Available online at www.sciencedirect.com

ScienceDirect

Journal of the Franklin Institute 360 (2023) 1635–1671

www.elsevier.com/locate/jfranklin

Automated learning of gravitational mass of elliptical galaxies

Dalia Chakrabarty[1]

Department of Mathematics, Brunel University London, Middlesex, Uxbridge UB8 3PH, UK

Received 23 June 2022; received in revised form 14 November 2022; accepted 19 December 2022
Available online 26 December 2022

Abstract

We present a 3-staged method for automated learning of the spatial density function of the mass of all gravitating matter in a real galaxy, for which, data exist on the observable phase space coordinates of a sample of resident galactic particles that trace the galactic gravitational potential. We learn this gravitational mass density function, by embedding it in the domain of the probability density function (*pdf*) of the phase space vector variable, where we learn this *pdf* as well, given the data. We generate values of each sought function, at a design value of its input, to learn vectorised versions of each function; this creates the training data, using which we undertake supervised learning of each function, to thereafter undertake predictions and forecasting of the functional value, at test inputs. We assume that the phase space that a kinematic data set is sampled from, is isotropic, and we quantify the relative violation of this assumption, in a given data set. Illustration of the method is made to the real elliptical galaxy NGC4649. The purpose of this learning is to produce a data-driven protocol that allows for computation of dark matter content in any example real galaxy, without relying on system- specific astronomical details, while undertaking objective quantification of support in the data for undertaken model assumptions.

© 2022 The Author(s). Published by Elsevier Ltd on behalf of The Franklin Institute.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

E-mail address: dalia.chakrabarty@brunel.ac.uk.
URL: https://www.brunel.ac.uk/people/dalia-chakrabarty

[1] No funding source is relevant to this work

https://doi.org/10.1016/j.jfranklin.2022.12.029

0016-0032/© 2022 The Author(s). Published by Elsevier Ltd on behalf of The Franklin Institute. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

1. Introduction

The mass distribution of all gravitating matter in an external galaxy, is a coveted system property in astronomy and cosmology. The density function of the mass of gravitational matter in the galaxy - i.e. of luminous as well as dark galactic matter - if learnt or estimated, will permit computation of the spatial distribution of dark matter in the galaxy. Ubiquity of the shape of the total gravitational mass density function continues to be questioned [1–3], while agreement exists on the importance of galactic mass distribution in constraining galaxy formation [1,4–6], and evolution [7,8]. Computation of the spatial distribution of dark matter in a galaxy, will indeed need to include astronomical models of the link between luminous galactic matter, and photometric information obtained from such matter in the galaxy. This is however more easily accomplished than the learning (or estimation) of the total, or gravitational mass density function of elliptical galaxies - which is relatively more difficult in elliptical galaxies than in disky galaxies. The latter systems being rotationally supported, the rotational velocity measurements of their resident particles, taken at different galactocentric radii, offer an estimate of the gravitational mass enclosed within the given radius [9]. For elliptical galaxies, no reliable universal parametric model exists to link light and total (or gravitational) mass, thereby compromising the ambition of learning gravitational mass density using only photometric information. It follows, that in an elliptical galaxy, it is in general best to avoid reliance on such a link when attempting learning or estimation of the total (or gravitational) mass density function.

In this paper, we focus on the learning of gravitational mass distribution of elliptical galaxies, without resorting to astronomical details of individual galaxies, except for observations of the observable phase space variables of a sample of resolved galactic particles that trace the galactic gravitational field - and hence are referred to as tracers. Thus, our methodology offers a black- boxed protocol in which data comprising such observations is input for a given galaxy, and the protocol outputs the cumulative (with galactocentric radius) gravitational mass in this galaxy, (along with its phase space probability distribution), under an assumption on the symmetry of the galactic phase space, where the validity of this assumption for the considered galaxy, is also quantified.

Almost all galactic mass modelling exercises that employ tracer kinematic data, resort to the Jeans Equation formalism. Within this framework, the galaxy is treated as autonomous and Hamiltonian. Adherence of the temporal evolution of the probability density function (*pdf*) of the galactic phase space vector to the Collisionless Boltzmann Equation (CBE) is used, to formulate a deterministic link between moments (and spatial derivatives of moments) of this *pdf* and the gravitational potential [10]. Such a link is useful, considering that one of the inputs in Jeans Equation, is the variance of the line-of-sight (LOS) component of the velocity vector - which is observed at some locations in the galaxy - and the gravitational potential function that is the desired output in this exercise. However, tacitly imposing an arbitrarily-chosen correlation structure on a sample of observations - sparsely sampled typically - is not a reliable way of learning the spatial distribution of a moment. Indeed, a persisting difficulty with conventional estimation of galactic mass distribution is the approximating of an unknown spatially-varying function with a discrete sample of pairs of the input variable of this function, and the corresponding, widely-uncertain value of the output variable. This problem is further compounded by seeking spatial derivatives of such an approximated æfunctionø. While the presence of such a sample can in principle allow for supervised learning of the sought function, the large error on the output variable will only learn the function (treated as random), as

highly uncertain, thereby offering meaningless spatial derivatives. Here by æmeaninglessg, we imply error bars so large, that the value of the derivative is rendered uninformative. This problem with the estimation of total galactic mass distribution, plagues both the usage of tracer kinematics (i.e. observations of phase space coordinates of individual galactic particles) in Jeans Equation formalisms, as well as observations of temperature of hot gas, modelled under assumed hydrostatic equilibrium [11,12]. Imposing a smoothness by hand, on such æapproximatedg functions, will naturally need to be justified; else the function is rendered *ad hoc* and the learnt/estimated mass distribution of the galaxy then stands compromised.

Additionally, given the observational limitations that typify this domain, data does not exist on everything other than the sought output, i.e. the galactic gravitational potential. This includes data on the parametrisation of anisotropy of the galactic phase space. Indeed, one major idiosyncratic difficulty with the Jeans Equation formalism, is the absence of information on how deviant the galactic phase space *pdf* is, from invariance to rotation, i.e. from phase space isotropy. This then triggers the need for specification of values of such deviation from phase space isotropy in a galaxy. Such values are fundamentally *ad hoc*, since it is the lack of information on the anisotropy parametrisation that motivates the need for its manual specification. The resulting galactic gravitational mass is then rendered arbitrary. Then from the above, it appears that an automated route to mass modelling would be one that: avoids computing spatial derivatives of functions that are learnt (or worse, approximated) as highly-uncertain; permits quantification of anisotropy of the galactic phase space; eschews reliance on photometric observations. In this paper, we advance a 3-staged protocol that undertakes the learning of the gravitational mass density function of a galaxy, while also learning the phase space *pdf* of the galaxy, using available noisy measurements on a small sample of the 3 observable components of the 6-dimensional phase space vector, treating the galaxy as autonomous and Hamiltonian [13], while assuming that within the spatial extent of the galaxy that we offer the density functions for, the galactic potential is central [14], and the galactic phase space *pdf* is an isotropic function of location and momentum vectors [15]. At the same time, we quantify how this assumption of isotropy is supported in the data, by computing a parametrisation of the departure of the system from this assumption, given the available data.

We do not possess training data on the 2 functions that we desire to learn, namely the gravitational mass density function and the phase space *pdf*. This would then negate the undertaking of supervised learning of these functions, it would seem. However, in the 1st-Stage of our work, we will generate such originally-absent training data, by discretising the relevant range of the domain variable relevant to each function, and learning the vectorised version of each function, given the data. This effectively offers the functional value over each design partition of the relevant range of values of the domain variable, i.e. offers that originally-absent training data on the gravitational mass density and phase space *pdf*. In the 2nd-Stage, we will implement these produced training sets, to learn each function, by modelling them as random realisations from an adequately chosen stochastic process, allowing for predictions and forecasting of values of the functions. Ultimately in the 3rd-Stage, we will check the assumption of isotropy in the data, and compute how anisotropic the galaxy is, given a data set. We will illustrate our 3-staged method on the data available on an observed sample of Planetary nebulae (PNe) and another sample of Globular Clusters (GCs), for the galaxy NGC4649. These empirical data sets on the 2 types of tracers in this galaxy, was shared with us by Dr. Kristin Woodley. Kinematics of GCs in this galaxy have been employed by [16–18], while kinematic information of PNe in this galaxy has been employed by Teodorescu et al. [19]. The data on GCs that we employ here, is a subsample of the bigger sample used by

other authors. Indeed, while we advocate our 3-staged protocol for the objective and automated mass modelling of a sample of galaxies, using kinematic information on tracer samples, it is crucially important to note that any other information - either as data or priors - if available to the astronomer for an individual galaxy in this sample, can definitely be incorporated within this framework that is Bayesian by nature. All through, inference is carried out using different Markov Chain Monte Carlo (MCMC) algorithms [20].

2. Method

The system is treated as autonomous, implying that the probability density function (*pdf*) of the phase space vector $\boldsymbol{W}$ does not bear an explicit dependence on time. Here $\boldsymbol{W} = (X_1, X_2, X_3, V_1, V_2, V_3)^T$, where the location of a galactic particle is $\boldsymbol{X} = (X_1, X_2, X_3)^T$, and its velocity vector is $\boldsymbol{V} = (V_1, V_2, V_3)^T$. In this system, we discuss dynamics per unit mass of a galactic particle, such that the state of any particle is specified by its velocity, along with its location. Let the phase space vector $\boldsymbol{W} \in \mathcal{W} \subseteq \mathbb{R}^6$, i.e. $\mathcal{W}$ is the phase space of the galaxy. Thus in this autonomous system, the phase space *pdf* is denoted $f_{\boldsymbol{W}}(x_1(t), x_2(t), x_3(t), v_1(t), v_2(t), v_3(t))$, though we will often drop the time-dependence of phase space coordinates from our notation, for brevity's sake. We express the location and velocity vectors in the basis $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3\}$ such that the line-of-sight is along $\boldsymbol{e}_3$, implying that X_1 and X_2 are the location coordinates in the plane of the sky; these location coordinates are observable. The component V_3 of the velocity vector is the speed with which the particle is moving along the line-of-sight; this is the third observable. Thus, X_3, V_1, V_2 cannot be observed.

The system gravitational potential at location $\boldsymbol{X}$ is denoted $\Psi(X_1, X_2, X_3)$; again, the lack of explicit time-dependence in the potential, owes to the autonomous nature of the system. Our treatment of this dynamical system having reached a stationary state maybe circumspect, but with observations available only at a snapshot - rather than online - we need to assume so much, in order to undertake any tractable analysis.

The system is also treated as Hamiltonian [14]; this is motivated by the collisionless nature of galaxies. Then as the system moves along a trajectory in phase space, the flow of phase is conserved, i.e.

$$\frac{df_{\boldsymbol{W}}(x_1(t), x_2(t), x_3(t), v_1(t), v_2(t), v_3(t))}{dt} = 0,$$

i.e., the Collisionless Boltzmann Equation (CBE) holds [9].

A corollary of the phase space *pdf* abiding by CBE is that $f_{\boldsymbol{W}}(\cdot)$ can be recast as a function of integrals of motion [9,21], such as $I_1, \ldots, I_n$, where $I_k : \mathcal{W} \longrightarrow \mathbb{R}$ in general, with $d(I_k)/dt = 0$, for $k = 1, \ldots, n$. Then it follows that phase space trajectories that are allowed in $\mathcal{W}$, have to lie at the intersection of n such sub-volumes of $\mathcal{W}$ within which the phase space coordinates abide by the constraints that $d(I_1(x_1, \ldots, v_3)/dt = 0, \ldots, d(I_n(x_1, \ldots, v_3)/dt = 0$. In other words, phase space trajectories in $\mathcal{W}$ have n constraints imposed on them, confining them to $6 - n$ degrees of freedom; recall that $\mathcal{W} \subseteq \mathbb{R}^6$. Then to be allowed even 1 degree of freedom, $n \leq 5$.

We recall that one such integral of motion is the energy of any galactic particle. Energy is given partly by the kinetic energy of the particle, and partly by its potential energy. As we perform the analysis per unit mass, the total energy per unit mass is the sum of half the

squared (Euclidean) norm of the velocity vector, and the gravitational potential at the location of the particle, in this galaxy.

We assume the galaxy to be spherically symmetric. Thus, the gravitational potential is central, i.e. a function of the components of the location vector, via its explicit dependence on galactocentric radius $R := \sqrt{X_1^2 + X_2^2 + X_3^2}$, i.e. its notation is updated from above as $\Psi(R)$. Then under spherical symmetry, Poisson Equation links the gravitational mass density $\rho(R)$ to the potential as:

$$\frac{1}{R^2}\frac{d}{dR}\left(R^2\frac{d\Psi(R)}{dR}\right) = -4\pi G\rho(R),$$

where G is the known Newton's Gravitational constant. Assuming the potential to be central is perhaps not a bad assumption, as long as we attempt our learning only at locations that are at most "moderately" distant from the centre of the galaxy. This is indeed not verifiable given the observations that offer values of X_1 and X_2, (along with V_3). One possible suggestion for a limit on the spatial extent under consideration, could be the smaller between the maximal galactocentric radius r_{max} that observations are available to, and a benchmark photometrically-relevant radius, such a 5 times the effective radius (r_{eff}). We use this convention and consider only results obtained within such a galactocentric radius

$$r_{gal} = \min[r_{max}, 5r_{eff}].$$

Let us suggest that $f_{\boldsymbol{W}}(\cdot)$ is recast as the one and only integral of motion of energy

$$\varepsilon(X_1, X_2, X_3, V_1, V_2, V_3) := \Psi(R) + \frac{V^2}{2},$$

where

$$V := \sqrt{V_1^2 + V_2^2 + V_3^2}; \quad R := \sqrt{X_1^2 + X_2^2 + X_3^2}.$$

Then it implies that $f_{\boldsymbol{W}}(X_1, X_2, X_3, V_1, V_2, V_3) = f_{\boldsymbol{W}}(\varepsilon(X_1,\ X_2, X_3, V_1, V_2, V_3)) \equiv f_{\boldsymbol{W}}(\Psi(R) + V^2/2)$.

Definition 2.1. A function of the 2 vectors $\boldsymbol{X} \in \mathcal{X} \subseteq \mathbb{R}^3$ and $\boldsymbol{V} \in \mathcal{V} \subseteq \mathbb{R}^3$ is isotropic, if the function is invariant to rotation $\mathbf{Q}\boldsymbol{X}$ and rotation $\mathbf{Q}\boldsymbol{V}$, for any 3×3-dimensional orthogonal matrix $\mathbf{Q}$.

It follows that function $f_{\boldsymbol{W}}(R, V)$ is an isotropic function of $\boldsymbol{X}$ and $\boldsymbol{V}$, for $R = \sqrt{X_1^2 + X_2^2 + X_3^2}; V = \sqrt{V_1^2 + V_2^2 + V_3^2}$, since

$$R = \| \boldsymbol{X} \| := \sqrt{\boldsymbol{X}^T\boldsymbol{X}} = \sqrt{(\mathbf{Q}\boldsymbol{X})^T(\mathbf{Q}\boldsymbol{X})},$$

given that for orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{3\times 3}$, $\mathbf{Q}^T\mathbf{Q} = I$; similarly for $V = \| \boldsymbol{V} \|$. Here $\| \cdot \|$ denotes Euclidean norm.

So the phase space *pdf*, if recast as a function of the sole integral of motion, energy, is an isotropic function of location and velocity, i.e. $f_{\boldsymbol{W}}(\varepsilon) \equiv f_{\boldsymbol{W}}(\Psi(R) + V^2/2)$ is an isotropic function of $\boldsymbol{X}$ and $\boldsymbol{V}$.

Remark 2.1. We model the phase space *pdf* as a function of energy of the considered galactic particle. So in our model, the phase space distribution of the system is isotropic.

2.1. 1st-stage: generating originally-absent training sets

Our aim here is to learn the phase space $pdf f_W(\varepsilon)$ and the gravitational mass density function $\rho(R)$.

To undertake the supervised learning of $\rho(R)$, one needs a training data set comprising pairs of design values of radius R, and the gravitational mass density computed at this design radius. We do not possess such a training set apriori. Any such training set built from a simulated model of the galaxy is irrelevant since it is the essential lack of information about the galaxy under consideration, that motivated our pursuit of the galactic gravitational mass density function and the phase space *pdf*. In other words, we do not have information about these pursued functions $\rho(R)$ and $f_W(\varepsilon)$. In lieu of such information, any constructed model of the gravitational mass density of the galaxy - constructed with the aim of simulating values of this function at design radii - is rendered arbitrary, i.e. irrelevant to the galaxy at hand. Similarly, we state that there exists no information-driven model of the phase space *pdf* of the considered galaxy; if information exists to constrain such a model *pdf*, we would not need to embark upon our learning exercise.

2.1.1. Information available; priors

It may however be that we possess information on some properties of these system functions, that we would like to input towards the learning of these functions, at design inputs. Such information comprises

- Non-negativity of both the phase space *pdf* and the gravitational mass density function, at all energy and radius.
- Monotone non-increasing nature of $\rho(R)$ with increasing values of R. This is motivated by the gravitationally bound nature of the system, assumed spherically symmetric. Then under the central (gravitational) potential $\Psi(R)$ we expect matter at radius $R = r$ to be more tightly packed than matter at $R = r^{/}$, for $r < r^{/}$.
- The phase space *pdf* integrates to 1, over all energy values.

We will incorporate each of these known pieces of information on properties of the 2 sought functions via the inference that we will undertake.

