Plausible Adversarial Attacks on Direct Parameter Inference Models in Astrophysics

Benjamin Horowitz
Department of Astrophysical Sciences
Princeton University
Princeton, NJ
bhorowitz@princeton.edu

Peter Melchior
Department of Astrophysical Sciences
Center for Statistics & Machine Learning
Princeton University
Princeton, NJ

Abstract

In this work we explore the possibility of introducing biases in physical parameter inference models from adversarial-type attacks. In particular, we inject small amplitude systematics into an mixture density networks tasked with inferring cosmological parameters from observed data. The systematics are constructed analogously to white-box adversarial attacks. We find that the analysis network can be tricked into spurious detection of new physics in cases where standard cosmological estimators would be insensitive. This calls into question the robustness of such networks and their utility for reliably detecting new physics.

1 Introduction

Within the physical sciences, there has been an explosion of interest in direct parameter inference models, i.e. models which seek to map directly from observed data to underlying physical parameters of interest [14, 11, 6, 16, 15]. Many of these papers make strong claims of superiority of these direct methods compared to standard analysis based on field-level maximum likelihood methods or classical summary statistics. However, the topic of robustness of these models has not been explored rigorously in the literature. Unknown systematics in experiments that are not included in the training data can, and likely will, significantly affect the inferred parameters in ways that are not apparent to users accustomed to standard analysis methods. The results from direct inference techniques may then be misinterpreted as detection of new physics.

In this work we present one particularly strong pathological example inspired by recent work aiming to map from observed cosmological density fields to underlying physical parameters [14, 17, 16, 15]. Standard power-spectra based methods are known to be sub-optimal because of the non-linear evolution of cosmological density due to gravitational evolution, which forces information from two-point statistics into higher order modes. However, these methods are less sensitive to anomalous noise patterns or systematics because they average over the cosmological fields. Neural networks, on the other hand, particularly those utilizing convolutional architectures, can extract non-linear information beyond what is possible with existing classical summary statistics [8]. We wonder how robustly they can perform this extraction.

As shown in [7], adversarial attacks on neural networks can happen in the "physical world", i.e. in natural images processed via a standard camera, without the need to exactly manipulate individual pixels. In the context of the physical sciences this leads to the question whether conceivable physical systematics, not just specifically crafted pathologies, could result in effective adversarial patterns.

In this work, we construct a parameter inference network trained on two dimensional projected dark matter fields. We then construct adversarial attacks via the methods discussed in [7]. We show

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

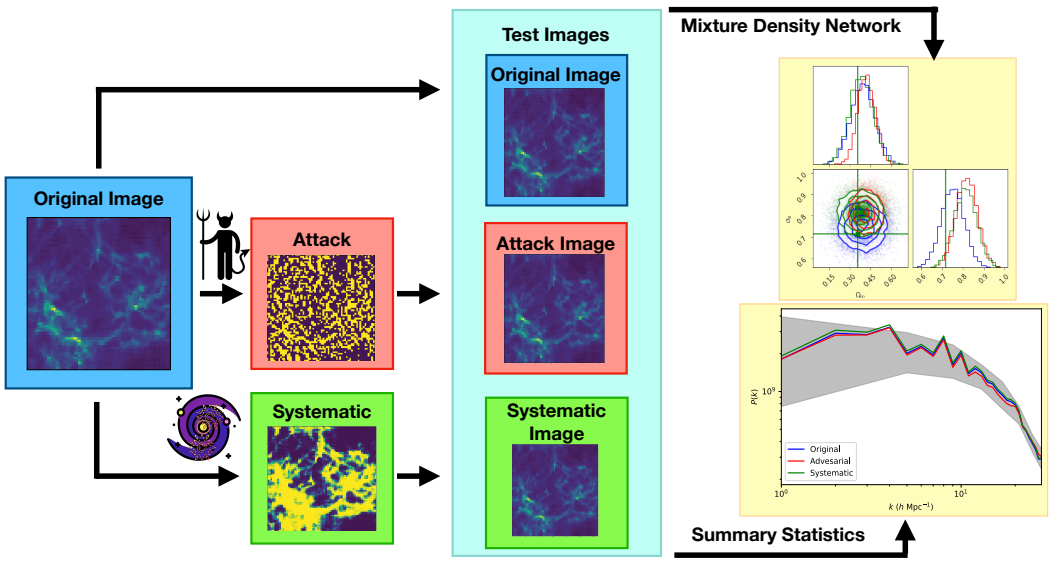

Figure 1: Overview of our workflow to generate and evaluate adversarial-type patterns. From our original image and trained model we can generate an adversarial attack, shown in red, as well as an unknown systematic example, shown in green. Both are added to our original image, leading to nearly imperceptible ($< 0.1\%$) changes to a powerspectrum analysis, however both result in significant shifts to the model's inferred cosmological values at a $2\,\sigma$ level.

