The Astrophysical Journal, 990:109 (21pp), 2025 September 10 https://doi.org/10.3847/1538-4357/adf1e3

Minimum-entropy Constraints on Galactic Potentials

Leandro Beraldo e Silva[1,2,3], Monica Valluri[1], Eugene Vasiliev[4], Kohei Hattori[1,5,6], Walter de Siqueira Pedra[7,8], and Kathryne J. Daniel[2]
[1] Department of Astronomy and Astrophysics, University of Michigan, Ann Arbor, MI, USA; lberaldoesilva@on.br, lberaldoesilva@gmail.com
[2] Steward Observatory and Department of Astronomy, University of Arizona, 933 N. Cherry Avenue, Tucson, AZ 85721, USA
[3] Observatório Nacional, Rio de Janeiro—RJ, 20921-400, Brasil
[4] University of Surrey, Guildford, Surrey GU2 7XH, UK
[5] National Astronomical Observatory of Japan, 2-21-1 Osawa, Mitaka, Tokyo 181-8588, Japan
[6] The Institute of Statistical Mathematics, 10-3 Midoricho, Tachikawa, Tokyo 190-8562, Japan
[7] University of São Paulo, Institute of Mathematics and Computer Sciences, Av. Trab. São Carlense 400, 13566-590, São Carlos, SP, Brazil
[8] BCAM—Basque Center for Applied Mathematics, Mazarredo, 14. 48009 Bilbao, Spain
Received 2024 October 2; revised 2025 July 1; accepted 2025 July 16; published 2025 September 2

Abstract

A tracer sample in a gravitational potential, starting from a generic initial condition, phase-mixes toward a stationary state. This evolution is accompanied by an entropy increase, and the final state is characterized by a distribution function (DF) that depends only on integrals of motion (Jeans' theorem). We present a method to constrain a gravitational potential assuming a stationary (phase mixed) sample by minimizing the entropy that the sample would have if it were allowed to phase-mix in trial potentials. This method avoids modeling the DF and is applicable to any sets of integrals. We provide expressions for the entropy of DFs depending on energy, $f(E)$, energy and angular momentum, $f(E, L)$, or three actions, $f(\boldsymbol{J})$, and investigate the bias and statistical uncertainties in their estimates. We show that the method correctly recovers the parameters for spherical and axisymmetric potentials. We also present a methodology to characterize the posterior probability distribution of the parameters with an approximate Bayesian computation, indicating a pathway for application to observational data. Using 10^4 tracers with 10%(20%) uncertainties in the 6D coordinates, we recover the flattening parameter q of an axisymmetric potential with $\sigma_q/q \sim 5\%(10\%)$. The python module for the entropy estimators, `tropygal`, is made publicly available.

Unified Astronomy Thesaurus concepts: Dark matter (353); Galaxy dynamics (591); the Milky Way (1054); Milky Way dark matter halo (1049); Milky Way mass (1058); Milky Way dynamics (1051)

1. Introduction

The gravitational potential is a fundamental aspect of any galaxy, determining its stellar orbits and, after all, their observed light distribution. In the Milky Way (MW), we can measure 6D coordinates for millions of stars with Gaia (T. Prusti et al. 2016) and spectroscopic surveys such as APOGEE (S. R. Majewski et al. 2017), LAMOST (X.-Q. Cui et al. 2012), GALAH (G. M. De Silva et al. 2015), and DESI-MWS (A. P. Cooper et al. 2023). With theoretical modeling, these data can be translated into a detailed picture of the Galaxy's mass distribution. Of particular interest is the MW's dark matter (DM) halo shape, which may constrain different scenarios for its composition (e.g., M. Valluri et al. 2022). Since this component is not directly observed, one needs to infer its mass distribution from stars' positions and kinematics.

A non-exhaustive list of methods to recover the underlying potential using a tracer sample includes: the virial theorem and its variants (F. Zwicky 1933; J. N. Bahcall & S. Tremaine 1981; L. L. Watkins et al. 2010), Jeans modeling (e.g., N. Rehemtulla et al. 2022), the "orbital roulette" (A. M. Beloborodov & Y. Levin 2004), the marginalization over an arbitrary number of distribution function (DF) components (J. Magorrian 2014), the generating-function method of S. Tremaine (2018), the minimization of the entropy of tidal streams (J. Peñarrubia et al. 2012; R. E. Sanderson et al. 2015), the "orbital probability density function" (pdf) method of J. Han et al. (2016) and Z. Li et al. (2024), orbital torus imaging (A. M. Price-Whelan et al. 2021), and the maximum-likelihood DF fitting (e.g., P. J. McMillan & J. Binney 2012; P. J. McMillan & J. J. Binney 2013; A. J. Deason et al. 2021).

In all of these methods, further assumptions are required in addition to the information in the observed data set. For instance, for tracers described by a DF, one needs to assume that they constitute a system in dynamical equilibrium. Otherwise, any potential is consistent with a DF describing a nonstationary system (P. J. McMillan & J. Binney 2012; G. M. Green et al. 2023). As another example, when modeling tidal streams, the equilibrium assumption is replaced by an equally strong one, that the debris were initially localized in phase-space.

From Jeans' theorem, the DF of a system in equilibrium can be written as a function of integrals of motion only, reducing the 6D phase-space to 3D or less (J. Binney & S. Tremaine 2008). For instance, isotropic spherical systems can be described by a DF $f = f(E)$, where E is the star's energy, while for anisotropic spherical systems, we can assume $f = f(E, L)$, where L is the magnitude of the angular momentum. In general, samples in realistic galactic potentials normally require three integrals of motion. In practice, this dimension reduction is fundamental for a more efficient use of data.

Assuming a DF that depends on fewer integrals than required (a dimension reduction too severe) delivers incorrect results. In contrast, assuming a DF depending on more

integrals than required is not the most efficient use of data since it does not reduce the dimensions as much as possible. Adopting three integrals is a good compromise between generality and efficiency.

Among all integrals, actions offer several advantages (despite the difficulties in estimating them in practice; see, e.g., J. L. Sanders & J. Binney 2016): the transformation from phase-space coordinates $(\boldsymbol{r}, \boldsymbol{v})$ to angle-action ones $(\boldsymbol{\theta}, \boldsymbol{J})$ is canonical and, thus, $d\boldsymbol{r}d\boldsymbol{v} = d\boldsymbol{\theta}d\boldsymbol{J}$; actions are adiabatic invariants, i.e., they are conserved under slow changes in the potential; angles are restricted to $[0, 2\pi)$; and a system in equilibrium (phase-mixed) is simply described by a pdf[9] in action space $F(\boldsymbol{J}) = (2\pi)^3 f(\boldsymbol{J})$. With angle-action variables, the Hamiltonian depends only on the momenta, $H = H(\boldsymbol{J})$, and the angle-coordinates increase linearly with time $\boldsymbol{\theta} = \boldsymbol{\Omega}t + \text{const}$, where $\boldsymbol{\Omega} = \partial H/\partial \boldsymbol{J}$. The dynamics is thereby reduced to that of "free particles."

In the action-based DF-fitting method developed by P. J. McMillan & J. Binney (2012) and P. J. McMillan & J. J. Binney (2013) and further applied and improved by, e.g., Y.-S. Ting et al. (2013), W. H. Trick et al. (2016), and K. Hattori et al. (2021), the tracer population is assumed to be in equilibrium, and characterized by a DF $f(\boldsymbol{J})$. The MW potential is constrained by fitting functional forms for both the total potential and the tracer DF. If the potential is the only function of interest, one further marginalizes over the DF parameters. For instance, K. Hattori et al. (2021) adopted a model with nine parameters for the potential and seven parameters for the DF, which are later marginalized over, similarly to other works employing this technique. A disadvantage of this method is that it assumes an analytic expression for the DF, which in reality is unknown.

The main goal of the current paper is to improve on this aspect, by not assuming any functional form for the DF (for other methods with this intent, see, e.g., J. Han et al. 2016; Z. Li et al. 2024 for spherically symmetric potentials). This avoids the overhead of fitting the DF parameters and possible biases introduced by the chosen DF. Information on the DF is obtained through nonparametric entropy estimates.

Consider a tracer sample in equilibrium, and described by an unknown DF $f(\boldsymbol{r}, \boldsymbol{v})$. As for any DF, we can define the so-called differential entropy as

$$S[f] \equiv -\int f \ln f \, d^6\boldsymbol{w}, \tag{1}$$

where $\boldsymbol{w} = (\boldsymbol{r}, \boldsymbol{v})$. This entropy is invariant for changes of variables, in particular to angle-action variables evaluated in any potential. In the correct potential where the sample is in equilibrium and in the absence of geometric cuts or other selection effects, the DF $f(\boldsymbol{r}, \boldsymbol{v}) = f(\boldsymbol{\theta}, \boldsymbol{J})$ is uniform in $\boldsymbol{\theta}$, whose phase-space volume is $(2\pi)^3$. The entropy associated with the angles is then maximum, and to keep S invariant, that associated with the actions must be minimum. This can be easily shown if $f(\boldsymbol{\theta}, \boldsymbol{J}) = \mathcal{F}(\boldsymbol{\theta})F(\boldsymbol{J})$, in which case the entropy is just the sum of the entropies in action and angle spaces—in particular, for the fully phase-mixed sample $\mathcal{F}(\boldsymbol{\theta}) = (2\pi)^{-3}$. In Appendix A, we show that a similar idea also applies to nonseparable DFs, which can always be separated in terms of conditional pdfs, $f(\boldsymbol{\theta}, \boldsymbol{J}) = \mathcal{F}(\boldsymbol{\theta}|\boldsymbol{J})F(\boldsymbol{J})$. We then conclude that the correct potential is recovered by minimizing a quantity involving the entropy of the marginal pdf $F(\boldsymbol{J})$ (see J. Magorrian 2014 for a simpler reasoning and an orbit-averaged interpretation).

This quantity is actually the entropy of the future final 6D DF describing the sample if it were allowed to phase-mix in each trial potential. This final DF would be a different (and unknown) function of actions in each trial potential. Since actions are conserved, we estimate this final entropy right away for each potential, with no need to wait for phase-mixing, and the true potential is the one with minimum entropy. We also show that the same method is applicable to any sets of integrals, provided they respect the symmetry requirements of the problem. While one might try to fit potentials by instead maximizing an entropy in angle-space, in Appendix B we discuss why this is not expected to work.

Our approach is related to the minimum-entropy estimates of semiparametric models (E. Wolsztynski et al. 2005), where the potential is the parametric part, and the pdf is the nonparametric one. In Section 2 we describe the general formalism, starting from the action-based DF-fitting and show how it is extended by our method. Section 3 presents the expressions for the entropy estimator in the assumption-free (6D) case and in cases where the DF is an (unknown) function of integrals of motion. Section 4 shows the physical basis of the method, investigates the bias and variance of the entropy estimates for DFs depending only on integrals, and applies a bias correction. In Section 5 we use a fixed sample that is phase-mixed in a given potential to illustrate that the future entropy of the sample (estimated using integrals in different potentials) is at minimum at the true potential. In Section 6 we demonstrate through actual fits that our method recovers the true parameters of a simple spherical potential, and of a flattened axisymmetric potential. We discuss our results in Section 7 and summarize in Section 8. The mathematical basis of the method is presented in Appendix A.

2. General Formalism

Assume a sample of N stars in dynamical equilibrium in a gravitational potential $\phi(\boldsymbol{r})$. Assume further that this is an unbiased sample of an unknown underlying DF f_0, which, as allowed by Jeans' theorem, is a function of integrals of motion in $\phi(\boldsymbol{r})$—we focus on actions $\boldsymbol{J}$, but other integrals can be used too. Our task is to use the 6D coordinates of these stars, assume a functional form for $\phi(\boldsymbol{r})$, and constrain its parameters.

To motivate the minimum-entropy method proposed in this work, we start presenting the maximum-likelihood DF-fitting formalism. In the DF-fitting method, one assumes functional forms for both the potential $\phi(\boldsymbol{r})$ and for the DF $f(\boldsymbol{J}|\boldsymbol{p})$ describing the tracer sample, where $\boldsymbol{p}$ encapsulates parameters of both the potential and the DF. The DF is assumed to describe the stationary state the given sample would achieve after phase-mixing in each trial potential. For simplicity, we assume a full-sky sample in the absence of any selection function or observational errors—the full treatment is presented by, e.g., P. J. McMillan & J. J. Binney (2013) and K. Hattori et al. (2021). In this case, the likelihood for a star to have coordinates $\boldsymbol{w}_i \equiv (\boldsymbol{r}_i, \boldsymbol{v}_i)$ is $f_i(\boldsymbol{J}_i|\boldsymbol{p})$, where $\boldsymbol{J}(\boldsymbol{w}|\phi)$ are actions, which depend on the potential, and $f(\boldsymbol{J}|\boldsymbol{p})$ is properly normalized. The sample joint likelihood is $\hat{\mathcal{L}} = \prod_{i=1}^{N} f_i$, and

[9] We reserve the term DF and the notation $f()$ to the probability density function (pdf) in 6D, and the term pdf and notation $F()$ to pdfs of integrals of motion.

the log-likelihood to be maximized is

$$\ln \hat{\mathcal{L}}(\boldsymbol{w}|\boldsymbol{p}) = \sum_{i=1}^{N} \ln f_i(\boldsymbol{J}_i|\boldsymbol{p}), \tag{2}$$

with trial potentials entering the fit through the actions. Note that Equation (2) can be seen as an estimate[10] of the "true" log-likelihood

$$\ln \mathcal{L} = -NH(f_0, f), \tag{3}$$

where

$$H(f_0, f) = -\int f_0 \ln f \, d\boldsymbol{w} \tag{4}$$

is the cross-entropy and $f_0 = f(\boldsymbol{J}|\boldsymbol{p}_0)$, with $\boldsymbol{p}_0$ being the true parameters. Note that $H(f_0, f)$ is minimum for $f = f_0$, illustrating that the likelihood is maximum at the true parameters (e.g., H. Akaike 1992).

The formalism above concerns the DF-fitting method where an analytic DF is assumed. It can be connected with the minimum-entropy method presented here as follows. As before, we consider the stationary state that the given sample would achieve after phase-mixing in each trial potential. For each of these stationary states, from Equation (1), the differential entropy of its DF can be estimated via Monte Carlo with a sample of f as

$$\hat{S} = -\frac{1}{N}\sum_{i=1}^{N} \ln \hat{f}_i(\boldsymbol{J}_i|\boldsymbol{p}), \tag{5}$$

where $\hat{f}_i$ is an estimate of $f(\boldsymbol{J}_i|\boldsymbol{p})$, as detailed in Section 3. Comparing Equations (2) and (5) might suggest writing

$$\ln \hat{\lambda}(\boldsymbol{p}) \equiv -N\hat{S}(\boldsymbol{p}) \tag{6}$$

for the "log-likelihood." However, despite appearances, $\ln \hat{\lambda}(\boldsymbol{p})$ is not an estimate of the log-likelihood, as can be seen by comparing Equation (1) with Equations (3)–(4). In other words, a log-likelihood would involve assuming a functional form for $f(\boldsymbol{J}|\boldsymbol{p})$ and estimating the cross-entropy between the true DF $f_0 = f(\boldsymbol{J}|\boldsymbol{p}_0)$ and trial DFs $f(\boldsymbol{J}|\boldsymbol{p})$. In contrast, $\ln \lambda(\boldsymbol{p})$ involves estimating the entropy of the (unknown) future DFs in trial potentials. Thus, $\ln \lambda(\boldsymbol{p})$ only corresponds to the log-likelihood at the best-fit model, i.e., $\ln \lambda(\boldsymbol{p}_0) = \ln \mathcal{L}(\boldsymbol{p}_0)$. Additionally, $\ln \hat{\lambda}$ is not a smooth function of the parameters as required for a log-likelihood estimate, but it is noisy, since it is based on estimates of the DF, rather than evaluating an analytical DF. However, as we demonstrate in practice in Section 6, and on mathematical grounds in Appendix A, on average, $\hat{S}$ has its minimum at $\boldsymbol{p}_0$, and it can be minimized to find the best-fit model—see Figure 1 for an illustration.

The maximum-likelihood principle is then replaced by a minimum-entropy one, where we minimize the "future entropy"—the entropy the sample would reach after phase-mixing in each trial potential. The DF describing these final states is always assumed to be a function of integrals, in accordance with Jeans' theorem. However, we do not need to assume any functional form for the DF, and in the remainder of this work, $\boldsymbol{p}$ encapsulates only parameters for the potential.

[10] For any quantity X, we denote its estimate by $\hat{X}$.

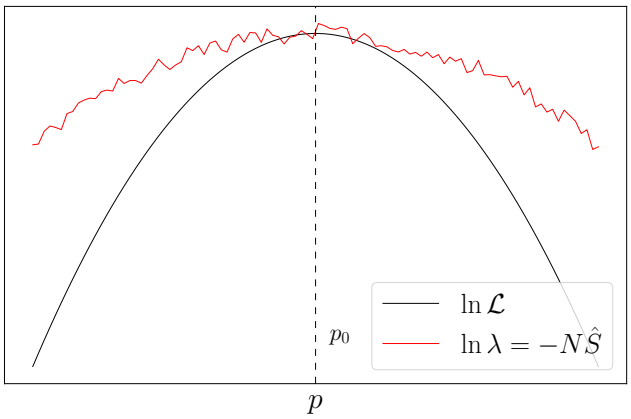

Figure 1. Illustrative comparison of the log-likelihood $\ln \mathcal{L}$ with the quantity used to find its maximum, $\ln \lambda$. Although being different quantities, on average they both peak at the same value p_0 and have the same value at the peak.

As illustrated in Figure 1, fluctuations in $\hat{S}$ can lead to misidentifying the best-fit model, and some smoothing is required to avoid that. In this paper, we estimate the entropy with the kth-Nearest-Neighbor method (k-NN. N. Leonenko et al. 2008a, 2008b), and we smooth out $\hat{S}$ by taking $k > 1$ (see Section 4.4). After identifying the best-fit model by minimizing the entropy of the DF, we perform an approximate Bayesian computation (ABC) to sample the posterior and get credible intervals for the parameters (Section 6). We remark that $\ln \lambda$ was just introduced above to motivate our minimum-entropy method with a conceptual link to the maximum-likelihood principle; but in practice, we simply minimize the future entropy of the sample, with no further mention of $\ln \lambda$.

Although in this paper we do not consider any selection effects or a realistic survey footprint with geometric cuts, these are fundamental aspects for the applicability of the method to real data. With real data, we do not have a sample of the DF $f(\boldsymbol{w})$ assumed in equilibrium. Rather, we have a sample of the DF

$$f_S(\boldsymbol{w}) = \frac{f(\boldsymbol{w})\mathbb{S}(\boldsymbol{r})}{A}, \tag{7}$$

where $\mathbb{S}(\boldsymbol{r})$ is the selection function encapsulating the footprint and spatial dependencies within it, and $A = \int f(\boldsymbol{w})\mathbb{S}(\boldsymbol{r}) d^6\boldsymbol{w}$ is a normalization constant. Substituting in Equation (1), we have

$$S = -\int \left(\frac{A}{\mathbb{S}(\boldsymbol{r})} f_S\right) \ln\left(\frac{A}{\mathbb{S}(\boldsymbol{r})} f_S\right) d^6\boldsymbol{w},$$

which is now a weighted differential entropy. This might make it difficult to estimate the entropy S, since the original estimators we discuss in Section 3 are intended to use samples of f. However, if the selection function $\mathbb{S}(\boldsymbol{r})$ is known, the estimation method can be adapted to provide S given samples of f_S (see J. Ajgl & M. Šimandl 2011).