Additionally, it may be possible that for a galaxy under consideration, priors exist on the shape of either, or both, of the sought functions. However, in our approach we advocate caution over priors motivated by astronomical theory, parameters of which are then fed generic values. We state this, backed by apprehension about a high level of diversity in a sample of galaxies observed within any observational programme; galactic properties are expected to be sensitive to the effect of the highly multivariate internal and external evolutionary influences - inclusive of non-linear dynamical effects - on the evolution of individual galaxies. This is particularly true for the phase space *pdf*, which does not need to follow any global parametric shape, and indeed, may violate the assumed isotropy that we model the *pdf* to abide by. Given that a galaxy is composed of multiple, dynamically interacting - in fact, differentially correlated - components, implies that the phase space *pdf* may not be a single or monolithic function of the integrals of motion, but may manifest different functional forms within distinct sub-volumes of the galactic phase space. There is no pressing motivation for the phase space *pdf* to be globally Normal or skewed-Normal in energy. In light of this discussion, we will

impose only weak priors on the sought gravitational mass density function $\rho(R)$, and the phase space *pdf* $f_{\boldsymbol{W}}(\varepsilon)$.

2.1.2. Embedding $\rho(R)$ in domain of pdf $f_W(\varepsilon)$

As seen above, expressing the phase space *pdf* as a function of energy, allows for the gravitational potential to be embedded in the support of the *pdf*, such that, the gravitational mass density $\rho(R)$ - which is deterministically computable, given $\Psi(R)$ - is effectively embedded in the support of the *pdf*. To summarise,

$$f_{\boldsymbol{W}}(X_1, X_2, X_3, V_1, V_2, V_3) = f_{\boldsymbol{W}}(\varepsilon) \equiv f_{\boldsymbol{W}}(\Psi(R) + V^2/2)$$

implies that the phase space *pdf* is

$$f_{\boldsymbol{W}}(\varepsilon) \equiv f_{\boldsymbol{W}}(\Psi(R), V) \equiv f_{\boldsymbol{W}}(\rho(R), V).$$

Let the data available for the galaxy under consideration, comprise N_{data} values of the observables X_1, X_2, V_3 of one type of galactic particles, where the noise on V_3 is also observed, with noise on X_1 and X_2 observations negligible, compared to the noise on V_3. In fact, the noise s_i on the i-th observation $v_3^{(i)}$ of V_3, is modelled as the standard deviation of the error density for V_3 observations, where we model this error density to be a Normal with a zero mean. Thus, the error on the observed value $v_3^{(i)}$ is $\epsilon_i \sim \mathcal{N}(0, s_i^2)$. We denote this available kinematic data of the N_{data} particles of a given type, as $\mathbf{D} = \{(x_1^{(i)}, x_2^{(i)}, v_3^{(i)}, s_i)\}_{i=1}^{N_{data}}$.

Now, assuming the N_{data} data points in $\mathbf{D}$ to be independent, likelihood of the model - of the sought gravitational mass density and phase space *pdf* - given the data $\mathbf{D}$, is expressed as the product of values of the probability density function $g_{\boldsymbol{U}}(\cdot)$ of the observable $\boldsymbol{U} := (X_1, X_2, V_3)^T$, (given the model mass density and phase space *pdf*), computed at each of the data points in $\mathbf{D}$. In other words, in the absence of measurement noise, likelihood is

$$\ell(\rho(\cdot), f_{\boldsymbol{W}}(\cdot)|\mathbf{D}) = \prod_{i=1}^{N_{data}} \frac{g_{\boldsymbol{U}}(x_1^{(i)}, x_2^{(i)}, v_3^{(i)}|\rho(\cdot), f_{\boldsymbol{W}}(\cdot))}{C(f_{\boldsymbol{W}}(\cdot), \Psi(\cdot))}, \tag{2.1}$$

where the *pdf* of the observable $\boldsymbol{U}$ can be computed from the phase space *pdf* by integrating out from the latter, all those phase space coordinates that are not observed and $C(\cdot, \cdot)$ is the normalisation of this *pdf* of the observables. Thus,

$$g_{\boldsymbol{U}}\left(x_1^{(i)}, x_2^{(i)}, v_3^{(i)}|\rho(\cdot), f_{\boldsymbol{W}}(\cdot)\right) = \int_{x_3^{(min,i)}}^{x_3^{(max,i)}} \int_{v_2^{(min,i)}}^{v_2^{(max,i)}} \int_{v_1^{(min,i)}}^{v_1^{(max,i)}}$$
$$f_{\boldsymbol{W}}\left(\Psi\left(\rho\left(\sqrt{(x_1^{(i)})^2 + (x_2^{(i)})^2 + x_3^2}\right)\right) + \left(v_1^2 + v_2^2 + (v_3^{(i)})^2\right)/2\right) dv_1 dv_2 dx_3,$$

where this density of the observables will need to be subsequently normalised by $C(f_{\boldsymbol{W}}(\cdot), \Psi(\cdot))$, which is defined as the integral over all values of the observables, i.e. this normalisation is:

$$C(f_{\boldsymbol{W}}(\cdot), \Psi(\cdot)) = \int g_{\boldsymbol{U}}(x_1, x_2, v_3|\rho(\cdot), f_{\boldsymbol{W}}(\cdot)) dx_1 dx_2 dv_3.$$

Recognising that values of X_1 and X_2 appear in the integrals relevant to RHS of the equation that defines $g_{\boldsymbol{U}}(\cdot|\cdot)$, via the term $X_1^2 + X_2^2$, we replace $(x_1^{(i)})^2 + (x_2^{(i)})^2$ in the integrand with $(x_p^{(i)})^2$, where we define

$$X_p^2 := X_1^2 + X_2^2.$$

Similarly, we replace $(v_1^{(i)})^2 + (v_2^{(i)})^2$ in the integrand with $(v_p^{(i)})^2$, where we define

$$V_p^2 := V_1^2 + V_2^2,$$

and $dv_1 dv_2$ by $2\pi v_p dv_p$, i.e. the double integral with respect to (w.r.t.) V_1 and V_2 is replaced by a single integral w.r.t. v_p, by invoking isotropy in velocity-space. Thus, the definition of $g_U(\cdot)$ reduces to:

$$g_U\left(x_1^{(i)}, x_2^{(i)}, v_3^{(i)} | \rho(\cdot), f_W(\cdot)\right) = 2\pi \int_{x_3^{(min,i)}}^{x_3^{(max,i)}} \int_{v_p^{(min,i)}}^{v_p^{(max,i)}} f_W\left(\Psi\left(\rho\left(\sqrt{(x_p^{(i)})^2 + x_3^2}\right)\right) + \left(v_p^2 + (v_3^{(i)})^2\right)/2\right) v_p dv_p dx_3. \quad (2.2)$$

In light of the introduction of X_p, as motivated just above, we clarify the normalisation of the *pdf* of the observables as

$$C(f_W(\cdot), \Psi(\cdot)) = 2\pi \int_{x_p=0}^{r_{max}} \int_{v_3=-\sqrt{-2\Psi(x_p)}}^{\sqrt{-2\Psi(x_p)}} g_U(x_p, v_3 | \rho(\cdot), f_W(\cdot)) x_p dx_p dv_3. \quad (2.3)$$

Here, the maximal value that V_3 can attain at a given value of X_p is obtained by recalling the definition of energy as sum of potential and kinetic energies. When energy attains the highest value (of 0), kinetic energy is maximal, at $x_3 = 0$, i.e. at potential $\Psi(x_p)$. This follows from $\Psi(x_p) < \Psi(\sqrt{x_p^2 + x_3^2})$, $\forall x_3 \neq 0$. Then the maximal value of V_3 is computed using this maximal kinetic energy at $V_1 = V_2 = 0$. Thus, the maximal value of V_3 is $\sqrt{0 - 2\Psi(x_p)}$. The minimal value of V_3 is the negative of this computed maximal value. The maximal value of X_p is the maximal radius r_{max}, observations of $\boldsymbol{U}$ till which, we consider in our learning.

However, it appears impossible to compute the likelihood introduced in Eq. (2.1), since calculation of $g_U(\cdot)$ appears impossible, given that learning of $f_W(\varepsilon)$ and $\rho(r)$ appears impossible. The last claim is due to the fact that training data set $\{(\varepsilon_j, f_W(\varepsilon_j))\}_{j=1}^{N_e}$ is unavailable, and training data $\{(r_k, \Psi(r_k))\}_{k=1}^{N_r}$ is also unavailable, $\forall N_e, N_r \in \mathbb{N}$. However, such training sets are pre-requisites for the supervised learning of the phase space *pdf* and gravitational mass density function. In lieu of these training sets - and with the aim of generating such training data sets - we learn "vectorised versions" of each of the sought functions, where we explain below, what we imply by vectorised version of a sought function.

We discretise the relevant interval ($[r_0, r_{max}]$) in the domain of the sought function $\rho(\cdot)$, into N_r partitions. Each such partition is referred to as an "R-bin". Then the j-th R-bin comprises $r \in [r_0 + (j-1)\delta_r, r_0 + j\delta_r)$, $(j = 1, \ldots, N_r)$, where we choose to use a constant width δ_r for all R-bins. So we use $\delta_r = (r_{max} - r_0)/N_r$. Then we approximate the function $\rho(r)$ as

$$\rho_j \equiv \rho(r), \ \ \forall r \in [r_0 + (j-1)\delta_r, r_0 + j\delta_r); \quad j = 1, \ldots, N_r.$$

We define the N_r-dimensional vector $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_{N_r})^T$, and replace our ambition of learning $\rho(r)$ - for now - by learning $\boldsymbol{\rho}$, i.e. each of the N_r "ρ-parameters" $\rho_1, \ldots, \rho_{N_r}$.

Similarly, we discretise the relevant range of normalised energy values ([-1, 0)), into N_e "ε-bins", such that we approximate the function $f_W(\varepsilon)$ as

$$f_k \equiv f_W(\varepsilon), \ \ \forall \varepsilon \in [-1 + (k-1)\delta_e, -1 + k\delta_e); \quad k = 1, \ldots, N_e,$$

where $\delta_e = 1/N_e$. Here by "normalised energy", we imply the variable $\tilde{\varepsilon}$ that takes value $\tilde{\varepsilon} \in [-1, 0)$, where we normalise ε to $\tilde{\varepsilon}$ as:

$$\tilde{\varepsilon} := \frac{\varepsilon}{-\Psi(0))}.$$

We define the N_e-dimensional vector $\boldsymbol{f} = (f_1, \ldots, f_{N_e})^T$. In fact, for every trial $\boldsymbol{f}$ - i.e. at every iteration of the inference, we normalise f_k with a global scale s.t. $f_k \leq 1$, $\forall k = 2, \ldots, N_e$. and learn $f_2, \ldots, f_{N_e}$, so that we can deterministically assign to f_1 the value such that $(f_1 + 2(f_2 + f_3 + \ldots + f_{N_e-1}) + f_{N_e})\delta_e/2 = 1$, i.e. the (trapezoidal implementation) of the Riemann sum approximation of the area under the phase space *pdf* equals 1. Thus, we learn $f_2, \ldots, f_{N_e}$ and treat f_1 as deterministically known, given the learnt "f-parameters" in every iteration of the adopted inferential scheme.

Thus, the Bayesian inference using MCMC that we will undertake, will make inference on the state space parameter vector $(\rho_1, \ldots, \rho_{N_r}, f_2, \ldots, f_{N_e})^T$, given the data $\mathbf{D}$. We summarise, that by vectorised-version of the sought $\rho(\cdot)$ and $f_{\boldsymbol{W}}(\cdot)$ functions, we imply the vectors $\boldsymbol{\rho}$ and $\boldsymbol{f}$, respectively. Learning the N_r ρ-parameters enables realisation of the pairs: $(r_1, \rho_1), \ldots, (r_{N_r}, \rho_{N_r})$, and learning the f-parameters leads to the pairs: $(\tilde{\varepsilon}_2, f_2), \ldots, (\tilde{\varepsilon}_{N_e}, f_{N_e})$. While the first set is the originally-absent training set that will allow for the supervised learning of the $\rho(R)$ function, the latter set is such a training data set for learning $f_{\boldsymbol{W}}(\varepsilon)$. Eq. (B.1), provides the means to compute the gravitational potential $\Psi(R)$, given a vectorised version of $\rho(R)$.

To learn the ρ-parameters and f-parameters, we recall the definition of the probability density of the observables, expressed in terms of the phase space *pdf* and in that, replace the sought functions with their respective vectorised forms. Thus, considering the projection of the phase space *pdf* into the space of observables - which is what the RHS of Eq. (2.2) manifests - over each ε-bin individually, and summing over all such energy partitions, this equation reduces to

$$g_{\boldsymbol{U}}(x_1^{(i)}, x_2^{(i)}, v_3^{(i)} | \boldsymbol{\rho}, \boldsymbol{f}) = 2\pi \sum_{j=1}^{N_e} \left[f_j \int_{v_p^{(min,i)}}^{v_p^{(max,i)}} \int_{x_3^{(min,i)}}^{x_3^{(max,i)}} v_p dv_p dx_3 \right], \tag{2.4}$$

where the integrals on the RHS represent the volume that the j-th ε-bin, i.e. energy-partition, occupies in the space of the unobservables. From the definition of energy, at a given $r^{(i)} = \sqrt{(x_p^{(i)})^2 + x_3^2}$, and given $v_3^{(i)}$, the maximal value of V_p is given as

$$\left(v_p^{(max,i)}\right)^2 = 0 - 2\Psi\left(r^{(i)}\right) - \left(v_3^{(i)}\right)^2,$$

i.e. $(v_p^{(max,i)}) = \sqrt{-2\Psi(r^{(i)}) - (v_3^{(i)})^2}$, while $(v_p^{(min,i)}) = -\sqrt{-2\Psi(r^{(i)}) - (v_3^{(i)})^2}$. Again, from the definition of energy, at given $x_p^{(i)}$ and $v_3^{(i)}$, the maximal value of X_3 is $x_3^{(max,i)}$ - root of the equation: $0 = \Psi\left(\sqrt{(x_p^{(i)})^2 + x_3^2}\right) + (v_3^{(i)})^2/2$. The minimal X_3 value is 0. The $g_{\boldsymbol{U}}(\cdot)$ computed this way from Eq. (2.4) is then normalised by the RHS of Eq. (2.3), which we refer to now as $C(\boldsymbol{f}, \boldsymbol{\Psi})$, in light of the vectorisation of the functions.

While Eq. (2.4) allows for computation of the *pdf* of the observables, conditional on the model parameters $\rho_1, \ldots, \rho_{N_r}, f_2, \ldots, f_{N_e}$ in the noise-free instance, noise in the measurement of V_3 exists, and we need to perform learning of our model parameters, while acknowledging such noise. As we state above, we model the noise on the i-th observation of V_3 to be the

$\epsilon_i \sim \mathcal{N}(0, s_i^2)$. We ignore the noise in the X_p measurement, in comparison to the noise in the measurement of V_3. Then in the presence of this noise, the *pdf* of the i-th datum given the model parameters, is updated to the convolution of $g_{\boldsymbol{U}}(x_1^{(i)}, x_2^{(i)}, v_3^{(i)}|\boldsymbol{\rho}, \boldsymbol{f})$ with the Gaussian in $v_3^{(i)}$, with parameters 0 and s_i^2, as mean and variance, respectively. This updated *pdf* of the i-th datum given the model parameters, is then to be normalised by the normalisation $C(\boldsymbol{f}, \boldsymbol{\rho})$ of the *pdf* of the observable given (vectorised) forms of the phase space *pdf* and gravitational mass density function. Thus, likelihood of the model parameters given data $\mathbf{D}$ that includes the observational noise in V_3, under the assumption of independent data points, is

$$\ell\big(\rho_1, \ldots, \rho_{N_r}, f_2, \ldots, f_{N_e}|\mathbf{D}\big) = \prod_{i=1}^{N_{data}} \frac{g_{\boldsymbol{U}}\left(x_1^{(i)}, x_2^{(i)}, v_3^{(i)}|\boldsymbol{\rho}, \boldsymbol{f}\right) * \mathcal{N}\left(v_3^{(i)}; 0, s_i^2\right)}{C(\boldsymbol{f}, \boldsymbol{\rho})}. \tag{2.5}$$

Having defined the likelihood of the model parameters - that are the ρ-parameters and the f-parameters - given the data $\mathbf{D}$, we can write the posterior *pdf* of these model parameters given this data, subsequent to the selection of priors on the model parameters. As stated above, we will motivate the case here, for weak priors, since we want to avoid biasing our inference, especially on the f-parameters, in this learning exercise in which the data has weaker informative influence on the f-parameters, over the ρ-parameters. In case the astronomer is blessed with information on the gravitational mass density and phase space *pdf* of a given galaxy, such information can be translated to more informative priors in such situations. Our weak priors $\pi_0(\theta)$ include Normal and truncated Normal priors on a model parameter $\theta \in \{\rho_1, \ldots, \rho_{N_r}, f_2, \ldots, f_{N_e}\}$, with a mean that is the chosen seed value of θ, and a variance that is large - namely, 3 to 10 times the variance used in the proposal density that trial values of this model parameter θ are proposed from, in any iteration of the undertaken MCMC-based inference. Our experimentation indicates robustness of the inferred results on $\rho_1, \ldots, \rho_{N_r}, f_2, \ldots, f_{N_e}$, to the choice of priors.

We choose a seed value for f_j by considering the "seed phase space *pdf*" to be uniform, such that the seed value of f_j is f_0, $\forall j = 2, \ldots, N_e$. Again, we choose a "seed gravitational mass density function", to bear pre-chosen forms, eg. an NFW form [2], such that ρ_k is computed using this chosen form, at radius $R = r_k$, $k = 1, \ldots, N_r$. We have experimented with different forms of the seed gravitational mass density function, and find our inferred results to be insensitive to the chosen seeds, i.e. each learnt parameter is inferred to be consistent within the learnt uncertainties on itself, (which given our MCMC-based inference, is the 95% Highest Probability Density credible region learnt for this parameter, given the data).

