Inferring CDM substructure with global astrometry beyond the power spectrum

Siddharth Mishra-Sharma
Massachusetts Institute of Technology
The NSF AI Institute for Artificial Intelligence and Fundamental Interactions
New York University
sm8383@nyu.edu

Abstract

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

1 Introduction

Although there exists plenty of evidence for dark matter (DM) on galactic scales and above, the DM distribution on sub-galactic scales is less well-understood and remains an active area of cosmological study. This distribution additionally correlates with and may provide clues as to the underlying particle physics nature of dark matter, highlighting its importance across multiple domains.

While larger dark matter clumps—subhalos—can be detected and characterized through their association with luminous tracers like bound stellar populations, subhalos with smaller masses $\lesssim 10^9\,\mathrm{M}_\odot$ are not generally associated with luminous matter, rendering their characterization challenging. Gravitational effects thus provide one of the few avenues to probe the distribution of these otherwise invisible subhalos. Gravitational lensing *i.e.*, the bending of light from a background source due to a foreground mass, is such an effect and has been proposed in various incarnations as a probe of the distribution of dark subhalos. Strong gravitational lensing, for example, has shown promise in probing substructure in galaxies outside of the Milky Way. *Astrometric lensing* has on the other hand recently emerged as a powerful of the dark matter population in our own Galaxy.

Astrometry refers to the precise measurement of the positions and motions of luminous objects like stars and other galaxies. Gravitational lensing of these background objects by a foreground mass, such as a dark matter subhalo, can imprint a characteristic pattern of motions on the measured kinematics (angular velocities and/or accelerations) of these objects. Ref. [1] introduced several methods to extract this imprint with the aim of characterizing the subhalo population in our Galaxy, including methods based on computing convolutions of the expected induced lensing signal, detecting local kinematic outliers, and computing two-point correlators on the observed astrometric field. Ref. [2] further proposed using the power spectrum of the astrometric field as an observable to extract the properties of a dark matter population.

Astrometric datasets are inherently high-dimensional, consisting of positions and angular velocities and/or accelerations of potentially millions of objects. Especially when the signal consists of the collective imprint of a large number of dark matter objects, characterizing the properties of the population involves *marginalizing* over all possible configurations of subhalos, rendering the

Preprint. Under review.

likelihood intractable and necessitating a reduction to data summaries like the power spectrum. While shown to be effective, such simplification can result in loss of information when the signal is non-Gaussian in nature. Systematic effects such as the existence of large-scale power expressed in the low-dimensional summary domain can further inhibit signal sensitivity.

The dawn of the era of precision astrometry, with the *Gaia* satellite delivering the most precise astrometric dataset to-date and surveys including the Square Kilometer Array and Roman Space Telescope set to further change the game, calls for methods that can extract more information from these datasets than possible using existing techniques. In this paper we propose a method that leverages recent advances in simulation-based inference and neural network architectures in order to characterized the subhalo population in our Galaxy.