In this paper, we consider ideal full-sky samples with no selection effects and set $A = \mathbb{S}(\boldsymbol{r}) = 1$. Having presented the general formalism, we now present expressions to estimate the entropy in general and in particular cases of DFs only depending on integrals of motion.

3. Entropy Estimators

We start by defining the entropy of a DF $f(\boldsymbol{w})$. Instead of Equation (1), we modify the entropy definition as

$$S \equiv -\int f \ln\left(\frac{f}{\mu}\right) d^6\boldsymbol{w}, \tag{8}$$

where μ is such that the argument in $\ln(f/\mu)$ is dimensionless; e.g., if $[f] = \text{length}^{-3}\,\text{velocity}^{-3}$, it is convenient to use coordinates normalized by their dispersions $\sigma_{w_1},\ldots,\sigma_{w_6}$, defining $w_1' = w_1/\sigma_{w_1},\ldots,w_6' = w_6/\sigma_{w_6}$ and setting $\mu = |\Sigma|^{-1}$, where $|\Sigma| = \sigma_{w_1}\ldots\sigma_{w_6}$. With $f'(\boldsymbol{w}') = |\Sigma| f(\boldsymbol{w})$, we have:

$$S = -\int f' \ln f'\, d^6\boldsymbol{w}' = -\int f \ln(|\Sigma| f)\, d^6\boldsymbol{w}. \tag{9}$$

For estimators with an isotropic kernel such as the k-NN discussed below, this normalization works as to "isotropize" the coordinates, whereas the entropy is made invariant by an appropriate change of variables. From Equation (9), $-\int f \ln f\ \mathrm{d}^6\boldsymbol{w} = S + \ln|\Sigma|$. Another advantage of the definition (8) is that it allows us to accommodate densities of states when using pdfs of integrals of motion, as shown below.

Equation (9) is the invariant entropy we start from in this section and from which we transform coordinates for the cases where the DF is a function of integrals of motion only. For a sample of N points, it can be estimated as

$$\hat{S} = -\frac{1}{N}\sum_{i=1}^{N} \ln \hat{f}_i', \tag{10}$$

where $\hat{f}_i'$ is an estimate of $f'(\boldsymbol{w}_i')$. In principle, any density estimator could be employed to estimate $f'(\boldsymbol{w}_i')$—see B. W. Silverman (1986) for a review on density estimates. However, for the particular purpose of estimating the entropy with Equation (10), a few estimators have been shown to be optimal (see, e.g., H. Joe 1989; P. Hall & S. C. Morton 1993; J. Beirlant et al. 1997; N. Leonenko et al. 2008a, 2008b)—for a comparison of different methods in N-body simulations, see L. Beraldo e Silva et al. (2017). The latter work demonstrated, in particular, a reasonable agreement of entropy estimates based on k-NN and kernel density estimates, and the high accuracy of the Fokker–Planck modeling of the collisional relaxation, later confirmed on rigorous theoretical grounds by J.-B. Fouvry et al. (2021). More recently, S. Modak & C. Hamilton (2023) used this estimator to study the eccentricity distribution of wide binaries.

Among the optimal methods, we use the k-NN estimator, which is fully nonparametric and fast, since the neighbors' identification can be optimized with kd-trees. This entropy estimator was introduced by L. F. Kozachenko & N. N. Leonenko (1987) for $k = 1$ and later generalized for any k. In this method, the plug-in density estimate is given by (see, e.g., N. Leonenko et al. 2008a, 2008b; G. Biau & L. Devroye 2015; T. B. Berrett et al. 2019, and references therein):

$$\hat{f}_i' = \frac{e^{\psi(k)}}{(N-1)\,V_d\,D_{ik}{}^d}, \tag{11}$$

where

$$V_d = \pi^{d/2}/\Gamma(d/2+1) \tag{12}$$

is the volume of the d-dimensional unit-radius hypersphere, $D_{ik} = \sqrt{(\boldsymbol{r}_i' - \boldsymbol{r}_k')^2 + (\boldsymbol{v}_i' - \boldsymbol{v}_k')^2}$ is the Euclidean phase-space distance of particle i to its kth nearest neighbor, and $\psi(x)$ is the digamma function.[11] For a sketch of a proof of convergence of this method for $k = 1$, see Appendix B of A. Charzyńska & A. Gambin (2015).

Equation (10) with Equation (11) plugged in is a proper entropy estimator in the sense that its bias and variance tend to zero for $N \to \infty$. For the bias, the convergence speed strongly depends on the dimension d and regularity of f (see G. Biau & L. Devroye 2015). Although the actual bias can depend on particular features of the pdf, it is typically smaller in lower dimensions, as we verify in Section 4.1. The expected variance scales as $\propto N^{-1}$, irrespective of the dimension (G. Biau & L. Devroye 2015), as we verify in Section 4.4.

Equation (11) contrasts with naively estimating the density as the number k of points, other than point i, in the hypersphere around point i, divided by its volume, which would introduce a nonvanishing bias for $N \to \infty$. A slightly better reasoning would provide better estimates, although not yet fully bias-corrected: since the kth neighbor is at the edge of the hypersphere, a small volume around it is approximately half inside and half outside the hypersphere. It should count as "half a neighbor" of i, estimating the pdf as

$$\hat{\mathrm{f}}_i = \frac{1}{N-1}\frac{k-1/2}{V_d\,D_{ik}{}^d}. \tag{13}$$

The entropy estimate based on Equation (13) differs from that based on Equation (11) by

$$\delta S = \ln(k-1/2) - \psi(k) = \ln(k-1/2) - \ln(k_{\rm eff}-1/2),$$

where $k_{\rm eff} = e^{\psi(k)} + 1/2$ is an "effective number of nearest-neighbors." For $k = 1, 2, 3, 4$, it is, respectively, $k_{\rm eff} \approx 1.06, 2.03, 3.02, 4.01$. For large k, $e^{\psi(k)} \approx k - 1/2 + \mathcal{O}(1/k)$, and $k_{\rm eff} \approx k$.

For two general distributions f_0 and f, we also re-define their cross-entropy as

$$H(f_0, f) \equiv -\int f_0 \ln\left(\frac{f}{\mu}\right) d^6\boldsymbol{w}. \tag{14}$$

Note that, in general, it is possible to estimate $H(f_0, f)$ even if the samples of f_0 and f have different sizes N and M, respectively. Equation (14) is estimated as

$$\hat{H} = -\frac{1}{N}\sum_{i=1}^{N} \ln \hat{\xi}_i', \tag{15}$$

where

$$\hat{\xi}_i' = \frac{e^{\psi(k)}}{M V_d\, D_{ik}{}^d}, \tag{16}$$

and D_{ik} is the distance between point i of the f_0-sample to its k-nearest neighbor *in the f-sample* (N. Leonenko et al. 2008b). We can interpret $\hat{\xi}_i'$ as an estimate of f at the point i of the f_0-sample. In this paper, we restrict to samples of equal sizes, so $M = N$, and normalize coordinates by typical dispersions of the f_0-sample. To explore the parameters' posterior distribution in Section 6.1, $f_0(\boldsymbol{J})$ will represent the (unknown) underlying

[11] In particular, $\psi(1) = -\gamma \approx -0.577$ (Euler-Mascheroni constant).

DF describing the sample in the best-fit potential and $f(\boldsymbol{J})$, the final (equilibrium) DF of the sample in each trial potential.

Equations (10) and (15), with Equations (11) and (16), respectively, plugged in, converge in probability to the true entropies under weak conditions on the underlying DFs (e.g., N. Leonenko et al. 2008a; G. Biau & L. Devroye 2015; D. Lombardi & S. Pant 2016). The python module `tropygal`[12] implements these entropy estimators, as well as a few galactic dynamics models with analytic DFs.

As explained in Section 1, the method developed here assumes the sample is phase-mixed in the true potential, and also considers the entropy the sample would have if evolved until phase-mixed in a trial potential. In the subsequent subsections, we present expressions for cases where the DF only depends on integrals of motion, as required by Jeans' theorem for phase-mixed samples. In the following, we denote $S_I = S[f(\boldsymbol{I})]$, i.e., the entropy of the DF when f is a function of integrals $\boldsymbol{I}$. Note that this differs from the entropy of the integrals' pdf $S[F(\boldsymbol{I})] = -\int F \ln F d\boldsymbol{I}$, as we show here and, in more detail, in Appendix A.

3.1. Isotropic Spherical System, f = f(E)

For isotropic spherical systems in equilibrium, we can write $f(\boldsymbol{w}) = f(E)$, where $E = v^2/2 + \phi(r)$ and $\phi(r)$ is the potential. In this case, Equation (9) reduces to

$$S_{\rm E} = -\int F(E) \ln\left[\frac{|\Sigma| F(E)}{g(E)}\right] dE, \tag{17}$$

where

$$F(E) = f(E) g(E) \tag{18}$$

is the pdf in energy space and

$$g[E|\phi(r)] = (4\pi)^2 \int_0^{r_m(E)} r^2 \sqrt{2[E - \phi(r)]}\, dr \tag{19}$$

is the density of states, with $r_m(E)$ being the radius where $\phi = E$. If σ_E is a typical energy dispersion, we define $E' = E/\sigma_E$, and estimate $S_{\rm E}$, Equation (17), as

$$\hat{S}_{\rm E} = -\frac{1}{N}\sum_{i=1}^{N} \ln\left[\frac{\hat{F}_i'(E_i')}{\mu(E_i)}\right], \tag{20}$$

where $\mu(E) = \sigma_E |\Sigma|^{-1} g[E|\phi(r)]$. We estimate $\hat{F}_i'(E_i')$, the energy pdf, with $d = 1$ and $D_{ik} = |E_i' - E_k'|$ in Equation (11). If it is convenient to write the density of states in terms of the normalized energy and angular momentum, we can replace $g[E|\phi(r)] = \sqrt{\sigma_E}\, g[E'|\phi(r)/\sigma_E]$.

3.2. Anisotropic Spherical System, f = f(E,L)

For anisotropic spherical systems with a DF $f(\boldsymbol{w}) = f(E, L)$, where $L = v_t r$ and $v_t^2 = v_\theta^2 + v_\varphi^2$ in spherical coordinates (r, θ, φ), Equation (9) reduces to

$$S_{\rm EL} = -\int F(E, L) \ln\left[\frac{|\Sigma| F(E, L)}{g(E, L)}\right] dE dL, \tag{21}$$

where the pdf for energy and angular momentum is

$$F(E, L) = f(E, L) g(E, L), \tag{22}$$

and the density of states is

$$g[E, L|\phi(r)] = 8\pi^2 L T_r[E, L|\phi(r)]. \tag{23}$$

The period of radial motion $T_r[E, L|\phi(r)]$ is given by

$$T_r[E, L|\phi(r)] = 2\int_{r_{\rm per}}^{r_{\rm apo}} \frac{dr}{\sqrt{2[E - \phi(r)] - L^2/r^2}}, \tag{24}$$

with $r_{\rm per}$ and $r_{\rm apo}$ being the peri- and apo-center distances. Defining $(E', L') = (E/\sigma_E, L/\sigma_L)$, we estimate

$$\hat{S}_{\rm EL} = -\frac{1}{N}\sum_{i=1}^{N} \ln\left[\frac{\hat{F}_i'(E_i', L_i')}{\mu(E_i, L_i)}\right], \tag{25}$$

where $\mu(E, L) = \sigma_E \sigma_L |\Sigma|^{-1} g[E, L|\phi(r)]$, and for the pdf, we plug in Equation (11) with $d = 2$ and $D_{ik} = \sqrt{(E_i' - E_k')^2 + (L_i' - L_k')^2}$. If desired, we replace $g[E, L|\phi(r)] = (\sigma_L^2/\sigma_E) g[E', L'|\phi(r')/(\sigma_L\sqrt{\sigma_E})]$, where $r' = (\sqrt{\sigma_E}/\sigma_L) r$.

3.3. Generic Integrable Potential, f = f(J)

For realistic galactic potentials, assuming that most orbits are regular or weakly chaotic, we may compute approximate actions with, e.g., the Stäckel approximation (J. Binney 2012). In this context, a system in dynamical equilibrium is described by a pdf in action space

$$F(\boldsymbol{J}) = (2\pi)^3 f(\boldsymbol{J}), \tag{26}$$

where $\boldsymbol{J}$ are three actions. Thus, Equation (9) reduces to

$$S_{\boldsymbol{J}} = -\int F(\boldsymbol{J}) \ln\left[\frac{|\Sigma| F(\boldsymbol{J})}{(2\pi)^3}\right] d\boldsymbol{J}. \tag{27}$$

The simplicity of Equation (27), in comparison to Equations (17)–(19) or Equations (21)–(24), illustrates the advantages of using action-based DFs instead of other integrals of motion. Defining new actions $\boldsymbol{J}'$ normalized by their dispersions $(\sigma_{J_1}, \sigma_{J_2}, \sigma_{J_3})$, we have

$$\hat{S}_{\boldsymbol{J}} = -\frac{1}{N}\sum_{i=1}^{N} \ln\left[\frac{\hat{F}_i'(\boldsymbol{J}_i')}{\mu}\right], \tag{28}$$

where $\mu = (2\pi)^3 \sigma_{J_1}\sigma_{J_2}\sigma_{J_3}|\Sigma|^{-1}$, and for the pdf, we plug in Equation (11) with $d = 3$ and $D_{ik} = \sqrt{|\boldsymbol{J}_i' - \boldsymbol{J}_k'|^2}$.

The same expressions apply to the cross-entropy estimates, Equations (15)–(16), mutatis mutandis.

Having presented the expressions in general and for DFs depending only on integrals of motion, in the next section we illustrate the physical basis of the method, as well as investigate the bias and fluctuation in these estimates. For that, we use a model with explicit expressions for $f(E)$, $g(E)$ and for the actions.

4. The Isochrone Model

To illustrate the accuracy of these entropy estimators and the physical basis of our method, we consider the isochrone

[12] The documentation and installation instructions can be accessed at https://tropygal.readthedocs.io/en/latest/.

model (M. Henon 1959), whose potential is

$$\phi(r) = -\frac{GM}{b}\frac{1}{1+\sqrt{1+(r/b)^2}}, \tag{29}$$

where M is the total mass, and b is the scale length. The DF of a self-consistent sample is (see J. Binney & S. Tremaine 2008; J. Binney & M. Petrou 1985)

$$f(E) = \frac{1}{\sqrt{2}(2\pi)^3(GMb)^{3/2}}\frac{\sqrt{\varepsilon}}{[2(1-\varepsilon)]^4} \times \left[27 - 66\varepsilon + 320\varepsilon^2 - 240\varepsilon^3 + 64\varepsilon^4 + 3(16\varepsilon^2 + 28\varepsilon - 9) \times \frac{\sin^{-1}\sqrt{\varepsilon}}{\sqrt{\varepsilon(1-\varepsilon)}}\right], \tag{30}$$

and the density of states, Equation (19), is

$$g(E) = (2\pi)^3\sqrt{GM}\,b^{5/2}\frac{(1-2\varepsilon)^2}{(2\varepsilon)^{5/2}}, \tag{31}$$

where $\varepsilon = -bE/(GM)$. The radial period is

$$T_r(E, L) = \frac{2\pi GM}{(-2E)^{3/2}}. \tag{32}$$

As for any spherical system, the azimuthal and latitudinal actions are $J_\varphi = L_z$ and $J_\theta = L - |L_z|$, respectively, and the radial action is

$$J_r = \frac{1}{\pi}\int_{r_{\rm per}}^{r_{\rm apo}} dr\sqrt{2E - 2\phi(r) - L^2/r^2}. \tag{33}$$

For the isochrone potential,

$$J_r = \frac{GM}{\sqrt{-2E}} - \frac{1}{2}(L + \sqrt{L^2 + 4GMb}). \tag{34}$$

4.1. Entropy Bias

We start evaluating the integral in Equation (17) numerically with Equations (30)–(31), from $E_{\rm min} = -0.5$ to $E_{\rm max} = -10^{-8}$, with $G = M = b = 1$. We take this as the true entropy value, $S_{\rm E,true}$ (thick solid gray line in the top panel of Figure 2). To compare with the entropy estimates, we generate self-consistent samples with different sizes N of this model with AGAMA (E. Vasiliev 2019), and integrate orbits for these samples for $50 \times \langle T_{\rm circ}\rangle$, where $T_{\rm circ}$ is the period of circular motion. Figure 2 (top panel) shows the entropy estimates $\hat{S}_{\rm 6D}$ (thin solid lines) at different times and for different N (colors), taking the nearest neighbor ($k = 1$). We recalculate $|\Sigma| = \sigma_{w_1}\ldots\sigma_{w_6}$, renormalizing the coordinates at each time with the appropriate change of variables in Equation (9). This provides better estimates than a fixed initial normalization, but the difference is small.

Since the initial sample is self-consistent with the potential, it is stationary, and $\hat{S}_{\rm 6D}$ should be conserved. We see that this is the case for all sample sizes, with larger fluctuations for smaller N. Furthermore, $\hat{S}_{\rm 6D}$ is significantly biased with respect to the true value, and this bias is time-independent, except for minor fluctuations. In the bottom panel, the hexagons show the relative bias $\delta S_{\rm 6D} = (\langle\hat{S}_{\rm 6D}\rangle_t - S_{\rm E,true})/S_{\rm E,true}$ as a function of N, where $\langle\hat{S}_{\rm 6D}\rangle_t$ is a time-average. Even for $N = 10^8$, $\hat{S}_{\rm 6D}$ has a relative bias of $\approx$1%.

