the normalizing flow distribution is efficient because very few samples are rejected. The computational time of importance sampling is negligible compared with α-VI; however, it produces a cleaner posterior estimation, as shown in Figure 3.

3.3. Implementation Details

In both exoplanet astrometry and black hole feature extraction problems, we use a Real-NVP model with 32 affine coupling layers as the variational density function (Dinh et al. 2016). Each affine coupling layer is composed of a neural neural network with three fully connected layers, where the width (the number of neurons) of each fully connected layer is 16 times the dimension of inferred parameters.

The VI optimization problem is solved using an Adam optimizer (Kingma & Ba 2014). In addition, we apply simulated annealing training (Huang et al. 2018) to avoid the normalizing flow from stopping exploration in early epochs, and subsequently converging to a poor local minimum. Instead of directly optimizing the α-divergence in Equation (7), we define an annealed objective function,

$$D_\alpha^i[q_\theta(x) \| p(x|y)] = \arg\min_\theta \frac{1}{\alpha - 1} \log \mathbb{E}_{x \sim q_\theta(x)}$$
$$\times \left\{ \exp\left[(1-\alpha)\left(\frac{1}{\beta_i} \log p(y|x) + \frac{1}{\beta_i} \log p(x) \right.\right.\right.$$
$$\left.\left.\left. - \log q_\theta(x) \right) \right] \right\}, \tag{10}$$

where i is the index of the optimization epoch, $\beta_i = \max\{1, \beta_0 \exp\left(-\frac{i}{\tau}\right)\}$, β_0 is the initial annealing weight, and τ is the decay period. Since the normalizing flow network used is initialized to an approximately random uniform distribution, the data likelihood ($\log p(y|x)$) and the prior likelihood ($\log p(x)$) are typically much larger than the entropy term ($\log q_\theta(x)$) at the initiation of training. β_j balances the values of these different terms so that the optimization becomes more numerically stable and the normalizing flow can gradually converge from a random uniform distribution to one that well approximates the posterior. In exoplanet astrometry, we use an initial annealing weight $\beta_0 = 10^4$, a decay period $\tau = 3000$ and in total 20,000 training epochs. In black hole feature extraction, the initial annealing weight, the decay period, the total number of epochs are $\beta_0 = 10^3$, $\tau = 3000$, and $epoch = 15,000$, respectively.

Appendix
More Comparison of KL-DPI and α-DPI

Both KL-DPI and α-DPI losses are approximated as stochastic forms when we solve for the normalizing flow

weights:

$$
\begin{aligned}
&D_{\mathrm{KL}}[q_\theta(x)\|p(x|y)] \\
&= \mathbb{E}_{x\sim q_\theta(x)}[-\log p(y|x) - \log p(x) + \log q_\theta(x)] \\
&\approx \sum_{n=1}^{N}[-\log p(y|g_\theta(z_n)) - \log p(g_\theta(z_n)) \\
&\quad - \log\left|\det\frac{dg_\theta(z_n)}{dz_n}\right| + \log\pi(z_n)], \\
&D_\alpha[q_\theta(x)\|p(x|y)] = \frac{1}{\alpha-1}\log\mathbb{E}_{x\sim q_\theta(x)} \\
&\quad \times \{\exp[(1-\alpha)(\log p(y|x) + \log p(x) - \log q_\theta(x))]\} \\
&\approx \frac{1}{\alpha-1}\log\sum_{n=1}^{N}\exp\{(1-\alpha)(\log p(y|g_\theta(z_n)) \\
&\quad +\log p(g_\theta(z_n)) + \log\left|\det\frac{dg_\theta(z_n)}{dz_n}\right|) - \log\pi(z_n)\},
\end{aligned}
$$
$$\text{(A1)}$$

where $z_n \sim \pi(\cdot)$ is an arbitrary sample from the base distribution, and N is the number of samples used for Monte Carlo approximation. Denoting the KL loss function of a particular sample as $L_\theta(z_n) = -\log p(y|g_\theta(z_n)) - \log p(g_\theta(z_n)) - \log\left|\det\frac{dg_\theta(z_n)}{dz_n}\right| + \log\pi(z_n)$, the gradient of the KL divergence and the α-divergence are, respectively

$$
\begin{aligned}
&\nabla D_{\mathrm{KL}}[q_\theta(x)\|p(x|y)] \approx \sum_{n=1}^{N}\nabla L_\theta(z_n) \\
&\nabla D_\alpha[q_\theta(x)\|p(x|y)] \approx \frac{1}{\alpha-1}\nabla\log \\
&\quad \times \left\{\sum_{n=1}^{N}\exp[-(1-\alpha)L_\theta(z_n)]\right\} \\
&\quad = \sum_{n=1}^{N}w_n\nabla L_\theta(z_n),
\end{aligned}
$$
$$\text{(A2)}$$

where $w_n = \mathrm{Softmax}\{[-(1-\alpha)L_\theta(z_n)]\}$. The derivative of α-divergence is a weighted version of the KL divergence's derivative gradient such that the samples that lead to a large loss have significantly less impact on gradient descent optimization in α-DPI when compared to KL-DPI. As a result, α-DPI allows the learned normalizing flow generative model to include a few "bad" samples as long as the resulting distribution $q_\theta(x)$ better captures the modes in the true posterior distribution.