Thus, the posterior *pdf* is given as

$$\pi\left(\rho_1, \ldots, \rho_{N_r}, f_2, \ldots, f_{N_e}|\mathbf{D}\right) \propto$$

$$\ell\big(\rho_1, \ldots, \rho_{N_r}, f_2, \ldots, f_{N_e}|\mathbf{D}\big)\pi_0(\rho_1)\ldots\pi_0(\rho_{N_r})\pi_0(f_2)\ldots\pi_0(f_{N_e}),$$

where any global scale in the definition of the posterior of the model parameters given the data, is irrelevant to the MCMC-based posterior sampling that we undertake towards making inference on the model parameters.

2.2. Inference on model parameters in the 1st-stage

We choose to employ the Metropolis-within-Gibbs algorithm [22], towards such inference, to appreciate the expectedly higher correlation amongst the ρ-parameters and amongst the f-parameters, compared to the correlation amongst ρ-parameter-f-parameter pairs. Thus, we update the ρ-parameters in the 1st block of an iteration, given the data $\mathbf{D}$, and then in the 2nd block of the iteration, at the updated $\boldsymbol{\rho}$ vector, we update the f-parameters, given the data.

In the 0-th iteration, the ρ-parameters and the f-parameters are assigned respective seed values, using the seed gravitational mass density form and the seed phase space *pdf* stated above. In the t-th iteration, ($t = 1, \ldots, N_{iter}$), let the current value of ρ_k be $\rho_k^{(t-1)}$ and the current value of f_j is $f_j^{(t-1)}$; $k = 1, \ldots, N_r$, $j = 1, \ldots, N_e$. In the 1st block of this t-th iteration, the proposed ρ_j, is sampled from a truncated Normal proposal density as

$$\rho_j^{(\star,t)} \sim \mathcal{TN}\left(\rho_j^{(t-1)}, 0, \infty, \sigma_j^2\right), \quad \text{for } j = N_r, \text{ and}$$

$$\rho_j^{(\star,t)} \sim \mathcal{TN}\left(\rho_j^{(t-1)}, \rho_{j+1}^{(\star,t)}, \infty, \sigma_j^2\right), \quad \text{for } 1 \leq j < N_r,$$

where the truncated Normal density with mean a, left truncation b, right truncation c, and variance d is depicted as $\mathcal{TN}(a, b, c, d)$. As can be appreciated from this proposal scheme suggested above, the ρ-parameters are proposed at the outermost radial bin first, and then the other ρ-parameters are sequentially proposed, as we move inwards, from the outermost radial bin. Thus, the constant jump-scale σ_j is used to propose ρ_j in any iteration.

The demand that a proposed ρ_j not fall below the recently-updated ρ_{j+1}, is implemented via proposing from the truncated Normal density that is left truncated at this minimally allowed value for ρ_j in any iteration. This allows for adherence of each ρ-parameter to the physically-motivated constraint of monotonic non-increasing with increasing radius, and at the same time, also satisfies positivity. The correlation amongst the ρ-parameters that is suggested via this monotonicity, once allowed to percolate to the learning of the ρ-parameters via the MCMC-based inference, renders the learning robust to small to moderately large changes in seeds and priors. Unlike ρ_j, $\forall j = 1, \ldots, N_r - 1$, there is no value that ρ_{N_r} can be deterministically known to be in excess of - other than 0, (since all ρ-parameters are non-negative). Thus, the uncertainty in our learnt value of ρ_{N_r} is typically the highest, amongst uncertainties on all other learnt ρ-parameters.

Then the ρ-parameters proposed in (the first block of) this tth iteration are accepted or not depending on whether the following acceptance criterion is obeyed:

$$\alpha_1\left(\boldsymbol{\rho}^{(\star,t)}, \boldsymbol{\rho}^{(t-1)}\right) \geq u,$$

where $U = u$, with $U \sim$ Uniform[0, 1], and

$$\alpha_1\left(\boldsymbol{\rho}^{(\star,t)}, \boldsymbol{\rho}^{(t-1)}\right) = \frac{\pi\left(\rho_1^{(\star,t)}, \ldots, \rho_{N_r}^{(\star,t)}, f_2^{(t-1)}, \ldots, f_{N_e}^{(t-1)} | \mathbf{D}\right)}{\pi\left(\rho_1^{(t-1)}, \ldots, \rho_{N_r}^{(t-1)}, f_2^{(t-1)}, \ldots, f_{N_e}^{(t-1)} | \mathbf{D}\right)} \times \frac{\Psi\left(\rho_1^{(t-1)}, \rho_2^{(\star,t)}, \infty, \sigma_1^2\right) \ldots \Psi\left(\rho_{N_r}^{(t-1)}, 0, \infty, \sigma_{N_r}^2\right)}{\Psi\left(\rho_1^{(\star,t)}, \rho_2^{(t-1)}, \infty, \sigma_1^2\right) \ldots \Psi\left(\rho_{N_r}^{(\star,t)}, 0, \infty, \sigma_{N_r}^2\right)}. \tag{2.6}$$

If the acceptance criterion is obeyed, we state that $\rho_j^{(t)} = \rho_j^{(\star,t)}$; else $\rho_j^{(t)} = \rho_j^{(t-1)}$; $j = 1, \ldots, N_r$. Thus, the ρ-parameters are updated in the first block of the tth iteration.

Then in the 2nd block of this iteration, at the updated $\boldsymbol{\rho}$ of $\rho_1^{(t)}, \ldots, \rho_{N_r}^{(t)}$, we update the f-parameters. We propose the kth f-parameter from a truncated Normal that is left truncated at 0 to ensure non-negativity of the parameter; has a mean that is the current value of the parameter, i.e. $f_k^{(t-1)}$; has a constant variance ν_k, $\forall k = 2, \ldots, N_e$. Thus,

$$f_k^{(\star,t)} \sim \mathcal{TN}\Big(f_k^{(t-1)}, 0, \infty, \nu_k^2\Big), \quad \text{for } k = 2, \ldots, N_e.$$

Then acceptance of this proposed value of f_k depends on adherence to the acceptance criterion:

$$\alpha_2\big(\boldsymbol{f}^{(\star,t)}, \boldsymbol{f}^{(t-1)}\big) \geq u,$$

and

$$\alpha_2\big(\boldsymbol{f}^{(\star,t)}, \boldsymbol{f}^{(t-1)}\big) = \frac{\pi\Big(\rho_1^{(t)}, \ldots, \rho_{N_r}^{(t)}, f_2^{(\star,t)}, \ldots, f_{N_e}^{(\star,t)} | \mathbf{D}\Big)}{\pi\Big(\rho_1^{(t)}, \ldots, \rho_{N_r}^{(t)}, f_2^{(t-1)}, \ldots, f_{N_e}^{(t-1)} | \mathbf{D}\Big)} \times \frac{\Psi\Big(f_2^{(t-1)}, 0, \infty, \nu_2^2\Big) \ldots \Psi\Big(f_{N_r}^{(t-1)}, 0, \infty, \nu_{N_r}^2\Big)}{\Psi\Big(f_2^{(\star,t)}, 0, \infty, \nu_1^2\Big) \ldots \Psi\Big(f_{N_r}^{(\star,t)}, 0, \infty, \nu_{N_r}^2\Big)}. \tag{2.7}$$

If the acceptance criterion is obeyed, we state that $f_k^{(t)} = f_k^{(\star,t)}$; else $f_k^{(t)} = f_k^{(t-1)}$; $k = 2, \ldots, N_e$. This way, we update the f-parameters in the 2nd block of the tth iteration.

2.3. What is N_r and N_e in 1st-stage?

We definitely do not wish to learn the number N_r of R-bins that the radial range $[r_0, r_{max})$ is partitioned into; neither do we want to learn the number N_e of ε-bins. This reluctance about treating N_r and N_e as variables stems from desired avoidance of an MCMC-based implementation in which the dimensionality of the state space vector $(\rho_1, \ldots, \rho_{N_r}, f_2, \ldots, f_{N_e})^T$ is a variable. Such an implementation will generally be Reversible Jump MCMC; we wish to avoid it since it is a cumbersome inferential tool, and its need for the application at hand is over-ridden by provisions in the data for deterministic choice of N_r and of N_e.

We choose N_r to be such that if the interval $[r_0, r_{max})$ is partitioned into these many R-bins, then there will be at least one datum in each such bin. At the same time, we appreciate that the larger is N_r, smaller is the error in approximating the function $\rho(R)$ with its vectorised-version $\boldsymbol{\rho}$.

We could invoke similar considerations to guide our choice of N_e, except that there is the additional complication that computation of value of the energy variable, requires input from the potential function. To acknowledge this, we first employ the N_r R-bins over which the frequency distribution of particle numbers is constructed, and treat this as proportional to the frequency distribution of the particle gravitational mass, i.e a rudimentary indicator of $\boldsymbol{\rho}$. We then employ this vectorised version of the gravitational mass density function, in the discretised Poisson Equation, to compute a rudimentary value of the vectorised version (referred to as $\boldsymbol{\Psi}_0 = (\Psi_{1,0}, \ldots, \Psi_{N_r,0})^T$) of the gravitational potential function, where $\Psi_{j,0}$ is defined over the jth R-bin; $j = 1, \ldots, N_r$. Then adding the ith observed value of $V_3^2/2$ to this preliminary indicator for the gravitational potential (vector), we generate N_{data} values

of energy ε. We normalise these values such that all such indicator values of energy lie in [-1,0). The histogram of the set of such computed values of energy is constructed using N_e bins, with bounds of 0 and -1, where N_e is chosen to ensure that no energy bin is bereft of a computed energy value. With a preliminary choice of N_e, a chain of MCMC is run, and the gravitational mass density vector $\boldsymbol{\rho}$ learnt from it, is again employed to learn the vectorised version of the gravitational potential, by inputting the learnt $\boldsymbol{\rho}$ into the discretised Poisson Equation. Normalised energy values computed by adding observations of $V_3^2/2$ are then again used to compute the energy histogram with N_e bins, over the interval [-1,0). If the number of data points within each energy bin is ≥ 1, then N_e is retained as the number of ε-bins in the learning exercise. We appreciate that the computational time rises super-linearly with increase in N_e, and therefore, in the first instance of choosing N_e to construct the indicative histogram of energy values, we choose the smallest N_e that satisfies the constraint that no ε-bin is empty of an input.

2.4. Learning & predicting sought gravitational mass density and phase space pdf*: 2nd-stage*

At the end of the 1st-Stage, we obtain the originally-absent training data sets:

$$\mathbf{D}_\rho := \{(r_1, \rho_1), \ldots, (r_{N_r}, \rho_{N_r})\},$$

and

$$\mathbf{D}_\varepsilon := \{(\tilde{\varepsilon}_2, f_2), \ldots, (\tilde{\varepsilon}_{N_e}, f_{N_e})\}.$$

In the 2nd-Stage, the aim is to perform supervised learning of the gravitational mass density function and the phase space *pdf*, using $\mathbf{D}_\rho$ and $\mathbf{D}_\varepsilon$, respectively. We do this by treating either function as random function, which is equivalent to saying that each unknown function is treated as if it attains a given form, with a probability. In other words, we undertake the Bayesian approach, in which we model a random structure with a probability distribution. Now, a probability distribution on a space of functions is of course a stochastic process.

Thus, we treat the random function $\rho(\cdot)$ as a random realisation from an adequately selected stochastic process, and we treat the random function $f_{\boldsymbol{W}}(\cdot)$ as a random realisation from a stochastic process as well. We aim to invoke generic stochastic processes to model the sought functions. Thus, processes that are such that the realised functions are constrained to abide by given equations, are not ideal. On the other hand, if these functions are considered to be generated by respective Gaussian Processes, ($\mathcal{GP}$s), then the only constraint that each function has to abide by, is that the joint probability distribution of a finite number of realisations of the function is Multivariate Normal, [23]. In other words, the we would need to set:

$$\rho(\cdot) \sim \mathcal{GP}(\mu_\rho(\cdot), K_\rho(\cdot,\cdot)) \text{ and } f_{\boldsymbol{W}}(\cdot) \sim \mathcal{GP}(\mu_f(\cdot), K_f(\cdot,\cdot)),$$

where $\mu_\rho(\cdot)$ and $K_\rho(\cdot,\cdot)$ are the mean and covariance functions of the $\mathcal{GP}$ that $\rho(\cdot)$ is a realisation of, and $\mu_f(\cdot)$ and $K_f(\cdot,\cdot)$ are the mean and covariance functions of the $\mathcal{GP}$ that $f_{\boldsymbol{W}}(\cdot)$ is a realisation of.

Then by definition of GPs, the joint of N_r realisations of $\rho(\cdot)$ - namely, $\rho_1, \ldots, \rho_{N_r}$ - is a Multivariate Normal, with a mean vector $\boldsymbol{\mu}_\rho$ and variance-covariance matrix

$$\boldsymbol{\Sigma}_\rho = [Cov(\rho_c, \rho_d)] \equiv [K_\rho(r_c, r_d)]; \;\; c, d \in \{1, \ldots, N_r\},$$

where the covariance between the pair ρ_c, ρ_d of ρ-parameters, is modelled as a declining function $K_\rho(\cdot,\cdot)$ of the difference between the inputs r_c and r_d, $\forall c, d \in \{1, \ldots, N_r\}$. Then

$K_\rho(\cdot,\cdot)$ is a function that parametrises the covariance of this Multivariate Normal density, and thereby the covariance structure of the $\mathcal{GP}$ that underlines the function $\rho(\cdot)$; we say that this covariance structure is parametrised with the covariance kernel $K(\cdot,\cdot)$. Many forms of $K(\cdot,\cdot)$ are possible [24]; we can for example choose the simple Square Exponential (or SQE) form of this kernel function. Under the SQE form of the kernel,

$$Cov(\rho_c, \rho_d) = K(r_c, r_d) := A \exp\left(-\frac{(r_c - r_d)^2}{\ell^2}\right),$$

where the amplitude $A > 0$ and the length scale $\ell \in \mathbb{R}$ are the hyperparameters of this kernel function that we learn from the data. The same values of the hyperparameters will suffice, for all elements of the covariance matrix, i.e. $\forall c, d \in \{1, \ldots, N_r\}$, as long as the $\rho(\cdot)$ function is continuous[2] However, given the training data $\mathbf{D}_\rho$ alone, we cannot be confident if the sought function is continuous; on the other hand, the distribution of the output variable in the training set, across the design input points, can indicate if the underlying function is not continuous. In other words, the underlying function may still not be continuous, though a finite discrete sample of input-output pairs from the function may suggest continuity. But if the training sample indicates lack of continuity, the function is not likely to be continuous. We will however proceed with the SQE covariance kernel for the sake of simplicity of computation - which we acknowledge, might compromise accuracy of predictions. We will check on this accuracy by predicting the value of $\rho(R)$ at test inputs.

As stated above, the joint of N_r realisations of $\rho(\cdot)$ is the Multivariate Normal ($\mathcal{MN}$) with parameters $\boldsymbol{\mu}_\rho$ and $\boldsymbol{\Sigma}_\rho$.

$$[\rho(r_1), \ldots, \rho(r_{N_r})] = \mathcal{MN}(\boldsymbol{\mu}_\rho, \boldsymbol{\Sigma}_\rho),$$

which given our vectorised learning in the 1st-Stage, is equivalent to stating that the joint

$$[\rho_1, \ldots, \rho_{N_r}] = \mathcal{MN}(\boldsymbol{\mu}_\rho, \boldsymbol{\Sigma}_\rho),$$

i.e. probability of data on the output of the sought function, conditional on $\boldsymbol{\mu}_\rho$ and $\boldsymbol{\Sigma}_\rho$, is the Multivariate Normal density with parameters $\boldsymbol{\mu}_\rho$ and $\boldsymbol{\Sigma}_\rho$, [25]. In fact, the only unknowns in the mean vector and covariance matrix are the amplitude and length scale (hyperparameters): A_ρ and ℓ_ρ, where $Cov(\rho_c, \rho_d) = K(r_c, r_d) = A_\rho \exp(-(r_c - r_d)^2/\ell_\rho^2)$, $\forall c, d \in \{1, \ldots, N_r\}$.

But, this conditional probability of data is the likelihood of the model parameters given the data, i.e. the likelihood is

$$\mathcal{L}(A_\rho, \ell_\rho|\mathbf{D}_\rho) = \frac{1}{\sqrt{|2\pi\, \boldsymbol{\Sigma}_{\boldsymbol{\rho}}|}} \exp\left(-\frac{(\boldsymbol{\rho} - \bar{\boldsymbol{\rho}})^T \boldsymbol{\Sigma}_\rho^{-1} (\boldsymbol{\rho} - \bar{\boldsymbol{\rho}})}{2}\right), \tag{2.8}$$

where $\bar{\boldsymbol{\rho}}$ is the empirical mean of the ρ-parameters learnt in the 1st-Stage. In our application, we actually standardise the data with the sample mean and standard deviation of the ρ-parameters; this renders $\boldsymbol{\Sigma}_\rho$ the correlation matrix, implying that A_ρ=1.