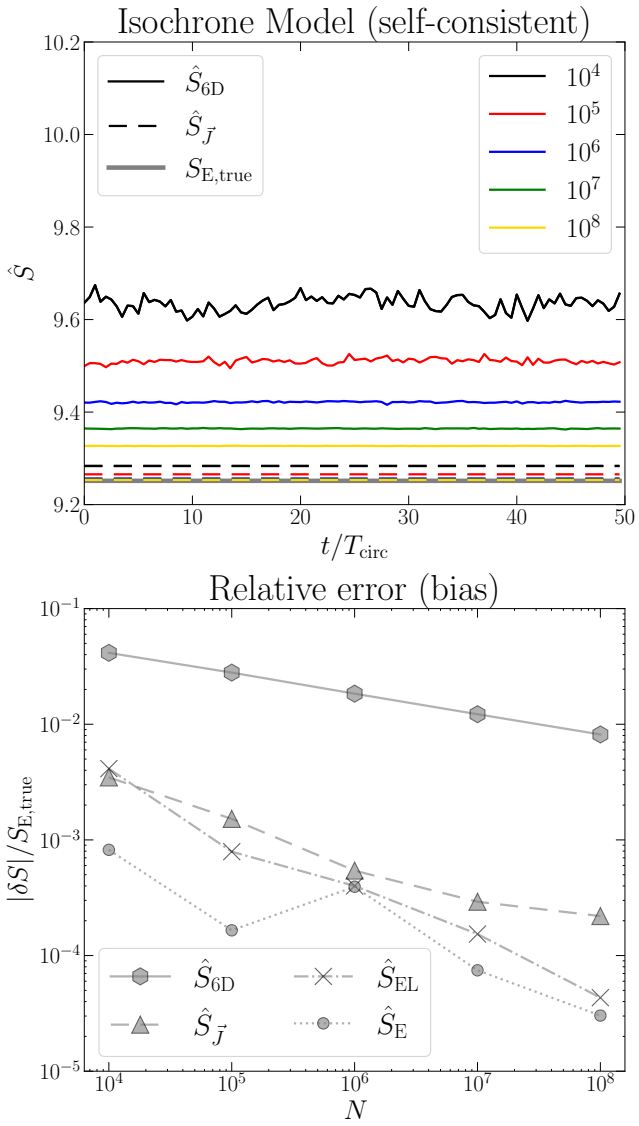

Figure 2. Top panel: entropy estimates in 6D (solid) and assuming the DF is an unknown function $f(\boldsymbol{J})$ (dashed) for self-consistent samples of the isochrone model, with different sample sizes (colors). The thick solid gray line shows the true value—numerical integral in Equation (17). Bottom panel: relative error (bias) of $\hat{S}_{\rm 6D}$, $\hat{S}_{\boldsymbol{J}}$, $\hat{S}_{\rm EL}$ and $\hat{S}_{\rm E}$. For a fixed sample size, estimates in lower dimensions are more accurate.

Figure 2 (top panel) shows the entropy estimates $\hat{S}_{\boldsymbol{J}}$, Equation (28), i.e., assuming the DF is an unknown function of the actions (dashed lines). Since these are conserved, we only estimate $S_{\boldsymbol{J}}$ at $t = 0$. We see that $\hat{S}_{\boldsymbol{J}}$ produces a much smaller bias, due to the dimension reduction from 6D to 3D in the practical estimates—but $S_{\boldsymbol{J}}$ is still the entropy of the 6D DF. The triangles in the bottom panel show that the bias stays below $\approx$1% even for $N = 10^4$. Crosses and dots show the relative bias for $\hat{S}_{\rm EL}$ and $\hat{S}_E$, respectively. These are estimated with Equation (20) for $S_{\rm E}$, i.e., assuming the DF is an unknown function $f = f(E)$, and Equation (25) for S_{EL}. We see that the bias is also significantly smaller than that of $\hat{S}_{\rm 6D}$, and it is generally smaller for lower dimensions, as expected.

Thus, we have shown that: $\hat{S}_{\rm 6D}$ is appropriately conserved in the self-consistent model, but it is biased with respect to the true value by $\delta S/S_{\rm E,true} \approx 5\%$ for $N = 10^4$, whereas in the space of integrals $\delta S/S_{\rm E,true} < 1\%$ for $N = 10^4$, and $\delta S/S_{\rm E,true} \lesssim 0.01\%$ for $N = 10^8$.

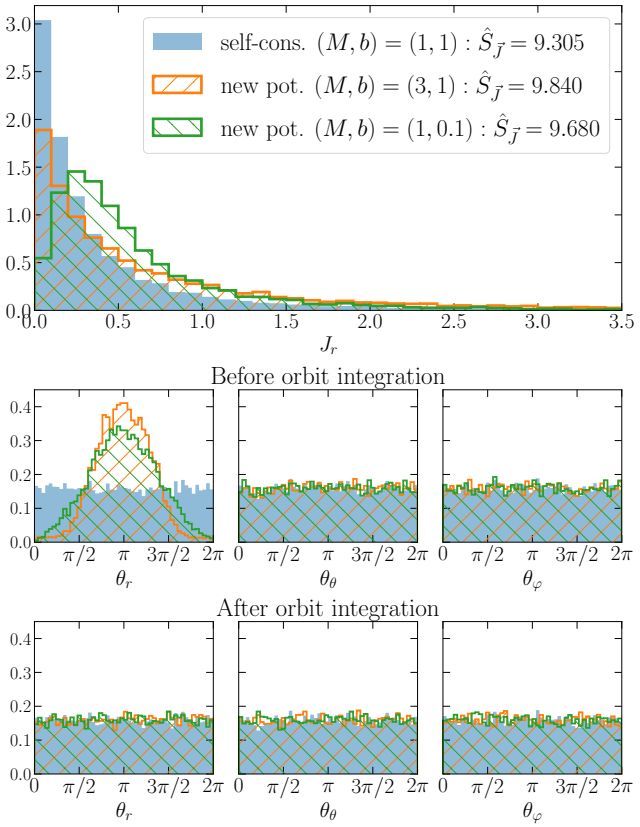

Figure 3. Top panel: histograms of the radial action J_r for a self-consistent sample of an isochrone model with $(M, b) = (1, 1)$, with J_r evaluated in this model ("self-cons.") and for $(M, b) = (3, 1)$ and $(M, b) = (1, 0.1)$. The distribution is broader in the new potentials. Middle panels: histograms of angle variables for the original sample in the self-consistent potential and in the other ones (where the original sample is not stationary). Bottom panels: the same as the middle panels, but after orbit integration in each of the new potentials. The final angle distributions are all uniform, as expected for phase-mixed samples. The legend shows final entropy values in each potential. All panels have 50 equally spaced bins.

4.2. Phase-mixing and Entropy Increase

Here we use the same initial sample and consider two new isochrone potentials with $(M, b) = (3, 1)$ and $(M, b) = (1, 0.1)$, in addition to the self-consistent one. Figure 3 (top panel) shows histograms of the radial actions J_r evaluated in these three potentials. We do not show histograms of J_φ or J_θ, since they do not depend on the potential and are identical in the three cases. The histogram is narrow in the original (self-consistent) potential and broader in the new ones. The middle panels comprise histograms of the angle variables evaluated in the three potentials—all panels in this figure have 50 equally spaced bins. As the sample is not phase-mixed in the new potentials, θ_r is not uniformly distributed in these cases.

We then integrate orbits for this initial sample for $50 \times \langle T_{\rm circ}\rangle$ in the two new isochrone potentials, which is long enough for the samples to relax. The bottom panels of Figure 3 show histograms of the final angles, as well as the initial ones in the self-consistent potential. As expected for phase-mixed samples, these are all equally uniform. In fact, Kolmogorov–Smirnov tests comparing the θ_r distribution in the self-consistent potential with the final ones in the new potentials result in statistic values $\sim$0.01, with p-values $\gtrsim 0.6$, largely failing to reject the equally uniform hypothesis.

Just as for the self-consistent sample, this uniformity does not require any coarse-graining, but is rather an objective fact.

We now show, before estimating the entropy, that the evolution of the original sample in new potentials, as illustrated in Figure 3, is necessarily accompanied by an entropy increase. We define the sample's initial entropy as $S[f_0] = -\int f_0 \ln f_0 \, d\boldsymbol{w} = -\int f_0 \ln f_0 \, d\boldsymbol{\theta} d\boldsymbol{J}$, where $f_0(\boldsymbol{w})$ is the initial DF and $\boldsymbol{w} = (\boldsymbol{r}, \boldsymbol{v})$. $S[f_0]$ is invariant for angle-actions evaluated in any potential. For the self-consistent potential, $f_0(\boldsymbol{\theta}, \boldsymbol{J}) = (2\pi)^{-3} F_0(\boldsymbol{J})$, and thus,

$$S_{\boldsymbol{J}}[f_0] = \ln(2\pi)^3 + S[F_0], \tag{35}$$

where $S[F_0] = -\int F_0(\boldsymbol{J}) \ln F_0 \, d\boldsymbol{J}$. Similarly, after phase-mixing in the new potential, the final DF is $f_{\rm final}(\boldsymbol{\theta}, \boldsymbol{J}) = (2\pi)^{-3} F_{\rm final}(\boldsymbol{J})$, and its entropy is

$$S_{\boldsymbol{J}}[f_{\rm final}] = \ln(2\pi)^3 + S[F_{\rm final}]. \tag{36}$$

Since the three samples have the same actions' distribution, except for J_r being broader in the new potentials (Figure 3), we see that $S[F_{\rm final}] > S[F_0]$, for broader pdfs have larger entropies. Thus, from Equations (35)–(36), $S_{\boldsymbol{J}}[f_{\rm final}] > S_{\boldsymbol{J}}[f_0]$, i.e., the phase-mixing of a nonrelaxed sample is necessarily accompanied by an entropy increase. This is confirmed by our estimates (legend). We emphasize that the practical entropy calculation only uses actions, while assuming that the final angle distribution will be uniform, as required by Jeans' theorem. For a given sample in any trial potential, we can estimate the final entropy right away, since actions are conserved.

To study the sample evolution in more detail, Figure 4 (top panel) shows entropy estimates using 6D coordinates at several time steps ($\hat{S}_{\rm 6D}$, solid lines) as well as $\hat{S}_{\boldsymbol{J}}$ (dashed lines) for the same initial sample evolved in the potential $(M, b) = (3, 1)$. Since the initial sample is not in dynamical equilibrium in the new potential, it responds to the higher mass developing a radially biased velocity anisotropy. The final DF is unknown, but it should respect Jeans' theorem, being a function $f(E, L)$, or $f(\boldsymbol{J})$. The thick solid gray line shows $\hat{S}_{\rm EL}$ for $N = 10^8$ in the new potential, which is the lower dimension allowed by the phase-mixed sample. Here we proceed exactly as previously to get $\hat{S}_{EL}$ for the self-consistent sample (Figure 2), the only difference being that energies are evaluated in a new potential. Since we have shown that $\hat{S}_{\rm EL}$ has a negligible bias for $N = 10^8$, we take this as the true final entropy, $\hat{S}_{\rm EL,true}$.

In addition to the biases with respect to the initial true entropy $S_{\rm E,true}$ (Figure 2), Figure 4 shows that the asymptotic values of $\hat{S}_{\rm 6D}$ ($t \to \infty$) in the new potential are also biased with respect to $\hat{S}_{\rm EL,true}$. On the other hand, $\hat{S}_{\boldsymbol{J}}$ is again much less biased, since it is estimated in a lower dimension space, while the angles' contribution to $\hat{S}_{\boldsymbol{J}}$ is $\ln(2\pi)^3$ (see Equation (36)). In both cases, the bias decreases for larger N (see the inset plot).

Figure 4 (bottom) shows $\Delta\hat{S}_{\rm 6D} = \hat{S}_{\rm 6D}(t) - \hat{S}_{\rm 6D}(0)$ (colors) and $\Delta S_{\rm true} = \hat{S}_{\rm EL,true} - S_{\rm E,true}$, the true entropy increase (thick gray). The final $\Delta\hat{S}_{\rm 6D}$ is similar for all sample sizes, approximately converging to $\Delta S_{\rm true}$. This confirms that the bias is nearly independent of time and is thus nearly eliminated by calculating entropy variations, as done by L. Beraldo e Silva et al. (2017, 2019a, 2019b).

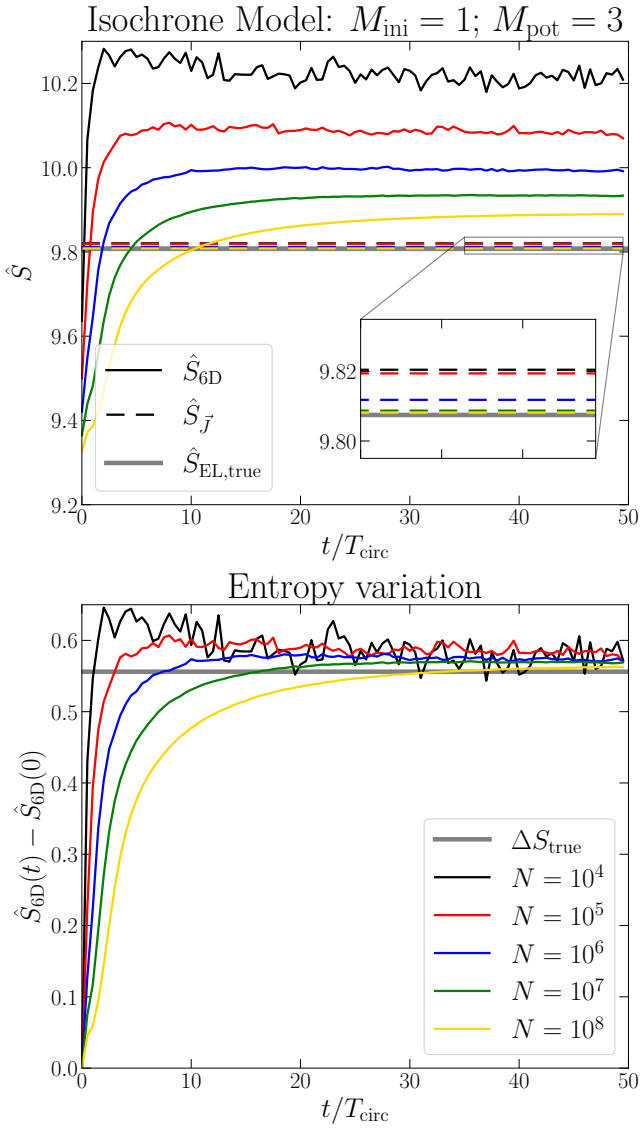

Figure 4. Top panel: entropy estimates in 6D (solid) and assuming $f = f(\boldsymbol{J})$ (dashed) for initial self-consistent samples of the isochrone model with $M = 1$, but integrated in (and $\boldsymbol{J}$ evaluated at) an isochrone potential with $M = 3$. The thick solid gray line shows the entropy for a phase-mixed system with $f = f(E, L)$ and $N = 10^8$, considered as the true final entropy. Bottom panel: entropy variation $\Delta\hat{S} = \hat{S}_{\rm 6D}(t) - \hat{S}_{\rm 6D}(0)$ for different sample sizes, which approximately converges to $\Delta S_{\rm true} = \hat{S}_{\rm EL,true} - S_{\rm E,true}$ for all samples.

4.3. Bias Correction

If the bias of $\hat{S}$ does not depend on the model parameters, it poses no problem for the minimum-entropy fits, since it only introduces an additive constant in $\hat{S}$. For a possibly model-dependent bias, we investigate it in more detail and test a prescription to suppress it.

It is known that taking the kth neighbor for larger k increases the bias in the entropy estimate, but decreases its variance, a manifestation of the bias-variance trade-off (e.g., L. Wasserman 2010). To investigate this, we generate 10^3 realizations of size-N self-consistent isochrone samples with $M = b = 1$ and calculate actions and $\hat{S}_{\boldsymbol{J}}$, Equation (28), for each realization, normalizing the actions in each one. Here we do not compare with $\hat{S}_{\rm 6D}$; thus, we do not normalize by $|\Sigma| = \sigma_{w_1} \ldots \sigma_{w_6}$, which would introduce unnecessary extra noise.

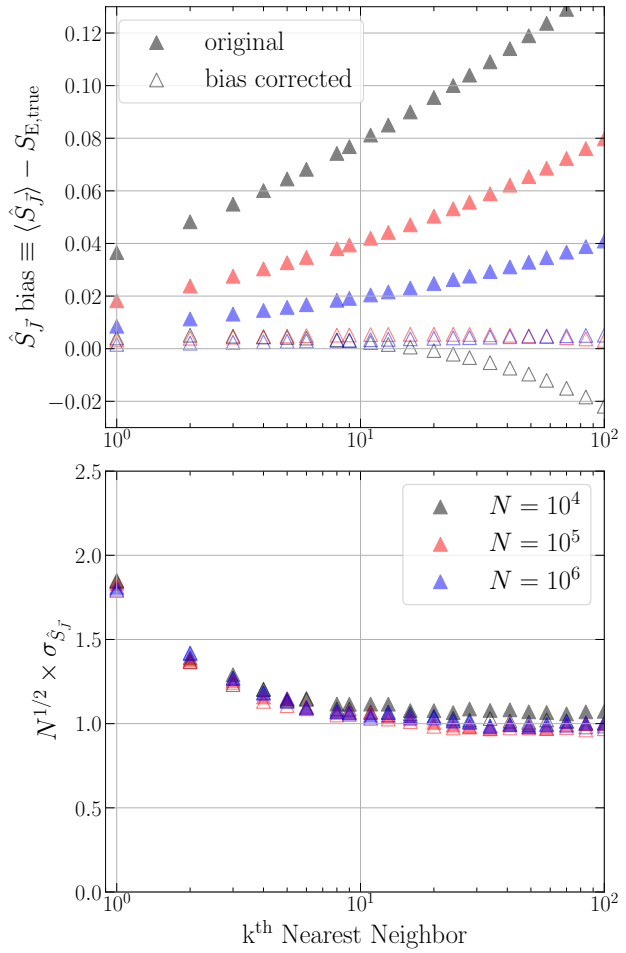

Figure 5. Bias (top panel) and fluctuation (bottom panel) of entropy estimates for self-consistent samples of the isochrone model. We see that the uncorrected bias (full triangles) increases with k, with the correction suppressing the bias. Empty blue and red triangles nearly overlap. The fluctuation $\sigma_{\hat{S}_{\boldsymbol{J}}}$ decreases with k, saturating at $\sigma_{\hat{S}_{\boldsymbol{J}}} \approx 1/\sqrt{N}$ for $k \approx 10$.

Figure 5 (top panel) shows the bias, i.e., the difference between the mean of the realizations and the true value, as a function of k for different sample sizes (full triangles). We confirm the increase in the bias for larger k, with $k = 10$ producing a $\sim 2\times$ larger bias than $k = 1$.

We investigate the correction of A. Charzyńska & A. Gambin (2015), who suggested that the bias is essentially due to points near the edges of the distribution support. For these points, the hypersphere around the point (defined by the distance D_{ik} to the kth neighbor) can have a fraction of its volume outside the support. This results in overestimating the volume, and Equation (11) underestimating the DF for these points. When plugged into Equation (10), this produces a positive bias, in accordance with our results (see Figures 2, 4, and 5). To compensate for this, A. Charzyńska & A. Gambin (2015) proposed to add the following correction to the entropy estimate:

$$C = \frac{1}{N}\sum_{i=1}^{N} \ln\left(\frac{|v(\boldsymbol{w}_i, D_{ik}) \cap \text{supp(W)}|}{|v(\boldsymbol{w}_i, D_{ik})|}\right), \tag{37}$$

where $v(\boldsymbol{w}_i, D_{ik})$ is the volume around point $\boldsymbol{w}_i$, which is drawn from W, in d-dimensions.