that there are classes of reasonable systematics that could exist below the noise level of existing experiments, and which existing analysis techniques are insensitive to.

2 Methodology

2.1 Simulated training data

To simulate idealized astronomical data, we run 10000 small box, particle mesh simulations spanning cosmological parameters $\Omega_M \in [0.05, 0.55]$ and $\sigma_8 \in [0.5, 1.01]$. For simplicity we use a uniform grid of values with spacing of 0.05 for each parameter value. We use FlowPM [10], a Tensorflow [1] GPU-based implementation of FastPM [4]. Other that Ω_M and σ_8, we hold all other cosmological parameters fixed at the Planck 2015 Best Fit Values [12]. For our simulations we use a box size of 128 h^{-1}Mpc side-length with particle resolution of 64^3. These are very coarse simulations by cosmological standards, but are able to capture key differences in the growth of structure caused from variations of the fundamental physics parameters. We use a 70-30 split for training and testing respectively.

2.2 Network architecture

For our model architecture to predict fundamental physics parameters, we will use a convolutional neural network, whose outputs are flattened and passed to a fully-connected mixture density network[2]. We will aim to predict the possible distributions of our target cosmological parameters (Ω_m and σ_8) by predicting the means, μ, standard deviations, σ, and relative weights, π of various Gaussian components, i.e. we assuming a form

$$p(x|y) = \sum_i^m \pi_i \phi(y|\vec{\theta}_i), \tag{1}$$

where ϕ are Gaussian distributions defined by parameters $\vec{\theta}_i = (\vec{\mu_i}, \vec{\sigma}_i)$. For this analysis we set $m = 1$, i.e. one Gaussian component, although we can easily generalize it to many mixture components. We assume a diagonal covariance for our parameters due to ease of optimization. We make use of the

symmetry of our simulated physical system by performing random rotations and translations during training to increase our networks robustness and effective training size. Training was performed on a Tesla V100-PCIE GPU with 32 GB of memory. For more information about our network architecture see our github repository here: `https://github.com/bhorowitz/adversarial-cosmology`.

2.3 Adversarial attack

We present two families of adversarial attacks on our network. The first is the "worst case", representing the smallest possible permutation resulting in the most significant change in our inferred cosmological parameters. To achieve this we use the Basic Iterative Method (BIM) [7], an extension of the Fast Gradient Sign Method [5]. This is an example of white-box attack where the attacker knows the full likelihood function, $J(\theta, x, y)$, with θ being the model's (trained) parameters, x the input image and y the simulated true parameters, respectively. The BIM iteration is given by

$$x_{t+1} = \text{Clip}(x_t + \alpha \times \text{Sign}(\nabla_x J(\theta, x_t, y))), \tag{2}$$

where Clip keeps all values below a value of 1. We run this process for ten steps with α value of 0.01. This method will construct the smallest change to the original image that results in the maximum effect on the inferred parameters.

2.4 Unknown Systematic attack

For a more realistic example we examine a perturbation which scales as a non-linear function of observed density field. This could, for example, reflect an incorrect calibration of a detector or some unforeseen small-scale hydrodynamical effect in galaxy clusters. We parameterize this model as

$$x_{sys} = x + \beta \times \frac{1}{1 + e^{ax+b}} \tag{3}$$

To find an "adversarial" example with this property we use a similar method as BIM but instead of studying the gradient with respect to the underlying field we find it with respect to the parameters of Equation 3. Tuning the parameters by hand, we find $a = 0.25$ and $b = -15.0$ provides a significant shift to the trained network's accuracy with minimal visual change. We choose β such that the total mean squared change induced by our unknown systematic is similar in amplitude to our adversarial attack, finding $\beta = 3.0$.

