Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **506,** 5721–5730 (2021)
Advance Access publication 2021 July 21

https://doi.org/10.1093/mnras/stab2049

Charting galactic accelerations: when and how to extract a unique potential from the distribution function

J. An ,[1] A. P. Naik ,[2] N. W. Evans[3]★ and C. Burrage[2]

[1]Center for Theoretical Astronomy, Korea Astronomy & Space Science Institute, 776 Daedeok-daero, Yuseong-gu, Daejeon 34055, Korea (South)
[2]School of Physics & Astronomy, University of Nottingham, University Park, Nottingham NG7 2RD, UK
[3]Institute of Astronomy, University of Cambridge, Madingley Rd., Cambridge CB3 0HA, UK

Accepted 2021 July 9. Received 2021 July 5; in original form 2021 May 29

ABSTRACT

The advent of data sets of stars in the Milky Way with 6D phase-space information makes it possible to construct empirically the distribution function (DF). Here, we show that the accelerations can be uniquely determined from the DF using the collisionless Boltzmann equation, providing the Hessian determinant of the DF with respect to the velocities is non-vanishing. We illustrate this procedure and requirement with some analytic examples. Methods to extract the potential from data sets of discrete positions and velocities of stars are then discussed. Following Green & Ting, we advocate the use of normalizing flows on a sample of observed phase-space positions to obtain a differentiable approximation of the DF. To then derive gravitational accelerations, we outline a semi-analytic method involving direct solutions of the overconstrained linear equations provided by the collisionless Boltzmann equation. Testing our algorithm on mock data sets derived from isotropic and anisotropic Hernquist models, we obtain excellent accuracies even with added noise. Our method represents a new, flexible, and robust means of extracting the underlying gravitational accelerations from snapshots of 6D stellar kinematics of an equilibrium system.

Key words: methods: analytical – methods: data analysis – Galaxy: fundamental parameters – Galaxy: kinematics and dynamics – galaxies: fundamental parameters – galaxies: kinematics and dynamics.

1 INTRODUCTION

In galactic astronomy, a fundamental problem is to extract the underlying gravitational potential from the kinematics of a tracer population. If stars are moving on circular orbits in a spherical potential, then the matching of the centrifugal force to the gravitational one gives the rotation curve, and by extension the potential. Elaborations of this basic idea to stellar streams have proved to be one of the most powerful methods available to us today (e.g. Lynden-Bell 1982; Johnston et al. 1999; Bowden, Belokurov & Evans 2015; Erkal et al. 2019; Malhan & Ibata 2019).

If the stellar population is not kinematically cold, the traditional way in which the problem is tackled is via the Jeans equations (Binney & Tremaine 2008, chapter 4). Given measurements of the second velocity moments and the density of the tracer population, the Jeans equations can be solved to yield the potential. There are numerous applications of this method both to the Milky Way (e.g. King et al. 2015; Bowden, Evans & Williams 2016; Nitschai, Cappellari & Neumayer 2020) and external galaxies (e.g. Cappellari 2008; Walker et al. 2009). Some studies have instead worked directly with the distribution function, fitting some assumed parametric form to the observed stellar data (e.g. Binney & Piffl 2015; Williams & Evans 2015; Posti & Helmi 2019). More rarely, the distribution function is constructed directly from the data, as in Kuijken & Gilmore (1989)'s numerical Abel inversion of the vertical tracer density. This though relies on the assumption that the vertical and in-plane dynamics are decoupled, and so is not of general applicability.

However, the advent of the *Gaia* satellite (Gaia Collaboration 2016) has made possible the empirical construction of the full phase-space distribution function for stellar populations in the Milky Way, and perhaps even for some of its satellite galaxies. The data now comprise the full positions and velocities of many millions of stars. The process of averaging to obtain the second velocity moments does not do justice to the richness of the data. Green & Ting (2020) recently raised the possibility of direct determination of the gravitational potential from the distribution function using the collisionless Boltzmann equation itself, which is the continuity equation satisfied by the distribution function in the 6D phase space of positions and velocities.

At every location in physical space, the collisionless Boltzmann equation provides a *single* constraint on the *three* unknown components of the gravitational force. Thus, it is unclear if the identification of a stationary distribution function is sufficient to specify the gravitational potential (modulo an additive constant) uniquely. So, the first aim of our paper is to establish the conditions under which the potential can be uniquely recovered, given the distribution function. The second aim of our paper is to provide a working algorithm to extract the potential. Whereas Green & Ting (2020) proposed a neural network, we instead utilize an efficient and accurate semi-analytic method, based on a direct solution of the collisionless Boltzmann equation. We demonstrate the efficacy of our method on mock data sets sampled from isotropic and anisotropic distribution functions of galaxy models, including the effects of errors.

★ E-mail: nwe@ast.cam.ac.uk

Published by Oxford University Press on behalf of Royal Astronomical Society

2 THE COLLISIONLESS BOLTZMANN EQUATION AND THE POTENTIAL

Here, we address the theoretical question that underlies all this work: namely, when is the potential uniquely specified by the distribution function? We prove a uniqueness theorem in Section 2.1 subject to certain conditions, and investigate the instances when the conditions are violated in Section 2.2.

2.1 Uniqueness theorem

If $F(\boldsymbol{p};\boldsymbol{x})$ is a phase-space distribution function (DF) in equilibrium in the static potential $\Phi(\boldsymbol{x})$, then it is an integral of motion of the Hamiltonian $H=\frac{1}{2}\sum_{j,k=1}^{3}g^{jk}p_jp_k+\Phi(\boldsymbol{x})$. Here $\boldsymbol{p}=(p_1,p_2,p_3)$ is the momentum component conjugate to a coordinate set $\boldsymbol{x}=(x^1,x^2,x^3)$ with the metric coefficients g_{ij} and its inverse g^{ij}. A mathematical representation of F being an integral of motion is given by the vanishing Poisson bracket of the integral F with the Hamiltonian H (An & Evans 2016), namely,

$$\begin{aligned}\{F,H\}&=\sum_{i=1}^{3}\left(\frac{\partial F}{\partial x^i}\frac{\partial H}{\partial p_i}-\frac{\partial F}{\partial p_i}\frac{\partial H}{\partial x^i}\right)\\&=\sum_{i=1}^{3}\left[\sum_{j=1}^{3}g^{ij}p_j\frac{\partial F}{\partial x^i}-\left(\frac{1}{2}\sum_{j,k=1}^{3}\frac{\partial g^{jk}}{\partial x^i}p_jp_k+\frac{\partial\Phi}{\partial x^i}\right)\frac{\partial F}{\partial p_i}\right]=0.\end{aligned}\tag{1}$$

Considered as a partial differential equation for F, this is equivalent to the (time-independent) collisionless Boltzmann equation (CBE).[1] Since equation (1) is a linear homogeneous equation for F, any function $F=f(J)$ is its solution if $J=J(\boldsymbol{p};\boldsymbol{x})$ is also a solution. That is, the CBE only describes a (necessary) condition for the DF to be stationary and cannot uniquely determine the DF for any given potential. In fact, physical considerations make it obvious that many different DFs can indeed be in equilibrium with the given potential.

On the other hand, if a stationary DF is known, the CBE may also be interpreted as a partial differential equation for the potential. Here the question is whether the given DF (or more generally an integral of motion) can determine a unique potential through the CBE. The CBE is linear in Φ (albeit non-homogeneous) and so there exists a gauge freedom such that, if Φ_0 is a particular solution, the function $\Phi_0+G(I)$, where $G(I)$ is an arbitrary function of a particular solution I to the homogeneous counterpart, also satisfies the same inhomogeneous differential equation. However, the potential $\Phi=\Phi(\boldsymbol{x})$ is a function of only the configuration-space coordinates, whereas the CBE is a partial differential equation in phase space. In other words, we must only consider the solutions that are also constant along any direction in momentum space; that is, the solution must also be subject to the constraints that $\partial\Phi/\partial p_1=\partial\Phi/\partial p_2=\partial\Phi/\partial p_3=0$. Are these then sufficient to uniquely determine the potential Φ for the given DF?