If learning both A_ρ and ℓ_ρ, we choose adequate priors on these variables, to then define their joint posterior probability density, given the data $\boldsymbol{D}_\rho$. We choose to work with Truncated Normal and Normal priors that are centred at the seed values of the variables, and variances that are typically 3 to 10 times that of the proposal density used in our MCMC-based inference on these unknowns. The seed values are chosen as 1, typically. The joint posterior *pdf* of

[2] In an upcoming contribution, Chakrabarty & Wang suggest a judicious continuity descriptor as globally Lipschitz.

the unknowns given data $\mathbf{D}_\rho$ is then proportional to the product of the likelihood (given in Eq. (2.8)) and the prior. We perform posterior sampling using Random Walk Metropolis Hastings, in which ℓ_ρ is proposed from a Normal proposal density with a mean given by the current value of ℓ_ρ and an experimentally chosen constant variance. A_ρ if learnt, is proposed from a Truncated Normal proposal density, which is assigned the current A_ρ to be the mean, and the constant variance of this proposal density is chosen through experimentation. The MCMC-based inference allows for the learning of the marginal posterior *pdf* on each learnt variable, using which, the 95% Highest Probability Density credible region (HPD) on each learnt variable is computed, [26].

We undertake the same route delineated above, to learn hyperparameters ℓ_f and A_f (if learnt), of the covariance structure of the $\mathcal{GP}$ that the phase space *pdf* is treated as a random realisation from. Again, the joint probability of the f-parameters learnt in the 1st-Stage, at design energy values, is a Multivariate Normal density, with mean $\boldsymbol{\mu}_f$ and variance-covariance matrix $\boldsymbol{\Sigma}_f = [Cov(f_c, f_d)] = [K_f(\varepsilon_c, \varepsilon_d)] = A_f \exp(-(\varepsilon_c - \varepsilon_d)^2/\ell_f^2)$, for $c, d \in \{2, \ldots, N_e\}$. In our application we typically standardise the f-parameters using the sample mean and standard deviation of the outputs in the training set $\mathbf{D}_f$, rendering $\boldsymbol{\Sigma}_f$ the correlation matrix and $A_f = 1$ then. We compute the 95% HPD on each learnt variable., given the training data $\mathbf{D}_f$.

2.5. Uncertainties in learnt training sets

We have discussed the learning of the functions $\rho(\cdot)$ and $f_{\boldsymbol{W}}(\cdot)$ given training data sets that were learnt in the 1st-Stage, as if ρ-parameters and f-parameters are learnt in the 1st-Stage without errors. This is not true. We in fact learnt each of these parameters with their respective 95% HPD. Thus, the correct representation of $\mathbf{D}_\rho$ is $\{(r_1, [\rho_1^{(min)}, \rho_1^{(max)}]), \ldots, (r_{N_r}, [\rho_{N_r}^{(min)}, \rho_{N_r}^{(max)}])\}$, where $[\rho_j^{(min)}, \rho_j^{(max)}]$ is the 95% HPD on the learnt ρ_j, $j = 1, \ldots, N_r$. Similarly, each f-parameter is learnt in the 1st-Stage with 95% HPD.

When we learn the hyperparameters of the covariance kernel that we invoke to parametrise the covariance matrix of the Multivariate Normal likelihood, we actually model the covariance matrix as $\boldsymbol{\Sigma}_X + \boldsymbol{D}_X$, for $X ='' \rho'', ``f''$, where $\boldsymbol{D}_\rho$ is a diagonal matrix, with diagonal elements of $((\rho_1^{(max)} - \rho_1^{(min)})/5)^2, \ldots, ((\rho_{N_r}^{(max)} - \rho_{N_r}^{(min)})/5)^2$ while $\boldsymbol{D}_f$ is a diagonal matrix, with diagonal elements of $((f_2^{(max)} - f_2^{(min)})/5)^2, \ldots, ((f_{N_e}^{(max)} - f_{N_e}^{(min)})/5)^2$. We treat the distribution of the uncertainty learnt on any parameter as approximated by a Normal, such that the width of the 95% HPD on the parameter is 5 times the standard deviation of this distribution of the "noise" that we in fact learn on this parameter. Thus, the variance-covariance matrix of the Multivariate Normal likelihood density, is augmented by a diagonal matrix, diagonals of which are the variances of the error distribution on each parameter, [23].

2.6. Prediction of gravitational mass density and phase space pdf

The ulterior motivation behind the learning of the gravitational mass density function and the phase pace *pdf* is to predict values of the gravitational mass density at test radii, $r_1^{(test)}, \ldots, r_{N_t}^{(test)}$ i.e radii that are not included as design radii $r_1, \ldots, r_{N_r}$ in the training data $\mathbf{D}_\rho$. We define $\boldsymbol{r}^{(test)} = (r_1^{(test)}, \ldots, r_{N_t}^{(test)})^T$. Let gravitational mass density at $R = r_q^{(test)}$ be $\rho_q^{test)}$. Then it follows from the joint probability of $\rho_1, \ldots, \rho_{N_r}, \rho_1^{(test)}, \ldots, \rho_{N_t}^{(test)}$ to be

Multivariate Normal, that the posterior predictive of $\rho_1^{(test)}, \ldots, \rho_{N_t}^{(test)}$ is also Multivariate Normal [23]:

$$[\rho_1^{(test)}, \ldots, \rho_{N_t}^{(test)} | \{r_c\}_{c=1}^{N_r}, \{r_q^{(test)}\}_{q=1}^{N_t}, \rho_1, \ldots, \rho_{N_r}, \ell_\rho, A_\rho] = \mathcal{MN}(\boldsymbol{\mu}_\rho^\star, \boldsymbol{\Sigma}_\rho^\star),$$

where

$$\boldsymbol{\mu}_\rho^\star = [K_\rho(r_q^{(test)}, r_c)]([K_\rho(r_c, r_d)] + \boldsymbol{D}_\rho)^{-1}\boldsymbol{\rho},$$

and

$$\boldsymbol{\Sigma}_\rho^\star = [K_\rho(r_q^{(test)}, r_p^{(test)})] - [K_\rho(r_q^{(test)}, r_c)]([K_\rho(r_c, r_d)] + \boldsymbol{D}_\rho)^{-1}[K_\rho(r_c, r_q^{(test)})],$$

where $c, d \in \{1, \ldots, N_r\}$; $p, q \in \{1, \ldots, N_t\}$; $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_{N_r})^T$; $K(\cdot, \cdot) = A_\rho \exp(-(\cdot - \cdot)^2/\ell_\rho^2)$. Thus, the mean value of the gravitational mass density is predicted at $R = r_q^{(test)}$ as the qth component of the $\boldsymbol{\mu}^\star$ vector defined above, with uncertainty on this prediction given as the (standard deviation that is) square root of the qth diagonal element of the matrix $\boldsymbol{\Sigma}_\rho^\star$ given above. This way, we predict values of the gravitational mass density function at a test radius.

The posterior predictive of values of the phase space *pdf* - conditional on the test input energies $\varepsilon_1^{(test)}, \ldots, \varepsilon_{N_s}^{(test)}$; the design energy values in learnt training set $\mathbf{D}_f$; the learnt f-parameters; and the learnt A_f, ℓ_f - is also Multivariate Normal. Mean and variance of this posterior predictive are closed-form and identified. In other words, the phase space *pdf* can be predicted in a closed-form way, with known uncertainty, at a test energy.

To summarise, the 2nd-Stage allows the prediction of the gravitational mass density at any radius, and the phase space *pdf* at any energy.

2.7. Testing for the assumption of isotropy in the data: 3rd-stage

In the 1st-Stage, we have performed the learning of the vectorised gravitational mass density function as $\boldsymbol{\rho}$, and the vectorised phase space *pdf* as $\boldsymbol{f}$, using the empirical or observed data $\mathbf{D}$. In the 2nd-Stage, we have performed the learning of the gravitational mass density function $\rho(R)$, and the phase space $pdf f_{\boldsymbol{W}}(\varepsilon)$, using the training data sets that comprise values of the respective vectorised function.

Both sets of learning in the previous two stages were undertaken under the assumption that the phase space that the tracer particles - observable phase space coordinates of which we use in our work - live in an isotropic phase space $\mathcal{W}$. Equivalently, we recall that the learning in the first 2 stages has been undertaken under the assumption that the phase space *pdf* is an isotropic function of the location vector $\boldsymbol{X}$ and velocity vector $\boldsymbol{V}$, i.e. this *pdf* is $f_{\boldsymbol{W}}(\| \boldsymbol{x} \|, \| \boldsymbol{v} \|, \boldsymbol{x} \cdot \boldsymbol{v})$, where $\| \cdot \|$ denotes Euclidean norm. This assumption is affected by expressing the support of the phase space *pdf* as energy ε, (or a function thereof); the modelled phase space *pdf* then adheres to the assumption that it is an isotropic function of $\boldsymbol{X}$ and $\boldsymbol{V}$, (since energy $= \Psi(R) + V^2/2$, where $\| \boldsymbol{X} \| = R$ and $\| \boldsymbol{V} \| = V$).

[27] advance a new Bayesian test of hypothesis to test for the null

$$H_0 : f_{\boldsymbol{W}}(X_1, X_2, X_3, V_1, V_2, V_3) = f_{\boldsymbol{W}}(\varepsilon) \text{ against}$$

$$H_1 : f_{\boldsymbol{W}}(X_1, X_2, X_3, V_1, V_2, V_3) \neq f_{\boldsymbol{W}}(\varepsilon).$$

Taking inspiration from the methodology presented therein, here we forward a parametrisation of how anisotropic the learnt phase space *pdf* vector is. This new parameter for quantifying

anisotropy in the learnt (vectorised) form of a phase space *pdf* is a divergence measure between the joint posterior probability density of the learnt parameters given the empirical data under consideration, and the posterior given the "generated data". We discuss such generated data, and the motivation behind the formulation of this parametrisation of departure from isotropy, before discussing implementation.

We generate a data set $\mathbf{D}^{(gen)}$ - comprising the same number (N_{data}) of observations as in the empirical data[4] $\mathbf{D}$, where said observations are on X_p, V_3 and the parametrisation of noise in the measurement of V_3, as the standard deviation S of the error density in V_3. Here $\mathbf{D}^{(gen)}$ is generated by sampling location and velocity coordinates from the (vectorised) state space *pdf* ($\boldsymbol{f}$) learnt under the assumption of isotropy, using empirical data, at the (vectorised) gravitational mass density ($\boldsymbol{\rho}$) that is simultaneously learnt using $\mathbf{D}$. So out of the 6 sampled phase space coordinates - sampled from the phase space space *pdf* learnt in the 1st-Stage, at the $\boldsymbol{\rho}$ learnt in the 1st-Stage - we retain only the N_{data} sampled values of the (X_p, V_3, S) triad, to serve as data points in $\mathbf{D}^{(gen)}$.[3]

Remark 2.2. Data $\mathbf{D}^{(gen)}$ is sampled from an isotropy-abiding phase space *pdf* that was learnt using empirical data $\mathbf{D}$, at gravitational mass density (and thereby potential) learnt using $\mathbf{D}$, under the assumption that the phase space *pdf* is isotropic. In other words, the data $\mathbf{D}^{(gen)}$ is sampled from an isotropic phase space *pdf*, unlike the empirical data $\mathbf{D}$, which might, or might not have been sampled from an isotropic galactic phase space *pdf*. It then follows that, $\boldsymbol{\rho}$ and $\boldsymbol{f}$ learnt under the assumption of an isotropic phase space, using data $\mathbf{D}^{(gen)}$, will be more "compatible", (or at least, as compatible) with the used data, than the $\boldsymbol{\rho}, \boldsymbol{f}$ learnt under the assumption of an isotropic phase space, using empirical data $\mathbf{D}$. So the difference between such a parametrised "compatibility" will inform on how much less such compatibility is with the empirical data $\mathbf{D}$, than with the generated data $\mathbf{D}^{(gen)}$. Here we parametrise "compatibility" of a learnt set of $\boldsymbol{\rho}, \boldsymbol{f}$ with a given data set, by the joint posterior *pdf* of the ρ-parameters and f-parameters that are learnt using the given data set.

So the modus operandi of our computation of the compatability of a learnt $\boldsymbol{\rho}; \boldsymbol{f}$ pair, with a given empirical data set, is

1. That we first learn $\rho_1, \ldots, \rho_{N_r}$ and $f_2, \ldots, f_{N_e}$ given the empirical data $\mathbf{D}$ that comprises N_{data} data points, using the learning scheme delineated under 1st-Stage, under the assumption that this data is sampled from an isotropic phase pace *pdf*. Let the joint posterior *pdf* of the sought parameters, given this empirical data be

   $$\pi(\rho_1, \ldots, \rho_{N_r}, f_2, \ldots, f_{N_e} | \mathbf{D}).$$

2. Then we sample N_{data} values of (X_p, V_3) - using Rejection Sampling - from the $\boldsymbol{f}$ that is learnt under the assumption of isotropy, using $\mathbf{D}$, at the (vectorised) potential that is computed using the $\boldsymbol{\rho}$ learnt using $\mathbf{D}$ under the same assumption, (in the 1st-Stage). These N_{data} samples constitute the generated data $\mathbf{D}_{gen}$.
3. We learn $\rho_1, \ldots, \rho_{N_r}$ and $f_2, \ldots, f_{N_e}$ given the generated data $\mathbf{D}_{gen}$, using the learning scheme delineated under 1st-Stage, under the assumption of isotropy. We use the same priors on each parameter, as we do in our learning undertaken with $\mathbf{D}$, and allow for

[3] We recall here that observations on X_1 and X_2 are condensed into values of $X_p := \sqrt{X_1^2 + X_2^2}$.

the chain to run for the same number of post-burnin iterations ($N_{iter} - N_{burnin}$). Let the joint posterior of the parameters given the generated data be

$\pi_{gen}(\rho_1, \ldots, \rho_{N_r}, f_2, \ldots, f_{N_e}|\mathbf{D}_{gen})$.

4. Compute the difference between posterior density of parameters learnt given empirical and generated data, obtained at each post-burnin iteration of the chains run (respectively) with $\mathbf{D}$ and $\mathbf{D}_{gen}$. We will employ a divergence measure between the posterior densities to compute this difference.

We use a divergence measure between the post-burnin values of the (logarithm of the) joint posterior $\pi^{(\cdot)}$ of $\boldsymbol{\rho}$ and $\boldsymbol{f}$ computed within the MCMC chains, run using the empirical data, and $\pi_{gen}^{(\cdot)}$ run using the generated data. (We recall that our MCMC-based inference readily offers the logarithm of this joint posterior, but replacing $\pi_{gen}^{(\cdot)}$ with its logarithm, and $\pi^{(\cdot)}$ with its logarithm, in the definition of the Kullbeck-Leibler divergence, does not make statistical sense. Neither does replacing the posterior values with their respective logarithms in the definition of Hellinger distance $=\sum_{t=N_b urnin}^{N_{iter}}(\sqrt{(\sqrt{\pi^{(t)}} - \sqrt{\pi_{gen}^{(t)}})})^2/\sqrt{2}$, [28]). So we simply use the sum over $t = N_{burnin}, N_{burnin} + 1, \ldots, N_{iter}$, of the difference between the logarithm of $\pi^{(t)}$ and the logarithm of $\pi_{gen}^{(t)}$ as the divergence measure $\delta(\pi, \pi_{gen})$ that we use in our work. In other words, the divergence between the computed $\log(\pi^{(t)})$ and $\log(\pi_{gen}^{(t)})$ is suggested as

$$\delta(\pi, \pi_{gen}) = \sum_{t=N_{burnin}+1}^{N_{iter}} \left(\frac{1 - \left(\frac{\log(\pi^{(t)})}{\log(\pi_{gen}^{(t)})} \right)}{N_{iter} - N_{burnin}} \right).$$

However, we do not know how to interpret a value of $\delta(\pi, \pi_{gen})$; we ask if a computed value of $\delta(\cdot, \cdot)$ can be considered such that we can reject the assumption of an isotropic phase space. While a computed $\delta(\pi, \pi_{gen})$ of 0 implies that the phase space that the (used) empirical data has been sampled from is isotropic, a non-zero $\delta(\pi, \pi_{gen})$ is indicative of such phase space being anisotropic. The strength of the effect - namely, anisotropy - is computable and interpretable in a comparative sense, i.e. we quantify how much more anisotropic the phase space *pdf* is, from which a given empirical data set (say, of size N) is sampled, as distinguished from the phase space *pdf* that underlines another data set (say, of size $N^{/}$).

3. Illustration of the 3-staged learning strategy on real galaxy NGC4649

In our empirical illustration, we use the kinematic data comprising observed values of R_p; of V_3; and of the measurement noise in V_3, (parametrised as S), of 269 PNe and 115 GCs in the elliptical galaxy NGC4649. These 2 datasets were shared with us by Dr. Kristin Woodley. From this data on the tracked PNe, we discard the observations of PNe that move with observed $|v_3| > 700\,\text{km s}^{-1}$. In fact, all PNe except one, appear in the original data set to bear a V_3 value in $[-650, 650]$ km s^{-1}; this motivates our treatment of the single PNe with $V_3 < -700$ km s^{-1} as an outlier that is omitted from the data that we work with. The single PNe with such high, absolute LOS speed in this data set, is depicted in red in the lower right panel of Fig. 1 that displays the plot of V_3 of the observed PNe against R_p. Observations of the remaining PNe are used in our work; these observations comprise the data set $\mathbf{D}_{PNe}$.

