Thesis Notes

Evan Saraivanov

October 20, 2022

Contents

Chapter 1

Cosmology

1.1 Introduction

With Hubble's discovery of the expanding universe, there has been great efforts to understand this expansion and the evolution of this expansion (and of the universe as a whole). A related idea, General Relativity has described the importance of the geometry of the universe. There are, in general, three possible geometries:

- A flat geometry which is equivalent to Euclidean space with zero curvature.
- An open geometry with constant negative curvature (Anti-de Sitter space).
- A Closed geometry with constant positive curvature (de Sitter space)

Given the relationship between curvature and the energy-momentum tensor given by the Einstein field equations, it seems reasonable to assume that any non-zero energy means the universe is not flat. Through quantum field theory (QFT) the Casimir effect shows that the vacuum has non-zero energy density, and through astronomical observation of distant galaxies it can be seen that our universe is flat. At first these two observations contradict each other. However, by introducing another term to the Einstein field equations, this discrepency can be resolved.

$$G_{\mu\nu} - \underbrace{\Lambda g_{\mu\nu}}_{\text{new term}} = T_{\mu\nu} \tag{1.1}$$

The constant Λ is called the *cosmological constant* which can absorb the contributions from the vacuum energy, allowing for a flat universe and non-zero vacuum energy density. A natural question to ask is 'what is the source of the cosmological constant?'

1.2 General Relativity

To align with my interest I will write this section from a mathematical perspective and give a physics interpretation after. In this section, I will derive Einstein's field equation. First starting with defining smooth structures on a manifold. Afterwards I move into curvature before using the Bianchi identity on the Riemann tensor to derive the equation governing General Relativity. I will also discuss some properties of Einstein's Field Equation.

1.2.1 Smooth Manifolds

The basic structure of our universe according to general relativity is that the universe is a smooth 4-manifold. This already contains a lot of information so lets break it down. To begin, we start with defining a topological manifold. Given a space X with topology τ, X is an n dimensional topological manifold if

- X is Hausdorff. For each $p, q \in X$ there exists neighborhoods $p \in U$, $q \in V$, $U, V \in \tau$ such that $U \cap V =$.

- X is second countable. There exists a countable basis for the topology τ.
- X is locally Euclidean. For all $U \in \tau$ there exists a homemorphism $\phi : U \to V$ such that $V \subset \mathbb{R}^n$.

The last condition is one that hints at a differentiable structure becuase of known calculus in $\mathbb{R}^n$. To solidify this, consider the pair (U, ϕ) called a chart. Given two charts (U, ϕ) and (V, ψ) with $U \cap V \neq$, we can create the transition map $\psi \circ \phi^{-1} : \phi(U \cap V) \to \psi(U \cap V)$. The two charts are smoothly compatible if $U \cap V =$ or the transition map is a homeomorphism.

1.2.2 Curvature

1.2.3 Einstein's Field Equation

Suppose we have a (semi-) Riemannian manifold M with metric g and tangent bundle TM. An *affine connection* is a map

$$\nabla : \Gamma(TM) \times \Gamma(TM) \to \Gamma(TM) \tag{1.2}$$

$$(X, Y) \mapsto \nabla_X Y \tag{1.3}$$

That is, it parallel transports the vector field Y along the connection ∇ in the direction of vector field X. From this, we can write the affine geodesic equation for a path $\gamma(t)$

$$\nabla_{\dot{\gamma}} \dot{\gamma}(t) = 0 \tag{1.4}$$

Thus, a geodesic is a path such that its tangent vector is parallel translated. Since we observe our world with coordinates, in physics it is often more instructive to work this out in a specific set of coordinates x^μ. Thus this can be written as

$$\ddot{\gamma}^\mu + \Gamma^\mu_{\rho\lambda} \dot{\gamma}^\rho \dot{\gamma}^\lambda \tag{1.5}$$