In general, the support's shape and the intersections in Equation (37) are unknown, and A. Charzyńska & A. Gambin (2015) proposed assuming a hyper-rectangular box for the

support and a hyper-cubic box for the volume $v(\boldsymbol{w}_i, D_{ik})$, although their analysis restricted to $k = 1$. For cubic boxes of side l_i, we correct for points such that $w_{j,i} > w_{j,\max} - l_i/2$, or $w_{j,i} < w_{j,\min} + l_i/2$, where $j = 1, \ldots, d$, and calculate the volume fractions of the cube inside the rectangular box. Concisely, it results in

$$C = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d} \ln\left[\min\left(\frac{w_{j,\max}}{l_i}, \frac{w_{j,i}}{l_i} + \frac{1}{2}\right) - \max\left(\frac{w_{j,\min}}{l_i}, \frac{w_{j,i}}{l_i} - \frac{1}{2}\right)\right]. \quad (38)$$

After a few experiments, we settled on a cube inscribed within the sphere of radius D_{ik}, i.e., $l_i = (2/\sqrt{d})D_{ik}$. Figure 5 (top panel) shows the corrected biases (empty triangles), which are smaller than the original ones by factors 5–15 (note that the empty blue and red triangles nearly overlap). The improvement is even better for larger k, where the bias is not larger than that of $k = 1$ (up to some k, beyond which the bias is overcorrected).

4.4. Entropy Fluctuation

The noise in the entropy estimate is theoretically expected to have amplitude $\sigma_{\hat{S}} \approx N^{-1/2}$ (G. Biau & L. Devroye 2015). Figure 5 (bottom) shows the fluctuations $\sigma_{\hat{S}_J}$, estimated as half the 16th–84th interpercentile range of the realizations, and multiplied by $N^{1/2}$. We confirm that $\sigma_{\hat{S}_J} \approx N^{-1/2}$, and we see that $\sigma_{\hat{S}}$ decreases with k, but it saturates at $k \approx 10$, reducing $\sigma_{\hat{S}_J}$ by a factor $\approx$2 in comparison to $k = 1$. Empty triangles show $\sigma_{\hat{S}_J}$ for the bias-corrected estimates, which are nearly identical to those of the uncorrected estimates.

In summary, we conclude that taking $k = 10$ suppresses the noise by a factor 2, and the correction proposed by A. Charzyńska & A. Gambin (2015) suppresses the bias without increasing the noise.

5. Minimum Entropy Illustrated

In Appendix A, we rigorously demonstrate why the entropy of a fixed sample is minimum in the correct potential, i.e., in the one where the sample is phase-mixed. In this section, we illustrate this with phase-mixed samples in a self-consistent isochrone model and in potentials of the hypervirial family (N. W. Evans & J. An 2005).

5.1. Isochrone Potential

We generate an initial sample of the isochrone model with $M = b = 1$, and sample size $N/0.7$, selecting the 70% most bound particles in the self-consistent potential, with a final sample of $N \approx 10^4$. This allows us to explore a larger set of models, since we restrict to models where all particles are bound. Note that this cut does not affect the method because the DF is still a function of integrals of motion only, and self-consistency is not required as we explicitly demonstrate below.

We calculate $\hat{S}_J$, i.e., the entropy that the sample would have after phase-mixing, on a grid of potentials (M, b), but in this exercise, we do not correct the bias discussed in Section 4.3. Figure 6 shows $\hat{S}_J$ values in the grid (M, b), using the nearest neighbor, $k = 1$ (top panel), and $k = 10$ (bottom panel). The magenta dots show the true parameters, and the white X's

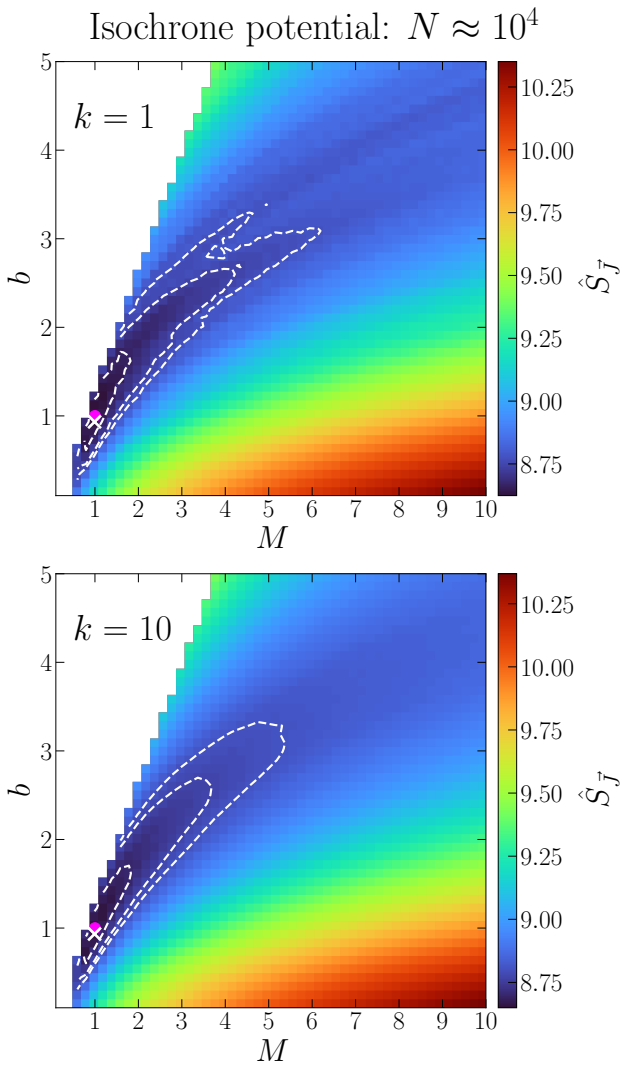

Figure 6. $\hat{S}_J$ values of a self-consistent sample of the isochrone model $(M, b) = (1, 1)$ (magenta dots) with actions evaluated on a grid of parameters (M, b) of the isochrone potential. Contours are percentile levels relative to the minima of $\hat{S}_J$ (white X's). $\hat{S}_J$ is estimated with the nearest neighbor, $k = 1$ (top panel), and $k = 10$ (bottom panel). As expected, a larger k smooths out the $\hat{S}_J$-surface (Section 4.4). $\hat{S}_J$ is minimum near the true potential where the sample is phase-mixed.

show the location of the minimum entropy. The minimum entropy is indeed very near the true values. We note, however, that its exact location depends on the sample realization. The white curves are illustrative contours of the 1st, 5th, and 10th percentiles of $\hat{S}_J$ (not credible contours). The wrinkles in the colors and contours in the top panel reveal the noise in $\hat{S}_J$ for $k = 1$, while for $k = 10$, the surface is much smoother, in agreement with Figure 5 (bottom panel).

5.2. Hypervirial Potentials

At this point, the reader might think that the identification of the potential with the minimum entropy depends on something special about the isochrone potential, or on having a self-consistent sample, as opposed to a generic stationary sample. To dispel this concern, we now use the same sample used before as initial conditions and integrate orbits in four different potentials of the hypervirial family (N. W. Evans & J. An 2005) characterized by

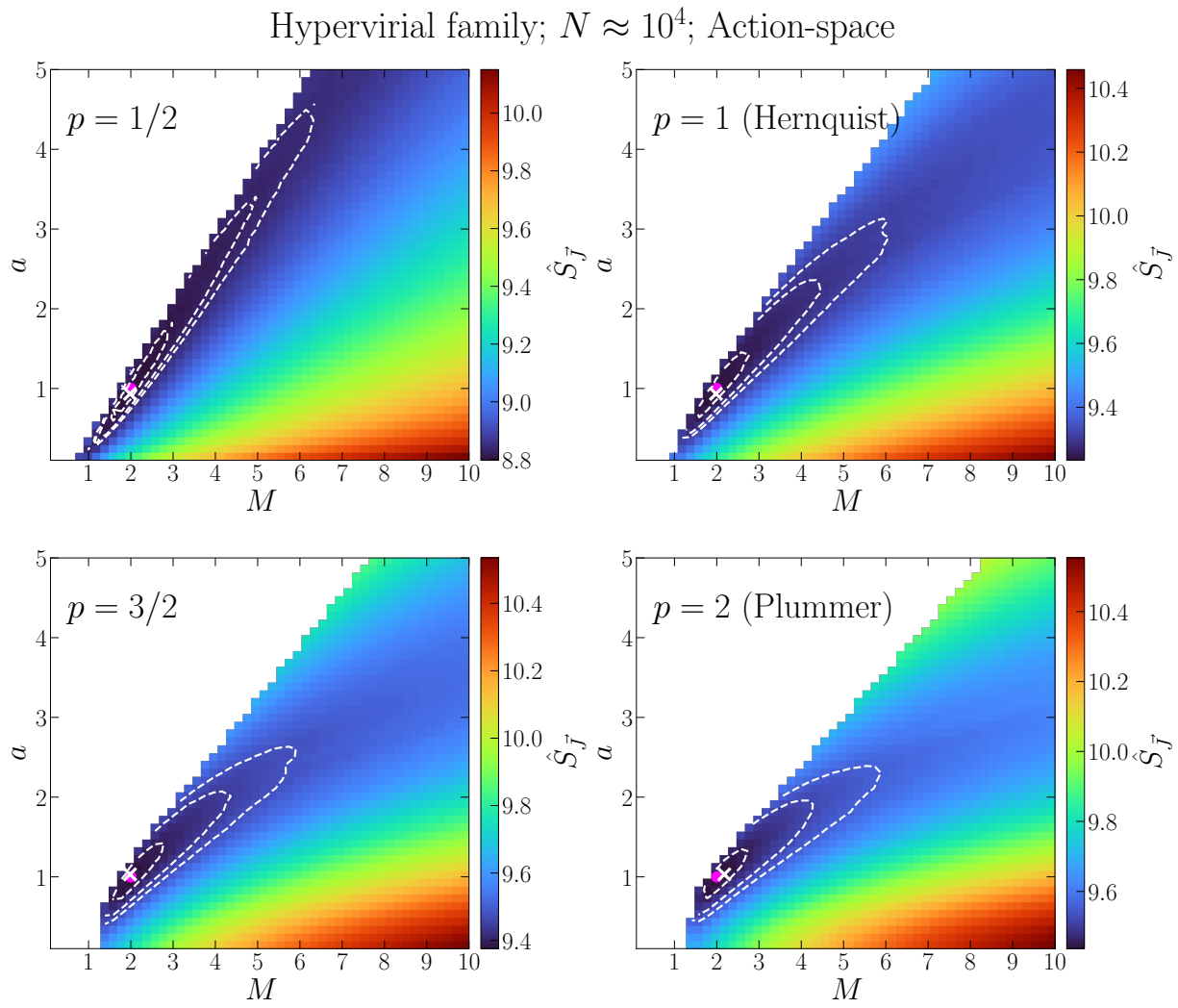

Figure 7. The $\hat{S}_J$-surface with actions calculated on a grid of parameters (M, a). Each panel represents a different potential of the hypervirial family of N. W. Evans & J. An (2005), where the same initial sample is phase-mixed. Magenta dots show the true values, with contours showing percentile levels relative to the minimum of $\hat{S}_J$ (white X). The entropy is estimated with the $k = 10$ nearest neighbor. For all models, $\hat{S}_J$ has its minimum near the correct parameters.

the potential/density pair

$$\phi(r) = -\frac{GM}{a}\frac{1}{[1 + (r/a)^p]^{1/p}}, \tag{39}$$

$$\rho(r) = \frac{(p+1)M}{4\pi a^3}\frac{(r/a)^{p-2}}{[1 + (r/a)^p]^{2+1/p}}, \tag{40}$$

where $0 < p \leqslant 2$ for the most physically interesting cases. These models have $\rho \sim r^{p-2}$ near the center and $\rho \sim r^{-(p+3)}$ in the outskirts, and have finite mass M. Their most interesting property is that they respect the virial theorem locally, in addition to the usual global one. We use these models for their simplicity and because they reduce to well-known models for $p = 1$ (L. Hernquist 1990), and $p = 2$ (H. C. Plummer 1911). We also explore the cases $p = 1/2$ (strong cusp) and $p = 3/2$ (weak cusp). We set $G = a = 1$, but $M = 2$ in order to have only bound orbits in all models. We integrate orbits for $100 \times \langle T_{\rm circ} \rangle$, which is enough for the samples to phase-mix in each of these four potentials. This creates, for each potential, a different equilibrium (phase-mixed) DF, with no explicit expression. Then, for each of these four phase-mixed samples, we calculate actions and $\hat{S}_J$ in trial potentials (M, a), with the corresponding parameter p fixed.

Figure 7 shows the entropy for these potentials. The minima (white crosses) lie near the true values (magenta dots), but once more, their exact locations depend on the particular data realization. This shows that the only requirement of the minimum-entropy method is that the sample is phase-mixed in the true potential, with self-consistency playing no special role. Let us emphasize that this procedure does not require knowing the sample's density or anisotropy profile, or its DF, but only assumes that the DF is an unknown function satisfying the Jeans' theorem, i.e., $f = f(\boldsymbol{J})$.

Figure 8 shows a similar picture, but with the entropy calculated in the space of energy and angular momentum (Equations (23)–(25)), with $T_r = 2\pi/\Omega_r$, where Ω_r is the radial frequency calculated with AGAMA. Once more, the entropy minima are close to the true values for all models. This $\hat{S}_{EL}$ is slightly noisier than $\hat{S}_J$, even though the former is defined in 2D and thus expected to have smaller noise. We suspect that this extra noise in $\hat{S}_{\rm EL}$ may be due to the numerical calculation of the radial period in the density of states (Equations (23)–(24)), further illustrating the advantages of actions.

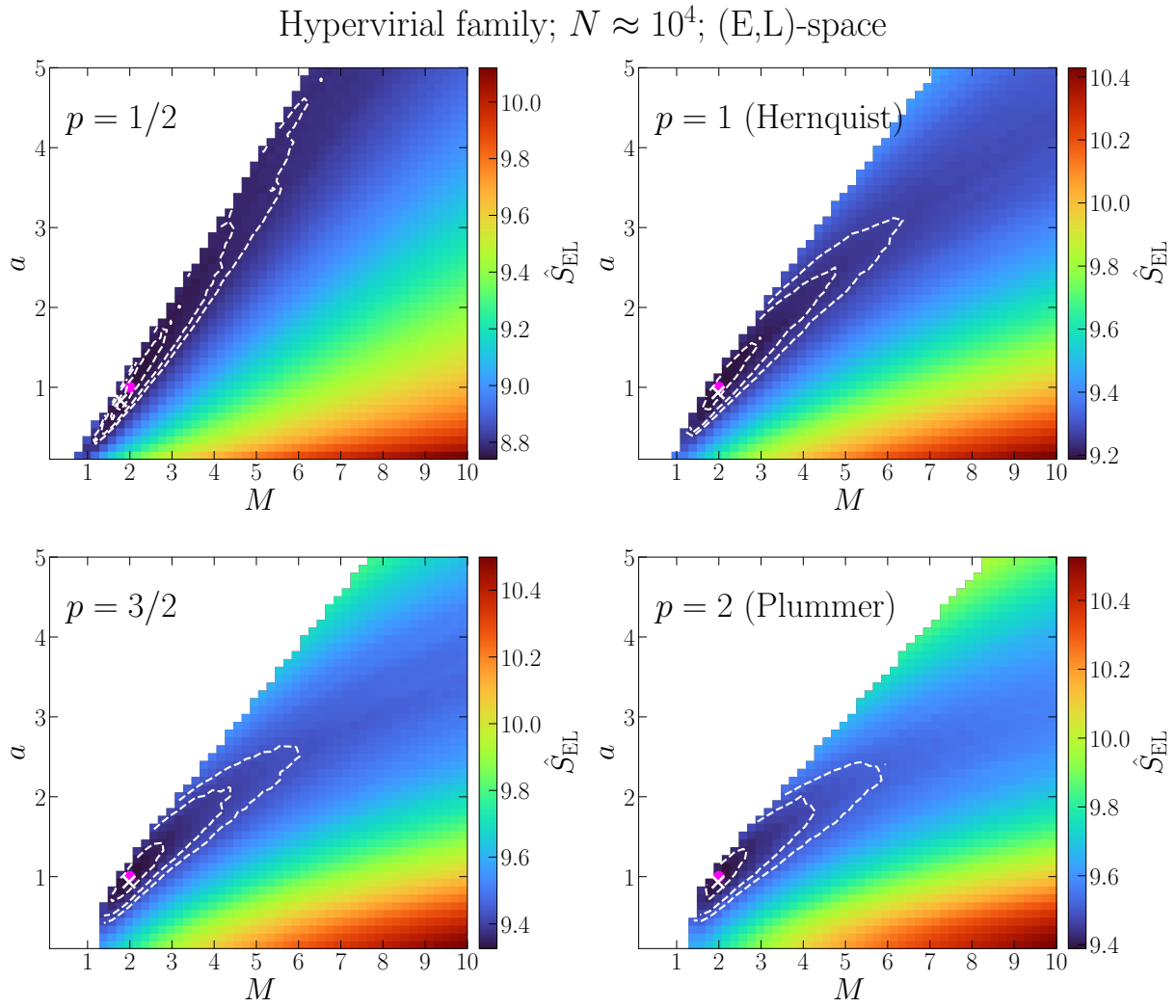

Figure 8. Similar to Figure 7, but calculated in the (E, L)-space. The minima of $\hat{S}_{E,L}$ (white X) again are close to the true values (magenta dots), but the $\hat{S}_{E,L}$-surface has more wrinkles, revealing a slightly larger noise.

6. Model Fitting

While Figures 6–8 may be seen as approximate fits, evaluating models in a grid can quickly become inefficient for models with larger numbers of parameters. Moreover, Figures 6–8 do not provide the odds ratios of different trial potentials. In this section, we perform the actual fits in two steps. We first use the downhill simplex ("Nelder-Mead"), as implemented in scipy, to minimize the entropy of the final DFs considered as unknown functions $f(\boldsymbol{J})$, with actions evaluated in trial potentials.

Having found the best-fit potential where the final (equilibrium) DF is an unknown function $f_0(\boldsymbol{J})$, we explore the posterior of the parameters of the potential. For that, one might want to use the Kullback–Leibler divergence (KLD) between $f_0(\boldsymbol{J})$ and a trial potential with final DF $f(\boldsymbol{J})$. We do not use the KLD as a direct estimate of posterior ratios, as done by R. E. Sanderson et al. (2015), but we present the main expressions in that approach for completeness. The KLD is defined as

$$D_{\rm KL}(f_0||f) \equiv \int f_0 \ln\left(\frac{f_0}{f}\right) d\boldsymbol{\theta} d\boldsymbol{J} = H(f_0, f) - S_0, \qquad (41)$$

where $S_0 = S_{\boldsymbol{J}}[f_0]$. For two distributions f and g in general, $D_{\rm KL}(f||g)$ can be seen as a directed distance from f to g. In fact, $D_{\rm KL}(f||f) = 0$, and it can be shown that $D_{\rm KL}(f||g) \geqslant 0$ (S. Kullback 1968). As for the entropy, the KLD can be estimated via a Monte Carlo using samples of f and f_0, with no explicit expressions for these DFs.

