Systematic Uncertainties and Data Complexity in Normalizing Flows

Yonatan Kahn
Department of Physics and Center for Artificial Intelligence Innovation
University of Illinois Urbana-Champaign
Urbana, Illinois 61801
yfkahn@illinois.edu

Sandip Roy
Department of Physics
Princeton University
Princeton, NJ 08544
sandiproy@princeton.edu

Jessie Shelton
Department of Physics
University of Illinois Urbana-Champaign
Urbana, Illinois 61801
sheltonj@illinois.edu

Victoria Tiki
Department of Physics
University of Illinois Urbana-Champaign
Urbana, Illinois 61801
vtiki2@illinois.edu

Abstract

Normalizing flows are a powerful technique for inferring probability distributions from finite samples, a highly relevant task across the physical sciences. Using two toy examples from astrophysics, we investigate the interplay between two different sources of uncertainty in normalizing flow analyses: varying the draws from the training distribution (data variance) versus varying the network initialization (initialization variance). We find that for sufficiently large training sets, initialization variance dominates for "simple" distributions while data variance dominates for more "complex" distributions, as measured by the Kullback-Leibler divergence. This suggests that normalizing flows trained on real-world datasets may (fortunately) be robust against initialization choices.

1 Introduction

Normalizing flows (1) (NFs) are flexible and invertible neural networks capable of density estimation, namely modeling complex and high-dimensional probability distributions from unlabeled training data comprising a finite sample from the target distribution. A particularly important distribution in astrophysics and astronomy is the classical 6-dimensional phase space distribution $f(\mathbf{r}, \mathbf{v})$ of stars in galaxies[1], from which many galactic properties can be inferred, such as the distribution of dark matter. Classical density estimation techniques for $f(\mathbf{r}, \mathbf{v})$ suffer from the curse of dimensionality, but the somewhat surprising generalization abilities of NFs (perhaps coming from implicit or explicit regularization (2)) have been shown to be a promising tool on simulated data (3; 4; 5; 6; 7). Recently, Ref. (8) applied this technique to a subset of data from the *Gaia* survey (9; 10) consisting of 5.8×10^6

[1] $\mathbf{r}, \mathbf{v}$ refer to three-dimensional spatial coordinates and three-dimensional velocities respectively.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

stars within 4 kpc of the Sun[2] for which full 6-dimensional phase space measurements are available, and inferred a local dark matter density of $0.47 \pm 0.05\ \mathrm{GeV/cm^3}$.

In the physical sciences, quantifying errors on measurements is paramount. The quoted error bars in Ref. (8) include numerous sources of uncertainty, including measurement uncertainty, statistical uncertainty from the finite training data, and fit uncertainty. The latter was measured empirically by training an ensemble of networks with different initializations on the same training data (which we refer to as *initialization variance*), with the result that the spread of the ensemble of trained distribution functions was negligible compared to the statistical uncertainty from training the same initialized network on different draws from a simulated data distribution (which we dub *data variance*). Given the flexibility and power of these NF networks in modeling physical datasets, understanding the relative size of the errors associated with initialization and data variance is critical. Underestimating the effects of the former can result in an underestimate in the final error associated with the physical inference. On the other hand, if initialization variance can be confidently estimated to be negligible for a sufficiently large training set, the results of the analysis can be considered more robust to the many arbitrary hyperparameter choices inherent in neural network analyses.

In this work, inspired by the results of Ref. (8), we begin an exploratory investigation into how the relative size of initialization and data variance depend on the target distribution and the size of the training set. To this end, we compute the variance of the validation error for two benchmark astrophysical datasets (a spherically symmetric Plummer sphere potential and an axisymmetric Miyamoto-Nagai disk potential), varying over the training dataset size. We also perturb these benchmark datasets by adding in a fraction of the total mass as an idealized stellar stream and varying the stream fraction, as a toy model for deviations from a "simple" base distribution. We find the somewhat surprising result that for the unperturbed base distributions, initialization variance seems to be comparable to or greater than data variance for any sufficiently large training set. As the stream fraction is increased in both cases, we observe a crossover where data variance begins to dominate for large training sets. We conjecture that the crossover may happen at a particular value of the Kullback-Leibler (KL) divergence of the perturbed distribution relative to the base distributions, a measure of data complexity. Our preliminary conclusion is that, fortunately, real-world datasets such as *Gaia* which are sufficiently complex may be generically robust against initialization variance. Throughout this paper, we work in units where the Newtonian gravitational constant $G = 1$.

2 Datasets and Methods

We use two benchmark datasets to quantify the initialization and data variances of our NFs, both of which are often used to model astrophysical data and have known phase space densities which can be carefully perturbed and studied. The first is a Plummer sphere potential (12), which has been used to model globular clusters (13), and a Miyamoto-Nagai (MN) disk potential (14), which has been used to model the disk potential of the Milky Way (14; 15). The Plummer sphere potential (Φ_{Plummer}) and density profile (ρ_{Plummer}), from which we sample to train our NFs, are as follows:

$$\Phi_{\mathrm{Plummer}}(r) = -(r^2 + a^2)^{-1/2}; \qquad \rho_{\mathrm{Plummer}}(r) = \frac{3}{4\pi}(r^2 + a^2)^{-5/2}, \tag{1}$$

where r is the radial distance from the center. The Plummer sphere is spherically symmetric and is characterised by a flat inner density profile (i.e. $\rho \sim r^0$) at distances $r \ll a$ and a slope of -5 at large radial distances, $r \gg a$. When sampling particle positions using ρ_{Plummer}, we set the scale radius $a = 1$ for simplicity. To sample the velocity phase space, we use the following phase space distribution $f(\mathbf{r}, \mathbf{v})$:

$$f(\mathbf{r}, \mathbf{v}) \propto \begin{cases} [-E(\mathbf{r}, \mathbf{v})]^{7/2}, & E(\mathbf{r}, \mathbf{v}) < 0 \\ 0, & E(\mathbf{r}, \mathbf{v}) \geq 0 \end{cases} \quad \text{where} \quad E(\mathbf{r}, \mathbf{v}) = \frac{1}{2}v^2 + \Phi(r). \tag{2}$$