In a curved spacetime, conservation of energy is written

$$\nabla^\mu T_{\mu\nu} = 0 \tag{1.6}$$

When deriving Einstein's equation, we want to find a divergence-less tensor that depends only on the geometry. The following procedure follows closely one would to find the classical Yang-Mills equation for gauge fields. The Riemann curvature is anti-symmetric in the first two lower indices, so the Riemann tensor is like a GL(TM) valued differential 2 form, and the bianchi identity is

$$d_\nabla R = 0 \tag{1.7}$$

Explicitely writing this becomes

$$\nabla_\alpha R^\alpha_{\beta\gamma\delta} + \nabla_\beta R^\alpha_{\gamma\alpha\delta} + \nabla_\gamma R^\alpha_{\alpha\beta\delta} = 0$$
$$\nabla_\alpha R^\alpha_{\beta\gamma\delta} + \nabla_\beta R^\alpha_{\gamma\alpha\delta} - \nabla_\gamma R^\alpha_{\beta\alpha\delta}$$
$$\nabla_\alpha R^\alpha_{\beta\gamma\delta} + \nabla_\beta R_{\gamma\delta} - \nabla_\gamma R_{\alpha\delta}$$

Multiplying by $g^{\beta\delta}$ and doing some relabelling/contractions of internal indices we find

$$\nabla^\alpha (R_{\gamma\alpha} - \frac{1}{2} g_{\gamma\alpha} R) \equiv \nabla^\alpha G_{\gamma\alpha} = 0 \tag{1.8}$$

There is one more divergencless tensor, the metric tensor $g_{\mu\nu}$, so we can write

$$G_{\mu\nu} - \Lambda g_{\mu\nu} = 8\pi\kappa T_{\mu\nu} \tag{1.9}$$

1.2.4 Gauge Choice

1.2.5 Scalar-Vector-Tensor Decomposition

Our universe, as described above, is a real valued 4 dimensional space, $\mathbb{R}^4$. Suppose that we can separate the universe into a spacial part and a temporal part $\mathbb{R}^4 \mapsto \mathbb{R} \times S$ where S is some 3-manifold. Under such a decomposition, the metric decomposes as (with comoving time as the time coordinate)

$$g = a^2(\tau) \left(g_{00} d\tau d\tau + g_{0i} d\tau dx^i + g_{ij} dx^i dx^j \right) \tag{1.10}$$

The three parts are as follows:

- g_{00} has degrees of freedom (DOF) of a scalar. This is the scalar portion of the decomposition.
- g_{0i} has DOF of a vector. This is the (co)vector portion of the decomposition.
- g_{ij} has DOF of a rank 2 tensor. This is the tensor portion of the decomposition.

The metric, however, is a special case of a symmetric rank 2 tensor. For antisymmetric tensors the only change is that rather than the product $dx^\mu dx^\nu$, we need the exterior product $dx^\mu \wedge dx^\nu$. This means the temporal component/scalar component is 0 for antisymmetric tensors. Since any rank 2 tensor can be decomposed to a symmetric and an antisymmetric part, these two decompositions are sufficient for decomposing any rank 2 tensor.

We can take this even further. Note that a vector can be decomposed into a divergence part and a dual part

$$v^i = g^{ij}\partial_j f + g^{ij} \underbrace{w_{jk} \star (dx^i \wedge dx^j)}_{\equiv \hat{w}_j} \tag{1.11}$$

$$\Rightarrow v^i = (\partial^j f + \hat{w}^j) \tag{1.12}$$

Also, any rank 2 tensor can be written as the sum of a trace and a traceless part.

1.3 ΛCDM: The Standard Model of Cosmology

1.3.1 The FLRW Metric and Dark Energy

In general, metrics allow one to attribute distances between points in a space as $d = g_{\mu\nu}x^\mu x^\nu$. The 'flat' metric for relativity is the *Minkowski metric* given by $\text{diag}(-1, 1, 1, 1)$, however, a metric to describe the expanding universe is given by the *FLRW metric* by $d^2 = t^2 - a^2(t)s^2$ with s^2 the standard Euclidean distance in $\mathbb{R}^3$. What is immediately appeearent from the FLRW metric is that spacial slices of the $d = 4$ spacetime remain curvature free under the expansion of the universe describe by the scale factor $a^2(t)$.