From Equation (3), we get (e.g., T. M. Cover & J. A. Thomas 2006)

$$D_{\rm KL}(f_0||f) = \frac{1}{N}(\ln \mathcal{L}_0 - \ln \mathcal{L}), \qquad (42)$$

where $\ln \mathcal{L}_0 = -NS_0$ is the expectation value of the log-likelihood of the best model. From Bayes' theorem:

$$P(\boldsymbol{p}|\boldsymbol{w}) = \frac{\mathcal{L}(\boldsymbol{w}|\boldsymbol{p})P(\boldsymbol{p})}{P(\boldsymbol{w})}, \qquad (43)$$

where $P(\boldsymbol{p}|\boldsymbol{w})$ is the parameters' posterior probability, $P(\boldsymbol{p})$ is their prior probability, and $P(\boldsymbol{w})$ is a normalization factor. For the best-fit model $\boldsymbol{p}_0$:

$$P(\boldsymbol{p}_0|\boldsymbol{w}) = \frac{\mathcal{L}(\boldsymbol{w}|\boldsymbol{p}_0)P(\boldsymbol{p}_0)}{P(\boldsymbol{w})}. \qquad (44)$$

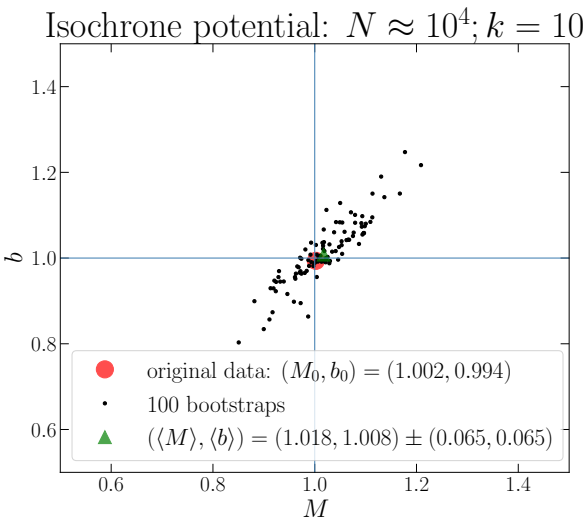

Figure 9. Fit results obtained with a sample of $N \approx 10^4$ in equilibrium (red dot) and bootstrap samples (black points). The green triangle shows the median of the best-fit parameters. In both cases, the true values (lines) are well recovered.

Taking the logarithm of Equations (43)–(44) and replacing in Equation (42), one can translate the KLD into ratios of posterior probabilities of different models. In particular, for flat priors, we have $P(\boldsymbol{p}) = P(\boldsymbol{p}_0)$.

However, this would make a point-wise comparison using a fixed sample evaluated in different models, which does not take into account the intrinsic uncertainties in the data-generation process. In other words, to get meaningful posteriors, one needs to recognize that the data set is just a particular realization of underlying unknown DFs f_0 and f. Not considering this and using KLD with a single sample would produce unrealistically tiny credible contours. On the other hand, neglecting the factor $1/N$ in Equation (42), as done by R. E. Sanderson et al. (2015), significantly overestimates the uncertainties.

We emphasize that we do not use the KLD as a direct translation of the posterior probability ratios, but we use it as a distance metric to explore the posterior probabilities in an ABC, as described below. In this way, each model is accompanied by a different data realization and uncertainties in the data-generation process are appropriately incorporated.

6.1. Fitting the Isochrone Potential

We generate a self-consistent sample of the isochrone model with $M = b = 1$, and sample size $N/0.7$, selecting the 70% most bound particles, with a final sample of $N \approx 10^4$. We assume that the final DF describing the sample in each trial isochrone potential, if orbits were integrated until phase-mixed, is an unknown function $f(\boldsymbol{J})$. We estimate $S_{\boldsymbol{J}}$ (Equation (28)), taking the kth neighbor with $k = 10$ and correcting for the bias as discussed in Section 4.4. We then minimize $\hat{S}_{\boldsymbol{J}}$.

To prevent trapping at local minima, we fit the data starting with initial parameters in a regular grid of 4×4 points, with $0.1 < M < 10$, and $0.1 < b < 5$. We only fit potentials with no unbound particle, setting $\hat{S}_{\boldsymbol{J}} = \infty$ otherwise. The best-fit potential, i.e., the one with smallest $\hat{S}_{\boldsymbol{J}}$ among all fits, is shown as a red dot in Figure 9.

Having found the best-fit potential, we generate 100 new data sets via bootstraps (randomly selecting N points with replacement), fitting the potential for each one. In principle, bootstrap samples might put a problem in the entropy estimate, since duplicated points would have zero-distance to the nearest neighbor. The solution, already implemented in `tropygal`, is treating repeated points as copies of the same point and neglecting copies in the search for neighbors.

In Figure 9, black points show results obtained for the bootstrap samples, and the green triangle is the median of the best-fit parameters. The parameters recovered with both the original sample and with the median of the samples are very near the true values, but we remark that they vary for different data realizations.

In Figure 10, the red dots show again the minimum-entropy best-fit potential for the original sample. Panel (a) shows the results for 10^4 bootstraps, with red contours representing percentiles 39.3 and 86.4 (1σ and 2σ equivalent contours in 2D). Panel (b) shows results for fits of 10^4 data sets generated assuming 10% Gaussian uncertainties in each of the 6D coordinates, i.e., by sampling from Gaussian error distributions centered on the original values (as one might do with observational data, and correlated uncertainties could be introduced through a covariance matrix). These results are biased toward higher masses due to particles in the high-energy tail of the Gaussians, which are unbound in lower-mass models that are thus rejected, an issue not present in the bootstrap samples.

Panels (a) and (b) represent the frequentist confidence contours on the parameters, i.e., without explicitly introducing their prior probabilities. For a Bayesian analysis, after finding a first estimate of the best-fit potential whose DF is $f_0(\boldsymbol{J})$, we characterize the potential's posterior probabilities. Without a bona fide likelihood, we cannot use traditional Markov Chain Monte Carlo sampling, but we resort to a simulation-based inference (see K. Cranmer et al. 2020, for a review). In particular, we perform an ABC (see M. A. Beaumont et al. 2002; S. Sisson et al. 2018; O. Martin et al. 2021), a sampling-rejection method that allows one to sample the posterior in problems where the likelihood is unknown or intractable (see, e.g., E. E. O. Ishida et al. 2015; C. Hahn et al. 2017, for applications in cosmology).

We use the Python package pyABC (Y. Schälte et al. 2022), which implements a sequential Monte Carlo ABC (S. A. Sisson et al. 2007). We start drawing η trial potentials from flat priors, with $0.1 \leqslant M \leqslant 5$ and $0.1 \leqslant b \leqslant 5$. For each trial potential, we generate a new data set, i.e., 6D coordinates $\boldsymbol{w}_i = (\boldsymbol{r}_i, \boldsymbol{v}_i)$ for $i = 1, \ldots, N$. These new coordinates are generated either by bootstrapping the original sample or by sampling from Gaussian error distributions as explained above. For each trial potential, we calculate actions for the associated N coordinates. These actions are then compared with the previously obtained actions of the original sample in the best-fit minimum-entropy potential. For this comparison, we estimate the KLD, Equation (41) (see, e.g., B. Jiang 2018, for its use in ABC). In practice, we calculate $\hat{S}_0$ once for the best-fit potential, and $\hat{D}_{\rm KL}(f_0||f)$ from the cross-entropy $\hat{H}(f_0, f)$, Equations (14)–(16), between actions in the best-fit potential and those at each trial potential. We take the median of all $\hat{D}_{\rm KL}(f_0||f)$ to set a distance threshold ϵ to be used in a next iteration. In each new iteration, we draw new trial potentials from a probability distribution built from weighted

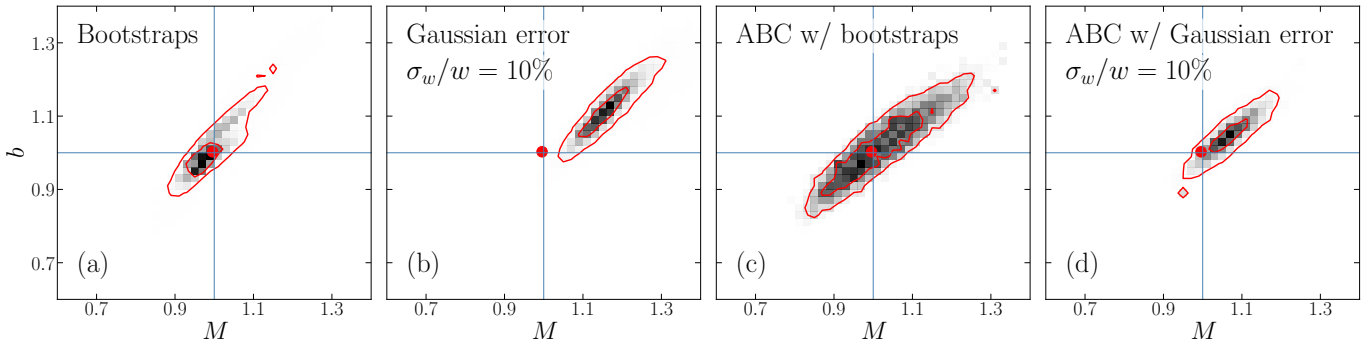

Figure 10. Fit results obtained for 10^4 data sets generated as perturbations of an equilibrium sample with $N = 10^4$ particles. The minimum-entropy fit obtained with the original sample is shown as a red dot. Red lines are 1σ and 2σ equivalent contours. Panel (a) shows fits for bootstrap samples; panel (b) shows fits re-sampling from Gaussian error distributions with relative uncertainties of 10% in each coordinate; panel (c) [(d)] shows the ABC posteriors assuming flat priors, with data sets generated as in panel (a) [(b)].

kernel density estimates of the previously accepted potentials (see S. A. Sisson et al. 2007, for details). Potentials are accepted if $\hat{D}_{\rm KL}(f_0||f) \leqslant \epsilon$, and each iteration finishes when η trial potentials are accepted.

The evolution of the sampling probability distribution toward the parameters' posterior is driven by an adaptive decrease in the distance threshold ϵ (calculated as the median of $\hat{D}_{\rm KL}(f_0||f)$ at each iteration), with a consequent decrease in the samples acceptance rate. It is possible to show that such sampling probability distribution iteratively converges to the parameters' posterior (see, e.g., M. A. Beaumont et al. 2002; S. A. Sisson et al. 2007; B. Jiang 2018). In practice, convergence is assumed after the distance threshold ϵ or the acceptance rate fall below a certain value, or the changes in the posterior become negligible.

We iterate pyABC until the acceptance rate or the distance threshold ϵ falls below 10^{-3}, requiring $\eta = 10^4$ accepted potentials in each iteration. Thus, the final iterations are slower due to the high number of rejected potentials. The KLD is strictly nonnegative, and in order to avoid negative KLD estimates due to noise for potentials near the best fit, we actually take $\max(\hat{D}_{\rm KL}(f_0||f), 10^{-6})$ as the distance metric.

Figure 10 shows the ABC results with data generated via bootstraps (panel (c)) and by sampling from Gaussian error distributions with relative errors of 10% for each coordinate (panel (d)). Note that the first estimate of the best-fit potential (red dots) was obtained with the single original sample. For the ABC with bootstraps, pyABC runs up to iteration 9, when the distance threshold quickly drops to $\epsilon = 10^{-6}$. This is due to our strategy to avoid the negative KLD estimates mentioned above. The larger contours in comparison to panel (a) are due to this premature truncation of the procedure, indicating that in this case, the KLD estimates are not precise enough to guarantee positive values for more iterations and better convergence. The true parameters are recovered with ∼3% errors and ∼10%–12% statistical uncertainties. Thus, these contours are conservative estimates of the true ones, and they encapsulate modeling uncertainties due to lack of better precision in the KLD estimates.

For the ABC with 10% Gaussian errors, pyABC runs up to iteration 19, with the acceptance rate steadily declining to $\approx 10^{-3}$. In this case, we have ∼5%–7% errors and ∼6% statistical uncertainties. This ∼5%–7% bias is still a manifestation of the Gaussian error distribution producing high-energy particles that exclude low-mass potentials, as seen in panel (b). However, since now the distance metric is anchored in the best-fit potential with DF $f_0(\boldsymbol{J})$ (not affected by the Gaussian re-sampling), this problem is alleviated and the bias is reduced with respect to panel (b).

6.2. Fitting an Axisymmetric Potential

We now use a halo-like sample to fit an axisymmetric modified version of the DM halo potential of P. J. McMillan (2017), where we introduce a flattening parameter q, i.e., the ratio between the minor and major axes. The potential is that associated with the density profile

$$\frac{\rho_{\rm DM}(\tilde{r})}{\rho_0} = \left(\frac{\tilde{r}}{r_s}\right)^{-\gamma}\left[1 + \left(\frac{\tilde{r}}{r_s}\right)\right]^{\gamma-3} \exp\left[-\left(\frac{\tilde{r}}{400\ \mathrm{kpc}}\right)^6\right], \tag{45}$$

where $\rho_0 = 8.53702 \times 10^6 M_\odot\ \mathrm{kpc}^{-3}$, $r_s = 19.5725$ kpc, $\gamma = 1$, $\tilde{r} = \sqrt{x^2 + y^2 + (z/q)^2}$, and $q = 0.7$. The exponential term is just a cutoff to assure a finite mass and to avoid numerical problems. In principle, this DM halo potential could be added to all of the other components of the P. J. McMillan (2017) potential (such as the thin and thick disks), even if we only fit the parameters of the former. However, in this case, the inner potential would be dominated by the baryonic components, and the number of star particles of our tracer sample (described below) in the outer regions would not be large enough to constrain the DM halo parameters. Therefore, in what follows, we use the DM halo potential only.

We use AGAMA to generate a spherical stellar halo sample of $N = 10^4$ particles (the tracers), with a broken power-law density profile given by

$$\rho_h(r) \propto \left(\frac{r}{r_h}\right)^{-2.5}\left[1 + \left(\frac{r}{r_h}\right)\right]^{-0.5} \exp\left[-\left(\frac{r}{300\ \mathrm{kpc}}\right)^3\right], \tag{46}$$

where $r_h = 25$ kpc, and the exponential term is again a cutoff at large radii to avoid numerical problems. We set the velocities such that this sample is stationary in our axisymmetric potential. Specifically, we first create a sphericalized version of the potential and initialize the isotropic DF using the Eddington inversion formula, then express this DF as a function of actions, embed it in the flattened potential, and sample positions and velocities of stars from the resulting system. This procedure is equivalent to adiabatically

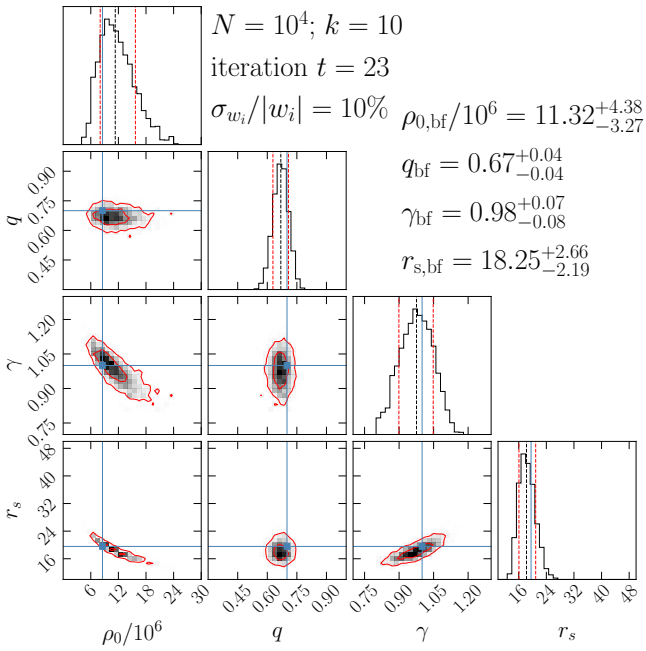

Figure 11. ABC results for $N = 10^4$ particles that phase-mixed in a flattened axisymmetric potential ($q = 0.7$). Note that 6D coordinates are assumed to have Gaussian error distributions with relative uncertainties $\sigma_{w_i}/|w_i| = 10\%$, for $i = 1 \ldots 6$. In red are shown the 1σ and 2σ equivalent contours. Vertical dashed lines show the 16th, 50th, and 84th percentiles. The true parameters (blue lines/dots) are well recovered. In particular, for the flattening parameter, $\sigma_q/q \sim 5\%$.

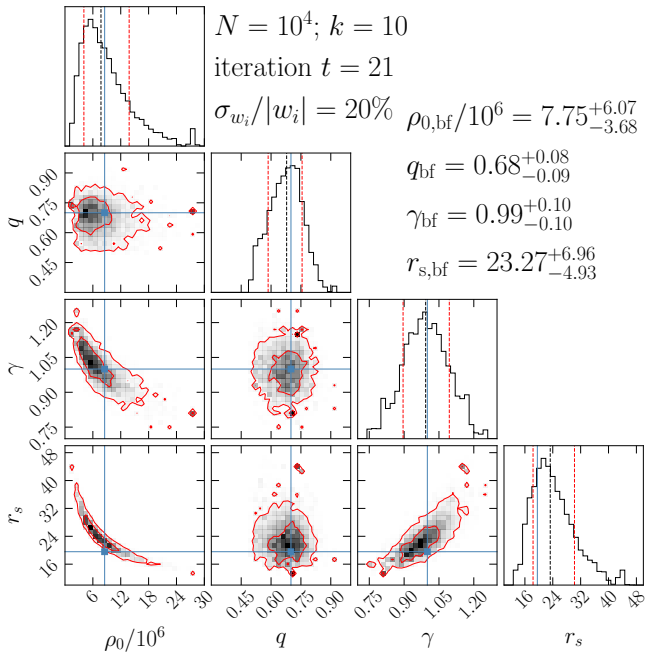

Figure 12. Similar to Figure 11, but now assuming each coordinate to have Gaussian uncertainties $\sigma_{w_i}/|w_i| = 20\%$. We see the worsening of the fit, but the true parameters are still overall well recovered. In particular, $\sigma_q/q \sim 10\%$.

deforming the potential from the initial (spherical) to the final (nonspherical) shape.

With this sample, assumed to be described by an unknown DF $f(\boldsymbol{J})$, we fit the potential parameters ρ_0, q, γ, and r_s. We estimate the actions $\boldsymbol{J} = (J_r, J_\varphi, J_z)$ in each trial potential through the Stackel fudge (J. Binney 2012) using AGAMA. As in Section 6.1, we first identify the best-fit potential minimizing $\hat{S}_{\boldsymbol{J}}$, Equation (28), starting in a grid of parameter values. We then use the actions in the globally best-fit potential as the "observed data" described by an unknown DF $f_0(\boldsymbol{J})$ to characterize the parameters' posterior in the ABC.

Once more, we run pyABC accepting $\eta = 10^4$ models in each iteration, generating a new size-N data sample for each potential. New data sets are generated by sampling from Gaussian error distributions of width $\sigma_{w_i}/|w_i| = 10\%$ for $i = 1, \ldots, 6$. Figure 11 shows the resulting corner plot. Once more, we run pyABC until the acceptance rate or the distance threshold falls below 10^{-3}, by which time the true parameters are well recovered. This suggests that this is a reasonable choice when dealing with observed data, where the true answer is unknown. In particular, the flattening parameter is recovered with relative uncertainty $\sim$5%.