We normalize the total mass of the sphere so each sampled particle has an individual mass of $1/n$ where n is the total number of particles. The gravitational potential of the MN disk is axisymmetric and given by

$$\Phi_{\mathrm{MN}}(R, z) = -(R^2 + (\sqrt{z^2 + b^2} + a)^2)^{-1/2}, \tag{3}$$

[2] 1 pc ≈ 3.3 light-years. Our Sun is located ≈ 8 kpc from the galactic center (11).

where R is the radial distance in the disk plane, z is the vertical distance, b characterises the disk height and a is the scale radius in the disk plane. We set $a = 1$ and $b = 0.1$ so the disk is thin. To sample the particle positions and velocities, we use the same method as Ref. (4), drawing the initial particle positions from the following double exponential density profile:

$$\rho_{\text{initial, MN}} \propto \exp\left(-\frac{-R}{a} - \frac{|z|}{b}\right). \tag{4}$$

The initial particle velocities are sampled from the distributions

$$v_R = 0.05\, v_c(R,z)\, \delta_1, \qquad v_T = v_c(R,z)(1 - 0.1|\delta_2|), \qquad v_z = 0.05\, v_c(R,z)\, \delta_3, \tag{5}$$

where v_R, v_T, v_z are the radial disk plane velocities, the tangential disk plane velocities and vertical velocities respectively, $\delta_i \sim \mathcal{N}(0,1)$, and $v_c(R,z)$ is the circular velocity at a given (R,z) position calculated as $\sqrt{r \frac{\partial \Phi_{\text{MN}}(R,z)}{\partial r}}$. Since this initial distribution is not in equilibrium, we also evolve the system of particles along orbits governed by the MN potential with the parameters $a = 1$, $b = 0.1$, using `galpy` (16). We integrate the particle orbits over a time period of approximately 100 orbital times assuming an orbital time period $2\pi a / v_c(R = a, z = 0)$. The final distribution of positions and velocities forms our training and validation datasets. Plots of samples from our distribution functions and more details on the orbit integration are given in App. A.1.

While these distributions are standard in astrophysical settings, real data distributions are often more complex and include non-equilibrium, dynamical structures like satellite galaxies and stellar streams. In order to perturb the two base distributions in a controlled manner, we also add a toy model of a stellar stream to the $z = 0$ plane. The number of particles in each distribution is quantified by $f_{\text{stream}} = M_{\text{stream}}/(M_{\text{base}} + M_{\text{stream}})$ where M_{stream} represents the total mass of all particles in the stream. We ensure that the total mass is normalized as $M_{\text{stream}} + M_{\text{base}} = 1$ and that all particles have the same mass. The positions and velocities of the stream particles are chosen from the following distributions:

$$\phi_{\text{stream}} \sim \text{Uniform}(0, \pi/4), \qquad \theta_{\text{stream}} \sim \mathcal{N}(0, \sigma_\theta = 0.01), \tag{6}$$

$$r_{\text{stream}} \sim \mathcal{N}(r = 10, \sigma_R = 0.01), \qquad v_{\text{stream}} \sim \mathcal{N}(v_c(r = 10)\,, \sigma_v = 0.01), \tag{7}$$

where ϕ, θ and r represent the azimuthal angle, zenith angle and radial distance respectively. We ensure the direction of the stream velocity v_{stream} is a circular orbit on the plane connecting the particle position to the origin. This choice of parameters ensures that the stream particles are reasonably separated spatially from the base distributions and are also aligned fairly closely with a perfect circular orbit in the $z = 0$ plane. We show visually the effect of perturbing the base distributions in App. A.1.

Our general training procedure is as follows. We implement a Masked Autoregressive Flow (MAF) using the `nflows` package (17) to model the different distributions, with six MAF blocks arranged in sequence. Each MAF block applies a Masked Affine Autoregressive Transformation to the 6-dimensional input. The transformation uses hidden layers with 32 features and 2 blocks per layer, utilizing a `GELU` activation function. To increase the expressiveness of the model, a permutation is applied between each MAF block. The base distribution of the flow is a standard normal distribution over 6 dimensions. Before training, each dataset is preprocessed by subtracting the mean and scaling by the standard deviation. The training is performed with the `Adam` optimizer (18) using a learning rate of 0.001, optimizing the negative log likelihood of the flow model. We implement early stopping with a patience of 50 epochs based on validation loss to prevent overfitting. The training data is split into batches of 1,024 particles, and the model is trained over a maximum of 1000 epochs, saving the model state when an improvement in validation loss is observed. In practice, we find our patience criterion is satisfied before reaching the maximum number of epochs.