In cosmology, there are other distances which can be more useful than the distance given by the FLRW metric. In the FLRW metric the distance between two points grows in time. We can avoid this by defining the *comoving distance* in which distances remain fixed through time. If we look at a coordinate function x^μ at $t = t_0$, at a later time the coordinate function can be written as $x^\mu \to a(t)x^\mu$, thus by dividing by the scale factor $a(t)$ we can define the comoving coordinates as

$$\chi = \int_{t_0}^{t} \frac{1}{a(t')} dt' \tag{1.13}$$

with the standard Minkowski metric. This can be taken a step further by determining how far light has travelled since $t = 0$

$$\eta = \int_{0}^{t} \frac{1}{a(t')} dt' \tag{1.14}$$

Since we can't see anything beyond this distance, it is often called the *comoving horizon.* There is one last useful distance to define, the *angular distance* which is inferred by the angle subtended by two objects. This relates distances to the geometry discuss in the first section where the measured distance will be the radial distance D

$$D_A = \begin{cases} R & K = 0 \\ R\sin(D/R) & K > 0 \\ R\sinh(D/R) & K < 0 \end{cases} \tag{1.15}$$

When describing the structure of the universe, I will make a few assumptions (which hold up to small perturbations):

- Homogeneity. The cosmology describing the universe does not depend on location.

- Isotropy. The cosmology describing the universe does not depend on location.

These two conditions for what is sometimes referred to as *the cosmological principle.* In general, they don't hold on small scales, however averaging over a sufficiently large distance these assumptions give an accurate description. The isotropy condition means that the universe should have 0 net momentum, and by assumming the universe is smooth the energy-momentum tensor can be written

$$T^{\mu}_{\nu} = \begin{pmatrix} -\mathcal{E} & 0 & 0 & 0 \\ 0 & \mathcal{P} & 0 & 0 \\ 0 & 0 & \mathcal{P} & 0 \\ 0 & 0 & 0 & \mathcal{P} \end{pmatrix} \tag{1.16}$$

The usual conservation law holds

$$\nabla_{\mu} T^{\mu}_{\nu} = 0 \Rightarrow \partial_t \mathcal{E} + \frac{\dot{a}}{a}(3\mathcal{E} + 3\mathcal{P}) = 0 \tag{1.17}$$

We can use the geodesic equation to examine how the energy of the massless particles evolves through time.

If we examine the 00 component of Einstein's Field Equation, the result is the *first Friedman equation* (note that ρ is shorthand for $\sum_i \rho_i$ and $R(G)$ to note that R depends on the geometry of spaceial slices)

$$H^2(a) + \frac{1}{a^2 R^2(G)} = \frac{8\pi G}{3}\rho \tag{1.18}$$

We can interpret the second term on the left (which is spacial curvature) as some density asociated with the curvature ρ_k. If we divide by ρ_{crit} at $z = 0$ / $a = 1$ we find the usual form.

$$\omega + \omega_k = 1 \tag{1.19}$$

The second Friedman equation comes from the trace of Einstein's equation.

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3P) \tag{1.20}$$

1.3.2 informal notes

perfect fluid approximation: treating galaxies as particles of a gas, the particles cluster at small scales and can the particle nature can be ignored. The fluid of galaxies has stress-energy

$$T = (\rho + p) u \otimes u + gp$$

with u the 4-velocity and g the metric tensor and p the pressure and ρ the mass-energy.