Figure 12 shows a similar plot, obtained with Gaussian error distributions with $\sigma_{w_i}/|w_i| = 20\%$ for each coordinate. We clearly see the worsening of the fit compared to Figure 11, but the true parameters are still recovered reasonably well. In this case, the flattening parameter is recovered with uncertainty $\sim$10%.

7. Discussion

7.1. Future Improvements

An ideal method to constrain a gravitational potential using the kinematics of a stellar sample should:

1. allow constraints on general mass distributions, including general axisymmetric and triaxial systems;
2. properly incorporate uncertainties and covariances in the data, providing not only best-fit values, but full probability distributions of the fit parameters;
3. avoid making any assumptions regarding the DF besides the requirements of Jeans' theorem;
4. be computationally efficient in order to handle samples with $\sim 10^4$–10^6 stars, typical of stellar halo samples, or stars within a globular cluster;
5. properly consider the survey's footprint and selection function;
6. handle incomplete information, e.g., samples missing line-of-sight velocities and/or distances.

We demonstrated that our method already satisfies items 1–4. Although we have not tested it for triaxial potentials, the only difficulty is to efficiently estimate actions in such potentials. With these actions at hand, one can also investigate triaxial systems with this method.

In Section 4, we discussed the bias and noise of the k-NN entropy estimator used in this paper. Although this estimator is good enough for most applications, our method would benefit from more precise and accurate estimates (see, e.g., D. Lombardi & S. Pant 2016; T. B. Berrett et al. 2019; Z. Ao & J. Li 2023 for recent works with this aim).

In Section 4.3, we showed that the bias correction proposed by A. Charzyńska & A. Gambin (2015) effectively suppressed the bias in the entropy estimates for self-consistent samples of the isochrone model. This correction assumes the sample's support is a parallelepiped defined by the extreme values of each coordinate. The typical action space of a self-consistent sample of an axisymmetric potential has a shape close to a tetrahedron with two perpendicular faces (see J. Binney & S. Tremaine 2008). The reason why this simple correction

worked so well in the self-consistent isochrone sample is probably that this tetrahedron support is not so different from the assumed parallelepiped support for most stars. For non self-consistent samples, and particularly for samples with sharp geometric cuts, the actual support in action space can be more complicated. For these cases, it will be important to implement bias corrections for samples with a general support.

In the DF-fitting method, where one assumes an analytical expression for the DF, selection effects due to geometric cuts are taken into account by the normalization factor $A = \int_{\mathcal{V}} f(\boldsymbol{w}|\boldsymbol{p})\mathbb{S}(\boldsymbol{w})d^6\boldsymbol{w}$, where $\mathcal{V}$ is the survey volume. This integral can be very complicated and time consuming, and its limited numerical accuracy is the main source of noise in these methods (P. J. McMillan & J. J. Binney 2013; K. Hattori et al. 2021). In the minimum-entropy method developed in this paper, we do not have an analytic DF, but the survey footprint can be accounted for by the fractional time each orbit spends in it (see Equation (A8)). This can be done either by generating a number of angle variables uniformly distributed in $[0, 2\pi)$ for each star and checking how many pairs $(\boldsymbol{\theta}, \boldsymbol{J})$ end up inside the footprint, or simply integrating orbits and directly counting the fractional time inside the footprint for each orbit. Other important improvements involve handling samples with missing data and unbound stars, such as hyper-velocity stars. In particular, we currently do not fit potentials with even a single unbound star, which tends to bias the estimates to higher masses if the original sample has high-energy stars in the correct potential (see Figure 10). An improved version of the method should handle (and penalize for) unbound stars to eliminate this bias.

7.2. Comparison with Other Methods

In the DF-fitting method (P. J. McMillan & J. J. Binney 2013; K. Hattori et al. 2021), as the traditional likelihood-based approach in general, the inference process is facilitated by having a smooth function to be maximized/minimized and by optimizing model evaluations. However, this relies on the assumed DF correctly describing the data, which is hard to guarantee in general, especially if deviations from equilibrium are expected. The minimum-entropy method, in conjunction with the ABC analysis, avoids that assumption while taking into account the uncertainties in the data-generation process. In other words, the possible bias introduced by assuming a DF in the DF-fitting method is, in the minimum-entropy method, traded off for statistical uncertainties that reflect our ignorance of the true DF.

J. Peñarrubia et al. (2012) proposed to constrain the Galactic potential by minimizing the entropy of the energy distribution of cold tidal streams. Their method assumes a narrow energy distribution (in the right potential), and that the probabilities for a star to be in a certain position and to have a certain energy are independent. Under a few approximations, they show that the entropy of the energy distribution should be minimum at the true potential and, assuming a Gaussian energy distribution, demonstrate that their method works for spherical potentials. Analogously, R. E. Sanderson et al. (2015) proposed to maximize the KLD between the action distribution and the product of its marginal distributions of stellar streams. In other words, they proposed to recover the true potential as the one maximizing the correlations between the three actions. Their estimate of the KLD does not require assuming a specific DF, but is performed on a fixed grid in action space. When used to evaluate the odds of different models, this requires rejecting points outside the grid of actions in the best-fit model. R. E. Sanderson et al. (2015) applied their method to spherical models, approximately recovering the true potential. S. Reino et al. (2021) later constrained an axisymmetric Stackel potential using data on a few streams, improving some aspects of this method. In particular, they used EnLink (S. Sharma & K. V. Johnston 2009), a metric-free density estimator, to estimate the KLD (but see also S. Reino et al. 2022 for an erratum).

While these methods use stellar streams, assume the samples are clustered in the space of integrals, and minimize the entropy of the integrals' pdf, our method instead targets smooth stellar populations, assumes they are phase-mixed, and minimizes the entropy of the full (6D) DF.

The orbital pdf method developed by J. Han et al. (2016) and its successor emPDF (Z. Li et al. 2024) propose recovering the underlying potential exploring Jeans' theorem but without specifying a DF, in a similar vein as the minimum-entropy method developed here. Their methods are currently restricted to spherical systems, but can be extended into action space in a more general geometry. Their underlying general principles and final expressions are similar to those we derive for the spherical case, although based on different physical arguments and developed independently. While Z. Li et al. (2024) focused on estimating the DF using kernel density estimates, our approach uses well-established recipes to estimate the differential entropy of a sample via k-NN (but a few other estimators can be used too).

Thus, in some sense, emPDF and the minimum-entropy method represent two different views of the same general principles. Nonetheless, we believe the general formalism developed in the current work illuminates not only fundamental aspects of any method to constrain mass distributions exploring Jeans' theorem, but also our picture of the evolution of collisionless systems toward stationary states.

7.3. Disequilibrium in the MW

Complicating the application of the minimum-entropy method to the MW is the kinematic perturbation from the Large Magellanic Cloud (LMC), currently near a pericentric passage (at a distance of $\approx$50 kpc; G. Besla et al. 2007). This perturbation is significant enough to produce a reflex motion of the MW disk and its inner halo ($\lesssim$30 kpc) toward the LMC past trajectory (D. Erkal et al. 2021; N. Garavito-Camargo et al. 2021; M. S. Petersen & J. Peñarrubia 2021; A. Byström et al. 2025). Thus, dynamical equilibrium cannot be assumed for the outer halo ($\gtrsim$30 kpc). However, if one wants to probe the outer halo still assuming dynamical equilibrium, a promising avenue is to try to "undo" or correct for the kinematical perturbation from the LMC (A. J. Deason et al. 2021; L. Correa Magnus & E. Vasiliev 2022). On the other hand, for the inner halo ($\lesssim$30 kpc), the assumption of equilibrium still seems reasonable.

7.4. Time Evolution versus Fixed Sample

In Section 3, we introduced expressions for the entropy of DFs that are functions of integrals of motion. On the one hand, one can think of Equations (17), (21), and (27) as the entropy the system would achieve if the same sample is allowed to evolve in each trial potential until it phase-mixes, with the

original DF evolving to another DF that depends only on integrals evaluated in that trial potential. In this case, minimizing Equations (17), (21), or (27) corresponds to minimizing the future entropy, for the sample will phase-mix if put in an incorrect potential, increasing the entropy—this generalizes the simpler reasoning of J. Magorrian (2014) for minimizing the entropy of an "orbit-averaged" DF. Interestingly, we do not need to wait for the time evolution, since integrals are conserved and can therefore be evaluated at the onset in each potential. The DF evolution is purely driven by the remaining variables (e.g., angles), which evolve to a uniform distribution in their respective supports.

On the other hand, in Appendix A we demonstrate that, for a fixed equilibrium sample of a DF $f(\boldsymbol{r}, \boldsymbol{v})$,

$$S_{\boldsymbol{I}} \geqslant S(f), \tag{47}$$

where $S(f)$ is the sample's invariant entropy and

$$S_{\boldsymbol{I}} = -\int F(\boldsymbol{I}) \ln\left[\frac{F(\boldsymbol{I})}{g(\boldsymbol{I})}\right] d\boldsymbol{I} \tag{48}$$

is the general form of Equations (17), (21), and (27), with $\boldsymbol{I}$ being integrals, $g(\boldsymbol{I})$ being the density of states, and $F(\boldsymbol{I}) = f(\boldsymbol{I})g(\boldsymbol{I})$. In this case, with no time evolution implied, the marginalization defining the integrals' pdf $F(\boldsymbol{I})$, Equation (A3), is considered even when the remaining variables are not uniformly distributed, i.e., considering the fixed sample as nonstationary in each of the trial potentials. Furthermore, in Appendix A we show that in the correct potential, where the remaining variables are uniformly distributed on their supports, $S_{\boldsymbol{I}} = S(f)$.

The two interpretations (future entropy and fixed sample) require minimizing the same quantity $S_{\boldsymbol{I}}$, showing that they are equivalent. Thus, for potentials where stationary states are synonymous with uniform distributions in the remaining variables (not true in exceptional cases such as the harmonic oscillator), the derivation of Equation (47)—see Appendix A—represents a demonstration of the second law of thermodynamics for collisionless gravitational systems in these potentials, i.e., of the inevitable entropy increase for a sample starting out of equilibrium, as illustrated in Figures 3 and 4.

In contrast, it is traditionally assumed that the macroscopic evolution of a collisionless system, i.e., in a smooth potential ϕ, is described by the Vlasov (or collisionless Boltzmann) equation

$$\frac{df}{dt} \equiv \frac{\partial f}{\partial t} + \boldsymbol{v} \cdot \frac{\partial f}{\partial \boldsymbol{r}} - \frac{\partial \phi}{\partial \boldsymbol{r}} \cdot \frac{\partial f}{\partial \boldsymbol{v}} = 0, \tag{49}$$

which implies entropy conservation. According to this view, the aforementioned entropy increase would result from coarse-graining, i.e., from losing information in fine-scale phase-space structures (e.g., D. Lynden-Bell 1967; S. Tremaine et al. 1986; W. Dehnen 2005; Y. Levin et al. 2014; U. Banik et al. 2022; L. Barbieri et al. 2022). Although the work of W. Dehnen (2005) is the closest to our interpretation, it still assumes that the underlying evolution is described by Equation (49) (while acknowledging that the extra-fine phase-space structures introduced by it are artificial), and that the evolution to a stationary state requires coarse-graining.

Since coarse-graining is subjective, as it depends on the scale one chooses to coarse-grain, the recovery of the gravitational potential by minimizing the entropy of a sample in equilibrium would be surprising if based on coarse-graining. Also surprising would be the agreement of entropy estimates when using different sets of integrals, since they involve distances and neighbors in very different spaces. Additionally, the equivalence of Equation (48) with the expected future entropy after phase-mixing in each trial potential might appear coincidental: what would be special about this coarse-grain scheme?

In line with L. Beraldo e Silva et al. (2019a, 2019b), here we argue that these entropy estimates recover all information that is available from a finite-N sample and thus do not operate a coarse-grain (see Section 4.2). In contrast, Equation (49) assumes the limit $N \to \infty$ (R. L. Dobrushin 1979) and implies the development of indefinitely fine phase-space structures, i.e., indefinitely large wavenumbers k in Fourier space. According to the Nyquist–Shannon theorem (H. Nyquist 1928; C. Shannon 1949), to a given size-N sample in d-dimensions, one can only associate unique functions with maximum wavenumber $k \lesssim N^{1/d}$. Functions with higher wavenumbers (sharper features) introduce information that is not contained in the sample. This constrains the finest structures allowed for a DF describing a real, i.e., finite-N, system (L. Beraldo e Silva et al. 2019a). Starting out of equilibrium, the system starts developing fine structures, i.e., the maximum wavenumber increases, until hitting the Nyquist–Shannon upper limit. After that, the system approaches a steady state described by a DF that is a function of integrals only, with the remaining variables uniformly distributed in their domains. The timescale for this collisionless relaxation is $\tau \lesssim 0.1N^{1/6}\tau_{\rm cr}$ (L. Beraldo e Silva et al. 2019b), i.e., a few crossing times $\tau_{\rm cr}$ for typical stellar samples. Thus, the system does not produce the extra-fine phase-space structures predicted by Equation (49). For recent related discussions in plasma physics, see V. Zhdankin (2022, 2023), R. J. Ewart et al. (2023), and M. L. Nastac et al. (2024).

In summary, real collisionless gravitational samples are finite-N and, because of this, phase-mix toward stationary states described by DFs depending only on integrals. Our method explores the objective entropy increase associated with this process.

8. Summary

We have presented a method to constrain the gravitational potential where a tracer sample is in dynamical equilibrium. It is based on the idea that, if put in a different potential, this sample would phase-mix, producing an entropy increase. The potential is then recovered by minimizing the future entropy of the sample with respect to the parameters of the potential. This entropy is estimated using integrals of motion, and the parameters of the potential enter the fit through these integrals.

We focused on actions, and demonstrated their advantages, including possible constraints on the MW's DM halo shape. Investigation of this particular problem will benefit from large spectroscopic surveys such as the DESI-MWS (A. P. Cooper et al. 2023) in tandem with Gaia. This method can be similarly applied to other integrals, such as energy and angular momentum, e.g., in the study of spherical systems like globular clusters (see Appendix A). Finally, in Appendix B, we discuss the possibility of recovering a potential by maximizing the samples' entropy in angle-space, concluding that this is not expected to work in general.

Acknowledgments

We thank the anonymous referee for providing a careful reading and constructive comments. L.B.e.S. thanks Wyn Evans, Josh Speagle, Chirag Modi, David Hogg, Bernardo Modenesi, Sergey Koposov, Zhaozhou Li, Carrie Filion, and the stellar halos group at U. of Michigan for useful discussions. M.V. and L.B.e.S. acknowledge the support of NASA ATP award 80NSSC20K0509 and U.S. National Science Foundation AAG grant AST-2009122, and M.V. acknowledges support from NASA ATP award 80NSSC24K0938. E.V. thanks Hans-Walter Rix and Kathryn Johnston for valuable comments, and acknowledges support from an STFC Ernest Rutherford fellowship (ST/X004066/1). K.H. is supported by JSPS KAKENHI grant Nos. JP24K07101, JP21K13965, and JP21H00053. W.d.S.P. is supported by CNPq (309723/2020-5). L.B.e.S. and K.J.D. acknowledge support from the Heising Simons Foundation grant No. 2022-3927. We respectfully acknowledge that the U. of Arizona is on the land and territories of Indigenous peoples. Today, Arizona is home to 22 federally recognized tribes, with Tucson being home to the O'odham and the Yaqui. We respect and honor the ancestral caretakers of the land, from time immemorial until now, and into the future.

Software: numpy (C. R. Harris et al. 2020), scipy (P. Virtanen et al. 2020), Agama (E. Vasiliev 2019), pyABC (Y. Schälte et al. 2022), `tropygal` (this work).

Appendix A
Mathematical Basis of the Minimum-entropy Method

For a DF separable in the space of angles-actions, $f(\boldsymbol{\theta}, \boldsymbol{J}) = \mathcal{F}(\boldsymbol{\theta})F(\boldsymbol{J})$, the entropy, Equation (1), is the sum of the respective subspaces' entropies, $S(f) = S(\mathcal{F}(\boldsymbol{\theta})) + S(F(\boldsymbol{J}))$. For stationary states, the angle distribution is uniform, $\mathcal{F}(\boldsymbol{\theta}) = (2\pi)^{-3}$, and thus $S(\mathcal{F}(\boldsymbol{\theta}))$ is maximum, in the potential where the sample is stationary. Since $S(f)$ is invariant for changes of variables, i.e., for angle-actions evaluated in any potential, $S(F(\boldsymbol{J}))$ is minimum in that potential.

Here we generalize this idea to nonseparable DFs. In fact, one can always separate pdfs in terms of conditional pdfs, e.g., $f(\boldsymbol{\theta}, \boldsymbol{J}) = \mathcal{F}(\boldsymbol{\theta}|\boldsymbol{J})F(\boldsymbol{J})$, where $\mathcal{F}(\boldsymbol{\theta}|\boldsymbol{J})$ is the conditional pdf of $\boldsymbol{\theta}$, given $\boldsymbol{J}$. Thus, loosely speaking, we have $S(f) = S(\mathcal{F}(\boldsymbol{\theta}|\boldsymbol{J})) + S(F(\boldsymbol{J}))$ and can recover the potential by minimizing $S(F(\boldsymbol{J}))$, since $S(\mathcal{F}(\boldsymbol{\theta}|\boldsymbol{J}))$ is maximum at any given action in the right potential. Below, we formalize this idea and generalize it to other integrals.

Given the DF $f(\boldsymbol{w})$, where $\boldsymbol{w} \equiv (\boldsymbol{r}, \boldsymbol{v})$, consider a random variable $X = (X_1, \ldots, X_n)$, with $X_i = X_i(\boldsymbol{w})$. Let F_X be the pdf of X; i.e., F_X is a positive and normalized function on $\mathbb{R}^n$. The expectation value of X is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} f(\boldsymbol{w})X(\boldsymbol{w})\, d^6\boldsymbol{w} = \int_{-\infty}^{\infty} F_X(\boldsymbol{x})\boldsymbol{x}\, d^n\boldsymbol{x}. \tag{A1}$$

Let $\boldsymbol{I} = (I_1, \ldots, I_m)$, with $m < 6$, be a second random variable, with $I_i = I_i(\boldsymbol{w})$—we will later make $\boldsymbol{I}$ be the integrals of motion, e.g., for spherical and isotropic systems, we set $I_1 = E$; for spherical and anisotropic ones, we set $(I_1, I_2) = (E, L)$; for angle-action variables, $(I_1, I_2, I_3) = \boldsymbol{J}$. In general, we require $\boldsymbol{I}$ to have the following property:

1. There is a smooth function $\Psi\colon \mathbb{R}^6 \to \mathbb{R}^6$, whose Jacobian matrix

$$(J_\Psi)_{ij} \doteq \frac{\partial \Psi_i}{\partial w_j}, \quad i, j = 1,\ldots,6$$

is nondegenerate (i.e., its determinant is nonvanishing), such that $I_k(\boldsymbol{w}) = \Psi_{6-m+k}(\boldsymbol{w})$, $k = 1, \ldots, m$. In other words, $\boldsymbol{I}$ corresponds to the last $m < 6$ coordinates of some change of variables Ψ in 6D.