In order to characterise the data variance σ_{data}, we randomly generate an initial sampling of N_{train} particles from the chosen phase space distribution, and initialize and train a single MAF, recording the minimum validation loss. We then repeat this process 50 times, keeping the initial weights of the MAF fixed but varying the random particle sampling for all runs. For each training set size N_{train}, we compute the standard deviation of these negative log likelihood losses, σ_{data}, from the sample variance of the ensemble. We repeat this process for different training set sizes. In all the calculations of validation loss and σ_{data}, we ensure that the same validation dataset containing 10^6 sample particles is used, regardless of N_{train}, to capture the performance of the MAF on a

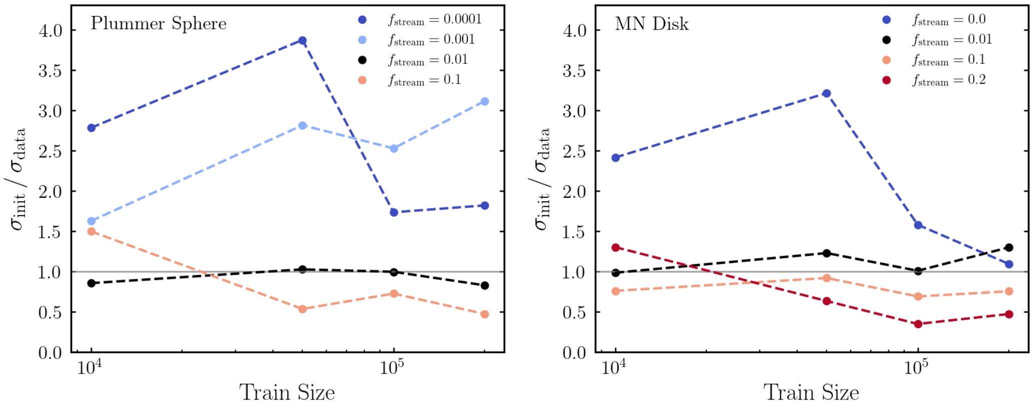

Figure 1: *Left panel*: Ratio of initialization variance to data variance for different training set sizes and different stream fractions ($f_{\rm stream}$) using the Plummer sphere as a base distribution. Each data point is the result of comparing 50 trained normalizing flows with varied initialization but fixed training dataset to 50 other trained normalizing flows with fixed initialization but varied datasets. *Right panel*: Same as left panel but using the MN disk potential as the base distribution. For both base distributions, increasing $f_{\rm stream}$ lowers the relative magnitude of $\sigma_{\rm init}$ compared to $\sigma_{\rm data}$. We plot the mean losses and their variances in App. A.2.

consistent dataset. We use a very similar method to calculate the initialization variance, $\sigma_{\rm init}$, with the only difference being that we fix the particle sampling for each $N_{\rm train}$ but vary the MAF weight initialization using the Kaiming uniform initialization (19). Thus, for each $N_{\rm train}$, we train our MAF 50 times with different random initializations, recording the minimum log likelihood loss. We then compute the standard deviation of these losses, $\sigma_{\rm init}$, from the sample variance of the ensemble. The separate mean and variances of the loss values for each normalizing flow ensemble are plotted and discussed in App. A.2.

To train all normalizing flows, we use a single A100 GPU. The total training time for 50 flows and $N_{\rm train} = 200,000$ is ~ 10 hours wall time. All computations were performed using the Princeton Research Computing resources at Princeton University which is a consortium of groups led by the Princeton Institute for Computational Science and Engineering (PICSciE) and the Office of Information Technology's Research Computing. A link to our code base is given in Appendix A.4.

3 Results and Discussion

In Fig. 1, we plot $\sigma_{\rm init}/\sigma_{\rm data}$, the ratio of initialization to data variance, for the Plummer sphere and MN disk base distributions. For the Plummer sphere (left panel), we find the initialization variance dominates over data variance for low $f_{\rm stream}$ values with ratios $\sigma_{\rm init}/\sigma_{\rm data} \gtrsim 2$ even at large training set sizes of 10^5 and greater, where for $f_{\rm stream} = 10^{-3}$ there are on average 100 stream particles in the training data. However, for greater stream fractions ($f_{\rm stream} \gtrsim 10^{-2}$), we see data variance becoming comparable to (and eventually dominant over) initialization variance. This is perhaps surprising because stream particles are now better sampled in the training data, which might be expected to reduce data variance. In the right panel of Fig. 1, we show the results of the introducing an idealised stream perturbation to the MN disk. The results qualitatively resemble those of the Plummer sphere: increasing the stream fraction reduces the relative magnitude of initialization variance against data variance. Specifically, for training sizes $\geq 10^5$ and for $f_{\rm stream} \gtrsim 0.1$, the ratio $\sigma_{\rm init}/\sigma_{\rm data} < 1$.

These results may indicate that for simple NF architectures like MAFs, initialization variance may dominate when the dataset is "smooth" and without significant substructure. For more complex datasets with variegated features, as is common in astrophysical settings, data variance may be more important in assessing the final errors associated with using NFs; viewed another way, the complexity of the distribution may render the NF fit more robust against model hyperparameters. One measure of the complexity of our perturbed datasets compared to their simpler, base distributions is the KL

Base Distribution	$\mathbf{f}_{\textbf{stream}}$	$\mathbf{D}_{\textbf{KL}}$
Plummer	10^{-1}	2.54
Plummer	10^{-2}	0.35
Plummer	10^{-3}	0.16
Plummer	10^{-4}	0.15
MN Disk	2×10^{-1}	5.88
MN Disk	10^{-1}	2.83
MN Disk	10^{-2}	0.24

Table 1: KL divergence for different base distributions and stream fractions.