Chapter 2

Connecting to Experiments: The CMB and Weak Lensing

2.1 The Cosmic Microwave Background

2.2 Weak Lensing

2.3 Testing ΛCDM

Chapter 3

Tension

As discussed before (where?), standard ΛCDM can be parameterized by 6 parameters, $(A_s, n_s, H_0, \tau, \Omega_b, \Omega_c)$. As usual, this is not the only parameterization of ΛCDM. A_s describes the amplitude of the matter power spectrum. By integrating the matter power spectrum up to a scale of $8h^{-1}$Mpc we can find a parameter denoted σ_8 defined as

$$\sigma_8^2 = \int_0^\infty P(k,r)\left(\frac{3j_1(kr)}{kr}\right)^2 \frac{k^2}{2\pi^2}dk\Bigg|_{r=8} \tag{3.1}$$

which can be described as the amplitude of matter fluctuations at the scale $8h^{-1}$Mpc. In addition, we can add up all of the matter densities to get

$$\Omega_m = \sum_i \Omega_i = 1 - \Omega_k \tag{3.2}$$

which describes the matter density. As it turn out, measurements of the early universe (CMB and high redshift observation) and late universe (galaxy observations and low redshift observation) disagree on measurements of H_0 and $S_8 = \sigma_8\sqrt{\Omega_m/0.3}$. These individual parameters are in tension between experiments, but the driving questions of this project is: how much tension is there between measurements of the entire ΛCDM parameters? Are there models that can resolve this tension? What is the best/most reliable way to measure the tension?

3.1 Tension Metrics

In a previous DES paper, the tension metrics are the following:

1. Bayesian evidence ratio given by

$$R = \frac{\mathcal{E}_{AB}}{\mathcal{E}_A\mathcal{E}_B} \tag{3.3}$$

   Where A and B are data sets. This method can be written many ways using bayes theorem, which will make it appearent that this metric depends heavily on the prior volume.

2. Bayesian suspiciousness. This metric attempts to remove the dependence on the prior volume by defining the suspiciousness as

$$\log S = \log R - \log I \tag{3.4}$$

   Of particular interest here is the new value I which is the information ratio, which is defined in terms of the KL divergence.

$$\log I = \mathcal{D}_A + \mathcal{D}_B - \mathcal{D}_{AB} \tag{3.5}$$

$$\mathcal{D} = \int \mathcal{P}\log\left(\frac{\mathcal{P}}{\Pi}\right) \tag{3.6}$$

3. The method of parameter difference Δ and n_σ.

4. Parameter difference in update form. Suppose you have two data sets A and B. The idea is to look at the difference in mean and covariance between data set A and data set $A + B$.

$$Q_{\text{UDM}} = (\mu_A - \mu_{A+B})^T (C_A - C_{A+B})^{-1} (\mu_A - \mu_{A+B}) \tag{3.7}$$

   Q_{UDM} will be χ^2 distributed with $\text{rank}(C_A - C_{A+B})$ degrees of freedom.

5. Goodness of fit degredation. This is similar to the previous, where it looks at how the goodness of fit of the model in data set A degrades after adding data set B. We have

$$Q_{\text{DMAP}} = 2\mathcal{L}_A(\hat{\theta}_A) + 2\mathcal{L}_B(\hat{\theta}_B) - 2\mathcal{L}_{A+B}(\hat{\theta}_{A+B}) \tag{3.8}$$

   with $\hat{\theta}$ being the parameter vector that maximizes the posterior, the maximum a posteriori. Again Q_{DMAP} is χ^2 distributed.

Metric 1: Parameter Difference

The idea behind this metric is simple: if two data sets largely agree, there difference posterior will be centered at 0, so the integral will be close to 0. To actually compute the integral we use the normalizing flow to learn the posterior and perform MCMC inegration. The integration error is determined using the Clopper-Pearson interval on the binomial distribution.

Metric 2: Eigentension

This metric is interesting. We start by diagonalizing the covariance matrix on one of our data sets. Then we take the ratio of the variance in the prior to the variance in the posterior and apply an ad hoc cut to determine which eigenmodes are well-measured. The idea is that we should not include poorly measured eigenmodes in our tension analysis because the difference is dominated by the prior rather than the data itself. Lastly we project the other data set onto the eigenmodes and perform the parameter difference metric on only the well-measured eigenmodes.