ORCID iDs

He Sun ⓘ https://orcid.org/0000-0003-1526-6787
Paul Tiede ⓘ https://orcid.org/0000-0003-3826-5648
Jason J. Wang ⓘ https://orcid.org/0000-0003-0774-6502
Sarah Blunt ⓘ https://orcid.org/0000-0002-3199-2888
Dimitri Mawet ⓘ https://orcid.org/0000-0002-8895-4735

References

Amari, S.-i., Ikeda, S., & Shimokawa, H. 2001, Advanced Mean Field Methods (Cambridge, MA: MIT Press), 241
Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123
Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33
Bate, M., Lodato, G., & Pringle, J. 2010, MNRAS, 401, 1505
Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393
Blunt, S., Nielsen, E. L., De Rosa, R. J., et al. 2017, AJ, 153, 229
Blunt, S., Wang, J. J., Angelo, I., et al. 2020, AJ, 159, 89
Bouman, C., & Sauer, K. 1993, ITIP, 2, 296
Broderick, A. E., Pesce, D. W., Tiede, P., Pu, H.-Y., & Gold, R. 2020, ApJ, 898, 9
Chael, A. A., Johnson, M. D., Bouman, K. L., et al. 2018, ApJ, 857, 23
Dinh, L., Krueger, D., & Bengio, Y. 2014, arXiv:1410.8516
Dinh, L., Sohl-Dickstein, J., & Bengio, S. 2016, arXiv:1605.08803
Event Horizon Telescope Collaboration, Akiyama, K., Alberdi, A., et al. 2019a, ApJL, 875, L4
Event Horizon Telescope Collaboration, Akiyama, K., Alberdi, A., et al. 2019b, ApJL, 875, L2
Event Horizon Telescope Collaboration, Akiyama, K., Alberdi, A., et al. 2019c, ApJL, 875, L6
Ferland, G. J., Porter, R. L., van Hoof, P. A. M., et al. 2013, RMxAA, 49, 137
Ferrer-Chávez, R., Wang, J. J., & Blunt, S. 2021, AJ, 161, 241
Gao, A., Castellanos, J., Yue, Y., Ross, Z., & Bouman, K. 2021, in Advances in Neural Information Processing Systems 34 (NeurIPS 2021) (Red Hook, NY: Curran Associates, Inc.), 34
Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. 2012, JMLR, 13, 723
Huang, C.-W., Tan, S., Lacoste, A., & Courville, A. 2018, arXiv:1809.01818
Kingma, D. P., & Ba, J. 2014, arXiv:1412.6980
Kingma, D. P., & Dhariwal, P. 2018, arXiv:1807.03039
Kullback, S., & Leibler, R. A. 1951, Ann.Math. Stat., 22, 79
Kuramochi, K., Akiyama, K., Ikeda, S., et al. 2018, ApJ, 858, 56
Lacour, S., Wang, J., Rodet, L., et al. 2021, A&A, 654, L2
Li, Y., & Turner, R. E. 2016, arXiv:1602.02311
Maire, A.-L., Rodet, L., Cantalloube, F., et al. 2019, A&A, 624, A118
Owen, A. B. 2013, Monte Carlo Theory, Methods and Examples, https://artowen.su.domains/mc/
Penny, W. D. 2012, Neuroimage, 59, 319
Raymond, A. W., Palumbo, D., Paine, S. N., et al. 2021, ApJS, 253, 5
Rényi, A. 1961, in Proc. of the Fourth Berkeley Symp. on Mathematical Statistics and Probability, 1: Contributions to the Theory of Statistics (Berkeley, CA: Univ. California Press), 547
Salomone, R., South, L. F., Drovandi, C. C., & Kroese, D. P. 2018, arXiv:1805.03924
Sanchez-Lengeling, B., & Aspuru-Guzik, A. 2018, Sci, 361, 360
Scharf, C., & Menou, K. 2009, ApJL, 693, L113
Schölkopf, B. 1997, PhD thesis, Oldenbourg München, Germany
Skilling, J., & Bryan, R. 1984, MNRAS, 211, 111
Speagle, J. S. 2020, MNRAS, 493, 3132
Sun, H., & Bouman, K. L. 2021, in Proc. of the AAAI Conf. on Artificial Intelligence Queries 35 (Palo Alto, CA: AAAI), 2628
Thompson, A. R., Moran, J. M., & Swenson, G. W. 2017, Interferometry and Synthesis in Radio Astronomy (Berlin: Springer Nature)
Van Erven, T., & Harremos, P. 2014, ITIT, 60, 3797
Vousden, W., Farr, W. M., & Mandel, I. 2016, MNRAS, 455, 1919
Wang, J. J., Graham, J. R., Pueyo, L., et al. 2016, AJ, 152, 97
Yu, Q., & Tremaine, S. 2001, AJ, 121, 1736
Zhang, C., Bütepage, J., Kjellström, H., & Mandt, S. 2018, ITPAM, 41, 2008