divergence.[3] The KL divergence between the perturbed distribution p and the base distribution q takes a particularly simple form if p is a mixture of q and a small (normalized) perturbation distribution α with weight ϵ, namely $p = (1-\epsilon)q + \epsilon\alpha$ with $\epsilon \ll 1$. In that case, we have to leading order in ϵ (denoting the integration measure $d^3r\, d^3v$ on phase space as $d\mu$ for brevity)

$$D_{\mathrm{KL}}(p||q) = \int d\mu\, ((1-\epsilon)q + \epsilon\alpha) \ln\left(\frac{(1-\epsilon)q + \epsilon\alpha}{q}\right) = \epsilon^2 \int d\mu\, \frac{(\alpha - q)^2}{2q} + \mathcal{O}\big(\epsilon^3\big), \tag{8}$$

where the linear term $\epsilon(\alpha - q)$ integrates to zero. We compute the KL divergence between the base and perturbed distributions numerically and list the values in Tab. 1 (calculation details are outlined in App. A.3). In Fig. 1, the approximate cross-over f_{stream} value for which the ratio $\sigma_{\mathrm{init}}/\sigma_{\mathrm{data}}$ becomes approximately unity occurs at $f_{\mathrm{stream}} = 0.01$ (black dots), for both the Plummer sphere and the MN disk. That corresponds to KL divergence values of 0.35 and 0.24, respectively. Thus, the cross-over point occurs at very similar KL divergence values, despite the base distributions being qualitatively different. The KL divergence between the base Plummer sphere and the MN disk, using the Plummer sphere as the reference distribution, is 8.16, an order of magnitude greater than the KL values where the cross-over occurs.

While this work has considered simple cases for the architecture and the astrophysical distributions used alongside a basic method to determine the relative importance of initialization error and data error, there are several avenues to extend this work. It would be interesting to see if our qualitative results hold with different functional forms of the perturbation distribution, including adding multiple streams, diffuse substructure meant to represent dwarf galaxies or the *Gaia* Sausage-Enceladus (20), as well as masks representing observer bias or artifacts from telescope slewing. These alterations would allow us to test our qualitative hypotheses on more realistic astrophysical datasets. Another important caveat is that we do not characterize the correlation between data variance and initialization variance in this work, although it is not hard to imagine that both quantities are correlated (21). With more compute, we also hope to test our conclusions to greater N_{train} to see if with high enough training set size, the ratios of initialization and data variances reach an asymptote or follow a different power law.

Acknowledgments

We thank Tilman Plehn, Christoph Weniger, Sung Hak Lim, and the participants of the IAIFI 2024 Summer Workshop for stimulating conversations. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of High Energy Physics, under Award Number DE-SC0023704. This work utilizes the computing resources of the HAL cluster (22) supported by the National Science Foundation's Major Research Instrumentation program, grant #1725729, as well as the University of Illinois Urbana-Champaign. The computations reported on in this paper were also partly performed using the Princeton Research Computing resources at Princeton University which is consortium of groups led by the Princeton Institute for Computational Science and Engineering (PICSciE) and Office of Information Technology's Research Computing.

[3] When we refer to the "base" distribution here, we do not mean the base Gaussian distribution used by our NFs for sampling. We specifically mean the "base" data distributions of the Plummer sphere and MN disk which are then perturbed by adding astrophysical substructure.

References

[1] I. Kobyzev, S. J. Prince and M. A. Brubaker, *Normalizing flows: An introduction and review of current methods*, *IEEE transactions on pattern analysis and machine intelligence* **43** (2020) 3964–3979.

[2] K. Cranmer, U. Seljak and K. Terao, *Machine Learning*, in *Review of Particle Physics*, ch. 41. PTEP, 2022. DOI.

[3] G. M. Green, Y.-S. Ting and H. Kamdar, *Deep Potential: Recovering the Gravitational Potential from a Snapshot of Phase Space*, *Astrophys. J.* **942** (2023) 26, [2011.04673].

[4] G. M. Green, Y.-S. Ting and H. Kamdar, *Deep Potential: Recovering the Gravitational Potential from a Snapshot of Phase Space*, **942** (Jan., 2023) 26, [2205.02244].

[5] J. An, A. P. Naik, N. W. Evans and C. Burrage, *Charting galactic accelerations: when and how to extract a unique potential from the distribution function*, **506** (Oct., 2021) 5721–5730, [2106.05981].

[6] A. P. Naik, J. An, C. Burrage and N. W. Evans, *Charting galactic accelerations – II. How to 'learn' accelerations in the solar neighbourhood*, *Monthly Notices of the Royal Astronomical Society* **511** (01, 2022) 1609–1621.

[7] M. R. Buckley, S. H. Lim, E. Putney and D. Shih, *Measuring Galactic dark matter through unsupervised machine learning*, *Mon. Not. Roy. Astron. Soc.* **521** (2023) 5100–5119, [2205.01129].

[8] S. H. Lim, E. Putney, M. R. Buckley and D. Shih, *Mapping Dark Matter in the Milky Way using Normalizing Flows and Gaia DR3*, 2305.13358.

[9] Lindegren, L., Klioner, S. A., Hernández, J., Bombrun, A., Ramos-Lerate, M., Steidelmüller, H. et al., *Gaia early data release 3 - the astrometric solution*, *A&A* **649** (2021) A2.

[10] Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., Brown, A. G. A., Vallenari, A., Babusiaux, C. et al., *The gaia mission*, *A&A* **595** (2016) A1.

[11] T. Camarillo, V. Mathur, T. Mitchell and B. Ratra, *Median statistics estimate of the distance to the galactic center*, *Publications of the Astronomical Society of the Pacific* **130** (8, 2017) .