3 Results

We train a model as described in Section 2.2 using the simulated training data described in 2.1. We stop training after 1000 epochs (approximately one hour), finding suitable accuracy on our test data-set is achieved with minimal marginal improvement in overall loss per additional epoch. We then use the methods described in Sections 2.3 and 2.4 to generate adversarial-type patterns which we add to our test images. We scale our systematic pattern to match the overall change in pixel value of the white-box adversarial pattern. We show this workflow in Figure 1.

For comparison, we also calculate the power spectra of the original, adversarial, and systematically altered fields. This is the most common way to extract cosmological information from observed density field data. We find that the adversarial-type patterns generate changes to the power spectra well below the intrinsic cosmological variability (grey bands in Figure 1), i.e. this method is robust to small scale unknown systematics. Meanwhile our model is, by construction, highly sensitive to this attack and results in significant nonphysical parameter shifts. We show additional examples of this procedure in Figure 2.

Beyond qualitative comparisons, we can compare the resulting distributions based on the average Kullback-Leibler (KL) Divergence. For our test sample, we calculate the KL divergence from the perturbed distribution to the unperturbed distribution (note that the KL Divergence is not symmetric). For our test sample, we find a mean KL divergence of 1.05 (0.98) for the distances adversarial (systematic) attack and the unperturbed example. Meanwhile, the mean symmetrized KL divergence between the adversarial and systematic attacks is 0.67. This indicates our resulting perturbed distributions are on significantly offset from the original distribution on average, while also being similar to each other.

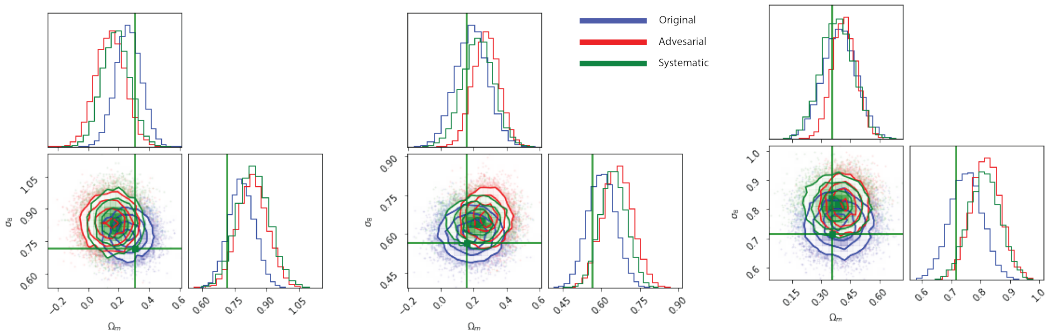

Figure 2: Shown are three random examples from our test data-set, passed through our mixture density network, showing the similarities between our adversarial attack and the mock systematic. While on average the mock systematic results in less of an offset, it still results in a significant bias that could cause a spurious detection. The green circle indicates the simulated truth.

4 Discussion

In this work we have constructed a direct parameter inference model for cosmological density fields, which is closely related to those discussed in the literature [14, 17, 16, 15]. We have shown that these types of neural network estimators are susceptible to adversarial attacks in ways traditional statistics are not. In addition, we have shown that there exists a space of reasonable physical systematics, which, while imperceptible to existing methods, lead to similar biases in cosmological parameters as white-box adversarial attacks. These systematics are closer in distribution, as measured by the KL divergence, to adversarial attacks than to the unperturbed distribution.

We conclude from these tests that there is reason for skepticism about cosmological results drawn from direct parameter inference models. A key feature of these models is that they derive cosmological information from small-scale features in the density fields. Great care is needed to ensure the robustness of such models to all conceivable new systematics, not just to those the community has investigated in the context of traditional analysis methods. Even if doing so leads to reduced constraining power [9, 13]. Going forward, we believe it should be standard practice to test parameter inference models with adversarial examples, or, even better, to inject adversarial examples during training to increase the robustness of the final models.

Limitations of work: In this study, we examined attacks on only one specific network architecture in one context. It is possible other networks would be more or less conducive to this method of attack. For example, neural flow-based models [3] have significantly fewer parameters which could make them more robust to adversarial-type systematics. We also did not explicitly explore examples of adversarial-type attacks to which the power spectrum would be sensitive but neural network models would be not. However, the sensitivity of power spectrum methods is well studied in existing astrophysics literature.