2 Model and inference

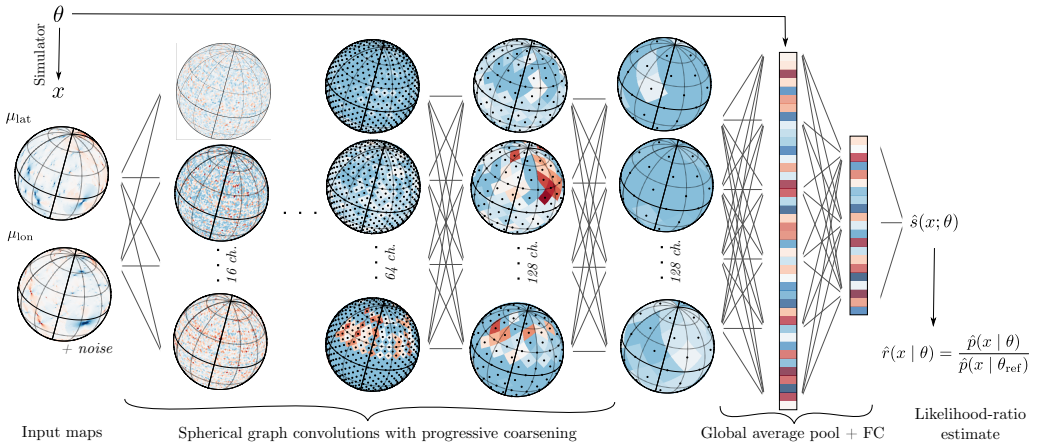

Figure 1: Caption

The forward model We consider a population of Navarro-Frenk-White (NFW) subhalos following a power-law mass function, $\mathrm{d}n/\mathrm{d}m \propto m^{\alpha}$, with $\alpha = -1.9$ as expected if the population is sources from a nearly scale-invariant distribution of primordial fluctuations in the ΛCDM scenario. The subhalo fraction $f_{\rm sub}$, quantifying the expected fraction of the mass of the Milky Way contributed by subhalos in the mass range 10^{-6}–$10^{10}\,\mathrm{M}_{\odot}$, is taken to be the parameter of interest. The fiducial scenario is taken to contain 150 subhalos in expectation between 10^{8}–$10^{10}\,\mathrm{M}_{\odot}$, corresponding to $f_{\rm sub} \simeq 0.2$ and motivated by results from the Aquarius simulation. The spatial distribution of subhalos is again modeled using results from the Aquarius simulation, following Ref. [].

Our dataset consists of the 2-dimensional velocity map of quasi-stellar objects (QSOs), also known as quasars which, owing to their large distances from us, are expected to have small intrinsic proper motions. We assume the velocity maps are discretized on a `HEALPix` grid with resolution parameter `nside=64`, the value in each pixel then quantifying the average velocity of quasars within that pixel. Given a subhalo with velocity $\mathbf{v}_l$, the expected induced velocity in a given pixel is

$$\boldsymbol{\mu}(\mathbf{b}) = 4G_{\rm N}\left\{\frac{M(b)}{b^2}\left[2\hat{\mathbf{b}}\left(\hat{\mathbf{b}}\cdot\mathbf{v}_l\right) - \mathbf{v}_l\right] - \frac{M'(b)}{b}\hat{\mathbf{b}}\left(\hat{\mathbf{b}}\cdot\mathbf{v}_l\right)\right\} \tag{1}$$

where $M(b)$ and $M'(b)$ are the projected mass and its gradient at impact parameter $\mathbf{b}$, given by the vector between the centers of the subhalo and pixel. An example of part of the induced velocity signal map on the celestial sphere, projected along the latitudinal and longitudinal directions, is shown in the leftmost column of Fig. 1.

In order to enable comparison with traditional approaches—which are generally not expected to be sensitive to a CDM subhalo population with next-generation astrometric surveys—we use an optimistic observational configuration corresponding to measure the proper velocities of 10^{9} quasars with expected noise $\sigma_{\mu} = 0.1\,\mu\mathrm{as\,yr}^{-1}$.

The power spectrum approach Ref. [2] introduced an approach for extracting the astrometric signal due to a dark matter population by decomposing the observed map into its *angular power spectrum*. The power spectrum is a summary statistic ubiquitous in astrophysics and cosmology, and quantifies the amount of correlation contained across different scales. In the case of data on a sphere, the basis of spherical harmonics is often used, and the power spectrum then encodes the correlation structure on different multipoles ℓ. The power spectrum effectively captures the *linear* component of the signal and, when the underlying signal is a Gaussian random field (exemplified by the Cosmic Microwave Background), captures all of the relevant information contained in the map(s).

The expected signal in the power spectrum domain can be computed analytically using the formalism described in Ref. [2], which we use here as a comparison point. While effective, reduction of the full astrometric map to its power spectrum results in loss of information; this can be seen from the fact that the signal on the leftmost column of Fig. 1 is far from Gaussian. Furthermore, the existence of systematic, unaccounted-for correlations on large angular scales (as measure on data from the *Gaia* satellite in Refs. []) introduced degeneracies with a putative signal and precludes their usage in the present context. This is especially true when *relative* astrometry is employed, and systematic variations between observed patches of the sky are present. For this reason multipoles $\ell < 10$ were discarded in Ref. [2], degrading the projected sensitivity.

Simulation-based inference with parameterized classifiers Recent advances in machine learning have enabled methods that aim to directly extract information from models defined through high-dimensional simulations; see Ref. [3] for a recent review. Here, we make use of parameterized classifiers in order to approximate the likelihood ratio associated with all-sky astrometric maps containing signatures of dark matter. Given a classifier that can distinguish between samples $x \sim p(x \mid \theta)$ and those from a fixed reference hypothesis $x \sim p(x \mid \theta_{\rm ref})$, the decision function output by the optimal classifier $s(x,\theta) = \frac{p(x|\theta)}{p(x|\theta)+p(x|\theta_{\rm ref})}$ is one-to-one with the likelihood ratio, $r(x \mid \theta) \equiv \frac{p(x|\theta)}{p(x|\theta_{\rm ref})} = \frac{s(x,\theta)}{1-s(x,\theta)}$.

