Dissertation

submitted to the

Combined Faculty of Mathematics, Engineering and Natural Sciences
of Heidelberg University, Germany

for the degree of

Doctor of Natural Sciences

Put forward by

Lennart Kai Röver

born in Lörrach, Germany

Oral examination: 22.01.2025

Statistical Inference in Cosmology

From Parameters to Learnable Functions

Referees: Prof. Dr. Björn Malte Schäfer
Prof. Dr. Tristan Bereau

Abstract

Inference tasks on non-Gaussian posterior distributions are commonly approached using Markov chain Monte Carlo. We draw an analogy to canonical partition functions defined as Laplace transforms of the Bayesian likelihood and prior. This allows to derive analytical expressions for cumulants of the posterior. At second order, we recover the conventional Fisher matrix formalism. We find a closed formula for cumulants of weakly non-Gaussian posteriors. Additionally, we use this formalism to construct physically motivated convergence criteria with clearly defined target values based on virialization, equipartition, and thermalization of the Markov chain. We successfully validate these approaches using a dark energy model applied to supernova data. To speed up forward simulation we use physics-informed neural networks (PINNs). They provide fast and accurate predictions of the luminosity distance for a given choice of parameters. Using the same architecture we perform a model-independent, parameter-free reconstruction of the Hubble function. The PINN uncertainties are quantified using a heteroscedastic loss and repulsive ensembles. Continuing in the vein of fast simulations, we construct the parallelized inflation solver Parallizis, based on the Madelung transformed perturbation equations. It provides a forward simulation from arbitrary inflaton potentials to the primordial power spectrum, while allowing for GPU parallelization.

Zusammenfassung

Das Markov-Chain-Monte-Carlo-Verfahren ist eine übliche Methode zur Inferenz auf nicht-Gaußschen A-posteriori-Verteilungen. Markov-Ketten können als Random Walk in einem, durch eine statistische Zustandssumme bestimmten, thermischen System verstanden werden. Hierbei ist die Zustandssumme als Laplace-Transformation der bayesschen Likelihood und A-priori-Verteilung definiert. Diese kann als kumulantenerzeugende Funktion genutzt werden. Aus den Kumulanten erster und zweiter Ordnung lässt sich wiederum der Fisher-Matrix-Formalismus herleiten. Darüber hinaus wird eine geschlossene Formel für Kumulanten schwach nicht-Gaußscher A-posteriori-Verteilungen konstruiert. Die so definierten Zustandssummen werden in der Folge genutzt, um, basierend auf der Virialisierung, Äquipartition und Thermalisierung von Markov-Ketten, Konvergenzkriterien mit klar definierten Zielwerten zu entwickeln. Anschließend werden Supernova Daten verwendet, um diese Ansätze erfolgreich auf ein Modell Dunkler Energie anzuwenden. Dabei wird die Konvergenz der untersuchten Markov-Ketten durch den Einsatz Physik-informierter neuronaler Netze (PINNs) beschleunigt, welche schnelle und präzise Vorhersagen der Leuchtkraftentfernung liefern. Diese Architektur wird verwendet, um eine modellunabhängige, parameterfreie Rekonstruktion der Hubble-Funktion zu erstellen. Hierbei werden die Unsicherheiten über heteroskedastische Verlustfunktionen und repulsive Ensembles quantifiziert. Ferner wird ausgehend von den Madelung-transformierten Mukhanov-Sasaki-Gleichungen eine GPU-parallelisierte Inflationssimulation vorgestellt. Diese bestimmt die primordialen Fluktuationen nach der Inflation ausgehend von einem beliebigen Inflatonpotential.

Contents

Preface

The research presented in this thesis covers the following four publications:

Röver et al. [2023a] Lennart Röver, Lea Carlotta Bartels and Björn Malte Schäfer,
Partition function approach to non-Gaussian likelihoods: formalism and expansions for weakly non-Gaussian cosmological inference,
Monthly Notices of the Royal Astronomical Society, Volume 523, Issue 2, (2023), arXiv:2210.03138 [astro-ph]

Röver et al. [2023b] Lennart Röver, Heinrich von Campe, Maximilian Philipp Herzog, Rebecca Maria Kuntz and Björn Malte Schäfer,
Partition function approach to non-Gaussian likelihoods: physically motivated convergence criteria for Markov chains,
Monthly Notices of the Royal Astronomical Society, Volume 526, Issue 1, (2023), arXiv:2305.07061 [asro-ph]

Röver et al. [2024] Lennart Röver, Björn Malte Schäfer, Tilman Plehn,
PINNferring the Hubble Function with Uncertainties,
to be published in SciPost, arXiv:2403.13899 [astro-ph]

Röver et al. [2024] Lennart Röver, Björn Malte Schäfer, Tilman Plehn,
Parallizis: Parallelized differentiable inflation solver,
in preparation.