(here is a good place to compare with metric 1)

Metric 3: Parameter Difference in Update Form

As discussed above, we can compute the parameter Q_{UDM} by

$$Q_{\text{UDM}} = (\mu_A - \mu_{A+B})^T (C_A - C_{A+B})^{-1} (\mu_A - \mu_{A+B}) \tag{3.9}$$

The difference of means is precisely the mean of the parameter difference distribution, and we are using the covariance $C_A + C_{A+B}$. Thus it is clear Q_{UDM} is χ^2 distributed with degrees of freedom given by $\text{rank}(C_A - C_{A+B})$. It is clear, however, that this metric relies on the parameter difference to be gaussian distributed because of its reliance on Q_{UDM} being χ^2 distributed. Despite this, we proceed anyway. With proper calibration this metric can be useful even for non-gaussian posteriors.

Metric 4: Goodness of Fit Degradation

Interpreting the Results

Given some probability P of a parameter shift, the following formula can give you the number of standard deviations if the probability shift comes from a gaussian distribution

$$n_\sigma = \sqrt{2}\text{Erf}^{-1}(P) \tag{3.10}$$

I have a notebook using two unit gaussian priors separated by a distance a. This example can be computed analytically.

$$\begin{aligned}
\mathcal{P}(\Delta\theta) &= \frac{1}{2\pi} \int\limits_{-\infty}^{\infty} e^{-\theta^2/2} e^{-(\theta-\Delta\theta)^2/2} d\theta \\
&= \frac{1}{2\pi} \cdot \sqrt{\pi} e^{-(\Delta\theta)^2/4} \\
&= \frac{1}{\sqrt{4\pi}} e^{-(\Delta\theta)^2/4}
\end{aligned} \tag{3.11}$$

The parameter difference posterior is a gaussian with standard deviation $\sqrt{2}$. The separation is fixed by a, hence the shift is $\mathcal{P}(a)$. Hence the shift probability is

$$\Delta = \int\limits_{-a}^{a} e^{-(\Delta\theta)^2/4} d\Delta\theta \tag{3.12}$$

Lets use the example $a = 2$. Then $n_\sigma = 2/\sqrt{2} = \sqrt{2}$. Using this we can work backwards to find Δ from a z-table to find $\Delta = 0.9207 - 0.0793 = 0.8414$.

3.1.1 DES v Planck Results

3.1.2 Building from previous results

3.2 Computing techniques

3.2.1 MCMC

Interestingly, MCMC algorithms have heavy analogies with statistical mechanics which are useful to demonstrate the concept. To examine this, lets first define what a Markov Chain is.

Definition 1. *A sequence $X_1, \ldots, X_n$ of random elements is a Markov Chain if the conditional distribution X_{n+1} depends only on X_n. The set in which X_i take values is called the state space of the chain.*

The Metropolis-Hastings Algorithm

Suppose we want to sample from a distribution $p(x)$. $p(x)$ can be high dimensional and is generally difficult to calculate (evidence is hard to compute since it requires integration over the entire parameter space). The goal is to use Markov Chains to sample from $p(x)$ without needing to compute the evidence. This will be represented as a path through state space until the chain reaches a stable point (stationary state).

We start with a proposal distribution $g(x_n)$. Sample from the proposal distribution to find the next state x_{n+1} with probability $g(x_{n+1}|x_n)$. This transition from state n to state $n+1$ must follow the *detailed balance condition*

$$p(x_n)g(x_{n+1}|x_n)A(x_n \to x_{n+1}) = p(x_{n+1})g(x_n|x_{n+1})A(x_{n+1} \to x_n) \tag{3.13}$$

where A is an *acceptance probability* which I will define more precisely later. Using Bayes' Theorem on $p(x)$, the evidence cancels out on each side, and thus the detailed balance condition can be simplified to only rely on the likelihood and prior of $p(x)$, which I will denote π and $\mathcal{L}$.