The classifier $d(x,\theta)$ in this case is a neural network that can work directly on the high-dimensional data x, and is parameterized by θ by including it as a feature in the neural network. In order to improve numerical stability and reduce dependence on the fixed reference hypothesis $\theta_{\rm ref}$, we follow Refs. [] and train a classifier to distinguish between data-sample pairs $\{x,\theta\} \sim p(x,\theta)$ and those from the marginal model $\{x,\theta\} \sim p(x)p(\theta)$ using the binary cross-entropy loss as the optimization objective.

Extracting information from high-dimensional datasets Out input dataset consists of a two-component velocity vector uniformly sampled on the sphere following the `HEALPix` representation. We use `DeepSphere`, a graph-based convolutional neural network (CNN) architecture tailored to data sampled on a sphere, and is able to leverage the hierarchical structure of the `HEALPix` representation. In particular, `DeepSphere` efficiently performs convolutions in the spectral domain, using a basis of Chebychev polynomials as the convolutional kernels.

Starting with 2 input channels representing the two orthogonal components of the velocity vector at resolution `nside=64`, we perform a graph convolution operation, increasing the channel dimension to 16, following by a batch normalization, ReLU nonlinearity, and a max pooling operation to coarsen the resolution to `nside=32`. Four more such convolutional layers are employed, increasing the channel by a factor of 2 at each step until a maximum of 128, with the final map having `nside=2`. At this stage, we average over the spatial dimension in order to encourage feature rotational invariance, outputting 128 features to which the parameter of interest $f_{\rm sub}$ is appended. This is passed through a fully-connected network with (1024, 128) hidden units outputting the classifier decision by applying a sigmoidal projection.

10^5 samples from the forward model are produced, and the parameterized likelihood-ratio estimator is trained for 50 epochs using the `AdamW` optimizer with initial learning rate 10^{-3} and weight decay 10^{-5}.

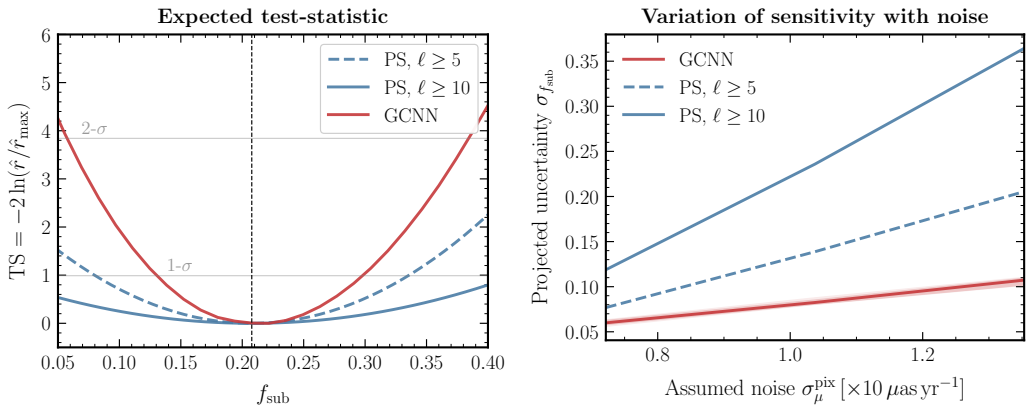

Figure 2: Caption

3 Results on simulated data

The left panel of Fig. 2 shows the expected test statistic (TS), defined as $\mathrm{TS} \equiv -2\ln(\hat{r}/\hat{r}_{\mathrm{max}})$, as a function of substructure fraction f_{sub} evaluated on test maps of the fiducial signal with $f_{\mathrm{sub}} \simeq 0.2$. Corresponding curves using the power spectrum approach are shown in blue, using minimum multipole thresholds of $\ell = 5$ (dashed) and $\ell = 10$ (solid). Thresholds corresponding to 1- and 2-σ discovery are shown as the horizontal grey lines. We see that sensitivity gains of over a factor of ~ 2 can be expected when using the machine learning approach presented here when compared to the traditional power spectrum approach.