Additional publications that the author participated in but was not first author of are:

Modak et al. [2023] Tanmoy Modak, Lennart Röver, Björn Malte Schäfer, Benedikt Schosser, Tilman Plehn,
Cornering Extended Starobinsky Inflation with CMB and SKA,
SciPost Phys. 15, 047 (2023), arXiv:2210.05698 [astro-ph]

Kuntz et al. [2023] Rebecca Maria Kuntz, Maximilian Philipp Herzog, Heinrich von Campe, Lennart Röver, Björn Malte Schäfer,
Partition function approach to non-Gaussian likelihoods: partitions for the inference of functions and the Fisher-functional,
Monthly Notices of the Royal Astronomical Society, Volume 527, Issue 3, (2024), arXiv:2306.17224 [astro-ph]

Herzog et al. [2024] Maximilian Philipp Herzog, Heinrich von Campe, Rebecca Maria Kuntz, Lennart Röver, Björn Malte Schäfer,
Partition function approach to non-Gaussian likelihoods: macrocanonical partitions and replicating Markov-chains,
The Open Journal of Astrophysics, Volume 7, Oct 25 2024, arXiv:2311.16218 [astro-ph]

During this time, the author of this thesis was a doctoral student at the Astronomisches Rechen-Institut (ARI) of Heidelberg University.

1 Introduction

The inflationary paradigm, first suggested by [Sato, 1981, Starobinsky, 1980, Guth, 1981], introduces exponentially accelerated expansion in the early Universe to cure the cosmological standard model ΛCDM of the horizon and flatness problems. This period of exponentially accelerated expansion is used to explain the experimentally observed, almost perfect isotropy of the Cosmic Microwave Background (CMB) [Bennett et al., 1992]. In addition, quantum fluctuation during this rapid expansion can be understood as the seeds of structure formation through the perturbation equations formulated in [Mukhanov et al., 1992, Maldacena, 2003].

The curvature fluctuations seeded in the early universe are later observed through temperature and polarization anisotropies in the CMB, [Smoot et al., 1992, Bennett et al., 1996, Akrami et al., 2020]. Additionally, they source density fluctuations in the matter power spectrum. These can in turn be observed through the galaxy power spectrum [Almeida et al., 2023]. Future probes of the primordial power spectrum include new CMB experiments such as LiteBIRD [Allys et al., 2023], CMB-S4 [Abazajian et al., 2022] and galaxy surveys such as Euclid [Mellier et al., 2024]. Additionally, HI intensity mapping offers a new avenue to measure the matter power spectrum in a larger redshift range [Bacon et al., 2020].

At redshifts around and below one, observation of type Ia supernovae provide evidence for another period of accelerated expansion [Riess et al., 1998, Perlmutter and et al., 2003, Perlmutter et al., 1999]. It is driven by dark energy with equation of state $w < -\frac{1}{3}$. While the cosmological constant in ΛCDM plays the role of dark energy with $w = -1$, there are compelling arguments for evolving dark energy [Wetterich, 1988, Ratra and Peebles, 1988, Linder, 2008, Tsujikawa, 2013, Mortonson et al., 2013]. Numerical applications often work with constant or linearly evolving equations of state [Chevallier and Polarski, 2001, Linder, 2003].

The dark energy equation of state can be reconstructed from measurements of the Hubble function $H(a)$. We focus on the supernova type Ia catalogs Union2.1 [Suzuki et al., 2012, Amanullah et al., 2010, Kowalski et al., 2008] and Pantheon+ [Scolnic et al., 2022] as redshift dependent probes of the Hubble function.

We develop methods to explore these two periods of accelerated expansion on supernova type Ia and CMB data. The first part of this thesis constructs a partition function approach to gain a physical intuition into the behavior of inference tasks focusing on Markov chain Monte Carlo (MCMC). To that end, chapter 2 provides an introduction to probability theory, sampling and Markov chains. Following that, chapter 3 explores partition functions constructed from the evidence of an inference task. Section 3.1.2 uses the partition function to generate cumulants of the posterior distribution. When a Gaussian approximation of the posterior is permissible, the Fisher formalism [Tegmark et al., 1997], based on the first two cumulants, fully captures its shape. This approach is applied throughout cosmology [Bassett et al., 2009, 2011, Coe, 2009, Elsner and Wandelt, 2012, Refregier et al., 2011, Amara and Kitching, 2011, Raveri et al., 2016]. For

weakly non-Gaussian posteriors the formalism can be extended, often using higher-order cumulants [Wolz et al., 2012, Giesel et al., 2021, Schäfer and Reischke, 2016, Sellentin et al., 2014]. For this type of distribution we derive an analytical approximation to cumulants of any order.