[12] H. C. Plummer, *On the Problem of Distribution in Globular Star Clusters: (Plate 8.)*, *Monthly Notices of the Royal Astronomical Society* **71** (03, 1911) 460–470, [https://academic.oup.com/mnras/article-pdf/71/5/460/2937497/mnras71-0460.pdf].

[13] J. Binney and S. Tremaine, *Galactic Dynamics: Second Edition*. Princeton University Press, October, 2008.

[14] M. Miyamoto and R. Nagai, *Three-dimensional models for the distribution of mass in galaxies.*, **27** (Jan., 1975) 533–543.

[15] D. A. Barros, J. R. Lépine and W. S. Dias, *Models for the 3d axisymmetric gravitational potential of the milky way galaxy - a detailed modelling of the galactic disk*, *Astronomy Astrophysics* **593** (9, 2016) A108.

[16] J. Bovy, *galpy: A python library for galactic dynamics*, *The Astrophysical Journal Supplement Series* **216** (Feb., 2015) 29.

[17] C. Durkan, A. Bekasov, I. Murray and G. Papamakarios, *nflows: normalizing flows in PyTorch*, Nov., 2020. https://doi.org/10.5281/zenodo.4296287.

[18] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 1412.6980.

[19] K. He, X. Zhang, S. Ren and J. Sun, *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification*, in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

[20] A. Helmi, C. Babusiaux, H. H. Koppelman, D. Massari, J. Veljanoski and A. G. A. Brown, *The merger that led to the formation of the Milky Way's inner stellar halo and thick disk*, **563** (Oct., 2018) 85–88, [1806.06038].

[21] B. Adlam and J. Pennington, *Understanding double descent requires a fine-grained bias-variance decomposition*, in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan and H. Lin, eds.), vol. 33, pp. 11022–11032, Curran Associates, Inc., 2020.

[22] V. Kindratenko, D. Mu, Y. Zhan, J. Maloney, S. H. Hashemi, B. Rabe et al., *Hal: Computer system for scalable deep learning*, in *Practice and Experience in Advanced Research Computing*, PEARC '20, (New York, NY, USA), p. 41–48, Association for Computing Machinery, 2020. DOI.

[23] Q. Wang, S. R. Kulkarni and S. Verdu, *A nearest-neighbor approach to estimating divergence between continuous random vectors*, in *2006 IEEE International Symposium on Information Theory*, pp. 242–246, 2006. DOI.

[24] F. Perez-Cruz, *Kullback-leibler divergence estimation of continuous distributions*, in *2008 IEEE International Symposium on Information Theory*, pp. 1666–1670, 2008. DOI.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel et al., *Scikit-learn: Machine learning in Python*, *Journal of Machine Learning Research* **12** (2011) 2825–2830.