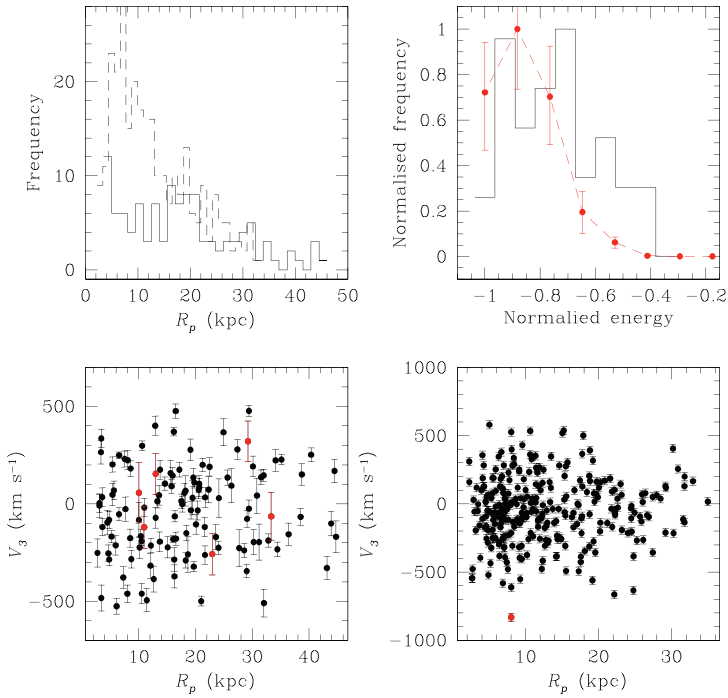

Fig. 1. *Right lower panel:* plot in black, of observed values of V_3 of the PNe, against R_p, for those tracked PNe, $|v_3|$ of which is ≤ 700 km s^{-1}; observations that disobey this constraint are in red. *Left lower panel:* same as in the right lower panel, except the data plotted here is for the tracked GCs. GCs with errors in the observation of V_3 in excess of 100 km s^{-1} are depicted in red. PNe and GCs, observations of which are in black, comprise the data sets $\mathbf{D}_{PNe}$ and $\mathbf{D}_{GC}$ that we perform our learning with. Histograms of the observed sample of R_p values in $\mathbf{D}_{PNe}$ and $\mathbf{D}_{GC}$, are depicted in broken and solid lines in the *top left panel*. In the *top right panel*, in black, we display the histogram of a rudimentary proxy for the energy variable as computed using the observed V_3 values, and the potential computed in Poisson Equation using a scaled frequency distribution of the observed R_p values. In broken lines, we join the f-parameters learnt using the 9 energy bins - suggested by this histogram. All depictions in red will appear to be in grey in the monochromatic version of the paper.

Again, from the full data set that consists of observations on GCs, we discard those observations that bear a measurement error $s > 100$km s^{-1}, where we recall that s_i is the value of this error of the measurement of V_3 of the i-th GC in the data set. GCs with such errors in V_3 are depicted in red in the plot of V_3 against R_p, shown in the lower left panel of Fig. 1. Observations of the remaining 115 GCs comprise the data $\mathbf{D}_{GC}$ that we use in our work. A "large" value of S, on comparable V_3 value, does render attaining convergence in the learning difficult. Hence we impose this arbitrary cutoff of 100 km s^{-1} on s. The sample size of $\mathbf{D}_{PNe}$ is about 2.33 times that of $\mathbf{D}_{GC}$.

For this galaxy, effective radius is suggested to be about 9.86 kpc [18]. Then 5 times the effective radius is ~ 50 kpc, which is in excess of the r_{max} in either data set that we use. Hence in this application, $r_{gal} \equiv r_{max}$. As motivated earlier in Section 2.3, we choose the partitioning of the radial interval $[r_0, r_{max})$ given a data set, keeping in mind that each R-bin should be ideally populated with at least one datum, as well as that a trade-off exists between increasing the number N_r of such R-bins that we partition the relevant range of

values of radius R into, and the computational effort (which increases as N_r^3). Then for the data $\mathbf{D}_{PNe}$, the optimal choice is for $r_0 = 2.2$ kpc, $r_{max} = 33$ kpc and R-bin width $\delta_r = 1.1$ kpc. For $\mathbf{D}_{GC}$, $r_0 = 3.44$ kpc, $r_{max} = 44.04$ kpc and R-bin width $\delta_r = 1.4$ kpc are suitable values. Histograms of the sample of observed values of R_p - as included in the data sets $\mathbf{D}_{PNe}$ and $\mathbf{D}_{GC}$ - are displayed in the top left panel of Fig. 1, in broken and solid lines, respectively. Indeed the observations of the GCs are so sparse at higher radii, that we cannot satisfy the desired property that the histogram of R_p of the observed GCs hosts ≥ 1 data point in each R-bin - a strict imposition of this desirable characteristic of the radial binning will lead to unsatisfactorily fewer R-bins, or truncation of the radial range that we can learn the gravitational mass density to. (Our latent aim is to extend this radial coverage to mimic $5R_{eff}$ of this galaxy, as closely as possible, where $5R_{eff}$ for NGC4649 is about 50 kpc).

In the top right panel of Fig. 1 we present the motivation behind choosing the number N_e of ε-bins for either data set. As stated in Section 2.3, one way to generate the value of energy is to use a rudimentary proxy for the energy variable, namely, the sum of the observed value of $V_3^2/2$ and of Ψ_0, where Ψ_0 is the vectorised version of the potential function, computed by using a scaled frequency distribution of the sample of observed R_p values in Poisson equation; the scaled frequency serves as a proxy for a rudimentary (vectorised) gravitational mass density. The histogram of the resulting (normalised) sample of such computed energy values is shown in solid lines in the figure. The figure includes only the energy-histogram for the GC data, for ease of visualisation. A chosen N_e of 9 serves the purpose of populating every ε-bin with at least one data point. This is found to be true for the PNe data as well. We corroborate the choice of $N_e = 9$ for the GC data, by running a chain with these many ε-bins, (and radial binning as discussed above), and note the (vectorised version of the) learnt (normalised) density over energy - i.e. the phase space *pdf* - to be as depicted with the filled circles and error bars, in red (or grey in the monochromatic print of the paper). The mean of the learnt f-parameters are joined with broken lines to aid visualisation. It is evident that a choice of $N_e = 9$ does not lead to any bin being rendered empty; density in the most sparsely populated bin is about 10^{-3} times that of the density in the most densely populated one.

3.1. Results from 1st-stage

Logarithm (to the base 10) of the components of the $\boldsymbol{\rho}$ vector that we learn using $\mathbf{D}_{PNe}$ are displayed in black, as plotted against the (logarithm of the) location of the corresponding R-bin, in the right panel of Fig. 2; the 95% HPD on each learnt ρ-parameter is overplotted on the mean value of the learnt ρ-parameter. The components of the $\boldsymbol{f}$ vector learnt using this data set, are plotted against the energy value of the corresponding ε-bin, where the energy value has been normalised by $-\Psi(0)$. This plot is depicted in the left panel of this figure. Components of $\boldsymbol{\rho}$ and $\boldsymbol{f}$ that are learnt using data $\mathbf{D}_{GC}$, are shown in red (or the grey in the monochromatic version of the paper) in the left and right panels, respectively.

Both chains that are run with the two data sets are started with a seed $\rho(R)$ function that is plotted in green in the right panel; this function is $\rho_{seed}(r) = K/(10 + r^2)^{1.5}$, where the constant $K = 10^{11}$, though other values of this constant that are 6 decades apart have been noted to yield the same results. Again, the seed for the learning of the f-parameters is a horizontal line, i.e. the seed $f(\varepsilon)$ function is a uniform density. This is depicted in green in the left panel.

We also include the logarithm of a scaled (truncated) Gaussian in broken lines in the left panel, where this fit curve is $0.3\mathcal{N}(-0.9, 0.175^2)$; this is the optimal fit to the learnt

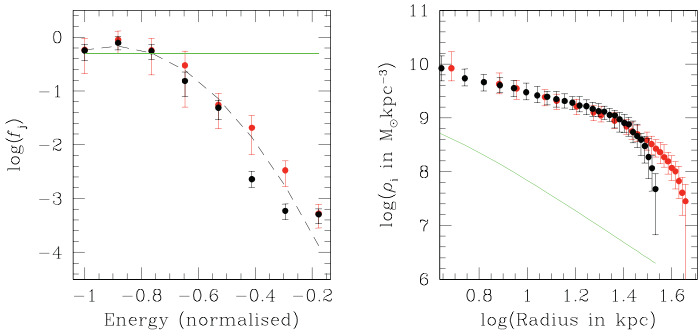

Fig. 2. *Right panel:* plot of logarithm (to the base 10) of the ρ-parameters learnt using $\mathbf{D}_{PNe}$ against logarithm of the radius, (depicted in black), while the ρ-parameters learnt using data $\mathbf{D}_{GC}$ are in red. The 95% HPDs on each learnt parameter is overplotted on the mean value of the learnt parameter that is depicted in filled circles. *Left panel:* logarithm of the f-parameters learnt using the data comprising the observations of the PNe and the GCs are plotted in black and red respectively, against the energy variable that is normalised with $-\Psi(0)$. The seed gravitational mass density function and the seed phase space *pdf* that we start each chain of MCMC with - with either data set - is depicted in green. In the broken lines we display a Gaussian (with mean of -0.9 and variance of 0.175^2) in this normalised energy variable, with this Gaussian scaled by 0.3. All depictions in red will appear to be in grey, and depictions in green in lighter grey, in the monochromatic version of the paper.

f-parameters given the two data sets. It is clear that a (truncated) Normal is not a good fit to all f-parameters learnt with either data set. A scaled Gaussian is a better fit but even then, it fits the f-parameters learnt at less negative energy values less well than parameters learnt at lower energies; a scaled truncated Normal is a better fit to the results learnt using the data on the GCs than results obtained using the data on PNe. So an important indicator of these results is that there is no apriori motivation behind modelling the galactic phase space density to be a truncated Normal.

Traces of the learnt ρ-parameters and f-parameters display convergence. We display traces of the ρ-parameters learnt using $\mathbf{D}_{PNe}$ in Fig. 3; these traces display convergence. Again, the f-parameters learnt using $\mathbf{D}_{GC}$ are shown in black in Fig. 4, with the traces of f-parameters learnt using $\mathbf{D}_{PNe}$ overplotted in green. Again, these traces also display convergence, offering confidence in our learning of the ρ-parameters and the f-parameters.

3.2. Learning the $\rho(\cdot)$ function and the phase space pdf *and predicting - implementation of the 2nd-stage*

Learning the ρ-parameters and the f-parameters using the data sets $\mathbf{D}_{GC}$ and $\mathbf{D}_{PNe}$ provides the training data sets $\mathbf{D}_{\boldsymbol{\rho}}^{(GC)}$; $\mathbf{D}_{f}^{(GC)}$ and $\mathbf{D}_{\boldsymbol{\rho}}^{(PNe)}$; $\mathbf{D}_{f}^{(PNe)}$, respectively. Then we employ these training sets to learn the gravitational mass density function and the phase space *pdf*, by modelling these functions as realisations from respective GPs. Covariance functions of the underlying GPs are kernel parametrised, and the (length scale and amplitude) hyperparameters of these kernels are learnt using these training sets that are generated in the 1st-Stage.

In Fig. 5 we see results from the learning of the multivariate Normal likelihood that results from the modelling of the sought $\rho(\cdot)$ and $f_{\boldsymbol{W}}(\cdot)$ functions with respective GPs, using the training data that are subsets of $\mathbf{D}_{\rho}^{(GC)}$ and $\mathbf{D}_{f}^{(GC)}$, respectively. Trace of the length scale hyperparameter of the covariance matrix of the Multivariate Normal likelihood is depicted for each functional learning, in the left panels of the figure. The training data used for the learning

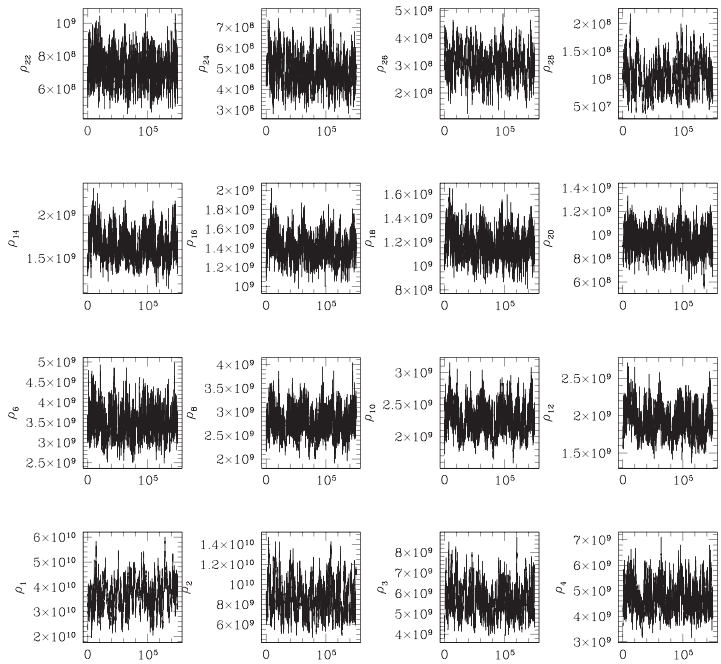

Fig. 3. Traces of 16 of the ρ-parameters learnt using data $\mathbf{D}_{PNe}$.

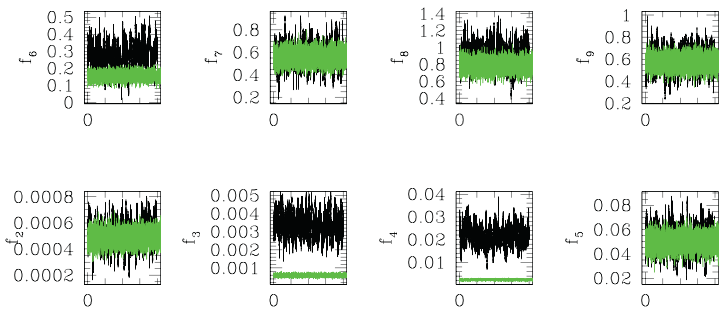

Fig. 4. Traces of the f-parameters learnt using data $\mathbf{D}_{GC}$ in black, with traces of the corresponding f-parameter learnt using data $\mathbf{D}_{PNe}$ overplotted in green (or grey in the monochromatic version of the paper).

of either function comprises ρ-parameters and f-parameters that are respectively plotted in black, in the right panels of Fig. 5, against logarithm of the radius, and against logarithm of the negative of values of the energy variable. The data points in either $\mathbf{D}_{\rho}^{(GC)}$ or $\mathbf{D}_{f}^{(GC)}$, that are not plotted in black in Fig. 5, are the test data points. We undertake prediction of the mean value of the learnt $\rho(\cdot)$ and $f_{\boldsymbol{W}}(\cdot)$ function, at the inputs of the test data points; such predicted values of the respective function are plotted in red, with 2.5 times the standard deviation in the value of the function at the given test input overplotted on either side of the predicted mean. (Here, we choose to use 5 times the predicted standard deviation as the width of the error bar, drawing motivation from the standard result that for Normally distributed variables, there is 95% probability for the variable value to lie within a interval of width 2.5

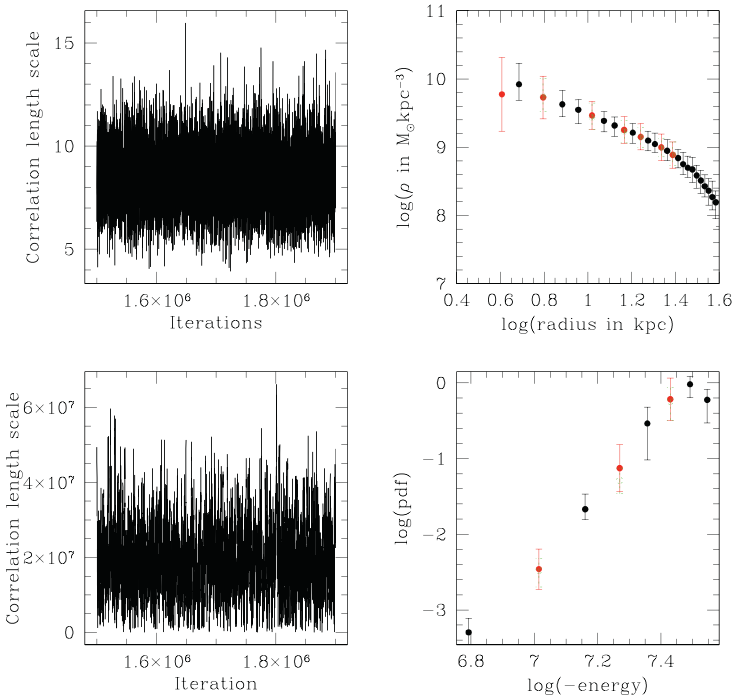

Fig. 5. Traces of the length scale hyperparameter ℓ_ρ, that is learnt using the training set that is a subset of $\mathbf{D}_\rho^{(GC)}$, is plotted in the top left panel. Again trace of the ℓ_f learnt when using a training set that is subset of $\mathbf{D}_f^{(GC)}$, is displayed in the lower left panel. In the top right panel, log of those values of $\rho(\cdot)$ that comprise the training data set that is a subset of $\mathbf{D}_\rho^{(GC)}$, are plotted against log of design radii. Predictions of the mean functional value at test radii, performed following the learning of this function, are shown in red filled circles, with an error bar of width given as 5 times the predicted standard deviation of the functional value. The true value of the function at the corresponding test radius is plotted in green. Forecasting is also performed at a radius that is more central in the galaxy, than the innermost training data point; this is shown in red. In the lower right panel, we plot predicted values of the learnt phase space *pdf*, at test energy values, using a training set that is a subset of $\mathbf{D}_f^{(GC)}$. The predicted mean values are shown in red circles with 5 times the predicted standard deviation overlaid in red. The true value of the *pdf* is overplotted in green broken lines. All depictions in red and green appear as grey and light grey in the black and white version of this paper.

times the standard deviation, symmetrically about the mean of this distribution). However, the true value of the function is known to us - as one of the outputs in $\mathbf{D}_\rho^{(GC)}$ or $\mathbf{D}_f^{(GC)}$, that we do not use as part of the training data employed towards the learning+prediction exercise that we undertake here. This known value of the $\rho(\cdot)$ or $f_{\mathbf{W}}(\cdot)$ function at a test radius/energy, is then overplotted on the predicted value of the function, in green. We also perform one forecasting - at a test radius that is less than the innermost design input. The uncertainty in our forecasting is expectedly higher than that of prediction. GP-based forecasting is also in general of inferior quality to prediction following GP-based learning; our results corroborate these expectations [29].