Let us suppose that a DF $F(\boldsymbol{p};\boldsymbol{x})$ is known to be stationary in the potential $\Phi_0(\boldsymbol{x})$. Then it follows that $\{F,H_0\}=0$ where $H_0=\sum_{i,j=1}^{3}\frac{1}{2}g^{ij}p_ip_j+\Phi_0$ or

$$\sum_{i=1}^{3}\left(\frac{\partial F}{\partial x^i}\sum_{j=1}^{3}g^{ij}p_j-\frac{1}{2}\frac{\partial F}{\partial p_i}\sum_{j,k=1}^{3}\frac{\partial g^{jk}}{\partial x^i}p_jp_k\right)=\sum_{i=1}^{3}\frac{\partial\Phi_0}{\partial x^i}\frac{\partial F}{\partial p_i}.\tag{2}$$

If there exists another potential Φ which the same F is also a stationary DF in, the potential Φ satisfies the CBE with F in equation (1) or equivalently equation (2) but with $\Phi_0\to\Phi$. Eliminating the common terms between two CBEs, we can construct a homogeneous linear partial differential equation for the difference $\Phi-\Phi_0$:

$$\frac{\partial(\Phi-\Phi_0)}{\partial x^1}\frac{\partial F}{\partial p_1}+\frac{\partial(\Phi-\Phi_0)}{\partial x^2}\frac{\partial F}{\partial p_2}+\frac{\partial(\Phi-\Phi_0)}{\partial x^3}\frac{\partial F}{\partial p_3}=0.\tag{3}$$

Here $\Phi-\Phi_0$ is a function of only the real-space component (x^1,x^2,x^3), whereas F is a function of phase space in general. Thus taking the partial derivative with respect to one of the momentum components results in the set of three differential equations,

$$\sum_{i=1}^{3}\frac{\partial(\Phi-\Phi_0)}{\partial x^i}\frac{\partial^2F}{\partial p_j\partial p_i}=0\qquad(\text{where }j=1,2,3).\tag{4a}$$

Since the Hessian matrix $[\partial_{p_i}\partial_{p_j}F]$ (where $\partial_{p_i}=\partial/\partial p_i$ and so on) is real symmetric, it is diagonalizable at least locally by a point-wise orthogonal transformation. In the local coordinate diagonalizing the Hessian (in which $\partial_{\tilde{p}_i}\partial_{\tilde{p}_j}F=0$ for $i\neq j$), equations (4a) reduce to

$$\lambda_1\frac{\partial(\Phi-\Phi_0)}{\partial q^1}=\lambda_2\frac{\partial(\Phi-\Phi_0)}{\partial q^2}=\lambda_3\frac{\partial(\Phi-\Phi_0)}{\partial q^3}=0.\tag{4b}$$

Therefore, if $\lambda_i=\partial^2_{\tilde{p}_i}F\neq0$ for a direction in the transformed coordinate, then $\partial(\Phi-\Phi_0)/\partial q^i=0$ along the conjugate coordinate direction associated with the non-zero eigenvalue λ_i. If m is the rank (i.e. the number of non-zero eigenvalues) of the Hessian, the difference $\Phi-\Phi_0$ is consequently an arbitrary function of $3-m$ functionally-independent functions $q^j=q^j(x^1,x^2,x^3)$, which are the coordinate functions corresponding to the eigenvectors associated with the null eigenvalues.

In particular, if the Hessian determinant

$$\det\left[\frac{\partial^2F}{\partial p_i\partial p_j}\right]=\det\left[\frac{\partial^2F}{\partial\tilde{p}_i\partial\tilde{p}_j}\right]=\lambda_1\lambda_2\lambda_3\neq0\tag{5}$$

is non-vanishing, then $\lambda_i\neq0$ for all i and $m=3$. Solving equations (4a) as a series of linear equations for $\partial(\Phi-\Phi_0)/\partial x^i$ then results in

$$\frac{\partial(\Phi-\Phi_0)}{\partial x^1}=\frac{\partial(\Phi-\Phi_0)}{\partial x^2}=\frac{\partial(\Phi-\Phi_0)}{\partial x^3}=0\ \Rightarrow\ \Phi=\Phi_0+C,\tag{6}$$

where C is an arbitrary constant; that is, the potential $\Phi(\boldsymbol{x})$ satisfying the CBE for a given DF, if it exists, is essentially unique up to an additive constant (resulting in the identical gravitational acceleration field). In other words, the non-vanishing Hessian of equation (5) is a sufficient condition for the uniqueness of the potential for a given stationary DF.

2.2 Are there physical DFs that do not specify a unique potential?

If the Hessian $[\partial_{p_i}\partial_{p_j}F]$ is singular, there exists a local momentum-space coordinate system $(\tilde{p}_1,\tilde{p}_2,\tilde{p}_3)$ such that the directional derivative of F in a fixed coordinate direction must be constant in momentum space. That is to say, the singularity condition indicates that at least one eigenvalue, which is the second-order partial derivative in the corresponding coordinate direction, must be zero (i.e. $\lambda_j=\partial^2_{\tilde{p}_j}F=0$ for $\exists j$). Since the coordinate can be chosen to be orthogonal so that all the second-order cross partial derivatives vanish ($\partial_{\tilde{p}_i}\partial_{\tilde{p}_j}F=0$ if $i\neq j$), there then exists a coordinate system in which all the second derivatives involving one particular coordinate should

[1] The CBE is actually derived from arguments based on number (or probability) conservation (see Binney & Tremaine 2008, section 4.1). In fact, F being an integral of motion may be interpreted as a consequence of the CBE.

be zero (i.e. $\partial_{\tilde{p}_i}\partial_{\tilde{p}_j}F = 0$ for $\forall i$ and $\exists j$). Therefore, the directional derivative of F in the same coordinate direction must be constant in momentum space; that is, $\partial_{\tilde{p}_j}F = k_0(\boldsymbol{x})$ for $\exists j$. In the original coordinates, this implies that $\sum_{i=1}^{3} k_i\partial_{p_i}F = k_0$ where k_i's are the constants in momentum space (but they are functions of the real-space positions) and at least one of $\{k_1(\boldsymbol{x}), k_2(\boldsymbol{x}), k_3(\boldsymbol{x})\}$ is non-zero. In fact, if there are two or more distinct potentials satisfying the CBE with the given DF, equation (3) further indicates that there exists $\{k_1, k_2, k_3\}$ such that $k_1^2 + k_2^2 + k_3^2 \neq 0$ and $(k_1\partial_{p_1} + k_2\partial_{p_2} + k_3\partial_{p_3})F = 0$.

In other words, if the function F is an integral of motion in two (or more) distinct – as in generating different gravitational accelerations – potentials, then there exists a fixed direction (k_1, k_2, k_3) in momentum space that is tangent to the level surfaces of the DF everywhere in momentum space. However, the integral curve of a constant vector is a straight line and momentum space is topologically equivalent to $\mathbb{R}^3$. Consequently, all the level surfaces of F have infinite extent and the inverse image of any real interval under F^{-1} in momentum space cannot have a compact support (unless empty). That is to say, such a function F is not integrable and cannot be a physical DF.

In light of this, we argue that the unique determination of the potential is a property related to the global behaviour in momentum space. That is to say, the CBE only describes the balance amongst the gradients of the DF and the external acceleration field in the local neighbourhood of a fixed phase-space location, whilst the external gravitational acceleration is shared in the whole momentum space at a fixed real-space position. By joining all the constraints on the acceleration fields coming from the CBE in different momentum-space locations (but at a fixed real-space position), we can narrow down to the unique acceleration. This fact is also demonstrated by the examples presented in the following section (Section 3) where a unique potential actually follows from insisting that the CBE holds for all values of the momentum components.

3 EXAMPLES

To gain insight into the steps needed to extract a unique potential from the CBE, we first look at some analytic examples.

3.1 Ergodic distributions: a unique potential

We start by examining the case of an ergodic DF $F = f(E)$ in a fixed potential $\Phi_0(\boldsymbol{x})$, where $E = \frac{1}{2}v^2 + \Phi_0$ is the specific energy and is known as a function of the phase-space coordinates. Here, no further assumption is made on the self-consistency of the system and so the potential need not be spherically symmetric (cf. An, Evans & Sanders 2017). In Cartesian coordinates, the CBE is then reducible to the differential equation on the difference between any two possible potentials,

$$\left(v_x\frac{\partial E}{\partial x} + v_y\frac{\partial E}{\partial y} + v_z\frac{\partial E}{\partial z} - \frac{\partial\Phi}{\partial x}\frac{\partial E}{\partial v_x} - \frac{\partial\Phi}{\partial y}\frac{\partial E}{\partial v_y} - \frac{\partial\Phi}{\partial z}\frac{\partial E}{\partial v_z}\right) f'(E)$$
$$= f'(E)\left(v_x\frac{\partial}{\partial x} + v_y\frac{\partial}{\partial y} + v_z\frac{\partial}{\partial z}\right)(\Phi_0 - \Phi) = 0. \tag{7}$$

Assuming that the DF in itself is not constant, that is, $f'(E) \neq 0$, then in order for this to hold everywhere in phase space,

$$\frac{\partial(\Phi - \Phi_0)}{\partial x} = \frac{\partial(\Phi - \Phi_0)}{\partial y} = \frac{\partial(\Phi - \Phi_0)}{\partial z} = 0. \tag{8}$$

Therefore $\Phi = \Phi_0 + C$ and the potential is unique (up to an additive constant).

3.2 Separable potentials with third integrals

If there exists a DF of the form $F = f(J)$ where $f'(J) \neq 0$ and J is a quadratic polynomial of p_i's such that $J = \sum_{i,j=1}^{3}\frac{1}{2}K^{ij}(\boldsymbol{x})p_ip_j + \sum_{i=1}^{3}X^i(\boldsymbol{x})p_i + \xi(\boldsymbol{x})$, then the resulting CBE in equation (1) reduces to a cubic polynomial equation on p_i's. This is of course the old 'ellipsoidal hypothesis' (see Chandrasekhar 1939; Camm 1941; Evans & Lynden-Bell 1991; and references therein). Assuming that the DF is stationary, the CBE should hold for any p_i's and so the coefficients to all the monomial terms ($p_ip_jp_k$, p_ip_j, and p_i etc.) must vanish identically. It is then found that the coefficients to the cubic and quadratic terms respectively only involve the tensor K^{ij} and the vector X^i, and the first-order partial differential equations resulting from setting them to be zero restrict the possible forms for K^{ij} and X^i (An 2013 and references therein). However if the DF is already given and known to be stationary, these conditions must hold automatically.