In general, cosmological posteriors are non-Gaussian and can be high dimensional. This behavior is captured by Markov chain Monte Carlo which has become an important tool for inference in cosmology starting with [Lewis and Bridle, 2002]. The samples of the Markov chain become representative of the underlying distribution only after burn-in [Roberts and Rosenthal, 2001, Tierney, 1994]. Their convergence is often quantified using the Gelman-Rubin criterion [Gelman and Rubin, 1992, Vats and Knudson, 2021]. Sampling processes in an MCMC algorithm can be understood as a random walk in the thermodynamic system described by the partition function based on the evidence. By constructing a partition function for samples from the Hamilton Monte Carlo algorithm [Duane et al., 1987, Neal, 2012] section 3.3 defines convergence criteria based on virialization, equipartition and thermal equilibrium.

In the second part of the thesis, chapter 4 gives a more thorough introduction to the physics of inflation and a brief introduction to the different data sets used throughout the thesis. In addition, it contains a brief introduction to neural networks, uncertainty estimation and physics-informed neural networks (PINNs) [Raissi et al., 2017, Piscopo et al., 2019, Araz et al., 2021, Li et al., 2021, Cuomo et al., 2022, Hao et al., 2022].

Continuing with the theme of inference, chapter 5 uses PINNs to reconstruct the Hubble function from supernova data. Section 5.2 describes an emulator for the forward simulation based on the differential equation governing the evolution of the luminosity distance. This emulator is used in section 3.3.5 to speed up the forward simulation needed to infer parameters using an MCMC approach. Additionally, section 5.3 introduces a method to find a neural network based, redshift-dependent reconstruction of the Hubble function. In this approach, the Hubble function is represented using a fully connected network equipped with uncertainty estimates, based on heteroscedastic loss functions [Le et al., 2005, Gal, 2016] and repulsive ensembles [D'Angelo and Fortuin, 2021]. The Hubble function and its uncertainties are reconstructed based on supernova type Ia distance moduli.

Cosmological inference of the primordial potential using MCMC methods requires fast simulations of either the angular power spectra in the CMB or the matter power spectrum. In current CMB experiments [Akrami et al., 2020] this task is performed using Boltzmann solvers such as CLASS [Blas et al., 2011] and CAMB [Lewis et al., 2000]. Chapter 6 introduces the GPU parallelized inflation solver PARALLIZIS based on the differential equation solver implemented in TORCHDIFFEQ [Chen et al., 2018]. Section 6.1 provides the theoretical groundwork for the differential equations used in the solver, identifying a constant in the perturbation equation for each mode. The implementation of the parallelized inflation solver is described in section 6.2. The last section in this chapter 6.3 describes an emulator approach to connect primordial power spectra to observations in cosmology.

Part I

Partition Functions in Inference

2 Probability Theory and Sampling

This chapter provides a rough introduction to the aspects of Bayesian statistics needed as a foundation for [Röver et al., 2023a,b]. More complete introductions can be found in [Sivia and Skilling, 2006] and [MacKay, 2003].

Throughout this chapter parameter tuples θ^μ and data tuples y^i are denoted as vectors with contravariant indices. Latin indices indicate objects in data space, while Greek indices denote objects in parameter space. With this convention, covariances are contravariant tensors. For example the data covariance is $C^{ij} = \langle y^i y^j \rangle - \langle y^i \rangle \langle y^j \rangle$ the corresponding covariant tensor denotes its inverse $C^{ij} C_{jk} = \delta^i_k$. A similar convention is used in parameter space for the Fisher matrix $F_{\alpha\beta}$.

Section 2.1 introduces relevant concepts from probability theory and information theory, while section 2.2 introduces moments, cumulants and their generating functions and section 2.3 gives a brief introduction to inference. Finally, section 2.4 gives a brief derivation of partition functions in probability theory.