We use the capacity for predicting at test inputs, to predict values of the learnt function $f_{\mathbf{W}}(\cdot)$ at multiple (100) values of the energy. The results are shown in Fig. 6. Logarithm (to the base 10) of the predicted value of the function is plotted against logarithm of the

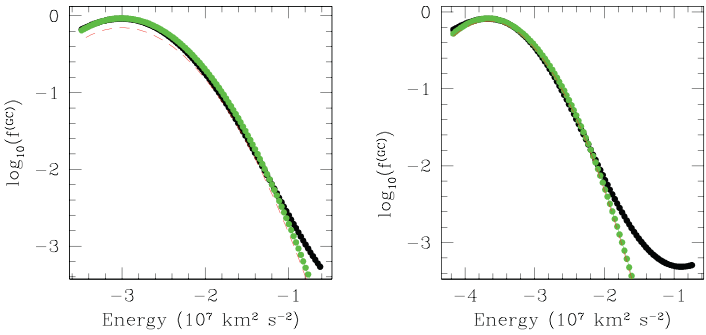

Fig. 6. *Right panel:* logarithm of the predicted values of the function $f_{\boldsymbol{W}}(\cdot)$ learnt using the training set that is a subset of the data $\mathbf{D}_f^{(PNe)}$, plotted in black against the logarithm of the negative of test energies. A scaled Gaussian is fit to these predicted values; it is then overlaid on these predicted functional values (in green or grey in the monochrome version of the paper). *Left panel:* similar to that in the right panel, but here, the predictions follow the phase space pdf learning undertaken with GC data.

negative of the test energy at which this functional value is computed. A truncated Normal density that is scaled by the factor γ is fit to the predicted functional values; the Gaussian with mean of about -3$\times 10^7$ km^2 s^{-2} and standard deviation of 0.5701$\times 10^7$ km^2 s^{-2}, scaled by a factor of γ=1.1786, is found to be the best fit to the values predicted using the training data that comprises observations of GCs, On the other hand the scaled Gaussian that best fits the predictions made after learning using PNe observations, is the Gaussian with a mean of about -3.67$\times 10^7$ km^2 s^{-2} and standard deviation of 0.5251$\times 10^7$ km^2 s^{-2}, scaled by a factor of γ=1.0856. It is seen that the incompatibility of the truncated Normal density with the learnt phase space *pdf* is not just in the demand for a non-unit scale factor γ, but also in the departure of this form from the *pdf* as learnt using the kinematic data.

3.3. Results from 3rd-stage

In Fig. 7, the lower panels display plots of the V_3 values of the PNe (in the right) and GCs (in the left), against the observed values of R_p of these particles. These are the data points in the data sets $\mathbf{D}_{PNe}$ and $\mathbf{D}_{GC}$, respectively, and are displayed in black circles. On these, the data points of the corresponding generated data set, are overplotted in red cross. The generated data $\mathbf{D}_{PNe}^{(gen)}$ is constructed by sampling X_1, X_2, V_3 from the $\boldsymbol{f}$ learnt using $\mathbf{D}_{PNe}$, at the potential computed with the $\boldsymbol{\rho}$ learnt using $\mathbf{D}_{PNe}$, in addition to the measurement error in V_3. Similarly, the generated data set $\mathbf{D}_{GC}^{(gen)}$ comprises sampled X_1, X_2, V_3 values, generated using Rejection Sampling, from the $\boldsymbol{f}$ learnt using $\mathbf{D}_{GC}$, at the potential computed with the $\boldsymbol{\rho}$ learnt using $\mathbf{D}_{GC}$, in addition to the measurement error S in the values of V_3. The top right panel shows the traces of the logarithm of the joint posterior $\pi^{PNe}(\cdot|\cdot)$ of the $\boldsymbol{f}$ and $\boldsymbol{\rho}$, given the empirical PNe data $\mathbf{D}_{PNe}$ (in black), learnt under the assumption of isotropy, and the posterior $\pi_{gen}^{(PNe)}(\cdot|\cdot)$ of the same parameters given the generated data $\mathbf{D}_{PNe}^{(gen)}$ (in red). A similar plot is displayed on the top left, in which the trace of the log of posterior $\pi^{(GC)}(\boldsymbol{\rho}, \boldsymbol{f}|\mathbf{D}_{GC})$ is plotted in black, while that of $\pi_{gen}^{(GC)}(\boldsymbol{\rho}, \boldsymbol{f}|\mathbf{D}_{GC})$ is plotted in red.

Using these traces of the joint posterior computed given the empirical data set and the corresponding generated data set, we compute the value of the divergence $\delta(\cdot,\cdot)$. We find that $\delta(\pi^{(PNe)}, \pi_{gen}^{(PNe)}) \approx 259.77$, while $\delta(\pi^{(GC)}, \pi_{gen}^{(GC)}) \approx 44.66$. Thus, we identify the empirical

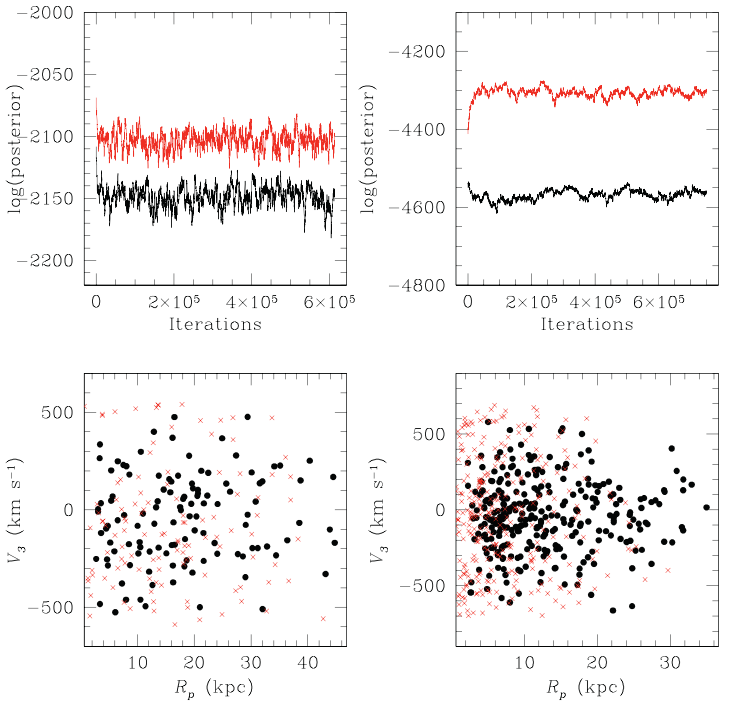

Fig. 7. *Lower right panel:* V_3 values in data $\mathbf{D}_{PNe}$ plotted in black, against values of $R_p = \sqrt{X_1^2 + X_2^2}$ in this data set. Values of V_3 are plotted in red against the corresponding value of R_p in the generated data set $\mathbf{D}_{PNe}^{(gen)}$. *Lower left panel:* same as in the figure included in the lower right panel, except here, the plots are of points in the set $\mathbf{D}_{GC}$ and the generated data of the same size, namely, $\mathbf{D}_{GC}^{(gen)}$. *Top right panel:* trace of the joint posterior of the ρ-parameters and the f-parameters, learnt using data $\mathbf{D}_{PNe}$ in black, while the joint posterior of these parameters learnt using $\mathbf{D}_{PNe}^{(gen)}$ is in red. *Top left panel:* same as in the figure in the top right panel, except here, the traces of joint posteriors of learnt parameters given $\mathbf{D}_{GC}$ and $\mathbf{D}_{GC}^{(gen)}$ are shown in black and red, respectively.

PNe data set $\mathbf{D}_{PNe}$ to deviate from the assumption that it has been sampled from an isotropic phase space *pdf*, more than the data comprising the GC observations, namely the data $\mathbf{D}_{GC}$.

Then indeed, the anisotropy parameter - defined as the ratio of the 2nd moment of the tangential component of the velocity vector to that of the radial component of the velocity vector, subtracted from 1, [10] - is likely to bear a higher value for the PNe data, over the GC data. However, our method does not offer a value of this anisotropy parameter. Our parametrisation of the anisotropy in the phase space that a given set of empirical data is sampled from, informs on the departure from an isotropic form, of the *pdf* of the phase space coordinates that live in this phase space.

Our parametrisation of anisotropy, offers information on the anisotropy of the phase space *pdf* that the empirical data $\mathbf{D}_I$ is sampled from, in comparison to the anisotropy of the *pdf* that another data $\mathbf{D}_{II}$ is sampled from. If however, we wish to offer information on data $\mathbf{D}_I$ in isolation - specifically on the anisotropy of the underlying *pdf* that $\mathbf{D}_I$ is sampled from - our identification of such anisotropy via the parameter $\delta(\pi^{(I)}, \pi_{gen}^{(I)})$ does not appear possible, (though a discussion of the same is suggested in the conclusive section). Thus, our method currently offers only a comparative quantification of anisotropy of the phase space *pdf* that

underlines a data set, in reference to the anisotropy borne in the density that underlines another data set.

Remark 3.1. It will be desirable to reconcile the computed $\delta(\cdot,\cdot)$ divergence measure, with the astronomically-motivated anisotropy parameter that is used in Jeans equation. Here, we have not been able to motivate a transformation of $\delta(\cdot,\cdot)$ that offers a meaningful connection with the anisotropy parameter. However, a possible means of achieving the same is discussed in Section 5.

4. Comments on results obtained from empirical illustration

In this section, we discuss the main takeaway from the implementation of the 3-staged strategy to learn the gravitational mass density function, along with the phase space *pdf*, of the galaxy NGC4649, using data on observable phase space coordinates of 2 different types of galactic particles.

4.1. Phase space pdf*s learnt using PNe data distinct from that learnt using GC data*

The phase space *pdf* that is learnt using observations of tracked PNe, is not consistent with the *pdf* learnt using the GC data, within the learnt 95% HPDs. This reinforces the result that the f-parameters learnt using the 2 different data sets in the 1st-Stage, are different within 95% HPDs. We see that the phase space *pdf* learnt with uncertainties of 95% HPDs, using the PNe and GC data sets, are distinct.

4.2. Distinction between phase space pdf*s that PNe and GC data are sampled from - implications*

The discrepancy between the learnt phase space *pdf*s feeds into the worry that the phase space *pdf* learnt using either data set - which we expect to interpret as the phase space *pdf* of the galaxy - is not consistently learnt for this galaxy, given the two data sets. One suggestion for a resolution to this worry is that it might be that the assumptions undertaken to permit the learning, are differently adhered to, under the two data sets, s.t. the inconsistent results are due to such differential obeying, (by the observed PNe and GC samples), of the undertaken assumptions about the galaxy having equilibrated - i.e. behaving as an autonomous dynamical system - and/or about the galactic phase space being isotropic.

Remark 4.1. If differential adherence to the assumption of isotropy is true,

- then it follows that the learnt $f_{\boldsymbol{W}}^{(GC)}(\cdot)$ and the learnt $f_{\boldsymbol{W}}^{(PNe)}(\cdot)$ functions are incorrect representations of the galactic phase space *pdf*.
- The outcome of our learning is that the phase space *pdf* that underlines the GC data is different - by departing less from the assumption of this *pdf* being isotropic - compared to the *pdf* that the PNe data are sampled from. Such differential departure from the assumption of isotropy confirms that the phase space *pdf* that the GC data are sampled from is distinct from the *pdf* that the PNe data are sampled from.

Hence the phase space of this galaxy is not a monolithic structure, but partitioned into - at least two - sub-volumes, the distributions of the phase space vectors in which are distinct. One

of these sub-volumes hosts the PNe and the other the GCs that are observationally tracked in this galaxy.

Thus, our learning adduces evidence towards this partitioned picture of the galactic phase space on NGC4649. The galactic phase space *pdf* is composed of (at least two) distinct basins of attractions, and the orbits of the galactic PNe population and the GC populations live in the respective basins. It then follows, that the observable phase space coordinates of the tracked PNe that comprise the data $\mathbf{D}_{PNe}$, are sampled from the phase space *pdf* that is generated by the orbital distribution in the basin of attraction of the galactic phase space, that includes orbits of the tracked PNe. Similarly, $\mathbf{D}_{GC}$ is sampled from the *pdf* that is generated by the orbital distribution of the basin of attraction that the tracked GCs are a part of. A galaxy - as a complex and multicomponent dynamical system - is likely to have a phase space that is marked by multiple attractors, and the current proposition of the same for the phase space *pdf* of NGC4649, is likely. Of course, if this is true, then the phase space *pdf* that the PNe data are sampled from - and therefore learnt with - will in general be unequal to the *pdf* that is learnt with the GC data, and one manifestation of this difference in the native phase space densities, will be in their differential anisotropies, in general.

4.3. Gravitational mass enclosed within a radius

Computing the gravitational mass that is enclosed within a given radius, by numerically integrating over the vectorised gravitational mass density function, supplemented with predicted and forecast density values, is not a robust computation. Different implementation of the Riemann sum - that approximates the integral

$$4\pi \int_{s=0}^{r} s^2 \rho(s) ds$$

that gives value $M(r)$ of the mass enclosed within radius $R = r$ - yields different $M(r)$. This owes to the very steep shape of the gravitational mass density function at low radii, compounded by the large (uniform) width of the R-bins that we use in our learning, given the available data.

We found a useful way of addressing the steepness of the density at low radii, by fitting a parametric function to the logarithm of the learnt ρ-parameters, plotted against log of the design radii. In the right panel of Fig. 8 we display the trend in the log of the uncertainty-included ρ-parameters learnt using the GCs data, against $\log(R)$. The inner few ρ-parameters betray a linear trend in this plot, i.e. a power-law relation is anticipated for the gravitational mass density function - between $\approx$ 4kpc and $\approx$ 29kpc. We realise that such linear fits to the learnt ρ-parameters in this radial interval can have the maximal and minimal slopes, as depicted by the broken straight lines (in red), in the right panel of Fig. 8. These fits then suggest the uncertainties in the mass values enclosed within the interval of $\approx$ [4, 29] kpc in this galaxy. To the maximal possible value of this computed enclosed mass in the computed uncertainty interval, we add the value of the gravitational mass that is distributed uniformly within the sphere of radius of about 4 kpc, at the corresponding uncertainty level, to produce the uncertainty-included gravitational mass values enclosed within 29 kpc. This is $[8.81 \times 10^{12}, 1.37 \times 10^{13}]$ $M_{\odot}$. where the mass enclosed within the inner 4 kpc of the galaxy lies in the interval $[1.11 \times 10^{12}, 3.61 \times 10^{12}]$ $M_{\odot}$.

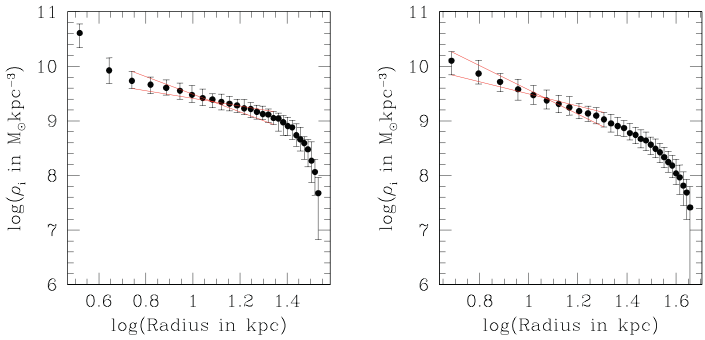

Fig. 8. *Left panel:* logarithm of the vectorised form of the gravitational mass density parameters - or $\boldsymbol{\rho}$ parameters - learnt with the empirical GC data, plotted against logarithm of radius. The maximally and minimally sloped straight lines fits to this data, in the radial interval of about 4.4 to 29 kpc, are shown in the broken red lines. *Right panel:* same as in the left panel, except $\boldsymbol{\rho}$-parameters learnt using the empirical PNe data are plotted. Linear fits to these data in the radial interval of about 3.3 to 22 kpc are shown in red lines.

From the ρ-parameters learnt using the PNe data, linear fits appear possible within the radial range of about [5.5, 22] kpc, to the values of logarithm of the learnt $\boldsymbol{\rho}$-parameters, and log of radius R. These maximally and minimally sloped linear fits are depicted in the left panel of Fig. 8. Adding the result on the gravitational mass enclosed within this radial interval, to the values relevant to radial intervals: between 5.5 kpc and $r_{min} = 3.3$ kpc for this data set; ≤ 3.3 kpc (by spreading mass uniformly at $R \leq r_{min}$), we get that uncertainty-included gravitational mass included within 22 kpc is $[4.6 \times 10^{12}, 1.29 \times 10^{13}]$ $M_\odot$. Gravitational mass enclosed with 3.3 kpc from learning done with this data set, is $[3.3 \times 10^{12}, 9 \times 10^{12}]$ $M_\odot$. Thus, the enclosed mass values learnt with both data sets concur within uncertainties.

When comparing results obtained with a given kinematic data set, but with different methods, we advocate comparison of the gravitational mass density functions - or values of the same learnt/estimated in the different methods at fixed radii - if the density is available. This is preferred to a direct comparison of learnt/estimated values of the gravitational mass enclosed within a given radius. A comparison of the gravitational mass density values helps avoid the accumulation of the uncertainties that render the enclosed mass values more uncertain; in a method like ours, in which the density is implemented to compute the mass, numerical integration over the uncertainty-included, non-linear density function leads to this uncertainty inflation. There is the additional uncertainty in the enclosed mass, stemming from the lack of information in the region inner to the inner-most radial bin.