On the other hand, setting the coefficients to the linear terms to be zero results in the set of three differential equations,

$$\sum_{i=1}^{3} g^{ij}\frac{\partial\xi}{\partial x^i} = \sum_{i=1}^{3} K^{ij}\frac{\partial\Phi}{\partial x^i} \qquad \text{(where } j = 1, 2, 3\text{)}. \tag{9}$$

If $\xi(\boldsymbol{x})$ is known, these can be considered as the coupled differential equations on the potential Φ. Provided that the matrix $[K^{ij}]$ is invertible (here also note that $K^{ij} = \partial_{p_i}\partial_{p_j}J$), equation (9) can be uniquely solved for $\partial\Phi/\partial x^i$ so that

$$\frac{\partial\Phi}{\partial x^i} = \sum_{j=1}^{3} K_{ij}^{-1}\left(\sum_{k=1}^{3} g^{jk}\frac{\partial\xi}{\partial x^k}\right) \qquad \text{(where } i = 1, 2, 3\text{)}, \tag{10}$$

where K_{ij}^{-1} is the matrix element of the inverse matrix of $[K^{ij}]$. In other words, if the local DF that is a function of a non-degenerate quadratic form of the canonical momenta is stationary, the gravitational acceleration is uniquely specified in the neighbourhood.

As a concrete example, suppose that there exists a stationary DF of the form $F = f(J)$ where

$$J = \frac{\ell^2 + a^2v_z^2}{2} + \xi(R, z); \qquad \xi = \frac{ka|z|}{[R^2 + (|z| + a)^2]^{1/2}} \tag{11}$$

with constants a and k. Here J is the third integral of the Kuzmin (1956) disc potential in the cylindrical polar coordinate (R, ϕ, z), and $\ell = \|\boldsymbol{\ell}\| = (\boldsymbol{\ell}\cdot\boldsymbol{\ell})^{1/2}$ is the magnitude of the specific angular momentum. Here, $\boldsymbol{\ell} = \boldsymbol{x}\times\dot{\boldsymbol{x}} = (R\hat{\boldsymbol{e}}_R + z\hat{\boldsymbol{e}}_z)\times(v_R\hat{\boldsymbol{e}}_R + v_\phi\hat{\boldsymbol{e}}_\phi + v_z\hat{\boldsymbol{e}}_z)$ and so follows that $\ell^2 = (zv_R - Rv_z)^2 + (R^2 + z^2)v_\phi^2$, whilst $(p_R, p_\phi, p_z) = (v_R, Rv_\phi, v_z)$. Provided $f'(J) \neq 0$, the CBE in the corresponding canonical phase-space coordinate $(p_R, p_\phi, p_z; R, \phi, z)$ then results in (here $r^2 = R^2 + z^2$)

$$p_R\frac{\partial\xi}{\partial R} + p_z\frac{\partial\xi}{\partial z} + z(Rp_z - zp_R)\frac{\partial\Phi}{\partial R} + \left[Rzp_R - (R^2 + a^2)p_z\right]\frac{\partial\Phi}{\partial z} - \frac{r^2p_\phi}{R^2}\frac{\partial\Phi}{\partial\phi}$$
$$= p_R\left(\frac{\partial\xi}{\partial R} - z^2\frac{\partial\Phi}{\partial R} + Rz\frac{\partial\Phi}{\partial z}\right)$$
$$+ p_z\left[\frac{\partial\xi}{\partial z} + Rz\frac{\partial\Phi}{\partial R} - (R^2 + a^2)\frac{\partial\Phi}{\partial z}\right] - \frac{r^2p_\phi}{R^2}\frac{\partial\Phi}{\partial\phi} = 0. \tag{12}$$

Since this holds for all (p_R, p_ϕ, p_z), we have $\partial\Phi/\partial\phi = 0$ and

$$z^2\frac{\partial\Phi}{\partial R} - Rz\frac{\partial\Phi}{\partial z} = \frac{\partial\xi}{\partial R} = -\frac{kaR|z|}{[R^2 + (|z| + a)^2]^{3/2}}; \tag{13a}$$

$$(R^2 + a^2)\frac{\partial\Phi}{\partial z} - Rz\frac{\partial\Phi}{\partial R} = \frac{\partial\xi}{\partial z} = \frac{z}{|z|}\frac{ka[R^2 + a(|z| + a)]}{[R^2 + (|z| + a)^2]^{3/2}}, \tag{13b}$$

where we have used $\partial|z|/\partial z = z/|z|$ (NB: $\partial\xi/\partial z$ at $z = 0$ does not exist). If $a \neq 0$, we can solve equations (13a,b) for $\partial\Phi/\partial R$ and $\partial\Phi/\partial z$

$$\frac{\partial\Phi}{\partial R} = \frac{1}{a^2}\left(\frac{R^2+a^2}{z^2}\frac{\partial\xi}{\partial R} + \frac{R}{z}\frac{\partial\xi}{\partial z}\right) = \frac{kR}{[R^2+(|z|+a)^2]^{3/2}}; \tag{14a}$$

$$\frac{\partial\Phi}{\partial z} = \frac{1}{a^2}\left(\frac{R}{z}\frac{\partial\xi}{\partial R} + \frac{\partial\xi}{\partial z}\right) = \frac{z}{|z|}\frac{k(|z|+a)}{[R^2+(|z|+a)^2]^{3/2}}, \tag{14b}$$

which satisfies the compatibility condition, $\partial_z(\partial_R\Phi) = \partial_R(\partial_z\Phi)$. Since $\Phi = \Phi(R, z)$ which follows $\partial\Phi/\partial\phi = 0$, equations (14a,b) can be directly integrated to yield a unique solution

$$\Phi = -\frac{k}{[R^2+(|z|+a)^2]^{1/2}} + C, \tag{15}$$

which recovers the axisymmetric potential of the Kuzmin disc up to an additive constant C.

3.3 Integrals of motion due to the symmetry of the potential

Let us consider the DF $F = f(\ell_z)$ where $\ell_z = \boldsymbol{\ell}\cdot\hat{\boldsymbol{e}}_z$ is the component of the specific angular momentum in a fixed (say, Cartesian z) direction. Technically any such function cannot be integrable over the whole phase space and so is unphysical. Nevertheless, the CBE merely requires F to be an integral of motion, and so is still applicable. Since $\ell_z = R^2\dot\phi = Rv_\phi$, the CBE in phase-space coordinates $(v_R, v_\phi, v_z; R, \phi, z)$ thus simplifies to

$$\begin{aligned}\dot R\frac{\partial F}{\partial R} + \dot v_\phi\frac{\partial F}{\partial v_\phi} &= \left[v_R\frac{\partial \ell_z}{\partial R} - \frac{1}{R}\left(v_Rv_\phi + \frac{\partial\Phi}{\partial\phi}\right)\frac{\partial\ell_z}{\partial v_\phi}\right]f'(\ell_z)\\ &= \left[\cancel{v_Rv_\phi} - \left(\cancel{v_Rv_\phi} + \frac{\partial\Phi}{\partial\phi}\right)\right]f'(\ell_z) = -\frac{\partial\Phi}{\partial\phi}f'(\ell_z) = 0.\end{aligned} \tag{16a}$$

Or given that $\ell_z = p_\phi$, the CBE in the canonical phase-space coordinate $(p_R, p_\phi, p_z; R, \phi, z)$ simply becomes

$$\dot p_\phi\frac{\partial F}{\partial p_\phi} = -\frac{\partial\Phi}{\partial\phi}f'(\ell_z) = 0. \tag{16b}$$

Provided $f'(\ell_z) \neq 0$, this implies that $\partial\Phi/\partial\phi = 0$, the general solution of which is any axisymmetric potential; that is, an arbitrary function $\Phi = \Phi(R, z)$ of two coordinate functions R and z. Also note $\partial F/\partial p_R = \partial F/\partial p_z = 0$ indicates that the only non-zero component of the Hessian $[\partial_{p_i}\partial_{p_j}F]$ is $\partial^2F/\partial p_\phi^2$ and so it follows that the rank of Hessian is 1 as long as $\partial^2F/\partial p_\phi^2 = f''(\ell_z) \neq 0$.

The result is independent of the choice of the coordinate, although the calculation may be more complicated. For example, in Cartesian coordinates, $\ell_z = xv_y - yv_x$ and so the Hessian becomes

$$\begin{bmatrix}\partial^2_{v_x}F & \partial_{v_x}\partial_{v_y}F & \partial_{v_x}\partial_{v_z}F\\ \partial_{v_y}\partial_{v_x}F & \partial^2_{v_y}F & \partial_{v_y}\partial_{v_z}F\\ \partial_{v_z}\partial_{v_x}F & \partial_{v_z}\partial_{v_y}F & \partial^2_{v_z}F\end{bmatrix} = f''(\ell_z)\begin{bmatrix}y^2 & -xy & 0\\ -xy & x^2 & 0\\ 0 & 0 & 0\end{bmatrix} \tag{17}$$

whose rank is still 1 if $f''(\ell_z) \neq 0$. The CBE on the other hand is

$$\left(\cancel{v_xv_y - v_yv_x} + \frac{\partial\Phi}{\partial x}y - \frac{\partial\Phi}{\partial y}x\right)f'(\ell_z) = 0 \tag{18a}$$

and so, unless $f'(\ell_z) = 0$, we have a homogeneous first-order linear partial differential equation on $\Phi(x, y, z)$,

$$\frac{\partial\Phi}{\partial x}y - \frac{\partial\Phi}{\partial y}x = 0. \tag{18b}$$

Utilizing standard techniques such as the method of characteristics, its general solution is found to be $\Phi = \Phi(x^2 + y^2, z)$, which is again an arbitrary axisymmetric function.