For random variables $\boldsymbol{I}$ with this property, we consider the conditional expectation in the sense of a "disintegration" of f with respect to $\boldsymbol{I}$, $\mathcal{F}(\cdot|\boldsymbol{I})_{\boldsymbol{I}\in\mathbb{R}^m}$ (for a friendly, yet thorough, introduction to this topic, see J. T. Chang & D. Pollard 1997). This is a family of pdfs such that for each $\boldsymbol{I}$, it gives the pdf $\mathcal{F}(\boldsymbol{z}|\boldsymbol{I})$ of the remaining variables $\boldsymbol{z} \in \mathbb{R}^{6-m}$. This pdf is properly normalized and is different from marginalizing over $\boldsymbol{I}$, or from simply taking f at fixed $\boldsymbol{I}$ values. Given that the pdf of the new variables $(\boldsymbol{z}, \boldsymbol{I})$ is $f(\Psi^{-1}(\boldsymbol{z}, \boldsymbol{I})) \cdot |J_{\Psi^{-1}}(\boldsymbol{z}, \boldsymbol{I})|$, and not $f(\Psi^{-1}(\boldsymbol{z}, \boldsymbol{I}))$ alone, the conditional probability with respect to $\boldsymbol{I}$ is explicitly given by:

$$\mathcal{F}(\boldsymbol{z}|\boldsymbol{I}) = \frac{f(\Psi^{-1}(\boldsymbol{z}, \boldsymbol{I}))|J_{\Psi^{-1}}(\boldsymbol{z}, \boldsymbol{I})|}{F(\boldsymbol{I})}, \tag{A2}$$

where

$$F(\boldsymbol{I}) = \int_{-\infty}^{\infty} f(\Psi^{-1}(\boldsymbol{z}', \boldsymbol{I}))|J_{\Psi^{-1}}(\boldsymbol{z}', \boldsymbol{I})|d^{6-m}\boldsymbol{z}' \tag{A3}$$

is the pdf of the random variable $\boldsymbol{I}$, i.e., the marginalization over the remaining variables $\boldsymbol{z}$. Moreover, as expected, F only depends on $\boldsymbol{I}$, and not on the particular choice of the transformation $\Psi\colon \mathbb{R}^6 \to \mathbb{R}^6$ for the remaining variables $\boldsymbol{z}$, since Equation (A3) marginalizes over them. This elementary remark is important later on in this appendix. Note that if $f = f(\boldsymbol{I})$, i.e., if it is uniform in $\boldsymbol{z}$, Equation (A3) reduces to Equations (18), (22), or (26) as particular cases.

With the change of variables $\boldsymbol{w} \to (\boldsymbol{z}, \boldsymbol{I})$ in Equation (A1), we get:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} f(\boldsymbol{w})X(\boldsymbol{w})d^6\boldsymbol{w} = \int_{-\infty}^{\infty} X(\Psi^{-1}(\boldsymbol{z}, \boldsymbol{I})) \times f(\Psi^{-1}(\boldsymbol{z}, \boldsymbol{I}))|J_\Psi(\boldsymbol{I}, \boldsymbol{z})|d^{6-m}\boldsymbol{z}d^m\boldsymbol{I},$$

and from Equation (A2), it results that

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} X(\Psi^{-1}(\boldsymbol{z}, \boldsymbol{I}))\mathcal{F}(\boldsymbol{z}|\boldsymbol{I})F(\boldsymbol{I})d^{6-m}\boldsymbol{z}d^m\boldsymbol{I} = \int_{-\infty}^{\infty} F(\boldsymbol{I})\left(\int_{-\infty}^{\infty} \mathcal{F}(\boldsymbol{z}|\boldsymbol{I})X(\Psi^{-1}(\boldsymbol{z}, \boldsymbol{I}))d^{6-m}\boldsymbol{z}\right)d^m\boldsymbol{I}.$$

In fact, the last equality is the formal *definition* of $\mathcal{F}(\cdot|\boldsymbol{I})_{\boldsymbol{I}\in\mathbb{R}^m}$ being the disintegration of the DF with respect to the random variable $\boldsymbol{I}$. Making $X = -\ln f$ and using Equation (A2), we get

$$\begin{aligned} S(f) = \mathbb{E}[-\ln f] &= \int_{-\infty}^{\infty} F(\boldsymbol{I})\left\{\int_{-\infty}^{\infty} \mathcal{F}(\boldsymbol{z}|\boldsymbol{I})\cdot [-\ln f(\Psi^{-1}(\boldsymbol{z}, \boldsymbol{I}))]d^{6-m}\boldsymbol{z}\right\}d^m\boldsymbol{I} \\ &= \int_{-\infty}^{\infty} F(\boldsymbol{I})\left\{\int_{-\infty}^{\infty} \mathcal{F}(\boldsymbol{z}|\boldsymbol{I})\cdot \left[-\ln\left(\frac{F(\boldsymbol{I})}{|J_{\Psi^{-1}}(\boldsymbol{I}, \boldsymbol{z})|}\mathcal{F}(\boldsymbol{z}|\boldsymbol{I})\right)\right]d^{6-m}\boldsymbol{z}\right\}d^m\boldsymbol{I}. \end{aligned}$$

Writing the logarithm of the product as the sum of logarithms, we get

$$S(f) = \int_{-\infty}^{\infty} F(\boldsymbol{I}) \left\{ \int_{-\infty}^{\infty} \mathcal{F}(\boldsymbol{z}|\boldsymbol{I}) \ln |J_{\Psi^{-1}}(\boldsymbol{I}, \boldsymbol{z})| d^{6-m}\boldsymbol{z} \right\} \times d^m\boldsymbol{I} - \int_{-\infty}^{\infty} F(\boldsymbol{I}) \ln F(\boldsymbol{I}) d^m\boldsymbol{I} - \int_{-\infty}^{\infty} F(\boldsymbol{I}) \left\{ \int_{-\infty}^{\infty} \mathcal{F}(\boldsymbol{z}|\boldsymbol{I}) \ln \mathcal{F}(\boldsymbol{z}|\boldsymbol{I}) d^{6-m}\boldsymbol{z} \right\} d^m\boldsymbol{I},$$

where for the second term on the right-hand side, we used the fact that $\int_{-\infty}^{\infty} \mathcal{F}(\boldsymbol{z}|\boldsymbol{I}) d^{6-m}\boldsymbol{z} = 1$. Hence,

$$S(f) = \mathbb{E}[\ln |J_{\Psi^{-1}}|] + S(F(\boldsymbol{I})) + \mathbb{E}_{\boldsymbol{I}}[S(\mathcal{F}(\cdot|\boldsymbol{I}))], \tag{A4}$$

where $\mathbb{E}_{\boldsymbol{I}}[S(\mathcal{F}(\cdot|\boldsymbol{I}))]$ denotes the expectation of the entropy

$$S(\mathcal{F}(\cdot|\boldsymbol{I})) \doteq -\int_{-\infty}^{\infty} \mathcal{F}(\boldsymbol{z}|\boldsymbol{I}) \ln [\mathcal{F}(\boldsymbol{z}|\boldsymbol{I})] d^{6-m}\boldsymbol{z} \tag{A5}$$

of the conditional pdfs. Note that these entropies define a random variable that only depends on $\boldsymbol{I}$. In particular, if Ψ is a canonical transformation ($|J_{\Psi^{-1}}(\boldsymbol{I}, \boldsymbol{z})| = 1$), from Equation (A4), we have $S(f) = S(F(\boldsymbol{I})) + \mathbb{E}_{\boldsymbol{I}}[S(\mathcal{F}(\cdot|\boldsymbol{I}))]$.

Equation (A4) is the main general result of this appendix. We show below that it justifies our minimum-entropy method for fitting galactic potentials. With this aim, it is convenient to make the following additional assumption on the variable transformation Ψ, and afterward, we show how it can be removed:

2. The Jacobian determinant $|J_{\Psi^{-1}}(\boldsymbol{I}, \boldsymbol{z})|$ only depends on $\boldsymbol{I}$.

In fact, given a partial transformation $\tilde{\Psi} : \mathbb{R}^{6-m} \to \mathbb{R}^{6-m}$, it is common to find a point transformation for the remaining variables such that the total new transformation Ψ: $\mathbb{R}^6 \to \mathbb{R}^6$ is even *canonical*, i.e., $|J_{\Psi^{-1}}(\boldsymbol{I}, \boldsymbol{z})| = 1$.

Suppose that, for all $\boldsymbol{I} \in \mathbb{R}^m$, the maximum allowed support of the pdfs $\mathcal{F}(\cdot|\boldsymbol{I})$ is some bounded region $\Omega_z(\boldsymbol{I})$ of $\mathbb{R}^{6-m}$. The region $\Omega_z(\boldsymbol{I})$ encodes the set of coordinates $\boldsymbol{z} \in \mathbb{R}^{6-m}$ corresponding to particles that, at fixed $\boldsymbol{I}$, are *not* forbidden to appear in the sample, e.g., for being unbound or for its coordinates lying outside the survey footprint.

If the coordinates $\boldsymbol{I} \in \mathbb{R}^m$ are constants of motion, one expects that the original DF $f(\boldsymbol{w})$ is stationary (phase-mixed), or, more generally, a cut of some stationary DF if, and only if, $f(\Psi^{-1}(\boldsymbol{z}, \boldsymbol{I}))$ is constant for $\boldsymbol{z}$ within the maximum allowed support $\Omega_z(\boldsymbol{I})$, at any fixed $\boldsymbol{I} \in \mathbb{R}^m$. Thus, here we tacitly use this property of the DF as equivalent to its stationarity. If condition 2 above is fulfilled, then from Equation (A2), for any fixed $\boldsymbol{I} \in \mathbb{R}^m$, as a function of $\boldsymbol{z}$, the conditional pdf $\mathcal{F}(\boldsymbol{z}|\boldsymbol{I})$ is proportional to $f(\Psi^{-1}(\boldsymbol{z}, \boldsymbol{I}))$. Thus, one can detect that the DF $f(\boldsymbol{w})$ is stationary, or a cut of a stationary DF, by showing that the *conditional* pdf $\mathcal{F}(\boldsymbol{z}|\boldsymbol{I})$ is constant for $\boldsymbol{z}$ in $\Omega_z(\boldsymbol{I})$, at any fixed $\boldsymbol{I} \in \mathbb{R}^m$. We show now that this is equivalent to our minimum-entropy principle.

A pdf supported on a fixed bounded region of $\mathbb{R}^{6-m}$ is uniform if, and only if, it has maximal entropy. In this case, $\mathcal{F}(\boldsymbol{z}|\boldsymbol{I}) = 1/V_z(\boldsymbol{I})$, where $V_z(\boldsymbol{I})$ is the volume of the maximum allowed support $\Omega_z(\boldsymbol{I})$ of $\mathcal{F}(\boldsymbol{z}|\boldsymbol{I})$, and from Equation (A5),

$$S(\mathcal{F}(\cdot|\boldsymbol{I})) = -\ln[1/V_z(\boldsymbol{I})] = \ln V_z(\boldsymbol{I}).$$

Thus, given a fixed DF f on $\mathbb{R}^6$, the expected value $\mathbb{E}_{\boldsymbol{I}}[S(\mathcal{F}(\cdot|\boldsymbol{I}))]$ in Equation (A4) is bounded from above by

$$\mathbb{E}_{\boldsymbol{I}}[\ln V_z(\boldsymbol{I})] = \int_{-\infty}^{\infty} F(\boldsymbol{I}) \ln V_z(\boldsymbol{I}) d^m\boldsymbol{I}$$

and $\mathbb{E}_{\boldsymbol{I}}[S(\mathcal{F}(\cdot|\boldsymbol{I}))]$ reaches this value when the $\mathcal{F}(\cdot|\boldsymbol{I})$ are uniform in their maximum allowed supports. Thus,

$$\mathbb{E}_{\boldsymbol{I}}[\ln V_z(\boldsymbol{I})] - \mathbb{E}_{\boldsymbol{I}}[S(\mathcal{F}(\cdot|\boldsymbol{I}))] \geqslant 0.$$

Using Equation (A4),

$$\mathbb{E}_{\boldsymbol{I}}[\ln V_z(\boldsymbol{I})] - S(f) + \mathbb{E}[\ln |J_{\Psi^{-1}}|] + S(F(\boldsymbol{I})) \geqslant 0$$
$$\mathbb{E}[\ln \tilde{V}] + S(F(\boldsymbol{I})) \geqslant S(f),$$

where $\tilde{V}(\boldsymbol{z}, \boldsymbol{I}) \doteq |J_{\Psi^{-1}}(\boldsymbol{z}, \boldsymbol{I})| V_z(\boldsymbol{I})$.

Therefore, by construction, the quantity $\mathbb{E}[\ln \tilde{V}] + S(F(\boldsymbol{I}))$ is bounded from below by $S(f)$ and reaches this value if, and only if, all $\mathcal{F}(\cdot|\boldsymbol{I})$ are uniform in their maximum allowed supports. By assumption 2, we have:

$$\mathbb{E}[\ln \tilde{V}] = \mathbb{E}_{\boldsymbol{I}}[\ln g(\boldsymbol{I})],$$

where $g(\boldsymbol{I}) = |J_{\Psi^{-1}}(\boldsymbol{I})| V_z(\boldsymbol{I})$ is the "density of states" at $\boldsymbol{I} \in \mathbb{R}^m$. Similar to $F(\boldsymbol{I})$, the density of states $g(\boldsymbol{I})$ is independent of the particular transformation Ψ in respect to the remaining variables $\boldsymbol{z}$. In fact, we can assume that there is a partial transformation $\tilde{\Psi}$ over the $\boldsymbol{z}$-coordinates such that the total transformation has $|J_{\Psi^{-1}}(\boldsymbol{z}, \boldsymbol{I})| = |J_{\Psi^{-1}}(\boldsymbol{I})|$,

$$g(\boldsymbol{I}) = |J_{\Psi^{-1}}(\boldsymbol{I})| V_z(\boldsymbol{I}) = \int_{\Omega_z(\boldsymbol{I})} |J_{\Psi^{-1}}(\boldsymbol{I})| d^{6-m}\boldsymbol{z} = \int_{\tilde{\Omega}_z(\boldsymbol{I})} |J_{\Psi^{-1}}(\boldsymbol{I})| \frac{|J_{\tilde{\Psi}^{-1}}(\boldsymbol{z}, \boldsymbol{I})|}{|J_{\Psi^{-1}}(\boldsymbol{I})|} d^{6-m}\boldsymbol{z} = \int_{\tilde{\Omega}_z(\boldsymbol{I})} |J_{\tilde{\Psi}^{-1}}(\boldsymbol{z}, \boldsymbol{I})| d^{6-m}\boldsymbol{z},$$

where $\tilde{\Omega}_z(\boldsymbol{I})$ is the maximum allowed support of the remaining variables under the second transformation. Thus, the density of states can be generalized for a transformation of coordinates that does *not* satisfy assumption 2 as

$$g(\boldsymbol{I}) \doteq \int_{\Omega_z(\boldsymbol{I})} |J_{\Psi^{-1}}(\boldsymbol{z}, \boldsymbol{I})| d^{6-m}\boldsymbol{z}. \tag{A6}$$

With this, we can finally relax condition 2. Hence, we proved, under the assumption 1 only, that the quantity

$$\mathbb{E}_{\boldsymbol{I}}[\ln g(\boldsymbol{I})] + S(F(\boldsymbol{I}))$$

is bounded from below by $S(f)$ and reaches this value if, and only if, $f(\Psi^{-1}(\boldsymbol{z}, \boldsymbol{I}))$ is constant for $\boldsymbol{z}$ in $\Omega_z(\boldsymbol{I})$.

Consider now a family $\Psi_{\boldsymbol{p}}$, $\boldsymbol{p} \in P$, of transformations of coordinates in $\mathbb{R}^6$ satisfying assumption 1, where $\boldsymbol{p}$ stands for generic parameters of the potential. We think of $\Psi_{\boldsymbol{p}}$ as the set of transformations leading to integrals of motion evaluated in all trial potentials. For some fixed $m < 6$, define $\boldsymbol{I}_{\boldsymbol{p}}$ by the last m components of $\Psi_{\boldsymbol{p}}$, as above. Suppose, as before, that, for all $\boldsymbol{p} \in P$ and $\boldsymbol{I} \in \mathbb{R}^m$, the maximum allowed support of the pdfs $\mathcal{F}_{\boldsymbol{p}}(\cdot|\boldsymbol{I})$ is some bounded region $\Omega_z(\boldsymbol{p}, \boldsymbol{I})$ of $\mathbb{R}^{6-m}$, which now can also depend on $\boldsymbol{p}$. If, for some $\boldsymbol{p}_0 \in P$, the DF f is stationary (phase-mixed), or a cut of a stationary DF, we can find this particular $\boldsymbol{p}_0$ by minimizing with respect to $\boldsymbol{p}$ the quantity

$$S_{\boldsymbol{I}} \doteq \mathbb{E}_{\boldsymbol{I}}[\ln g_{\boldsymbol{p}}(\boldsymbol{I})] + S(F(\boldsymbol{I})) = -\int F(\boldsymbol{I}) \ln \left[\frac{F(\boldsymbol{I})}{g_{\boldsymbol{p}}(\boldsymbol{I})} \right] d^m\boldsymbol{I}. \tag{A7}$$

We emphasize that $S_{\boldsymbol{I}}$ incorporates the density of states $g_{\boldsymbol{p}}(\boldsymbol{I})$ and thus differs from $S(F(\boldsymbol{I}))$. We now show particular cases in

terms of energy, angular momentum, and actions, making contact with Section 3.

A.1. Spherical and Isotropic Systems

For spherically symmetric systems with isotropic velocities, we use spherical coordinates, in terms of solid angles ω_r and ω_v, with Jacobian determinant $\partial(\boldsymbol{r}, \boldsymbol{v})/\partial(r, v, \omega_r, \omega_v) = r^2v^2$. For a given central potential $\phi_p(r)$, the energy being $E = v^2/2 + \phi_p(r)$, the Jacobian determinant for $(r, E) \rightarrow (r, v)$ is $\partial(r, v)/\partial(r, E) = 1/\sqrt{2(E - \phi_p(r))} = 1/v$. Thus,

$$|J_{\Psi_p}| = \frac{\partial(\boldsymbol{r}, \boldsymbol{v})}{\partial(r, E, \varpi_r, \varpi_v)} = \frac{\partial(\boldsymbol{r}, \boldsymbol{v})}{\partial(r, v, \varpi_r, \varpi_v)} \frac{\partial(r, v)}{\partial(r, E)} = r^2\sqrt{2(E - \phi_p(r))}.$$

Let $r_m(E)$ be the maximum radius for a particle with energy E. From Equation (A6), the density of states at fixed E is:

$$g_p(E) = \int_{(r,\omega_r,\omega_v)\in\Omega_z(E)} r^2\sqrt{2(E - \phi_p(r))}\, dr d\omega_r d\omega_v = (4\pi)^2 \int_0^{r_m(E)} r^2\sqrt{2(E - \phi_p(r))}\, dr.$$

From Equation (A7), our minimum-entropy principle translates into minimizing, with respect to the parameters $\boldsymbol{p}$,

$$S_{\rm E} \doteq -\int F(E) \ln\left[\frac{F(E)}{g_p(E)}\right] dE,$$

where $F(E)$ is the pdf for the energy (see Equation (17)). Note that not only g_p, but also $F(E)$ depends on $\boldsymbol{p}$, via ϕ_p.