4.4. Coincidence of gravitational mass density functions learnt using 2 data sets & possible implications

It is noted that even while the phase space *pdf*s learnt with the two data sets are distinct, the gravitational mass density functions overlap within the learnt 95% HPDs. We note that in NGC4649, data on distinct particle types imply distinct phase space *pdf*s - and therefore distinct moments of the learnt phase space *pdf*s - while implying the coincident gravitational mass density functions. This is similar to what [30] note for this galaxy, when they refer to the possibility that “the PNe and GCs trace different kinematics”.

It then follows that gravitational mass distribution in this galaxy cannot be computed as produced self-consistently. Such is clear from consideration of the Jeans equation, which results from computing the 0th and 2nd order moments of the phase space *pdf*, using the Collisionless Boltzmann Equation or CBE, to connect spatial derivatives of such moments to the radially-cumulative gravitational mass distribution $M(R)$. So in general we expect that inputting unequal phase space *pdf*s in Jeans equation will imply distinct mass distributions. In contrast, that coincident mass distributions are produced from the inputting of the distinct *pdf*s for this galaxy, appears to imply that the gravitational mass distribution does not follow self-consistently from the phase space *pdf*. One - perhaps less interesting - possibility is that the consistency that is noted between the cumulative gravitational mass values computed using the ρ-parameters learnt given the data set on the two types of tracer particles, is only an artefact of the largeness of the uncertainties on the computed mass values, caused by the size of the 95% HPDs learnt on the ρ-parameters.

4.5. Model checking

A check of how good the model and results are, in the available empirical data, can be answered by checking for overlap between such empirical data, and data that is generated using the models learnt given such empirical data. In other words, we can perform model checking in our work, via our consideration of the generated data above. If the learnt models and results are compatible with the given data, then data that is generated from the learnt models, will concur with the empirical data. If such generated data does not concur with the empirical data, then the model assumptions could be wrong, and/or results of the analysis could be wrong.

We have already generated data from our learnt model of the phase space *pdf*, at the respective learnt gravitational mass density function - both learnt given an empirical data on one type of tracer particles in this galaxy. We have found (in the 3rd-Stage) that the generated and empirical GC data are closer to each other, than the generated and empirical PNe data. In other words, our results obtained with the GC data are less circumspect, than our results reported using the PNe data.

4.6. Non-normal nature of learnt pdf*s*

A Normal approximation for the phase space *pdf* is in fact worse when we perform the learning with PNe data than GC data - with the scaled Normal form unable to fit $f_{\boldsymbol{W}}(\cdot)$ learnt with the PNe data for energies $\gtrsim$ -2$\times 10^7$ km^2 s^{-2}, while this parametric form deviates from the function learnt using the GC data for energies in excess of about -1$\times 10^7$ km^2 s^{-2}, approximately.

4.7. Anisotropy and non-normality

One interesting observation that we have noted above is that the Normal is not a good fit to the phase space *pdf* learnt with either observed data set. In fact, it is a scaled (truncated) Gaussian that is an approximate fit to the mean value of the *pdf* predicted at a given energy, subsequent to our learning of the parameters that specify the Gaussian Process - a sample path of which is the sought *pdf*, as per our modelling strategy.

Additionally, we find that such a scaled Gaussian is a worse fit to the *pdf* results that are learnt using the PNe data than results obtained using the GC data. The fit is worse at higher energies. At the same time we recall that we have learnt the phase space *pdf* that the empirical PNe data are sampled from to be more anisotropic than the phase space *pdf* that the empirical GC data is sampled from. This motivates curiosity on whether *departure from (scaled) Normality at higher energies, is the cause/effect for the phase space pdf - that underlines the PNe orbital distribution in the galaxy - to be anisotropic*. From the point of view of the Central Limit Theorem, there is no a priori reason for this distribution to be Normal as the galactic particles interact gravitationally, i.e. particles are not mutually independent, but there exists correlation between the phase space variable vector of one particle and another. So we take away the lesson that a (truncated) Normal description of a phase space *pdf* that is learnt/predicted using empirical data on a certain type galactic particles, is incorrect, but adopt a data-driven answer to the question that is italicised above. A future simulation study is suggested to explore the connection between non-Normality and anisotropy of a learnt phase space *pdf*.

4.8. Comparison of results on modelling of the gravitational mass distribution in NGC4649

Using kinematic data of nearly 300 PNe, measured using the FORS2 Cassegrain spectrograph of the ESO Very Large Telescope unit 1 (Antu), [19] report that the mass of the dark matter halo component in their model, within 3 times the effective radius of this galaxy, is 4×10^{11} $M_\odot$, which is æalmost one-half of the total massg of about 1.15×10^{12} $M_\odot$ within $3R_{eff}$ in their model. They state this total mass to be similar to that estimated using globular cluster kinematic data; observations from XMM-Newton; and Chandra observations. We recall that effective radius for this galaxy is suggested to be about 9.86 kpc, [18]. Cambell [31] reports a mass of about 1.2×10^{12} $M_\odot$ to slightly in excess of 2×10^{12} $M_\odot$, as enclosed within about 4.3 times the effective radius, using tracer kinematics data input to different models, the anisotropy and form of the potential of which are varied using pre-chosen values of parameters that distinguish different potential forms, (and anisotropy), from each other. In our model-free learning of the gravitational mass density function, as stated above, the mass enclosed within 29 kpc, using GC data lies in the interval $[8.81 \times 10^{12}, 1.37 \times 10^{13}]$ $M_\odot$, while that within 22 kpc, learnt using the PNe data is in $[4.6 \times 10^{12}, 1.29 \times 10^{13}]$ $M_\odot$. We have stated in Section 4.3 why it is less inaccurate to compare values of gravitational mass density function learnt/estimated across different methods, that the enclosed mass. Das et al. [30] state that "averaging all the GCs velocity dispersions" estimated using GC kinematic data in the NMAGIC method, yields enclosed mass values - enclosed within an annular region extending from a radius of about 21.8 kpc to about 36.9 kpc - that "correspond to the values fit by Shen & Gebhardt (2010)", where Shen and Gebhardt [32] offer an enclosed mass estimate of about 10^{12} $M_\odot$. Das et al. [30] advance that their results indicate that "it is possible that the PNe and GCs trace different kinematics" in NGC4649. We find the gravitational mass density learnt using the PNe and GC kinematic data sets to be consistent within the learnt 95% HPDs, though the phase space pdfs that we learn using the data on the two types of tracers, are distinct.

5. Conclusion

This paper offers a 3-staged protocol to learn the gravitational mass density, and phase space probability density function in real galaxies, using noisy, small-sample kinematic data that are available, on galactic particles that trace the galactic gravitational field. The formulation and implementation of said protocol adopt the approach that any such galaxy is a sample point in a statistical sample; the only galaxy-specific information invoked within this implementation is the accessed kinematic data. Thus, the method allows for automated - and yet reliable - learning of the distribution of gravitational mass in a galaxy, without this result being offered as predicated upon details of the (parametric) model employed to model the mass distribution in the galaxy. In addition, the phase space *pdf* is learnt, while the major assumption of phase space isotropy that is made in the learning achievable within this approach, is tested within the available data.

It is proposed that the method be converted into a black- box for astronomers usage; this will be addressed in a future contribution. The methodology discussed here, can be generalised to include radial and energy bins the widths of which are logarithmic, than constants.

An added advantage of the method is that it can accommodate results on summaries of the mass distribution learnt/estimated using other techniques, and that too, the astronomer can impose their confidence on such a summary, within the prior structure used in the learning of the ρ-parameters. For example, an estimate of the mass enclosed within a given radius may be available for the considered galaxy, though the astronomer may be cautious about the usage of this enclosed mass value, given their lack of conviction regarding the technique used to attain this enclosed mass value. Then the relevant sum over all the relevant ρ-parameters can be computed at every iteration of the MCMC chain that is run to learn the parameters, and a Gaussian-shaped prior *pdf* on this sum is designed, with a mean given by this measured enclosed mass, while the prior variance is maintained as an adequately-chosen large value, to reflect the weak belief in the centring of the current enclosed mass on this measured enclosed mass. On the other hand, if information on such enclosed mass is obtained from a different technique - say lensing measurements, if available for this galaxy - then the astronomer may have stronger faith in the available enclosed mass. Then the (Normal) prior of the relevant sum of the ρ-parameters is designed as centred at this given enclosed mass, with a smaller value of the prior variance, compared to that used in the previous example. Using such extra information - if available - will guide the learning of the parameters better than if such information is not used in the learning.

From a purely inferential point of view, this application offers a clear example of how physically-motivated constraints of positivity and monotonicity can be imposed on the sought functions, purely through MCMC.

The ultimate aim of supervised learning of the spatial density function of the gravitating mass of all matter in the galaxy - as well as of learning the *pdf* of the phase space vector variable - appears unattainable, given that the training data that is the requisite for such learning is not available at the outset. Said training set would comprise pairs of design value of the domain variable of the sought function, and values of the function computed at this design input. So the first stage of our protocol is dedicated to the generation of the originally-absent training sets - undertaken by embedding the gravitational mass density in the support of the phase space *pdf*, within a vectorised approach to each sought function. Using the thus generated training sets, we then learn the gravitational mass density function and the phase space *pdf*, given the tracer particle kinematic data at hand. The generation of the training sets

are undertaken within the model assumptions that includes isotropy of the phase space that the empirical kinematic data is sampled from; at the last stage of this protocol, we quantify this departure of the phase space that such an empirical data set is sampled from, from invariance to rotation, i.e. to isotropy. We illustrate the method on the 2 empirical data sets that are available to us, for the galaxy NGC 4649.

Learning with these 2 data sets in this example galaxy has indicated that the phase space of this galaxy has perhaps not equilibrated in its evolution and/or the galactic phase space is split into distinct sub-volumes that are not fully mixed, at the time when said data sets were observed, with each of the 2 available empirical data sets sampled from a distinct sub-volume of this galaxy. Further to the above, self-consistent solutions for the gravitational potential may not be possible in this galaxy. Neither result is surprising given the complex, multi-component nature of the dynamical system that a galaxy is.

A future endeavour is planned, to undertake reliable and with-uncertainty prediction of the anisotropy parameter of a newly observed galaxy, by learning the functional relationship between the anisotropy parameter, and the divergence measure $\delta(\cdot,\cdot)$ - that informs on how deviant the observed galaxy is from the assumption of phase space isotropy. An in-depth simulation study is anticipated, such that we sample synthetic data sets $\mathbf{D}_1,\ldots,\mathbf{D}_n$ from n distinct known phase space *pdf*s and learn the sought ρ-parameters and f-parameters in each of these cases. We then compute the $\delta(\cdot,\cdot)$ divergence measure for each case, between the joint posterior *pdf* of all parameters given the empirical data and generated data sets, in each of the n cases. Thus, for the i-th synthetic empirical data set $\mathbf{D}_i$ - that is sampled from a phase space *pdf* model ascribed the anisotropy parameter β_i - we now know $\delta(\cdot,\cdot)$, $\forall i=1,\ldots,n$. Then using the $\{(\delta(\cdot,\cdot),\beta_i)\}_{i=1}^{n}$ training set, we can learn the relationship between the anisotropy parameter and this divergence measure. That way, for a future galaxy for which $\delta(\cdot,\cdot)$ is computed - as delineated in the method presented here - we can predict what its anisotropy parameter is.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Learning vectorised gravitational mass density and phase space *pdf* functions using synthetic data

We present results of learning the vectorised gravitational mass density function, (i.e. the vector $\boldsymbol{\rho}$), and the vectorised phase space *pdf*, (i.e. $\boldsymbol{f}$), using synthetic data sets that are simulated from respective densities. In fact, we simulate a data set $\mathbf{D}_{Iso}=\{(x1^{(i)},x2^{(i)},v3^{(i)},si)\}_{i=1}^{N_{data}}$, with 270 observations of $R_p=\sqrt{X_1^2+X_2^2}$; V_3; and error in the observed V_3, from a basal phase space *pdf* $f_{Iso}(\boldsymbol{x},\boldsymbol{v})$. Here, this *pdf* is an isotropic function of the location variable $\boldsymbol{X}$, and velocity $\boldsymbol{V}$, where this known isotropic density $f_{Iso}(\boldsymbol{x},\boldsymbol{v})$ - that is a function of energy $\varepsilon=\Psi(r)+v^2/2$ - is defined using the basal potential $\Psi(R)$, which we model as a Plummer potential. This isotropic basal phase space *pdf* $f_{Iso}(\boldsymbol{x},\boldsymbol{v})$ is proportional to $\exp(-\varepsilon/2\sigma_0^2)$ and $\Psi(R)=M_0/\sqrt{R^2+R_c^2}$ with the arbitrarily chosen values of parameter M_0 set to 4×10^{11} $\mathrm{M}_\odot$; of R_c to 1 kpc; of σ_0 to 219 km s^{-1}. G is Newton's Universal Gravitational constant that is known. The true gravitational mass density function is then the

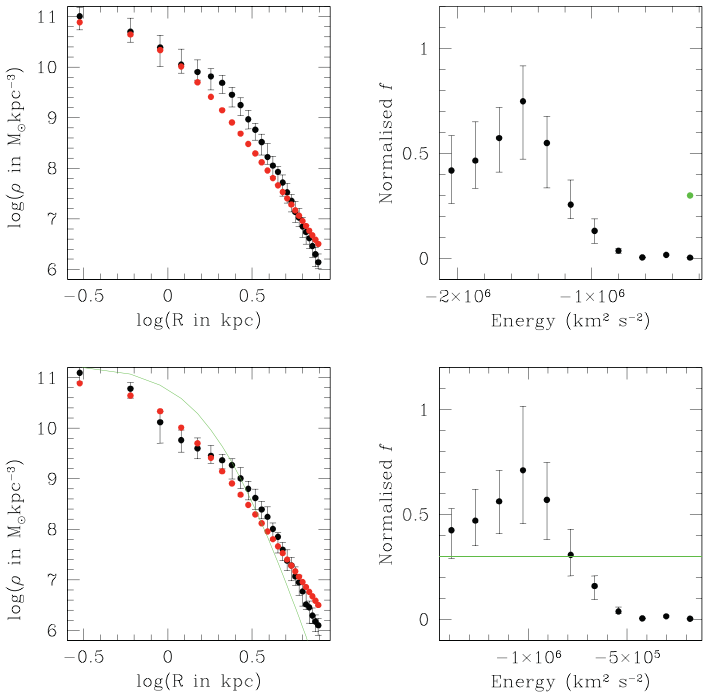

Fig. A1. *Left lower panel:* plot of the ρ-parameters learnt using the synthetic kinematic data set $\mathbf{D}_{Iso}$ that is simulated from an isotropic phase space pdf at the Plummer potential. The true Plummer gravitational mass density function computed using the used Plummer potential in the basal model, is depicted in red, (or grey in the monochrome version of the paper), at radial bins used in the learning. Mean of the learnt ρ-parameters is in black filled circles and the error bars are the learnt 95% HPDs. The MCMC chain used for this learning is initialised with ρ-parameters that are computed as values of the ǣseedg density function - depicted in green (or light grey in the monochrome version) - at the corresponding radial bin. The functional form of this seed density function and the Plummer density are discussed in the text. *Right lower panel:* plot of the f-parameters learnt using this data $\mathbf{D}_{Iso}$. The seed *pdf* is depicted in green (or light grey). *Left upper panel:* as in the lower left panel, except these ρ-parameters are learnt using the data $\mathbf{D}_{Iso}^{(gen)}$ that is generated from the simulated galactic model, the gravitational potential of which is computed using the ρ-parameters learnt using $\mathbf{D}_{Iso}$ and the phase space pdf of which is represented by the f-parameters learnt using $\mathbf{D}_{Iso}$. True values of the ρ-parameters are in red (or grey). *Right upper panel:* f-parameters learnt using data $\mathbf{D}_{Iso}^{(gen)}$.

Plummer density $GM_0/\sqrt{(R_c^2+R^2)^3}$. The data is sampled s.t. the observed values of R_p lie in the interval [0,8] kpc. Results of learning the vectorised gravitational mass density, i.e. the $\boldsymbol{\rho}$ vector, and the vectorised phase space *pdf* $\boldsymbol{f}$ - using the data $\mathbf{D}_{Iso}$ - are presented in the bottom left and right panels respectively, of Fig. A.9. To undertake this learning we use N_r =27, with R-bins that are 0.3 kpc wide each. In this learning, we use $N_e = 12$. These binning details are primarily motivated - as discussed in Section 3 - to ensure that each R-bin and ε-bin has at least 1 observation within it. Consistency between the learnt and true values of the ρ-parameters is indicated in this figure.