The right panel of Fig. 2 shows the scaling of expected sensitivity on substructure fraction f_{sub} with assumed noise per quasar, keeping the number of quasars fixed (red band, showing 1-σ variation over test datasets) compared to the power spectrum approach (blue lines). A far more favorable scaling of the machine learning approach is seen compared to the power spectrum approach, suggesting that it may be disproportionately advantageous in the low signal-to-noise regimes that are generally most relevant for dark matter searches.

4 Conclusions and outlook

In this paper we have leveraged recent advances in simulation-based inference and neural network architectures in order to introduce a method to analyze astrometric datasets over large regions of the sky, with the aim of inferring the signatures

1. Population hasn't been observed
2. Data is expensive
3. Break degeneracies
4. Use of more complex architectures
5. Other scenarios and the acceleration spectrum, smaller FOV

Code and data used for reproducing the results presented in this paper is available at https://github.com/smsharma/sbi-astrometry.

Broader Impact

We acknowledge the importance of considering the ethical implications of scientific research in general, and machine learning research in particular, as well as of placing both the process and output of scientific research in a broader societal context. We do not believe the present work presents any issues in this regard.

Acknowledgments and Disclosure of Funding

SM would like to thank the Center for Computational Astrophysics at the Flatiron Institute for their hospitality while this work was being performed. This work was performed in part at the Aspen Center for Physics, which is supported by National Science Foundation grant PHY-1607611. The participation of SM at the Aspen Center for Physics was supported by the Simons Foundation. SM is supported by the NSF CAREER grant PHY-1554858, NSF grants PHY-1620727 and PHY-1915409, and the Simons Foundation. This work made use of the NYU IT High Performance Computing resources, services, and staff expertise. This research has made use of NASA's Astrophysics Data System. This research made use of the Astropy [4, 5], `HEALPix` [6, 7], IPython [8], Jupyter [9], Matplotlib [10], NumPy [11], Pyro [12], PyTorch [13], SciPy [14], and Seaborn [15] software packages.

References

[1] K. Van Tilburg, A.-M. Taki, and N. Weiner, *"Halometry from Astrometry,"* JCAP **07**, 041 (2018), arXiv:1804.01991 [astro-ph.CO].

[2] S. Mishra-Sharma, K. Van Tilburg, and N. Weiner, *"Power of halometry,"* Phys. Rev. D **102**, 023026 (2020), arXiv:2003.02264 [astro-ph.CO].

[3] K. Cranmer, J. Brehmer, and G. Louppe, *"The frontier of simulation-based inference,"* Proc. Nat. Acad. Sci. **117**, 30055 (2020), arXiv:1911.01429 [stat.ML].

[4] T. P. Robitaille *et al.* (Astropy), *"Astropy: A Community Python Package for Astronomy,"* Astron. Astrophys. **558**, A33 (2013), arXiv:1307.6212 [astro-ph.IM].

[5] A. Price-Whelan *et al.*, *"The Astropy Project: Building an Open-science Project and Status of the v2.0 Core Package,"* Astron. J. **156**, 123 (2018), arXiv:1801.02634.

[6] K. M. Gorski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelman, *"HEALPix - A Framework for high resolution discretization, and fast analysis of data distributed on the sphere,"* Astrophys. J. **622**, 759 (2005), arXiv:astro-ph/0409513 [astro-ph].

[7] A. Zonca, L. Singer, D. Lenz, M. Reinecke, C. Rosset, E. Hivon, and K. Gorski, *"healpy: equal area pixelization and spherical harmonics transforms for data on the sphere in python,"* Journal of Open Source Software **4**, 1298 (2019).

[8] F. Pérez and B. E. Granger, *"IPython: a system for interactive scientific computing,"* Computing in Science and Engineering **9**, 21 (2007).

[9] T. Kluyver *et al.*, *"Jupyter notebooks - a publishing format for reproducible computational workflows,"* in *ELPUB* (2016).

[10] J. D. Hunter, *"Matplotlib: A 2D graphics environment,"* Computing In Science & Engineering **9**, 90 (2007).

[11] C. R. Harris *et al.*, *"Array programming with NumPy,"* Nature **585**, 357 (2020).

[12] E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. A. Szerlip, P. Horsfall, and N. D. Goodman, *"Pyro: Deep Universal Probabilistic Programming,"* J. Mach. Learn. Res. **20**, 28:1 (2019).

[13] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, in *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019) pp. 8024–8035.

[14] P. Virtanen *et al.*, *"SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python,"* Nature Methods **17**, 261 (2020).

[15] M. Waskom *et al.*, "mwaskom/seaborn: v0.8.1 (september 2017)," (2017).