Similarly if a stationary DF (or rather an integral of motion) of the form $F = f(\ell^2)$ is available, the CBE in the canonical phase-space coordinate $(p_r, p_\theta, p_\phi; r, \theta, \phi)$ inherited from the spherical polar coordinate (r, θ, ϕ) is reducible to

$$\begin{aligned}&\left[\frac{p_\theta}{r^2}\frac{\partial\ell^2}{\partial\theta} + \left(\frac{p_\phi^2\cos\theta}{r^2\sin^3\theta} - \frac{\partial\Phi}{\partial\theta}\right)\frac{\partial\ell^2}{\partial p_\theta} - \frac{\partial\Phi}{\partial\phi}\frac{\partial\ell^2}{\partial p_\phi}\right]f'(\ell^2)\\ &\quad= 2\left[-\frac{p_\theta}{r^2}\frac{p_\phi^2\cos\theta}{\sin^3\theta} + \left(\frac{p_\phi^2\cos\theta}{r^2\sin^3\theta} - \frac{\partial\Phi}{\partial\theta}\right)p_\theta - \frac{\partial\Phi}{\partial\phi}\frac{p_\phi}{\sin^2\theta}\right]f'(\ell^2)\\ &\quad= -2\left(p_\theta\frac{\partial\Phi}{\partial\theta} + \frac{p_\phi}{\sin^2\theta}\frac{\partial\Phi}{\partial\phi}\right)f'(\ell^2),\end{aligned} \tag{19a}$$

which follows as $\ell^2 = r^2(v_\theta^2 + v_\phi^2) = p_\theta^2 + p_\phi^2/(\sin^2\theta)$ and $(p_\theta, p_\phi) = (rv_\theta, rv_\phi\sin\theta)$. Assuming $f'(\ell^2) \neq 0$, this is equivalent to

$$p_\theta\frac{\partial\Phi}{\partial\theta} + \frac{p_\phi}{\sin^2\theta}\frac{\partial\Phi}{\partial\phi} = r\left(v_\theta\frac{\partial\Phi}{\partial\theta} + \frac{v_\phi}{\sin\theta}\frac{\partial\Phi}{\partial\phi}\right) = 0. \tag{19b}$$

If $f(\ell^2)$ is a non-constant integral of motion, equation (19b) should hold everywhere in phase space (i.e. for any p_θ and p_ϕ) and so

$$\frac{\partial\Phi}{\partial\theta} = \frac{\partial\Phi}{\partial\phi} = 0 \;\Rightarrow\; \Phi = \Phi(r). \tag{20}$$

Hence the general solution is any spherically symmetric potential. As for the rank of the corresponding Hessian, we observe that the rank of the matrix $[\partial_{p_i}\partial_{p_j}\ell^2]$ is 2 (independent of the coordinate system) with the radial vector being the eigenvector associated with a null eigenvalue (note $\partial\ell^2/\partial p_r = 0$ in the spherical polar coordinate). In addition the radial vector is also in the null space of the matrix $[(\partial_{p_i}\ell^2)(\partial_{p_j}\ell^2)]$, thanks again to $\partial\ell^2/\partial p_r = 0$. Hence, for any $f(\ell^2)$, the radial vector is in the null space of the Hessian matrix;

$$\left[\frac{\partial^2F}{\partial p_i\partial p_j}\right] = f'(\ell^2)\left[\frac{\partial^2\ell^2}{\partial p_i\partial p_j}\right] + f''(\ell^2)\left[\left(\frac{\partial\ell^2}{\partial p_i}\right)\left(\frac{\partial\ell^2}{\partial p_j}\right)\right]. \tag{21}$$

In other words, the Hessian is singular and its rank is at most 2.

Since any axisymmetric or spherical potential admits the integral of motion ℓ_z or ℓ^2, it is not an unexpected result that $F = f(\ell_z)$ or $f(\ell^2)$ only constrains the associated symmetry of the potential and cannot specify the unique potential. The above examples, however, demonstrate that such integrals of motion also fail the necessary condition of having a non-singular Hessian in momentum space. Furthermore, we also observe that $f(\ell_z)$ and $f(\ell^2)$ are not actually integrable in momentum space. That is to say, $f(\ell_z)$ is independent of p_R and p_z, but both components are unbounded, and so the integral of any non-negative $f(\ell_z)$ over the whole momentum space is infinite (unless it is identically zero). A similar argument can also be made for $f(\ell^2)$ and the component p_r. We have argued in Section 2.2 that this is not an accident, but that there is a logical connection between the singular Hessian and the non-integrability.

4 ALGORITHMS FOR EXTRACTING THE GRAVITATIONAL ACCELERATION

Suppose that stationary DF F is known. How then can we extract the gravitational accelerations? First consider the CBE in an arbitrary curvilinear orthogonal coordinate (in which the line element is $\mathrm{d}s^2 = h_1^2\mathrm{d}x_1^2 + h_2^2\mathrm{d}x_2^2 + h_3^2\mathrm{d}x_3^2$) rearranged to be

$$\frac{\partial F}{\partial v_1}\frac{\partial\Phi}{h_1\partial x_1} + \frac{\partial F}{\partial v_2}\frac{\partial\Phi}{h_2\partial x_2} + \frac{\partial F}{\partial v_3}\frac{\partial\Phi}{h_3\partial x_3} = S; \tag{22a}$$

$$S=\sum_{i=1}^{3}v_i\frac{\partial F}{h_i\partial x_i}+\sum_{i,j=1}^{3}v_j\left(v_j\frac{\partial\ln|h_j|}{h_i\partial x_i}-v_i\frac{\partial\ln|h_i|}{h_j\partial x_j}\right)\frac{\partial F}{\partial v_i},\tag{22b}$$

where $v_i=h_i\dot{x}_i$ is the velocity component projected on the orthonormal frame. Next let us observe that $\boldsymbol{\nabla}\Phi$ is constant at all the velocity-space points, given a fixed real-space position. Hence the subset of equations (22a) sampled over the range of velocity space at a fixed position results in an overdetermined (assuming there are more than three sampling points) system of linear equations on $(\partial\Phi/\partial x_1,\partial\Phi/\partial x_2,\partial\Phi/\partial x_3)$. Technically, we only need samples at three different velocity-space points so as to uniquely determine the local gravitational acceleration, provided that the three vectors $\boldsymbol{\nabla}_{\boldsymbol{v}}F$ at the three sampled points – where $\boldsymbol{\nabla}_{\boldsymbol{v}}=(\partial_{v_1},\partial_{v_2},\partial_{v_3})$ is the gradient operator in velocity space – are mutually linearly independent. In fact, the non-singular Hessian of F as discussed in Section 2 guarantees the existence of such three points in velocity space (and so is a sufficient condition for the unique determination of the potential).

On physical grounds, the overdetermined system of equations (22a) resulting from more than three velocity-space points at a single spatial location should be consistent and must possess a unique solution. However, due to the uncertainties in the data, the exact solution may not be necessarily found with the actual set of equations in practice. Instead, the problem should be approached by methods such as least-square, that is, minimizing

$$\sum_{\text{sample}}\frac{1}{\varsigma^2}\left(\frac{\partial F}{\partial v_1}\frac{\partial\Phi}{h_1\partial x_1}+\frac{\partial F}{\partial v_2}\frac{\partial\Phi}{h_2\partial x_2}+\frac{\partial F}{\partial v_3}\frac{\partial\Phi}{h_3\partial x_3}-S\right)^2,\tag{23}$$

where S is as defined in equation (22b), and the summation is over a suitably chosen sample of velocities with the weights ς^{-2}. Finding the extrema with respect to $\boldsymbol{\nabla}\Phi=(\partial\Phi/\partial x_1,\partial\Phi/\partial x_2,\partial\Phi/\partial x_3)$ is then equivalent to solving the set of linear equations

$$\sum_{i=1}^{3}A_{ij}\frac{\partial\Phi}{h_i\partial x_i}=\sum_{\text{sample}}\frac{S}{\varsigma^2}\frac{\partial F}{\partial v_j}\qquad(\text{where } j=1,2,3),\tag{24a}$$

$$\text{where}\quad A_{ij}=\sum_{\text{sample}}\frac{1}{\varsigma^2}\frac{\partial F}{\partial v_i}\frac{\partial F}{\partial v_j},\tag{24b}$$

which is basically the set of standard normal equations. Therefore, provided the matrix $[A_{ij}]$ defined as in equation (24b) is invertible, $\boldsymbol{\nabla}\Phi$ that minimizes equation (23) at the same position can be found through a matrix inversion.