Multiple choices of the seed, or the initial form of the gravitational mass density function were used; these all led to consistent values (within the learnt error bars) of each ρ-parameter. The traces of the parameters displayed in Fig. A.10, indicate convergence, suggesting that the chain is irreducible (and aperiodic), which indicates lack of dependence on the initial

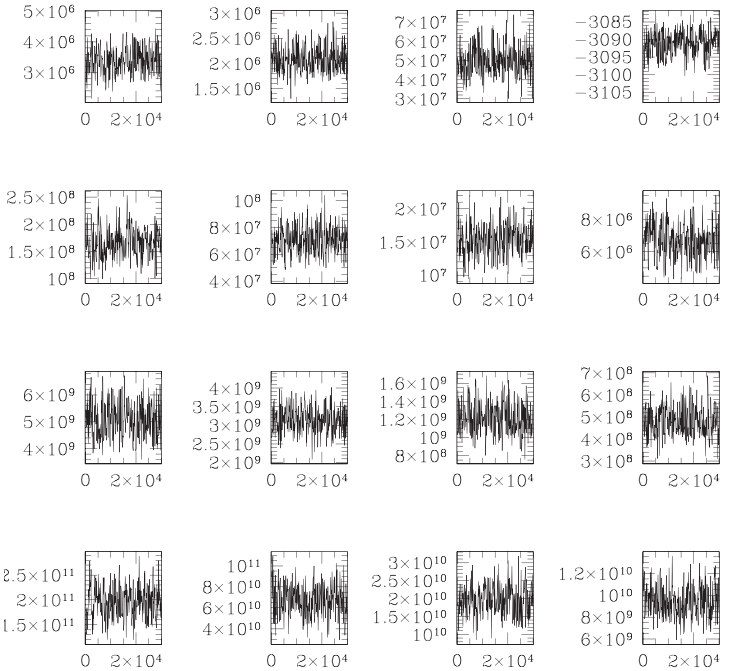

Fig. A2. Traces of various ρ-parameters that are learnt using the data $\mathbf{D}_{Iso}^{(gen)}$. The top right panel depicts the trace of the joint posterior of all the parameters that are sought, given this data.

choices of each parameter that we attempt learning, given the data $\mathbf{D}_{Iso}$. The seed gravitational mass density displayed in Fig. A.9 to learn the ρ-parameters given data $\mathbf{D}_{Iso}$, is the function $1.8 \times 10^{11}/(1+R^2/4)^5$. The seed phase space *pdf* used to learn the f-parameters given data $\mathbf{D}_{Iso}$, is chosen to be a uniform density with amplitude 0.3. The ρ-parameters and f-parameters learnt using the simulated data $\mathbf{D}_{Iso}$ under the assumption that the phase space is isotropic, are used to sample the generated data set $\mathbf{D}_{Iso}^{(gen)}$. The ρ-parameters and f-parameters learnt using this generated data set - again under the assumption of an isotropic phase space *pdf* - are displayed in the top panels of Fig. A.9. The seeds for learning the components of the $\boldsymbol{\rho}$ and $\boldsymbol{f}$ vectors using the generated data set $\mathbf{D}_{Iso}^{(gen)}$, are the respective vectors learnt using the simulated data $\mathbf{D}_{Iso}$. Other forms of the seeds, including those similar to the aforementioned functional form, are also used. Again, the learnt parameters are then consistent within the 95% HPDs, with those displayed in the top panels. This is only to be expected since the traces of the learnt parameters display convergence, implying that the chain is aperiodic and irreducible. In other words, the chain bears the ability to move to any part of the state space, having started from any other point in state space [20].

The proposal of each ρ-parameter is undertaken to ensure adherence to the monotonicity and positivity constraints on any such parameter, as discussed in Section 2.2. Similarly, the proposal of the f-parameters follow the discussion in that section. The prior on each parameter is chosen to be a Normal prior, the mean of which is the seed value of the parameter and the standard deviation of which is about 2.5 times the scaled seed value. This scale is the same for all ρ-parameters, which is different from the scale that is relevant in the learning of all the f-parameters.

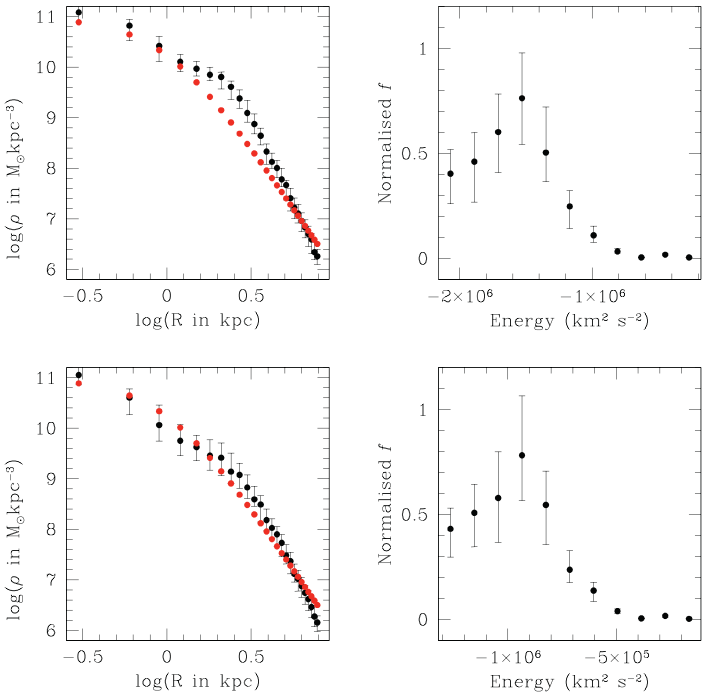

Fig. A3. As in Fig. A.1, except the ρ-parameters and f-parameters in the *left lower panel* and *right lower panel* respectively, are learnt using the synthetic data $\mathbf{D}_{Aniso}$ that is simulated from an anisotropic phase space *pdf* at the Plummer potential. ρ-parameters and f-parameters in the left upper panel and right upper panel respectively, are learnt using the data $\mathbf{D}_{Aniso}^{(gen)}$ that is generated by sampling data points from the f-parameters learnt using $\mathbf{D}_{Aniso}$ at the potential computed using the ρ-parameters learnt using this data.

Again, we simulate 270 data points on R_p, V_3, with observational error on V_3, by simulating from a basal phase space $pdff_{Aniso}(\boldsymbol{x}, \boldsymbol{v})$ that is an anisotropic function of the location variable $\boldsymbol{X}$ and velocity variable $\boldsymbol{V}$. This anisotropic basal phase space *pdf* is $f_{Aniso}(\boldsymbol{x}, \boldsymbol{v}) \propto \exp(-\varepsilon/2\sigma_0^2)\exp(-L_z^2/(R_a^2\sigma_0^2))$ and the basal potential is $\Psi(R) = -GM_0/\sqrt{R^2 + R_c^2}$. The model parameters M_0, R_c, σ_0 are as used in the case in which data was simulated from an isotropic phase space *pdf*, while the parameter R_a is set as 4 kpc. Results of learning the ρ-parameters and f-parameters using data $\mathbf{D}_{Aniso}$ are depicted in Fig. A.11. The proposal and priors used to run the MCMC chain using this data are as used when learning given data $\mathbf{D}_{Iso}$. Seeds for the $\boldsymbol{\rho}$ and $\boldsymbol{f}$ vectors in the learning with $\mathbf{D}_{Aniso}$ are also the same as those used when learning with data $\mathbf{D}_{Iso}$. The $\boldsymbol{\rho}$ and $\boldsymbol{f}$ vectors learnt using $\mathbf{D}_{Aniso}$ are input to a rejection sampling algorithm, and another data set - called $\mathbf{D}_{Aniso}^{(gen)}$ - comprising 270 number of the observables, is generated.

To predict the level of anisotropy of the phase space of the galaxy under consideration, we also computed the divergence measure $\delta(\cdot,\cdot)$ between the (logarithm) of the joint posterior probability density $\pi_{Iso}(\cdot|\mathbf{D}_{Iso})$ of all ρ-parameters and f-parameters given data $\mathbf{D}_{Iso}$, and the joint $\pi_{Iso}(\cdot|\mathbf{D}_{Iso}^{(gen)})$ given the generated data $\mathbf{D}_{Iso}^{(gen)}$. We then compare this value of the divergence measure to the same computed given $\mathbf{D}_{Aniso}$ and $\mathbf{D}_{Aniso}^{(gen)}$. We find that $\delta(\pi_{Iso}, \pi_{Iso}^{(gen)}) \approx 0.03237$, while $\delta(\pi_{Aniso}, \pi_{Aniso}^{(gen)}) \approx 0.1046$.

Appendix B. Computing gravitational potential from gravitational mass density

The gravitational potential $\Psi(R)$ at radius R, is computed by inputting the gravitational mass density function $\rho(R)$ in Poisson equation. It is easier to appreciate this computation of the potential using the gravitational mass $M(R)$ that is enclosed within the sphere of radius R, i.e. $M(R) = \int_{x=0}^{R} 4\pi\rho(x)x^2dx$ and G is the known (Universal Gravitational) constant. Then for the N_r number of radial bins used in our learning - with each bin of width δ_r - at $R = r$, $\Psi(r) = \dfrac{-GM(r)}{r}$, where

$$
\begin{aligned}
M(r) &= \sum_{s=1}^{t} \frac{4\pi}{3}\left[s^3\delta_r^3 - (s-1)^3\delta_r^3\right]\rho_s + \frac{4\pi}{3}\left[R^3 - (t\delta_r)^3\right]\rho_{t+1}, \\
&\quad \text{for } r \in [t\delta_r, (t+1)\delta_r), \\
M(r) &= \sum_{s=1}^{N_r} \frac{4\pi}{3}\left[s^3\delta_r^3 - (s-1)^3\delta_r^3\right]\rho_s, \ \text{ for } r \geq N_r\delta_r \\
M(r) &= \frac{4\pi}{3}[r^3]\rho_1, \ \text{ for } r \in [0, \delta_r].
\end{aligned} \tag{B.1}
$$

References

[1] K.-H. Chae, M. Bernardi, A.V. Kravtsov, Modelling mass distribution in elliptical galaxies: mass profiles and their correlation with velocity dispersion profiles, Mon. Not. R. Astron. Soc. 437 (4) (2013) 3670–3687, doi:10.1093/mnras/stt2163.

[2] J.F. Navarro, C.S. Frenk, S.D.M. White, A universal density profile from hierarchical clustering, Astrophys. J. 490 (2) (1997) 493–508, doi:10.1086/304888.

[3] D. Lynden-Bell, Statistical mechanics of violent relaxation in stellar systems, Mon. Not. R. Astron. Soc. 136 (1) (1967) 101–121, doi:10.1093/mnras/136.1.101.

[4] M. Hilz, T. Naab, J.P. Ostriker, How do minor mergers promote inside-out growth of ellipticals, transforming the size, density profile and dark matter fraction? Mon. Not. R. Astron. Soc. 429 (4) (2013) 2924–2933, doi:10.1093/mnras/sts501.

[5] R.-S. Remus, A. Burkert, K. Dolag, P.H. Johansson, T. Naab, L. Oser, J. Thomas, The dark halo—spheroid conspiracy and the origin of elliptical galaxies, Astrophys. J. 766 (2) (2013) 71, doi:10.1088/0004-637x/766/2/71.

[6] A. Agnello, N.W. Evans, A.J. Romanowsky, Dynamical models of elliptical galaxies I. Simple methods, Mon. Not. R. Astron. Soc. 442 (4) (2014) 3284–3298, doi:10.1093/mnras/stu959.

[7] C. Marsden, F. Shankar, M. Bernardi, R. Sheth, H. Fu, A. Lapi, The weak dependence of velocity dispersion on disk fractions, mass-to-light ratio and redshift: implications for galaxy and black hole evolution, Mon. Not. R. Astron. Soc. 510 (2021), doi:10.1093/mnras/stab3705.

[8] H. Domnguez Snchez, M. Bernardi, J.R. Brownstein, N. Drory, R.K. Sheth, Galaxy properties as revealed by MaNGA I. Constraints on IMF and M*/L gradients in ellipticals, Mon. Not. R. Astron. Soc. 489 (4) (2019) 5612–5632, doi:10.1093/mnras/stz2414.

[9] J. Binney, S. Tremaine, Galactic Dynamics, 13, Princeton University Press, 2011.

[10] J. Binney, Dynamics of elliptical galaxies and other spheroidal components, Annu. Rev. Astron. Astrophys. 20 (1982) 399–429, doi:10.1146/annurev.aa.20.090182.002151.

[11] N.N. Patra, A self-consistent hydrostatic mass modelling of pressure-supported dwarf galaxy Leo T, Mon. Not. R. Astron. Soc. 480 (4) (2018) 4369–4378, doi:10.1093/mnras/sty2167.

[12] P.R. Capelo, P. Natarajan, P.S. Coppi, Hydrostatic equilibrium profiles for gas in elliptical galaxies, Mon. Not. R. Astron. Soc. 407 (2) (2010) 1148–1156, doi:10.1111/j.1365-2966.2010.16962.x.

[13] K. Vogtmann, A. Weinstein, V. Arnol'd, Mathematical methods of classical mechanics, Graduate Texts in Mathematics, Springer New York, 1997. https://books.google.co.uk/books?id=Pd8-s6rOt_cC

[14] H. Goldstein, C. Poole, J. Safko, Classical Mechanics, Addison Wesley, 2002. https://books.google.co.uk/books?id=tJCuQgAACAAJ

[15] P. Kelly, Solid mechanics part I: an introduction to solid mechanics. Solid mechanics lecture notes, https://pkel015.connect.amazon.auckland.ac.nz/SolidMechanicsBooks/.

[16] T. Bridges, K. Gebhardt, R. Sharples, F.R. Faifer, J.C. Forte, M.A. Beasley, S.E. Zepf, D.A. Forbes, D.A. Hanes, M. Pierce, The globular cluster kinematics and galaxy dark matter content of NGC 4649 (M60), Mon. Not. R. Astron. Soc. 373 (1) (2006) 157–166, doi:10.1111/j.1365-2966.2006.10997.x.

[17] V. Pota, J.P. Brodie, T. Bridges, J. Strader, A.J. Romanowsky, A. Villaume, Z. Jennings, F.R. Faifer, N. Pastorello, D.A. Forbes, A. Campbell, C. Usher, C. Foster, L.R. Spitler, N. Caldwell, J.C. Forte, M.A. Norris, S.E. Zepf, M.A. Beasley, K. Gebhardt, D.A. Hanes, R.M. Sharples, J.A. Arnold, A SLUGGS and Gemini/GMOS combined study of the elliptical galaxy M60: wide-field photometry and kinematics of the globular cluster system, Mon. Not. R. Astron. Soc. 450 (2) (2015) 1962–1983, doi:10.1093/mnras/stv677.

[18] H.S. Hwang, M.G. Lee, H.S. Park, S.C. Kim, J.-H. Park, Y.J. Sohn, S. Lee, S.-C. Rey, Y.-W. Lee, H. Univ, Kasi, Y. Univ, C.N. Univ, The globular cluster system of m60 (NGC 4649). ii. kinematics of the globular cluster system, Astrophys. J. 674 (2008) 869–885.

[19] A.M. Teodorescu, R.H. Mendez, F. Bernardi, J. Thomas, P. Das, O. Gerhard, Planetary nebulae in the elliptical galaxy NGC 4649 (m 60): kinematics and distance redetermination, Astrophys. J. 736 (1) (2011), doi:10.1088/0004-637X/736/1/65.

[20] C. Robert, G. Casella, Monte Carlo statistical methods, Springer Texts in Statistics, Springer New York, 2005. https://books.google.co.uk/books?id=HfhGAxn5GugC

[21] H.E. Kandrup, Invariant distributions and collisionless equilibria, Mon. Not. R. Astron. Soc. 299 (4) (1998) 1139–1145, doi:10.1046/j.1365-8711.1998.01865.x.

[22] J. Geweke, H. Tanizaki, Bayesian estimation of state-space models using the metropolishastings algorithm within Gibbs sampling, Comput. Stat. Data Anal. 37 (2) (2001) 151–170, doi:10.1016/S0167-9473(01)00009-3.

[23] C. Rasmussen, C. Williams, Gaussian processes for machine learning, Adaptive Computation and Machine Learning Series, MIT Press, 2005. https://books.google.co.uk/books?id=Tr34DwAAQBAJ

[24] S. Roberts, M. Osborne, M. Ebden, S. Reece, N. Gibson, S. Aigrain, Gaussian processes for time-series modelling, Philos. Trans., Math. Phys. Eng. Sci. 371 (2013) 20110550, doi:10.1098/rsta.2011.0550.

[25] A. Gut, Probability: a graduate course, Springer Texts in Statistics, Springer New York, 2013. https://books.google.co.uk/books?id=9TmRgPg-6vgC

[26] M.-H. Chen, Q.-M. Shao, Monte carlo estimation of Bayesian credible and HPD intervals, J. Comput. Graph. Stat. 8 (1) (1999) 69–92. http://www.jstor.org/stable/1390921

[27] D. Chakrabarty, A new Bayesian test to test for the intractability-countering hypothesis, J. Am. Stat. Assoc. 112 (518) (2017) 561–577, doi:10.1080/01621459.2016.1240684.

[28] S. Banerjee, A. Basu, S. Bhattacharya, S. Bose, D. Chakrabarty, S.S. Mukherjee, Minimum distance estimation of milky way model parameters and related inference, SIAM/ASA J. Uncertain. Quantif. 3 (1) (2015) 91–115, doi:10.1137/130935525.

[29] G. Corani, A. Benavoli, M. Zaffalon, Time series forecasting with gaussian processes needs priors, in: Y. Dong, N. Kourtellis, B. Hammer, J.A. Lozano (Eds.), Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track, Springer International Publishing, Cham, 2021, pp. 103–117.

[30] P. Das, O. Gerhard, R.H. Mendez, A.M. Teodorescu, F. de Lorenzi, Using NMAGIC to probe the dark matter halo and orbital structure of the X-ray bright, massive elliptical galaxy, NGC 4649, Mon. Not. R. Astron. Soc. 415 (2) (2011) 1244–1258, doi:10.1111/j.1365-2966.2011.18771.x. 1105.3478

[31] A. Cambell, Globular Cluster Kinematics and Dark Mat- Ter Content of NGC 4649, Department of Physics, Engineering Physics and Astronomy, Queens University Kingston, Ontario, Canada, 2011. In conformity with the requirements for the degree of Master of Science

[32] J. Shen, K. Gebhardt, The supermassive black hole and dark matter halo of NGC 4649 (m60), Astrophys. J. 711 (1) (2010) 484.