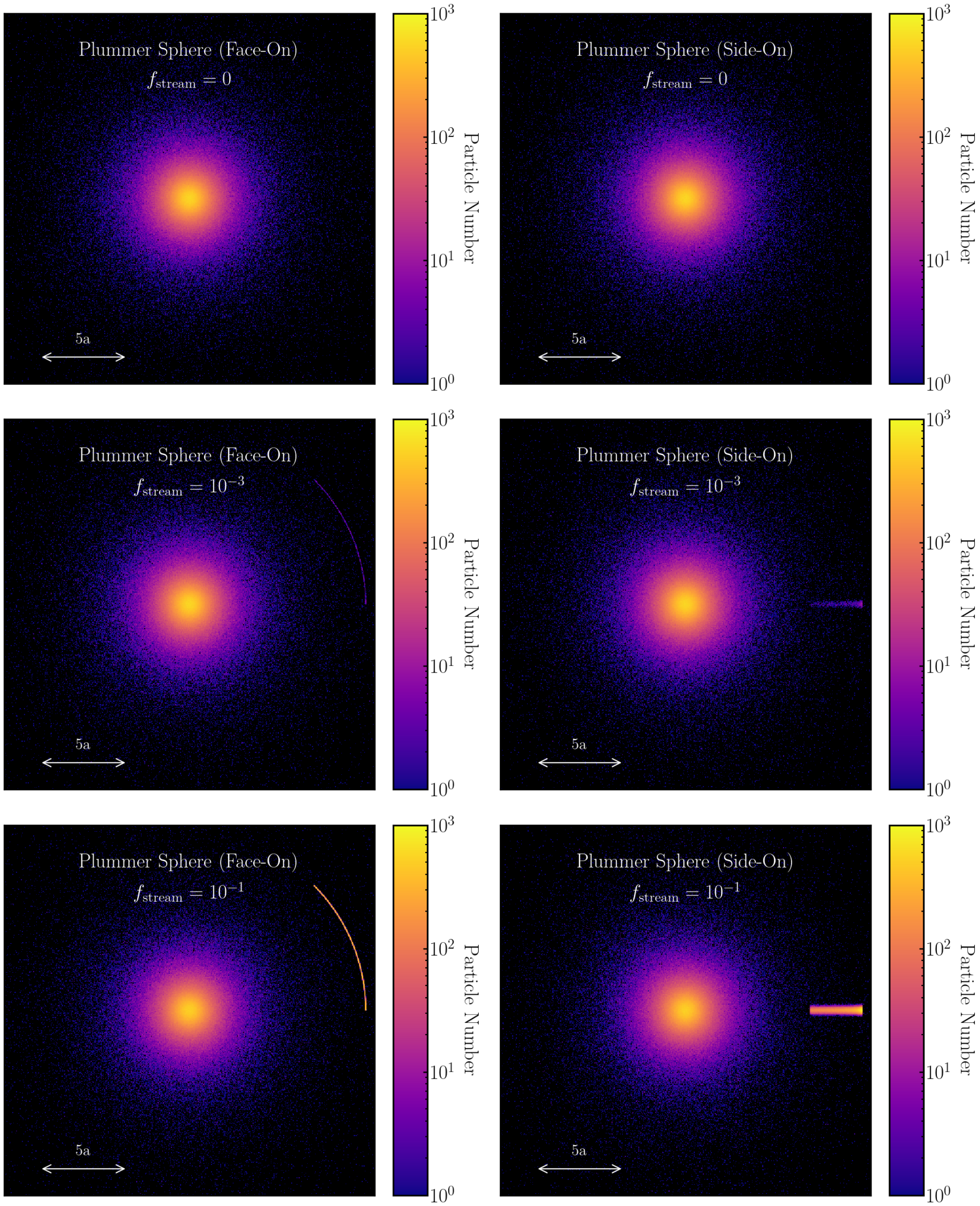

Figure 2: *Columns*: The Face-On or Side-On views of the Plummer sphere distribution but with different stream fractions ($f_{\rm stream} \in \{0,\ 0.001,\ 0.1\}$). *Rows*: The visualisations of a Plummer sphere with an idealised stream perturbation of fraction $f_{\rm stream}$, as described in Sec. 2. We include the scale bar for a distance of five scale radii (i.e. $5a$) although we set $a = 1$ for simplicity. The base Plummer sphere distribution is perfectly spherically symmetric but the stream perturbation is not.

A Appendix

A.1 Base and Perturbed Distributions

In order to visualise the base and the perturbed distributions, we display the base Plummer sphere distribution and the base Miyamoto-Nagai (MN) disk distributions with different perturbed stream fractions in Figs. 2 and 3 respectively. The main differences between both distributions is that the Plummer distribution is spherically symmetric while the MN disk distribution obeys cylindrical