Alternatively one may also attempt to minimize equation (23) summed over data points ranging in a region of space, in order to get the potential as an optimizing functional solution. In principle, this can be done with a suitably chosen parametric function for the potential or non-parametrically (pixelized or otherwise), which is closer to the implementation proposed by Green & Ting (2020) to recover the potential. After reconstructing the DF from the discrete data set via normalizing flows, Green & Ting (2020) characterized the potential as an optimized feed-forward neural network minimizing the cost function, which is defined similarly to equation (23) but with the absolute value instead of the square and also includes the penalty for the negative density. This procedure combines the determination of the local accelerations and their integration into the potential as one single optimization problem. None the less, the actual physical constraints due to the CBE are in the form of an algebraic relation on the local acceleration and so the measurements of the accelerations at different spatial locations should in principle be independent (except for possible systematic correlations relating to the determination of the DF).

5 EFFECTS OF DISEQUILIBRIUM

If the stellar system is not in equilibrium, its DF $F(\boldsymbol{p};\boldsymbol{x};t)$ by definition, is no longer an integral of motion. Provided that the collisional effects are negligible, the evolution of the DF is still governed by the CBE, but the CBE now must include explicit time dependence; $D_tF=\partial_tF+\{F,H\}=0$, where D_tF is the (Lagrangian) phase-space convective derivative and $\partial_tF=\partial F/\partial t$ is the (Eulerian) time rate of change of F at a fixed phase-space coordinate, whilst $\{F,H\}$ is the same as equation (1). We observe that the argument in Section 2.1 still holds for the time-dependent CBE as long as ∂_tF is also a known quantity. In particular, equation (22a) maintains the same form but the right-hand side additionally includes the ∂_tF term ($S\to S+\partial_tF$), and so the determination of the acceleration is still possible if ∂_tF's are known throughout phase space. However ∂_tF is impossible to measure directly within a practical time-scale barring few exceptional situations – by contrast, if $\boldsymbol{\nabla}\Phi$ is known independently, ∂_tF may instead be determined using the CBE. If ∂_tF is considered as unknown, the system of equations in equation (22a) becomes under-constrained and the problem is technically insoluble without some additional restrictive assumptions on the behaviours of $\boldsymbol{\nabla}\Phi$ or ∂_tF.

Nevertheless, we may still infer effects due to the system not being in equilibrium. If the time derivatives are neglected when not warranted, that will introduce a systematic bias. Notably, the linear system of equations in equation (22a) would then not necessarily be consistent even if all the phase-space derivatives of F are known exactly. Whilst equation (24a) still has a unique solution despite the system of equations in equation (22a) being inconsistent, the resulting solution is actually offset by the 'sample average' of ∂_tF.That is to say, if $\partial\Phi^{\mathrm{s}}/\partial x_i$ is the solution of inverting equation (24) with $\partial_tF=0$ (whereas $\partial\Phi/\partial x_i$ is the true gravitational acceleration component), then

$$\frac{\partial\Phi^{\mathrm{s}}}{h_i\partial x_i}=\frac{\partial\Phi}{h_i\partial x_i}-\sum_{j=1}^{3}A_{ij}^{-1}T_j,\ \text{where}\ T_j=\sum_{\text{sample}}\frac{1}{\varsigma^2}\frac{\partial F}{\partial v_j}\frac{\partial F}{\partial t}\tag{25}$$

and A_{ij}^{-1} is the matrix element of the inverse matrix of $[A_{ij}]$ in equation (24b). This follows from the fact that $\partial\Phi/\partial x_i$ is actually the solution of equation (24) with $S\to S+\partial_tF$. If we insert back the solution (equation 25) into equation (22a) and consider the departure from the equality at each sample point, then (with $B_i=\sum_{j=1}^{3}A_{ij}^{-1}T_j$)

$$\sum_{i=1}^{3}\frac{\partial\Phi^{\mathrm{s}}}{h_i\partial x_i}\frac{\partial F}{\partial v_i}-S=\frac{\partial F}{\partial t}-\sum_{i=1}^{3}B_i\frac{\partial F}{\partial v_i}.\tag{26}$$

In other words, the residuals consist of the time derivative ∂_tF and the projection of the bias (i.e. B_i) on to $\boldsymbol{\nabla}_{\boldsymbol{v}}F$. We note that B_i's are unknown but fixed constants and so the last term is also considered as $\boldsymbol{\nabla}_{\boldsymbol{v}}F$ projected on to a fixed (albeit unknown) direction, which behaves in a predictable systematic pattern. Consequently it would be a smoking gun for a system in disequilibria if the observed residual on each sample point exhibits a systematic behaviour over velocity space not consistent with a projection of $\boldsymbol{\nabla}_{\boldsymbol{v}}F$ on to a fixed direction.

6 IMPLEMENTATION

Given a known DF, equation (24) furnishes us with a way to calculate gravitational accelerations, under the assumption of equilibrium. We now wish to test this technique on a mock data set.

Here, we demonstrate a complete pipeline from a 6D stellar kinematics data set to a map of accelerations. This will necessitate an additional step in the procedure, that is, obtaining the underlying DF of the data. Whereas a conventional approach might have involved fitting the data with a well-motivated analytic DF (e.g. Binney & Piffl 2015; Williams & Evans 2015; Posti & Helmi 2019), we instead follow the philosophy of Green & Ting (2020) and construct a non-parametric DF directly from the data.

Our method can thus be summarized as follows:

(i) Employing a normalizing flow technique, we reconstruct a non-parametric DF from the mock data.
(ii) With this reconstructed DF in hand, we exploit equation (24) to calculate accelerations.

This exercise serves mainly as a proof of concept. In a subsequent paper (Naik et al., in preparation), we shall apply the same methodology to local stellar kinematics, with a view towards mapping the acceleration field (and thence the distribution of matter) in the solar neighbourhood.

It is worth noting that an acceleration field calculated with our method is not guaranteed to be physical, in the sense that it might show negative divergences (i.e. negative mass densities) or non-zero curls (i.e. non-conservative force). We view this feature as an advantage: the existence of such non-Newtonian accelerations can serve as a valuable *post hoc* test of our method. If they are found to be robust, they might hint at disequilibrium features or non-gravitational force (even modified gravity). On the other hand, the requirements for non-negative divergences and vanishing curls can be imposed *a priori* if so desired, by adding penalty terms to the loss function used to train the normalizing flow. These non-Newtonian accelerations are then still possible in principle, but heavily suppressed.

6.1 Ergodic models

We consider a simple galaxy halo model in which the DF self-consistently generates both the potential and the density. We generate a mock 6D data set using this DF, and then attempt to derive the underlying acceleration field from the mock data. For this model, we adopt the spherical Hernquist (1990) profile, specified by the potential-density pair

$$\Phi(r) = -\frac{GM}{r+a}; \qquad \rho(r) = \frac{M}{2\pi}\frac{a}{r(r+a)^3}, \tag{27}$$

where M and a are respectively the galaxy mass and scale radius. The isotropic (ergodic) DF for this model is given by[2]

$$F = f(E) = \frac{1}{8\sqrt{2}\pi^3(GMa)^{3/2}} \times \left[\frac{3\sin^{-1}\sqrt{\epsilon}}{(1-\epsilon)^{5/2}} + \frac{\sqrt{\epsilon}\,(1-2\epsilon)(8\epsilon^2-8\epsilon-3)}{(1-\epsilon)^2}\right], \tag{28}$$

where $\epsilon = -Ea/GM \geq 0$ (here, $-E$ is the specific binding energy of a star). In this case, the phase-space gradients of F are determined solely by the gradients of the energy E.

A visualization of the isotropic DF, for $M = 10^{10}$ $\mathrm{M}_\odot$ and $a = 5$ kpc, is given in the left-hand panel of Fig. 1. There is a clear curve above which the DF is everywhere zero: viz. the escape velocity $v_{\rm esc} = \sqrt{2GM/(r+a)}$. With this DF, we employ a Markov Chain Monte Carlo (MCMC) technique to sample a mock 6D data set with 10^6 stars. For this, we use the affine-invariant ensemble sampler implemented in the software package EMCEE (Foreman-Mackey et al. 2013). A density plot of this mock data set is shown in the second panel of Fig. 1.

From this mock data set, we now want to learn the underlying DF by means of a normalizing flow technique (Rezende & Mohamed 2015). Normalizing flows are a relatively new probability density estimation technique, and the basic principle behind them is rather straightforward: a simple base distribution such as a Gaussian is subject to a series (or 'flow') of complex (but bijective and invertible) transformations into a target distribution. The parameters of these transformations are then optimized so as to give a target distribution that closely resembles the data. More detailed descriptions of the technique are given in the article by Rezende & Mohamed (2015) first describing normalizing flows, and the recent review articles by Kobyzev, Prince & Brubaker (2020) or Papamakarios et al. (2021).

Despite taking a single Gaussian as the starting point, a flow with sufficiently flexible transformations (and sufficiently many of them) is able to mimic arbitrarily complex multimodal data distributions. In practice, even rather minimalist flow architectures are capable of achieving great complexity (see e.g. Kingma & Dhariwal 2018 for an impressive application of flows in image generation).