$$\pi(x_n)\mathcal{L}(x_n)g(x_{n+1}|x_n)A(x_n \to x_{n+1}) = \pi(x_{n+1})\mathcal{L}(x_{n+1})g(x_n|x_{n+1})A(x_{n+1} \to x_n) \tag{3.14}$$

$$\Rightarrow \frac{A(x_n \to x_{n+1})}{A(x_{n+1} \to x_n)} = \frac{\pi(x_{n+1})\mathcal{L}(x_{n+1})g(x_n|x_{n+1})}{\pi(x_n)\mathcal{L}(x_n)g(x_{n+1}|x_n)} \equiv R_{n,n+1} \tag{3.15}$$

This allows us to define the acceptance probability as

$$A(x_n \to x_{n+1}) = \min(1, R_{n,n+1}) \tag{3.16}$$

This probability is used to determine whether the chain moves to x_{n+1} or stays at x_n. The chain converges when it reaches a stationary state.

There are a few properties that can be observed for this algorithm:

- Having an asymmetrical proposal $g(x)$ can allow for faster convergence of the chain.
- The initial sampling may not accurately reflect samples for $p(x)$. This is regarded as the 'burn-in' and is generally discarded from the samples.
- MCMC Sampling loses sampling power for multi-modal distributions.

3.2.2 Data Emulators

Traditional methods of computing likelihoods (e.g. COSMOSIS) require an immense number of CPU hours, and since we need tens of thousands of chains, we cannot rely on these traditional methods. To accelerate likelihood computation, and thus the MCMC process, we employ emulators which are neural networks that map the cosmological parameters to data vectors. The map is highly non-linear and thus a straightforward analytic mapping is not known. The neural network architecture used is similar for each experiment.

LSST Emulator

Unfourtunatly, we do not have access to the training samples for this emulator. We can however approximate the training region. First, we must define what tempered MCMC means. As discuss in the previous section, an MCMC decides wheather to accept or reject a point based on the detailed balance condition. In our case, we want to ensure the training samples cover a wide range in the parameter space, so we can modify the posterior by raising it to a power T called the tempering factor. The detailed balance condition becomes

$$\frac{A(x_n \to x_{n+1})}{A(x_{n+1} \to x_n)} = \left[\frac{\pi(x_{n+1})\mathcal{L}(x_{n+1})}{\pi(x_n)\mathcal{L}(x_n)}\right]^T \frac{g(x_n|x_{n+1})}{g(x_{n+1}|x_n)} \tag{3.17}$$

It is believed the samples were generated via MCMC with a posterior tempering of 0.5. This gives us the following distribution

Planck C_ℓ Emulator

Fortunately the training samples for COSMOPOWER are available for download from google drive.

3.2.3 Normalizing Flows

The method of normalizing flows (MAF) implemented here uses Masked Autoencoders (MADE) to construct the flow. Suppose we have an input to the flow x_i. The output of the map is $y_i = \mu(x_{1:i-1}) + \sigma(x_{1:i-1})x_i$. The μ and σ are found using neural networks which recieve masked inputs $x_{1:i-1} = (x_1, \ldots, x_{i-1}, 0, \ldots, 0)$. Since the input only depends on the first $i-1$ inputs, the normalizing flow is *autoregressive* and the Jacobian is triangular.

The implementation in tensorflow uses *bijectors* which implements a local diffeomorphism between a manifold M and a target manifold N (which are our parameter spaces), i.e. $\phi : M \to N$ such that ϕ is differentiable and injective. In tensorflow it has three operations, Forward, Inverse, and log_deg_jacobian, which are exactly the three we want. By constructing a bijector for each masked input, the full normalizing map can be constructed.