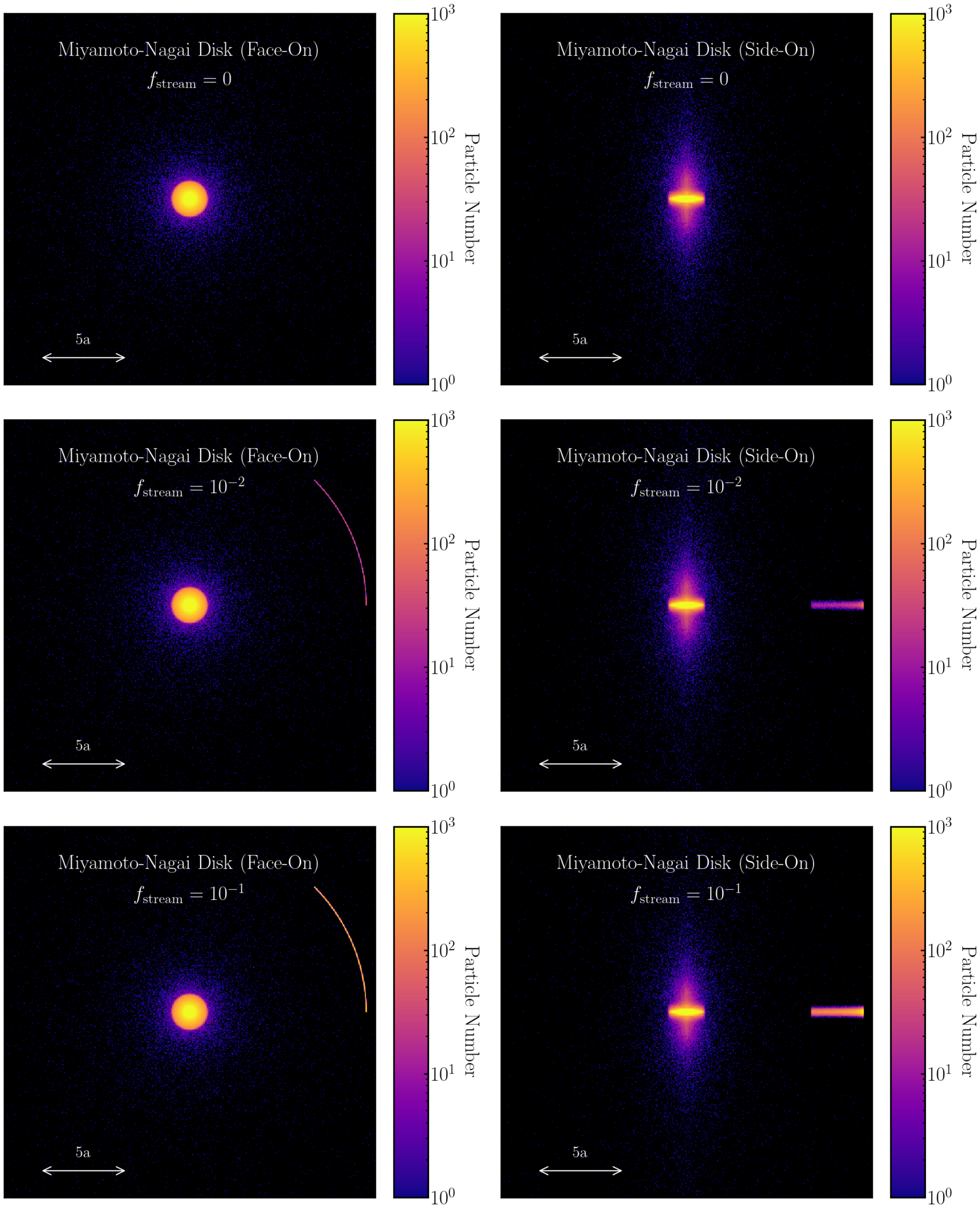

Figure 3: *Columns*: The Face-On or Side-On views of the Miyamoto-Nagai disk distribution but with different stream fractions ($f_{\mathrm{stream}} \in \{0,\, 0.01,\, 0.1\}$), similar to the format of Fig. 2. *Rows*: The visualisations of a Miyamoto-Nagai disk with an idealised stream perturbation of fraction f_{stream}. As in Fig. 2, we include the scale bar for a distance of five scale radii (i.e. $5a$) although we set $a = 1$ for simplicity. Unlike the Plummer sphere, the Miyamoto-Nagai disk is cylindrically symmetric but not spherically symmetric.

symmetry. The Plummer sphere is also more extended than the MN disk. Note that to equilibrate the MN disk, we chose a reference orbital timescale of $2\pi a/v_c(R = a, z = 0)$ over which to integrate the initial particle positions and velocities because most particles are contained within the scale radius and thus have shorter orbital times. The chosen timescale ensures that most of the particles in the system are being evolved over 100 orbital periods.

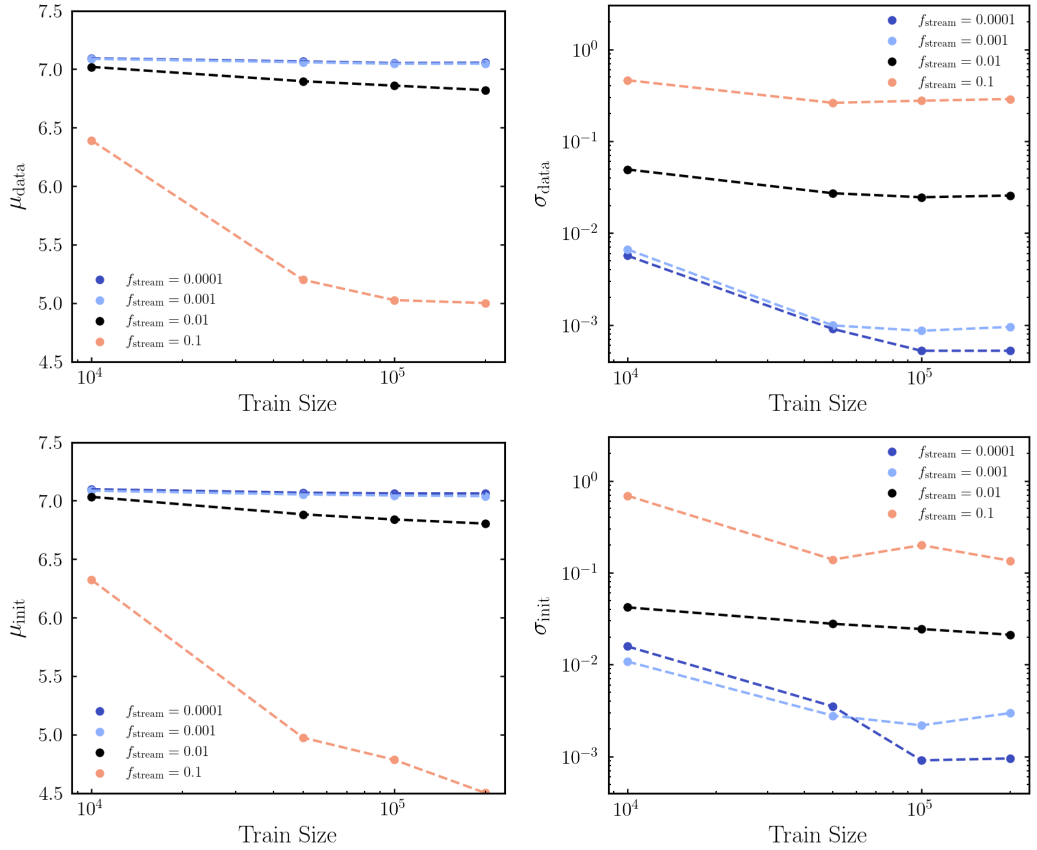

Figure 4: *Top row*: The mean (μ_{data}) and standard deviation (σ_{data}) of the validation loss for normalizing flows of fixed initial weights but varied training data samplings of the Plummer sphere distribution, varied over training set size and stream fraction. *Bottom row*: The mean (μ_{init}) and standard deviation (σ_{init}) of the validation loss for normalizing flows with varied initial weights but fixed training data samplings of the Plummer sphere, varied over training set size and stream fraction.

A.2 Model Losses and Variances

In this subsection, we present the absolute loss values of the models as well as the initialization and data variances in order to show the convergence of the models. Fig. 4 displays the mean losses μ and the standard deviation of the losses σ for the ensembles of normalizing flows trained on the Plummer sphere while Fig. 5 shows the same metrics for the MN disk. For each data distribution individually, all models achieve comparable losses to the base distributions even when additional substructure is added in the form of idealized stellar streams. The overall mean loss values broadly agree (i.e. $\mu_{\text{data}} \approx \mu_{\text{init}}$) for fixed training size, so the relation between $\sigma_{\text{init}}/\sigma_{\text{data}}$ appears to be robust. The main differences between both distributions are that the mean loss values (μ_{init} and μ_{data}) for the MN disk are lower than those of the Plummer sphere. This is likely because the MN disk's phase-space density is centrally concentrated, allowing the normalizing flow to map this compact distribution into the central, high-probability regions of the latent Gaussian space. Despite this difference, the normalizing flow ensembles trained on both distributions exhibit the same trend in $\sigma_{\text{init}}/\sigma_{\text{data}}$.