Impact Statement: This work shows potential limitations of a large (and growing) body of work found throughout the physical sciences which attempts to directly map from observed data to underlying physical parameters. We believe our work should encourages authors of such works to explore the robustness of their models, including potentially incorporating adversarial training. It also elucidates this issue for the broader scientific community to encourage well-founded skepticism on any claimed physical discoveries based on this class of models.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng.

TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv e-prints*, page arXiv:1603.04467, Mar. 2016.

[2] C. M. Bishop. Mixture density networks. 1994.

[3] B. Dai and U. Seljak. Translation and rotation equivariant normalizing flow (TRENF) for optimal cosmological analysis. , 516(2):2363–2373, Oct. 2022.

[4] Y. Feng, M.-Y. Chu, U. Seljak, and P. McDonald. FASTPM: a new scheme for fast simulations of dark matter and haloes. , 463(3):2273–2286, Dec. 2016.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and Harnessing Adversarial Examples. *arXiv e-prints*, page arXiv:1412.6572, Dec. 2014.

[6] C. K. Khosa, L. Mars, J. Richards, and V. Sanz. Convolutional neural networks for direct detection of dark matter. *Journal of Physics G Nuclear Physics*, 47(9):095201, Sept. 2020.

[7] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *arXiv e-prints*, page arXiv:1607.02533, July 2016.

[8] A. Lazanu. Extracting cosmological parameters from N-body simulations using machine learning techniques. , 2021(9):039, Sept. 2021.

[9] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii. Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning. *arXiv e-prints*, page arXiv:1704.03976, Apr. 2017.

[10] C. Modi, F. Lanusse, and U. Seljak. FlowPM: Distributed TensorFlow implementation of the FastPM cosmological N-body solver. *Astronomy and Computing*, 37:100505, Oct. 2021.

[11] M. Ntampaka, J. ZuHone, D. Eisenstein, D. Nagai, A. Vikhlinin, L. Hernquist, F. Marinacci, D. Nelson, R. Pakmor, A. Pillepich, P. Torrey, and M. Vogelsberger. A Deep Learning Approach to Galaxy Cluster X-Ray Masses. , 876(1):82, May 2019.

[12] Planck Collaboration. Planck 2015 results. XIII. Cosmological parameters. , 594:A13, Sept. 2016.

[13] A. Raghunathan, S. M. Xie, F. Yang, J. Duchi, and P. Liang. Understanding and Mitigating the Tradeoff Between Robustness and Accuracy. *arXiv e-prints*, page arXiv:2002.10716, Feb. 2020.

[14] S. Ravanbakhsh, J. Oliva, S. Fromenteau, L. C. Price, S. Ho, J. Schneider, and B. Poczos. Estimating Cosmological Parameters from the Dark Matter Distribution. *arXiv e-prints*, page arXiv:1711.02033, Nov. 2017.

[15] H. Shao, F. Villaescusa-Navarro, P. Villanueva-Domingo, R. Teyssier, L. H. Garrison, M. Gatti, D. Inman, Y. Ni, U. P. Steinwandel, M. Kulkarni, E. Visbal, G. L. Bryan, D. Angles-Alcazar, T. Castro, E. Hernandez-Martinez, and K. Dolag. Robust field-level inference with dark matter halos. *arXiv e-prints*, page arXiv:2209.06843, Sept. 2022.

[16] F. Villaescusa-Navarro, J. Ding, S. Genel, S. Tonnesen, V. La Torre, D. N. Spergel, R. Teyssier, Y. Li, C. Heneka, P. Lemos, D. Anglés-Alcázar, D. Nagai, and M. Vogelsberger. Cosmology with One Galaxy? , 929(2):132, Apr. 2022.

[17] F. Villaescusa-Navarro, B. D. Wandelt, D. Anglés-Alcázar, S. Genel, J. Manuel Zorrilla Matilla, S. Ho, and D. N. Spergel. Neural Networks as Optimal Estimators to Marginalize Over Baryonic Effects. , 928(1):44, Mar. 2022.

Checklist:

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]

(c) Did you discuss any potential negative societal impacts of your work? [No] We could not think of any particular negative societal impacts for this physical science focused work.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] We provide these things in an attached github repository: `https://github.com/bhorowitz/adversarial-cosmology`.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Results were not explicitly checked with respect to different random seed, but this was not a focus of this work.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See line 56.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]