An intuitive way to think of normalizing flow is as a reparameterization. Consider our posterior as a manifold X. The manifold is constructed via the parameterization with charts $\phi_i : X \to U_i \subset \mathbb{R}^n$ for i in indexing set I. The charts determine the atlas $A = \bigsqcup_{i\in I}(U_i, \phi_i)$. The goal is to determine a new parameterization for X given by the charts $\psi_j : X \to V_j \subset \mathbb{R}^n$ in which samples in X are gaussian distributed. Let $Y \subset X$ be given by $Y = \phi_i^{-1}(U_i) \cap psi_j^{-1}(V_j)$. Then the transition function is given by $\tau : \phi_i(Y) \to \psi_j(Y)$. To ensure our parameterizations are diffeomorphic, τ is required to be differentiable and its inverse differentiable (in fact, we are working with smooth reparameterizations so the transition function and its inverse should be smooth).

The next question one may have is 'what do flows have to do with this?' Above, the concept of normalizing flows is described as a reparameterization, but there is another good description involving flows. Again, let X be our paramter space, the graph $(\theta, \mathcal{P}(\theta))$, and TX the tangent bundle of X. Each MAF can be though of as a finite sequence of vector fields $v_i : X \to TX$. Each point is moved along its integral curve determined by each v_i. What is great about this definition is that it immediately implies an MAF is a diffeomorphism.

Proposition 1. *Suppose $v : X \to TX$ is a vector field on X and $x \in X$. Then for any $\epsilon > 0$ there exists a unique function $f : B[-\epsilon, \epsilon] \to X$ such that f is an integral curve of with infinitessimal generator v and with $f(0) = x$.*

Proof. Of course, given a vector $v_x : U \to T_xX$ with $x \in U$, for it to be in the tangent space T_pX it must have a curve with $f(0) = x$ and $f'(0) = v_x$. This follows from the definition of T_xX.

Suppose there are two such curves, α and β both mapping from $I = (-\epsilon, \epsilon)$ to X and with $\alpha(0) = x = \beta(0)$ and both with $\alpha'(0) = v_x = \beta'(0)$. Then they both solve the system of autonomous ODE's for $i = 1, \ldots, n$

$$\begin{aligned} \dot{\gamma}^i(t) &= v^i(\gamma(t)) \\ \gamma^i(0) &= c^i \end{aligned} \tag{3.18}$$

This can be solved by $\gamma^i(t) = \int\limits_{U \subset X} v^i(\gamma(t))\text{vol.}$. Since v is a tangent vector field, there exists a curve ϕ with $\phi' = v$, so the solution is $\gamma^i(t) = \phi^i(t) + C^i$. The integration constants are fixed by the initial condition $C^i = -c^i$, hence any two functions satisfying this system of ODE's, including α and β, are equivalent for all $t \in I$. □

Autoregressive models have serval pitfalls that we would like to avoid in this project. The main one is that autoregressive models are sensitive to the order of the input, however our result does not have any dependence.

To give a motivation, I will begin with a 2 dimensional example. In this example I used COBAYA to generate samples from two distributions. One is a pure gaussian centered at $(1/2, 1/2)$ with covariance $0.005I$ (I is the identity matrix). The other is a circle of radius 1 with points generated by a gaussian based on its distance from the circle. In other words, it is a gaussian centered at $x^2 + y^2$ with a mean of $x^2 + y^2 = 1$ and a standard deviation 0.02.

The first step is to find the parater difference distribution. Wrap the samples in the shorter chain until it is the same length as the longer chain, then take the difference of the two chains. In this example we get the following.

Now we can do the normalizing flow. As a general rule-of-thumb, the network will consist $2d$ MAFs each of 2 hidden layers with $2d$ hidden units each, with d the dimension of the distribution. This will generally give an expressive model without introducing error from the inability to train all of the models parameters in a reasonable time. These parameters are tunable to whatever the user will decide, and some experimentation with these parameters is needed to ensure the best results. In addition, we allow the parameters to be arbitrarily permuted between each MAF since the resulting probability of shift should be independent of parameterization.