Another class of density estimation technique capable of emulating arbitrarily complex data sets is kernel density estimation. The advantages of flow-based techniques over kernel-based techniques are two-fold. First, flows are less susceptible to over/under-fitting data (Both & Kusters 2019). The second advantage is more context-dependent. Kernel-based techniques typically require no training beyond simply loading the kernels into memory, and perhaps some tuning of the kernel-width parameter. However, given a data set of size N, evaluating the kernel density then essentially requires the computation of N kernel functions, which can be costly as N grows large. Flows do require a training procedure, the cost and duration of which depend on the flow architecture and the size and complexity of the data set in question. However, given a trained flow, evaluating the probability density function is then a mere matter of computing a single Gaussian and a small number of transformations, regardless of N. In summary, kernel densities are cheap to train but expensive to evaluate, while flow densities are expensive to train but cheap to evaluate. In our context, we need to train a density estimator only once to learn the DF, but would then like to evaluate it many times, e.g. for the sums in equation (24). This would therefore suggest flows over kernels.

Another notable aspect of normalizing flows is that the target distribution is guaranteed to be a well-behaved probability distribution, i.e. positive everywhere and normalized to unity. The positivity requirement is met straightforwardly by working in log-space, but the normalization requirement is more exacting: it restricts the space of usable transformations to bijective and invertible functions. This space is then restricted further by the desire for computational efficiency. Different normalizing flow techniques differ primarily in the details of these transformations, as well as the base distributions and flow architectures.

[2] Here the normalization uses $\int \mathrm{d}^3\boldsymbol{x}\,\mathrm{d}^3\boldsymbol{v}\,F = 1$ (cf. Binney & Tremaine 2008, equation 4.1), whereas equation (17) of Hernquist (1990) follows $\int \mathrm{d}^3\boldsymbol{x}\,\mathrm{d}^3\boldsymbol{v}\,F = M$.

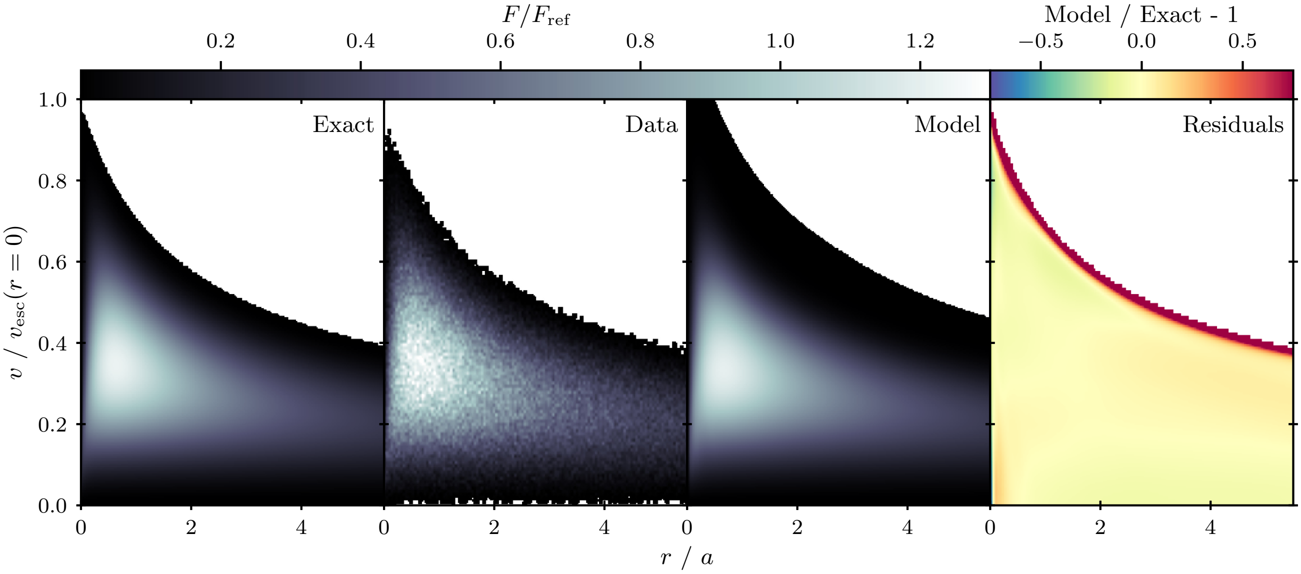

Figure 1. The isotropic Hernquist DF ($M = 10^{10}$ M$_\odot$, $a = 5$ kpc), projected into r-v space. Note that the absolute values of the DF are not of immediate interest, so in each case we have divided by a reference value, given by the *exact* DF evaluated at $r = a$, $v = 0.5v_{\rm esc}(r = a)$. *Left:* The exact DF given by equation (28). *Second panel:* A histogram of our mock data set. *Third panel:* The normalizing flow reconstruction of the DF. *Right:* Fractional residuals comparing the reconstructed and exact DFs. This figure illustrates that the normalizing flow technique successfully learns the isotropic Hernquist DF.

We differ from Green & Ting (2020) in that we employ 'masked autoregressive flows' (MAFs; Papamakarios, Pavlakou & Murray 2017). This choice is motivated by the benchmarking of a number of normalizing flow algorithms. We train an ensemble of 30 MAFs, each with eight transformations along the flow, each transformation being a neural network with one hidden layer of 64 units. We use the implementation of MAFs in the publicly available software package NFLOWS.[3]

The MAFs are trained on the mock data, and thus learn a non-parametric DF that closely resembles the data. This learned DF is shown in the third panel of Fig. 1. It is worth emphasizing that, whilst this plot is in two dimensions, the MAFs are trained using 6D data and learn a 6D DF. The plotted values here are taken from a 2D slice through this 6D DF, with $y = z = v_y = v_z = 0$ (so that $x = r$, $v_x = v$). The rightmost panel of Fig. 1 shows fractional residuals, i.e. $F_{\rm model}/F_{\rm exact} - 1$. Encouragingly, the residuals are less than 5 per cent throughout most of phase space. In other words, our algorithm is successfully able to reproduce the isotropic Hernquist DF.

One apparent qualification to this success is the region near the $v_{\rm esc}$-curve, where the DF is consistently overestimated. The $v_{\rm esc}$-curve represents a hard edge in the Hernquist DF, and even very flexible non-parametric density estimation schemes can struggle to reproduce such a hard edge. However, this need not be a cause for concern, for the following reason: if we progress to step (ii) of our method and attempt to derive acceleration at a given spatial location using this learned DF, the right-hand side of equation (24) requires us to choose a number of points in velocity space. At this stage, we are free to choose whichever velocities we like, and we can thus choose to steer well clear of this region near $v_{\rm esc}$, which we term a 'zone of avoidance'. Of course, in real-world applications, one might not know the exact value of $v_{\rm esc}$, but one can always make an educated guess (e.g. Williams et al. 2017; Deason et al. 2019).

Equation (24) requires the spatial and velocity derivatives of the DF to calculate accelerations. We therefore check if our technique accurately recovers not just the DF, but also its derivatives. Here, a compelling benefit of the normalizing flow technique is that the learned DF is everywhere exactly differentiable, irrespective of the complexity of the flow architecture. Thus, we can efficiently calculate exact derivatives, obviating the need for potentially noisy finite difference schemes.

Fig. 2 compares the first derivatives $\partial F/\partial x$ and $\partial F/\partial v_x$ of the exact and reconstructed DFs, evaluated on a 2D (x, v_x) plane in phase space. Inspecting the residuals in the lower panels of Fig. 2, it is apparent that the MAFs are rather successful at accurately recovering the gradients of the DF; the residuals are less than 10 per cent throughout most of phase space.

As seen in Fig. 1, there is a problematic region of larger residuals near $v_{\rm esc}$. In addition to this, two more such regions are apparent. First, the $\partial F/\partial v_x$ residuals grow rather large in the immediate vicinity of $v_x = 0$. This is the peak of the 1D v_x-distribution, and so the nearby gradients are small and susceptible to mis-estimation. Secondly, the $\partial F/\partial x$ residuals show similar issues around $x = 0$. The same arguments hold here, perhaps exacerbated by the power-law cusp in the Hernquist model. For calculating accelerations, the first problem can be avoided as in the $v_{\rm esc}$ case, i.e. by sampling velocities that avoid the region around $v_x = 0$ (likewise v_y, v_z). However, in the second region around $x = 0$, the residuals appear to be consistently large throughout velocity space, suggesting that our calculated accelerations at these small very radii will be biased.

With these points in mind, we now progress to step (ii) of our method, and derive accelerations from our learned DF using equation (24). Here, we take 50 points along the x-axis, and at each of these points we sample 10^3 velocities for the sums on the right-hand side of equation (24). We perform this sampling by calculating the escape speed $v_{\rm esc}$ at each spatial point, then uniformly sampling 10^4 speeds between 0 and 0.9 $v_{\rm esc}$. Random directions are then chosen from the unit sphere. Finally, we randomly subsample 10^3 velocities from this set, avoiding the region around $v_i = 0$.