A.3 KL Divergence Between Base and Perturbed Distributions

In order to quantify the differences between the base distributions and the perturbed distributions in Tab. 1, we compute the KL divergence with the base distribution as the reference. Mathematically, this amounts to the assigning the base distribution as q and the perturbed distribution as p in the KL divergence expression $D_{\text{KL}}(p||q) = \int dr^3\, dv^3\, p \ln{(p/q)}$ where we integrate over six-dimensional phase space (three spatial coordinates and three velocity coordinates). In practice, doing such a

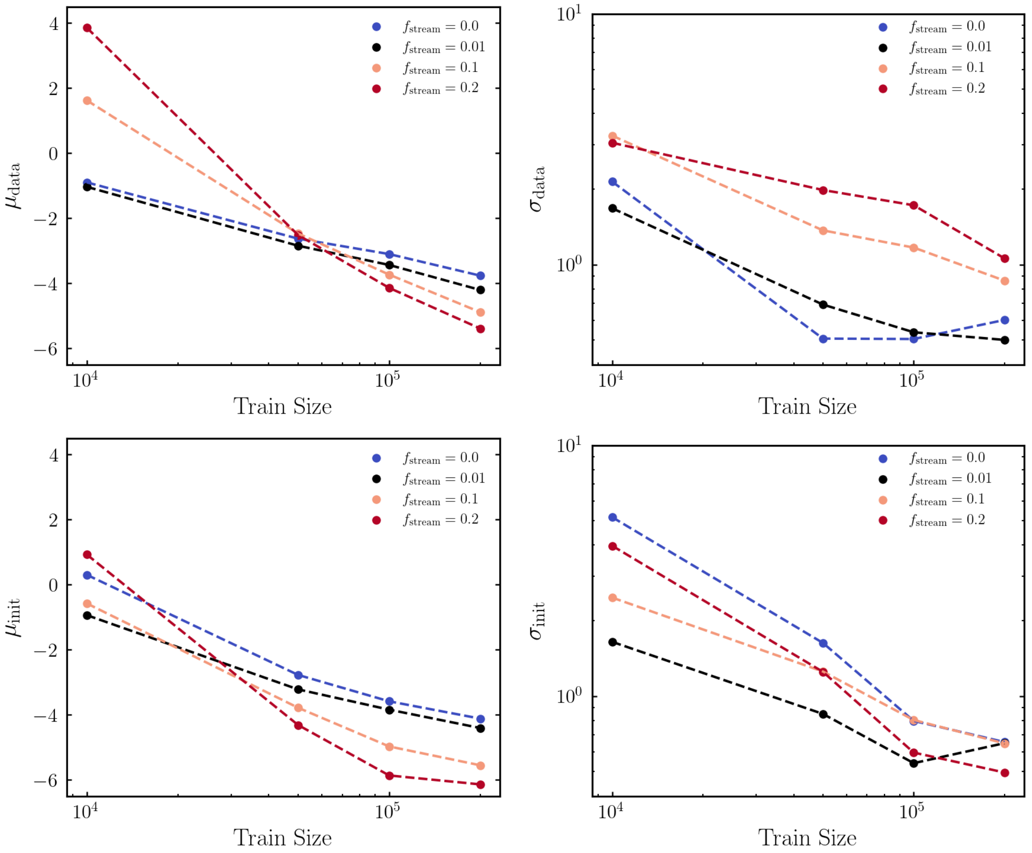

Figure 5: *Top row*: The mean ($\mu_{\rm data}$) and standard deviation ($\sigma_{\rm data}$) of the validation loss for normalizing flows of fixed initial weights but varied training data samplings of the Miyamoto-Nagai (MN) disk distribution, varied over training set size and stream fraction. *Bottom row*: The mean ($\mu_{\rm init}$) and standard deviation ($\sigma_{\rm init}$) of the validation loss for normalizing flows with varied initial weights but fixed training data samplings of the MN disk, varied over training set size and stream fraction.

calculation is numerically challenging so we use the k-nearest neighbors method to approximate the distributions p and q based on sampled particle coordinates and positions, using the following formula (23; 24):

$$D_{\rm KL}(p \parallel q) \approx \frac{d}{N_p} \sum_{i=1}^{N_p} \left(\log \frac{s_k(x_i)}{r_k(x_i)} \right) + \log \frac{N_q}{N_p - 1} \qquad (9)$$

where $d = 6$ (the data dimensionality), N_p is the number of datapoints sampled in the perturbed distribution p, N_q is the number of datapoints sampled in the base distribution q, $s_k(x)$ is the k-th nearest neighbor distance for the base distribution q and $r_k(x)$ is the k-th nearest neighbor distance for the perturbed distribution p. For our calculations, we use $k = 8$, $N_p = N_q = 10^6$ and the numerical computation of s_k, r_k is done with the publicly-available `sklearn` package (25).

A.4 Link to code package

All the code we use to generate the datasets, initialize and train the MAFs, and to visualise and calculate KL divergence values can be found at the following `github` repository: `https://github.com/roy-physics/normalising_flow_uncertainties.git`.