A.2. Spherical and Anisotropic Systems

For a spherical system with anisotropic velocity distribution, we let it depend on v_t and v_r, the tangential and radial velocities, respectively. The phase-space coordinate $(\boldsymbol{r}, \boldsymbol{v})$ is a function of r, the solid angle ω_r, v_r, and v_t, as well as a planar angle φ_v referring to the tangent direction of the velocity; i.e., we use cylindrical coordinates for the velocity $\boldsymbol{v}$, with its vertical axis along $\boldsymbol{r}$. For this transformation of coordinates, we have $\partial(\boldsymbol{r}, \boldsymbol{v})/\partial(r, v_r, v_t, \omega_r, \varphi_v) = r^2v_t$. With the angular momentum $L = rv_t$, and $v^2 = v_t^2 + v_r^2$, we have $v_r = \pm\sqrt{2(E - \phi_p(r)) - L^2/r^2}$. Thus, the Jacobian determinant of the transformation $(E, L) \rightarrow (v_r, v_t)$ is $\partial(v_r, v_t)/\partial(E, L) = \mp 1/[r\sqrt{2(E - \phi_p(r)) - L^2/r^2}]$. Hence,

$$|J_{\Psi_p}| = \frac{\partial(\boldsymbol{r}, \boldsymbol{v})}{\partial(r, E, L, \varpi_r, \varphi_v)} = \frac{\partial(\boldsymbol{r}, \boldsymbol{v})}{\partial(r, v_r, v_t, \varpi_r, \varphi_v)} \frac{\partial(v_r, v_t)}{\partial(E, L)} = \mp\frac{L}{\sqrt{2(E - \phi_p(r)) - L^2/r^2}}.$$

From Equation (A6), the density of states in this case is $g_p(E, L) = 8\pi^2 L T_r(E, L)$ (see Equations (23) and (24)). As before, from Equation (A7), the minimum-entropy principle refers to minimizing

$$S_{\rm EL} \doteq -\int F(E, L) \ln\left[\frac{F(E, L)}{g_p(E, L)}\right] dE dL,$$

where $F(E, L)$ is the joint pdf for the energy and angular momentum (see Equation (21)).

A.3. Generic Integrable Potentials: Action Variables

If $\Psi_{\boldsymbol{p}}$ is a canonical transformation, $|J_{\Psi_{\boldsymbol{p}}^{-1}}| = 1$. For instance, if $\Psi_{\boldsymbol{p}}$ refer to action-angle variables, $\boldsymbol{I_p}$ being actions, then, in a full-sky survey, i.e., in the absence of any geometric cuts, from Equation (A6):

$$g_{\boldsymbol{p}}(\boldsymbol{I}) = \tilde{V}_{\boldsymbol{p}}(\boldsymbol{I}) = (2\pi)^3.$$

From Equation (A7), our minimum-entropy principle is equivalent to minimizing

$$S_{\boldsymbol{J}} \doteq -\int F(\boldsymbol{J}) \ln\left[\frac{F(\boldsymbol{J})}{(2\pi)^3}\right] d\boldsymbol{J},$$

where $F(\boldsymbol{J})$ is the joint pdf for the actions (see Equation (27)).

More generally, in the presence of geometrical cuts,

$$g_{\boldsymbol{p}}(\boldsymbol{I}) \rightarrow g_{\boldsymbol{p}}(\boldsymbol{I}) A_{\boldsymbol{p}}(\boldsymbol{I}),$$

where the random variable $0 < A_{\boldsymbol{p}} \leqslant 1$ depends only on $\boldsymbol{I}$ (integrals) and refers to the portion of the remaining variables corresponding to stars lying within the survey footprint, at fixed integral. Hence, in the presence of geometric cuts and when using actions, our minimum-entropy principle is equivalent to minimizing

$$S_{\boldsymbol{J}} = -\int F(\boldsymbol{J}) \ln\left[\frac{F(\boldsymbol{J})}{(2\pi)^3 A_{\boldsymbol{p}}(\boldsymbol{J})}\right] d\boldsymbol{J}. \quad \text{(A8)}$$

Appendix B
Could We Maximize the Entropy in Angle-space?

Since the angle distribution is uniform for a phase-mixed sample, one might try to recover the potential by maximizing an entropy using angles. In Section 2, we motivated our method by connecting the maximum-likelihood principle with a minimum-entropy one. This already suggests *minimizing* an entropy using integrals, as opposed to *maximizing* one using the remaining variables. Since these live in higher dimensions for the cases $f(E)$ and $f(E, L)$, it would not be helpful to use those variables, so for this discussion, we focus on angles and actions, both of which live in $d = 3$. We show why we do not expect a maximum entropy in angle-space to work.

As we demonstrate in Appendix A, writing the DF as $f(\boldsymbol{\theta}, \boldsymbol{J}) = \mathcal{F}(\boldsymbol{\theta}|\boldsymbol{J})F(\boldsymbol{J})$, we get $S(f) = S(F(\boldsymbol{J})) + \mathbb{E}_{\boldsymbol{J}}[S(\mathcal{F}(\cdot|\boldsymbol{J}))]$, where we have set $|J_{\Psi^{-1}}| = 1$ in Equation (A4), $F(\boldsymbol{J})$ is the action's pdf, and $S(\mathcal{F}(\cdot|\boldsymbol{J}))$ is the entropy of the conditional pdfs (Equation (A5)). The maximum value of this last term, achieved in the potential where the sample is phase-mixed, is $\mathbb{E}[\ln V_{\boldsymbol{\theta}}]$, where $V_{\boldsymbol{\theta}}(\boldsymbol{J}) = (2\pi)^3 A(\boldsymbol{J})$ is the volume of the angles' support (density of states). In a full-sky survey, $A(\boldsymbol{J}) = 1$, and $0 < A(\boldsymbol{J}) < 1$ in the presence of geometric cuts. This maximum value depends on the potential and needs to join the optimization. Since we can calculate $V_{\boldsymbol{\theta}}(\boldsymbol{J})$ for each model, and $S(f)$ is invariant,

the correct potential is recovered if, and only if, S_J is minimum (see Equation (A8)).

On the other hand, in trying to constrain the potential by maximizing the entropy in angle-space, we would separate the DF as $f(\boldsymbol{\theta}, \boldsymbol{J}) = \mathcal{F}(\boldsymbol{J}|\boldsymbol{\theta})F(\boldsymbol{\theta})$, which implies

$$S(f) = S(F(\boldsymbol{\theta})) + \mathbb{E}_{\boldsymbol{\theta}}[S(\mathcal{F}(\cdot|\boldsymbol{\theta}))],$$

where now $F(\boldsymbol{\theta})$ is the angles' pdf, and $S(\mathcal{F}(\cdot|\boldsymbol{\theta}))$ is the entropy of the conditional pdfs. Although not easily justified, we could conjecture that this last term is minimized in the potential where the sample is phase-mixed, and that, as before, this depends on the potential and should thus join the optimization. However, in this case, we do not know what this value should be and do not know which exact quantity to maximize.

In principle, one might try simply maximizing the entropy of the marginal pdf, $S(F(\boldsymbol{\theta}))$, but we show two examples suggesting that this would fail. Let us consider an admittedly artificial (1+1)D toy model with DF

$$f(\theta, J) = \frac{1}{\pi}[\delta(\theta \leqslant \pi)\delta(J \leqslant 0) + \delta(\theta > \pi)\delta(J > 0)],$$

where $\delta(\mathcal{P}) = 1$ when $\mathcal{P}$ is true, and $\delta(\mathcal{P}) = 0$ otherwise, and $-1/2 \leqslant J \leqslant 1/2$. In words, for negative actions, half of the angle maximum allowed support $\Omega_z(J) = [0, 2\pi]$ is uniformly distributed, and for positive actions, the other half is. At fixed J, this DF is not constant as a function of θ in $\Omega_z(J)$. Nevertheless, the marginal $F(\theta) = \int f(\theta, J)\, dJ$ is uniform, and $F(\theta)$ has maximum entropy. Thus, maximizing $S(F(\theta))$ generally fails to reject nonstationary DFs.

As a second example, let $\Omega_z(J) = [0, 4\pi|J|]$ for $-1/2 \leqslant J \leqslant 1/2$ and $\Omega_z(J) = [0, 2\pi]$ when $|J| > 1/2$. This choice is to be understood as a "toy model" in the presence of a geometric cut. Define the DF by

$$f(\theta, J) = \frac{1}{\pi}\delta(\theta \in \Omega_z(J))\delta(|J| \leqslant 1/2).$$

This DF $f(\theta, J)$ is now uniform in angles. However, the marginal pdf in this case is not uniform:

$$F(\theta) = \int f(\theta, J) dJ = \frac{1}{\pi}\left(1 - \frac{\theta}{2\pi}\right),$$

and thus, its entropy is not maximum, and maximizing $S(F(\theta))$ also generally fails to detect stationary DFs.

ORCID iDs

Leandro Beraldo e Silva https://orcid.org/0000-0002-0740-1507
Monica Valluri https://orcid.org/0000-0002-6257-2341
Eugene Vasiliev https://orcid.org/0000-0002-5038-9267
Kohei Hattori https://orcid.org/0000-0001-6924-8862
Walter de Siqueira Pedra https://orcid.org/0000-0002-8014-4076
Kathryne J. Daniel https://orcid.org/0000-0003-2594-8052

References

Ajgl, J., & Šimandl, M. 2011, IFAC Proc., 44, 11991
Akaike, H. 1992, Information Theory and an Extension of the Maximum Likelihood Principle (Berlin: Springer), 610
Ao, Z., & Li, J. 2023, Artificial Intelligence, 322, 103954
Bahcall, J. N., & Tremaine, S. 1981, ApJ, 244, 805
Banik, U., Weinberg, M. D., & van den Bosch, F. C. 2022, ApJ, 935, 135
Barbieri, L., Di Cintio, P., Giachetti, G., Simon-Petit, A., & Casetti, L. 2022, MNRAS, 512, 3015
Beaumont, M. A., Zhang, W., & Balding, D. J. 2002, Genet, 162, 2025
Beirlant, J., Dudewicz, E. J., Györfi, L., & Van der Meulen, E. C. 1997, Int. Journal of Mathematical and Statistical Sciences, 6, 17
Beloborodov, A. M., & Levin, Y. 2004, ApJ, 613, 224
Beraldo e Silva, L., de Siqueira Pedra, W., Sodré, L., Perico, E. L. D., & Lima, M. 2017, ApJ, 846, 125
Beraldo e Silva, L., de Siqueira Pedra, W., & Valluri, M. 2019a, ApJ, 872, 20
Beraldo e Silva, L., de Siqueira Pedra, W., Valluri, M., Sodré, M., & Bru, J.-B. 2019b, ApJ, 870, 128
Berrett, T. B., Samworth, R. J., & Yuan, M. 2019, The Annals of Statistics, 47, 288
Besla, G., Kallivayalil, N., Hernquist, L., et al. 2007, ApJ, 668, 949
Biau, G., & Devroye, L. 2015, Lectures on the Nearest Neighbor Method (1st ed.; Berlin: Springer)
Binney, J. 2012, MNRAS, 426, 1324
Binney, J., & Petrou, M. 1985, MNRAS, 214, 449
Binney, J., & Tremaine, S. 2008, Galactic Dynamics: Second Edition, Princeton Series in Astrophysics (Princeton, NJ: Princeton Univ. Press)
Byström, A., Koposov, S. E., Lilleengen, S., et al. 2025, MNRAS, 542, 560
Chang, J. T., & Pollard, D. 1997, Statistica Neerlandica, 51, 287
Charzyńska, A., & Gambin, A. 2015, Entrp, 18, 13
Cooper, A. P., Koposov, S. E., Allende Prieto, C., et al. 2023, ApJ, 947, 37
Correa Magnus, L., & Vasiliev, E. 2022, MNRAS, 511, 2610
Cover, T. M., & Thomas, J. A. 2006, Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing) (New York: Wiley)
Cranmer, K., Brehmer, J., & Louppe, G. 2020, PNAS, 117, 30055
Cui, X.-Q., Zhao, Y.-H., Chu, Y.-Q., et al. 2012, RAA, 12, 1197
De Silva, G. M., Freeman, K. C., Bland-Hawthorn, J., et al. 2015, MNRAS, 449, 2604
Deason, A. J., Erkal, D., Belokurov, V., et al. 2021, MNRAS, 501, 5964
Dehnen, W. 2005, MNRAS, 360, 892
Dobrushin, R. L. 1979, Functional Analysis and Its Applications, 13, 115
Erkal, D., Deason, A. J., Belokurov, V., et al. 2021, MNRAS, 506, 2677
Evans, N. W., & An, J. 2005, MNRAS, 360, 492
Ewart, R. J., Nastac, M. L., & Schekochihin, A. A. 2023, JPlPh, 89, 905890516
Fouvry, J.-B., Hamilton, C., Rozier, S., & Pichon, C. 2021, MNRAS, 508, 2210
Garavito-Camargo, N., Besla, G., Laporte, C. F. P., et al. 2021, ApJ, 919, 109
Green, G. M., Ting, Y.-S., & Kamdar, H. 2023, ApJ, 942, 26
Hahn, C., Vakili, M., Walsh, K., et al. 2017, MNRAS, 469, 2791
Hall, P., & Morton, S. C. 1993, Annals of the Institute of Statistical Mathematics, 45, 69
Han, J., Wang, W., Cole, S., & Frenk, C. S. 2016, MNRAS, 456, 1003
Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Natur, 585, 357
Hattori, K., Valluri, M., & Vasiliev, E. 2021, MNRAS, 508, 5468
Henon, M. 1959, AnAp, 22, 126
Hernquist, L. 1990, ApJ, 356, 359
Ishida, E. E. O., Vitenti, S. D. P., Penna-Lima, M., et al. 2015, A&C, 13, 1
Jiang, B. 2018, in Proc. of the 21st Int. Conf. on Artificial Intelligence and Statistics, Vol. 84, ed. A. Storkey & F. Perez-Cruz (PMLR), 1711, https://proceedings.mlr.press/v84/jiang18a.html
Joe, H. 1989, Annals of the Institute of Statistical Mathematics, 41, 683
Kozachenko, L. F., & Leonenko, N. N. 1987, Probl. Peredachi Inf., 23, 9
Kullback, S. 1968, Information Theory and Statistics (New York: Dover)
Leonenko, N., Pronzato, L., & Savani, V. 2008a, AnSta, 36, 2153
Leonenko, N., Pronzato, L., & Savani, V. 2008b, Tatra Mt. Math. Publ., Smolenice, Slovakia, 39, 265, https://hal.science/hal-00322783
Levin, Y., Pakter, R., Rizzato, F. B., Teles, T. N., & Benetti, F. P. C. 2014, PhR, 535, 1
Li, Z., Han, J., Wang, W., et al. 2025, MNRAS, 538, 1442
Lombardi, D., & Pant, S. 2016, PhRvE, 93, 013310
Lynden-Bell, D. 1967, MNRAS, 136, 101
Magorrian, J. 2014, MNRAS, 437, 2230
Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2017, AJ, 154, 94
Martin, O., Kumar, R., & Lao, J. 2021, Bayesian Modeling and Computation in Python (Boca Raton, FL: CRC Press), https://books.google.com/books?id=0UtSEAAAQBAJ
McMillan, P. J. 2017, MNRAS, 465, 76
McMillan, P. J., & Binney, J. 2012, MNRAS, 419, 2251
McMillan, P. J., & Binney, J. J. 2013, MNRAS, 433, 1411
Modak, S., & Hamilton, C. 2023, MNRAS, 524, 3102
Nastac, M. L., Ewart, R. J., Sengupta, W., et al. 2024, PhRvE, 109, 065210
Nyquist, H. 1928, TAIEE, 47, 617
Peñarrubia, J., Koposov, S. E., & Walker, M. G. 2012, ApJ, 760, 2
Petersen, M. S., & Peñarrubia, J. 2021, NatAs, 5, 251

Plummer, H. C. 1911, MNRAS, 71, 460
Price-Whelan, A. M., Hogg, D. W., Johnston, K. V., et al. 2021, ApJ, 910, 17
Prusti, T., de Bruijne, J. H. J., Brown, A. G. A., et al. 2016, A&A, 595, A1
Rehemtulla, N., Valluri, M., & Vasiliev, E. 2022, MNRAS, 511, 5536
Reino, S., Rossi, E. M., Sanderson, R. E., et al. 2021, MNRAS, 502, 4170
Reino, S., Rossi, E. M., Sanderson, R. E., et al. 2022, MNRAS, 512, 4455
Sanders, J. L., & Binney, J. 2016, MNRAS, 457, 2107
Sanderson, R. E., Helmi, A., & Hogg, D. W. 2015, ApJ, 801, 98
Schälte, Y., Klinger, E., Alamoudi, E., & Hasenauer, J. 2022, JOSS, 7, 4304
Shannon, C. 1949, PIRE, 37, 10
Sharma, S., & Johnston, K. V. 2009, ApJ, 703, 1061
Silverman, B. W. 1986, Density Estimation for Statistics and Data Analysis (London: Chapman and Hall)
Sisson, S., Fan, Y., & Beaumont, M. 2018, Handbook of Approximate Bayesian Computation (Boca Raton, FL: CRC Press)
Sisson, S. A., Fan, Y., & Tanaka, M. M. 2007, PNAS, 104, 1760
Ting, Y.-S., Rix, H.-W., Bovy, J., & van de Ven, G. 2013, MNRAS, 434, 652
Tremaine, S. 2018, MNRAS, 477, 946
Tremaine, S., Henon, M., & Lynden-Bell, D. 1986, MNRAS, 219, 285
Trick, W. H., Bovy, J., & Rix, H.-W. 2016, ApJ, 830, 97
Valluri, M., Chabanier, S., Irsic, V., et al. 2022, arXiv:2203.07491
Vasiliev, E. 2019, MNRAS, 482, 1525
Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, NatMe, 17, 261
Wasserman, L. 2010, All of Statistics : A Concise Course in Statistical Inference (Berlin: Springer)
Watkins, L. L., Evans, N. W., & An, J. H. 2010, MNRAS, 406, 264
Wolsztynski, E., Thierry, E., & Pronzato, L. 2005, SigPr, 85, 937
Zhdankin, V. 2022, PhRvX, 12, 031011
Zhdankin, V. 2023, JPhA, 56, 385002
Zwicky, F. 1933, AcHPh, 6, 110