[3] NFLOWS: normalizing flows in PYTORCH, doi:10.5281/zenodo.4296287

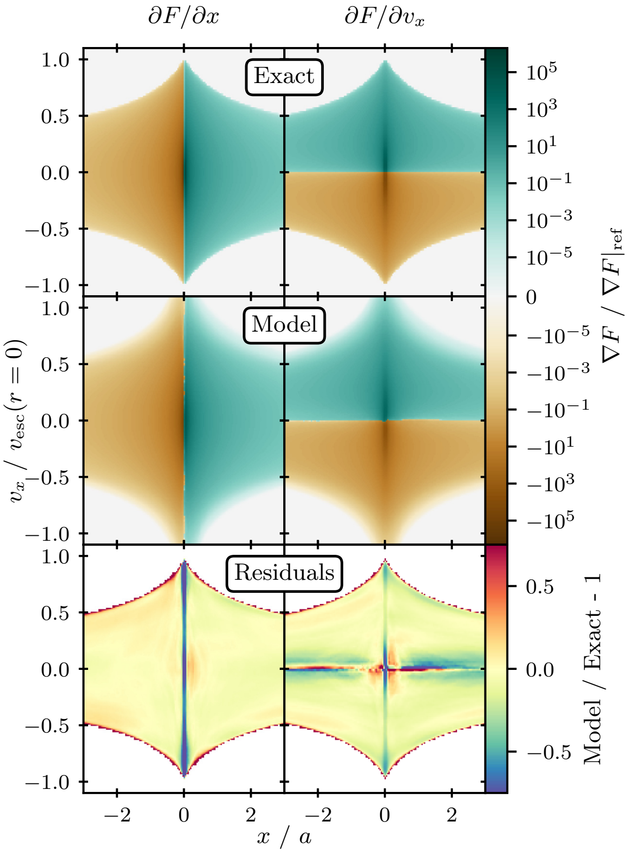

Figure 2. First derivatives of the isotropic Hernquist DF. The left column of three panels shows the spatial derivative $\partial F/\partial x$, whilst the right column gives the velocity derivatives $\partial F/\partial v_x$. The derivatives are everywhere divided by a reference value, evaluated as in Fig. 1. *Top row:* Exact derivatives computed by differentiating equation (28). *Middle row:* Derivatives of the flow-reconstructed DF. *Bottom row:* Fractional residuals. This figure demonstrates that the normalizing flow technique not only recovers the DF but also its gradients, which are required for measuring accelerations, cf. equation (24).

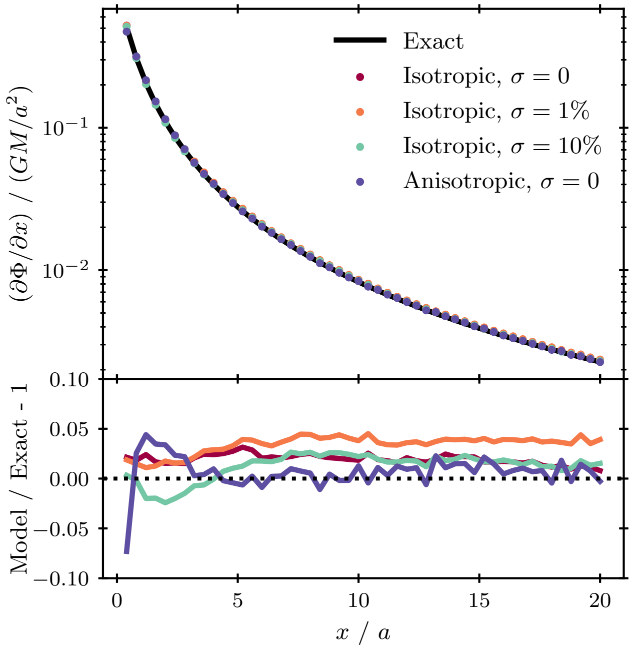

Figure 3. Accelerations in the Hernquist model. The solid black line in the upper panel shows the exact accelerations along the x-axis, whilst the points show the accelerations derived by applying equation (24) to the non-parametric DF learned by the normalizing flows. The different colours show results for flows trained on different data sets, as labelled in the legend. The lower panel shows fractional residuals. This figure illustrates that our method successfully derives accelerations from a 6D snapshot of kinematic data.

After performing this sampling, we have 10^3 points in phase space at which we evaluate equation (24) for each spatial location. The results of this are shown as the 'Isotropic, $\sigma = 0$' curve in Fig. 3. It is clear that the method derives the accelerations in the isotropic Hernquist model very well. The fractional residuals shown in the lower panel indicate an accuracy everywhere at the level of 3 per cent or better.

6.2 Anisotropic models

We repeat this exercise using a simple anisotropic DF for the Hernquist model (Baes & Dejonghe 2002; Evans & An 2005, 2006)

$$F = f(E, \ell) = \frac{3}{4\pi^3 GMa}\frac{\epsilon^2}{\ell}. \tag{29}$$

Now, the DF depends on the magnitude of the angular momentum $\ell = rv_\mathrm{t}$ (here $v_\mathrm{t}^2 = v_\theta^2 + v_\phi^2$) as well as the (dimensionless) binding energy ϵ. As before, we sample one million positions and velocities from this DF, then feed this data to an ensemble of MAFs.

Fig. 4 is the anisotropic analogue of Fig. 1, and shows the exact DF, a density plot of the mock data, the learned DF, and the fractional residuals. As the DF is not isotropic, we do not show the DFs projected into (r, v) space, but rather into (v_r, v_t) or radial versus tangential velocity space at fixed position ($r = a$). Consequently, the 'Data' panel does not show the full data set as in Fig. 1, but only the stars within a small radial slice around $r = a$.

The residuals in the anisotropic DF are generally larger than in the isotropic case, but none the less reasonably small, ~5–10 per cent. Moreover, there seems to a be an additional zone of avoidance here beyond those already discussed in the isotropic case, around $v_\mathrm{t} = 0$. The source of the large errors here can be seen directly from the form of the DF (equation 29): $v_\mathrm{t} = 0$ means $\ell = 0$, so the DF diverges. The probability distribution remains well-behaved, but the MAFs none the less struggle to reproduce the sharp rise in probability density at small v_t.

Despite these foibles, the accelerations are still well recovered in the anisotropic case. These are shown as the points labelled 'Anisotropic, $\sigma = 0$' in Fig. 3. Indeed, the residuals here are comparable to the isotropic case.

One aspect of our procedure worth emphasizing is that successful calculation of accelerations relies on the judicious choice of velocity samples, steering clear of the 'zones of avoidance' in which the DF and its gradients are poorly estimated. We have seen above that the existence and locations of zones can vary from context to context, and so it might be difficult to know *a priori* where they are for any given real stellar population. This is a potential drawback to our method, but it can be readily circumvented by performing tests on mock data sets.

6.3 Effect of errors

As a final test, we assess the potential impact of observational errors by adding Gaussian noise to the isotropic data set, at the 1 per cent

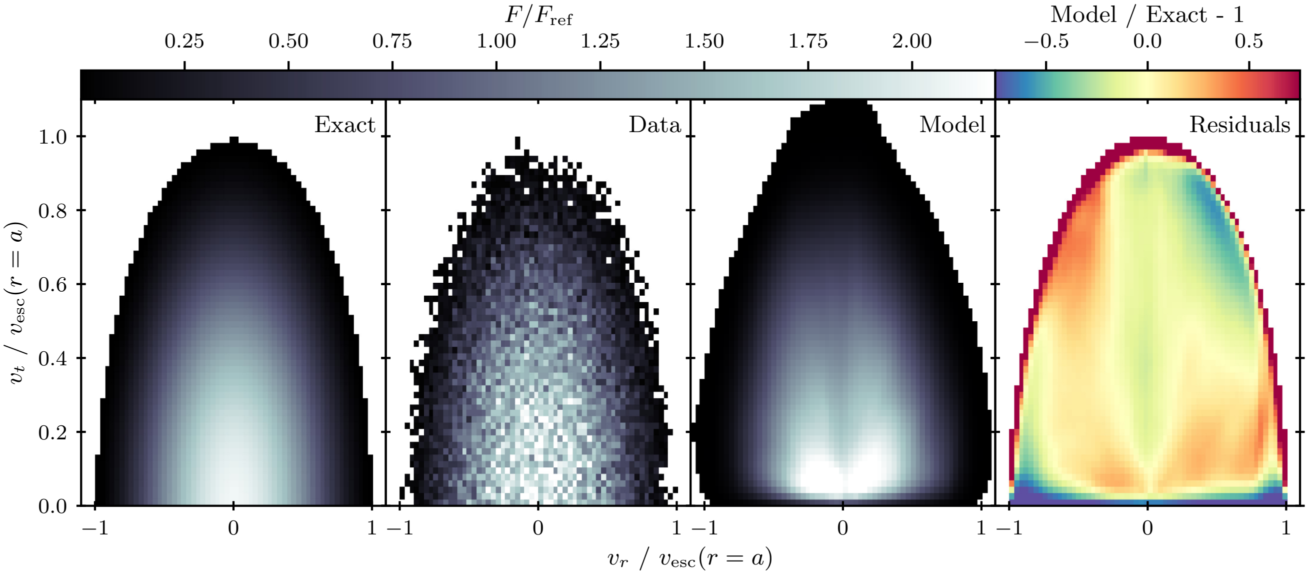

Figure 4. The anisotropic Hernquist DF, projected into v_r-v_t space at fixed position ($r = a$). The four panels carry the same meanings as in their isotropic analogues in Fig. 1, although some differences are discussed in the text. The normalizing flow technique is also successful at recovering the anisotropic Hernquist DF, albeit with larger residuals than in the isotropic case.

and 10 per cent level. The results of this trial are also shown in Fig. 3, alongside the original results for the noiseless data set. Based on this test, it appears that random errors of this magnitude have no appreciable adverse impact on the calculation of accelerations, with residuals still at the per cent level. The application of our method to real data is therefore unlikely to be limited by statistical error.

Going beyond our simple test, there is a natural way to propagate observational errors in our method: when training an ensemble of MAFs on the data, each MAF could be provided with a slightly different data set from which to learn, generated from a different realization of the error distribution. Each member of the ensemble will then have a different learned opinion about the acceleration at a given spatial location, and the spread of these values will incorporate observational errors.

7 CONCLUSIONS

The phase-space distribution (DF) for the stars in the Milky Way is an obvious way to organize the new data sets comprising of nearby stars in the full 6D phase-space coordinates. One question that follows is what information the DF actually contains about the overall properties of the Galaxy. We have proved that, if the stationary DF of a population is known locally in the neighbourhood of a fixed real-space position, then the gravitational acceleration at that location can be uniquely determined from the phase-space gradients of the DF, using the collisionless Boltzmann equation (CBE) under the assumption of dynamical equilibrium. A sufficient condition for this to be true is that the Hessian of the DF with respect to the momenta does not vanish (see equation 5).

In practice, once the CBEs are set up locally at more than three independent phase-space points sharing the real-space coordinates, we have an over-determined system of linear equations on the potential gradients, which can be solved via techniques, such as the least square and normal equations. A practical prescription of how to do this is provided in equation (24).

In light of this finding, we address the question as to how to empirically reconstruct the DF suitable for the local measurements of the gravitational acceleration. Recent developments in machine learning techniques offer great promise in this regard. In particular, Green & Ting (2020) proposed that the DF of stars can be reconstructed from samples of discrete positions and velocities via the method of normalizing flows and the underlying potential can be recovered from this empirical DF. We examine this suggestion by devising tests derived from isotropic and anisotropic Hernquist models using masked autoregressive flows to build the DF. Once built, direct solution of the overconstrained linear equations for the accelerations (equation 24) is highly efficient, and preferable to use of a neural network (cf. Green & Ting 2020). The accelerations are everywhere well reproduced with samplings of ∼1000 velocities at any given position. One caveat here is the existence of regions of velocity space in which the DF is poorly estimated, which need to be avoided in the sampling. Tests with the addition of Gaussian noise at the 1 per cent or 10 per cent level suggest that the method is stable against errors of this magnitude.

There are a number of evident applications of this method, some of which we are actively pursuing. For example, if we reconstruct the velocity distributions of a homogeneous (in equilibrium) stellar population in the solar neighbourhood from the sample of the nearby stars (e.g. Gaia Collaboration 2021), it is possible to measure the local gravity at the sun's position due to the Galactic potential (Naik et al., in preparation). This has implications both for the measurement of the local dark matter density and for tests of alternative theories of gravity. Equally, the method is potentially applicable to the data sets of Milky Way halo stars to measure the mass of the Milky Way and its escape speed.

One assumption underlying the implementation of our method is that of dynamical equilibrium. Incorrectly assuming $\partial F/\partial t = 0$ leads to an additive bias in the derived accelerations that is linear in $\partial F/\partial t$. In addition, disequilibrium can manifest itself through the system of equations (22a) sampled at many different velocity space positions being inconsistent with a single value of $\partial\Phi/\partial x_i$ (after accounting for observational uncertainties), or equation (24) resulting in different values of the acceleration for distinct choices of samples.

There is now a significant body of evidence suggesting the existence of disequilibria in the Milky Way disc (e.g. Antoja et al. 2018; Schönrich & Dehnen 2018; Salomon et al. 2020), which will need to be carefully considered in future applications of our technique to local stellar kinematics. Banik, Widrow & Dodelson (2017) find the bias in inferred accelerations to be at the 10 per cent level if such systematic perturbations are ignored. So, it is interesting to explore whether the pattern of residuals at a sampling point has a systematic behaviour over velocity space that may be a tell-tale signature of departures from equilibrium (cf. Li & Widrow 2021 for a somewhat similar idea).

It is also worth remarking that the first step of our outlined procedure, i.e. learning the DF with normalizing flows, is entirely assumption-free. Given this learned DF, one could then study the non-equilibrium structures themselves. These non-equilbrium structures imprinted in the stellar kinematics are much more than merely sources of systematic error: perturbations to a system can reveal insights about the system itself. For example, Widmark et al. (2021) has shown that the shape of the Gaia phase spiral can be used to constrain the local gravitational potential.

To summarize, our method bypasses many of the assumptions that have been traditionally adopted in studies of galactic dynamics, and represents an efficient, flexible, and data-driven means of extracting underlying gravitational accelerations from snapshots of stellar kinematics.

ACKNOWLEDGEMENTS

We thank Gregory Green and Yuan-Sen Ting for useful discussions, as well as the anonymous referee for a very useful report. Aneesh Naik and Clare Burrage are supported by a Research Leadership Award from the Leverhulme Trust. Clare Burrage is also supported by a Royal Society University Research Fellowship.

DATA AVAILABILITY

The code, mock data sets, and models used in this paper have all been made publicly available at https://github.com/aneeshnaik/HernquistFlows.

REFERENCES

An J., 2013, MNRAS, 435, 3045

An J., Evans N. W., 2016, ApJ, 816, 35

An J., Evans N. W., Sanders J. L., 2017, MNRAS, 467, 1281

Antoja T. et al., 2018, Nature, 561, 360

Baes M., Dejonghe H., 2002, A&A, 393, 485

Banik N., Widrow L. M., Dodelson S., 2017, MNRAS, 464, 3775

Binney J., Piffl T., 2015, MNRAS, 454, 3653

Binney J., Tremaine S., 2008, Galactic Dynamics, 2nd edn. Princeton Univ. Press, Princeton

Both G.-J., Kusters R., 2019, preprint (arXiv:1912.09092)

Bowden A., Belokurov V., Evans N. W., 2015, MNRAS, 449, 1391

Bowden A., Evans N. W., Williams A. A., 2016, MNRAS, 460, 329

Camm G. L., 1941, MNRAS, 101, 195

Cappellari M., 2008, MNRAS, 390, 71

Chandrasekhar S., 1939, ApJ, 90, 1

Deason A. J., Fattahi A., Belokurov V., Evans N. W., Grand R. J. J., Marinacci F., Pakmor R., 2019, MNRAS, 485, 3514

Erkal D. et al., 2019, MNRAS, 487, 2685

Evans N. W., An J., 2005, MNRAS, 360, 492

Evans N. W., An J. H., 2006, Phys. Rev. D, 73, 023524

Evans N. W., Lynden-Bell D., 1991, MNRAS, 251, 213

Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, 306

Gaia Collaboration, 2016, A&A, 595, A1

Gaia Collaboration, 2021, A&A, 649, A6

Green G. M., Ting Y.-S., 2020, in Machine Learning & the Physical Sci., Workshop at the 34th Conf. on Neural Inf. Processing Syst. (NeurIPS ML4PS). p.12(arXiv:2011.04673)

Hernquist L., 1990, ApJ, 356, 359

Johnston K. V., Zhao H., Spergel D. N., Hernquist L., 1999, ApJ, 512, L109

Kingma D. P., Dhariwal P., 2018, in Advances in Neural Inf. Processing Syst. Vol. 31 (NeurIPS 2018). preprint (arXiv:1807.03039)

King III C., Brown W. R., Geller M. J., Kenyon S. J., 2015, ApJ, 813, 89

Kobyzev I., Prince S. J. D., Brubaker M. A., 2020, IEEE Trans. Pattern Analysis & Machine Intelligence, early access (arXiv:1908.09257)

Kuijken K., Gilmore G., 1989, MNRAS, 239, 571

Kuzmin G. G., 1956, AZh, 33, 27

Li H., Widrow L. M., 2021, MNRAS, 503, 1586

Lynden-Bell D., 1982, Obser., 102, 202

Malhan K., Ibata R. A., 2019, MNRAS, 486, 2995

Nitschai M. S., Cappellari M., Neumayer N., 2020, MNRAS, 494, 6001

Papamakarios G., Pavlakou T., Murray I., 2017, in Advances in Neural Inf. Processing Syst. Vol. 30 (NIPS 2017). preprint (arXiv:1705.07057)

Papamakarios G., Nalisnick E., Rezende D. J., Mohamed S., Lakshminarayanan B., 2021, J. Machine Learning Res., 22, 57(arXiv:1912.02762)

Posti L., Helmi A., 2019, A&A, 621, A56

Rezende D. J., Mohamed S., 2015, in Proc. Machine Learning Res. Vol. 37, the 32nd Int. Conf. on Machine Learning (ICML 2015). p.1530(arXiv:1505.05770)

Salomon J.-B., Bienaymé O., Reylé C., Robin A. C., Famaey B., 2009, A&A, 643, A75

Schönrich R., Dehnen W., 2018, MNRAS, 478, 3809

Walker M. G., Mateo M., Olszewski E. W., Peñarrubia J., Evans N. W., Gilmore G., 2009, ApJ, 704, 1274

Widmark A., Laporte C., de Salas P. F., Monari G., 2021, A&A, in press (arXiv:2105.14030)

Williams A. A., Evans N. W., 2015, MNRAS, 454, 698

Williams A. A., Belokurov V., Casey A. R., Evans N. W., 2017, MNRAS, 468, 2359

This paper has been typeset from a TEX/LATEX file prepared by the author.