2.1 Basic probability theory

The basic framework of probability theory is centered on the concept of a measure space $\Omega = (X, \mathcal{A}, P)$. In this triple, X is the set of all possible outcomes a random variable can take. The second entry is a σ-algebra on this set, containing all measurable subsets of X. Third, P is the probability measure that assigns a probability to each of the subsets $A \in \mathcal{A}$. In the context of this thesis, the set of possible outcomes is the n dimensional real numbers, $X = \mathbb{R}^n$. The Borel algebra $\mathcal{B}(\mathbb{R}^n)$ constitutes a σ-algebra on this set. The measure is then set by different probability densities $p(x)$ as $p(x)\mathrm{d}x$. They fulfill the relation

$$\int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1, \quad p(x) > 0. \tag{2.1}$$

The probability of an event A can be recovered as

$$P(A) = \int_A p(x)\mathrm{d}x. \tag{2.2}$$

Probability densities transform under variable transformations $x \to y$ such that $p(x)dx = p(y)dy$.

In information theory, the Shannon information content of an outcome x is defined as

$$I(x) = -\ln(p(x)). \tag{2.3}$$

The expected information content of a probability distribution is the Shannon entropy

$$S = -\int dx p(x) \ln p(x). \tag{2.4}$$

When the outcome of a process is very certain the entropy is small. When all possibilities are equally likely the entropy is maximized.

Joint probability densities describing the probability of two separate events admit to the following relations. The probability density for one event can be recovered through marginalization

$$p(x) = \int_{-\infty}^{\infty} p(x, y) dy. \tag{2.5}$$

They can be expressed using conditional probabilities $p(x|y)$ describing the probability of x given y is true as

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x). \tag{2.6}$$

This relation gives rise to Bayes theorem

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}. \tag{2.7}$$

In Bayesian parameter inference, for a review in cosmology see [Trotta, 2008], Bayes theorem is used to recover probability distributions of parameters θ given data points y. Usually, it takes the form

$$p(\theta|y, \mathcal{M}) = \frac{\mathcal{L}(y|\theta, \mathcal{M})\pi(\theta|\mathcal{M})}{p(y|\mathcal{M})}. \tag{2.8}$$

Typically, the probability distribution of interest is the posterior $p(\theta|y, \mathcal{M})$. It describes the probability distribution of the parameters in a model $\mathcal{M}$ given the data. The likelihood $\mathcal{L}(y|\theta, \mathcal{M})$ describes how likely it is to generate the data points for a given set of parameters in the model. The prior $\pi(\theta|\mathcal{M})$ describes the knowledge or assumptions prior to conducting the experiment. The evidence is recovered from the expression $p(y|\mathcal{M}) = \int \mathcal{L}(y|\theta, \mathcal{M})\pi(\theta|\mathcal{M})d\theta$. It is used in model selection tasks to compare the probabilities that different models $\mathcal{M}_1$ and $\mathcal{M}_2$ give rise to the observed data points. For applications in cosmology see [Jaffe, 1996, Trotta, 2007, Schosser et al., 2024].

For most of this thesis, the likelihood and sometimes the prior are assumed to be part of the exponential family. They can be expressed as

$$\begin{aligned} \mathcal{L}(y|\theta) &= \frac{1}{N_{\mathcal{L}}} \exp\left(-\frac{1}{2}\chi^2(y|\theta)\right) \\ \pi(y) &= \frac{1}{N_{\pi}} \exp\left(-\frac{1}{2}\phi(y)\right) \end{aligned} \tag{2.9}$$

respectively. Here N_p are the respective normalizations and $\chi^2(y|\theta)$ is a function of the parameters and the data points. The dependence on the model is omitted for brevity. This results in a posterior that is also part of the exponential family

$$p(\theta|y) \propto \exp\left(-\frac{1}{2}(\chi^2(y|\theta) + \phi(y))\right). \tag{2.10}$$

2.2 Cumulants and moments

Given a probability density $p(x)$ the expected outcome of some function $f(x)$ of the random variable can be computed as

$$\langle f(x)\rangle_p = \int_{-\infty}^{\infty} f(x)p(x)dx. \tag{2.11}$$

For a set of samples from this probability distribution $\{x_i\}_{i=1,\dots N}$ the expectation value can be approximated as

$$\langle f(x)\rangle_p \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i). \tag{2.12}$$

The moments of a probability distribution are defined as

$$m_n = \int_{-\infty}^{\infty} x^n p(x)dx. \tag{2.13}$$

They can be generated from a moment-generating function

$$M(t) = \langle e^{tx}\rangle = \sum_{n=0}^{\infty} \frac{t^n\langle x^n\rangle}{n!}, \tag{2.14}$$

where the n-th moment is recovered from the moment-generating function as

$$m_n = \left.\frac{\partial^n M(t)}{\partial t^n}\right|_{t=0}. \tag{2.15}$$

Cumulants κ_n are constructed as an alternative to moments using the cumulant generating function

$$K(t) = \log\langle e^{tx}\rangle = \sum_{n=0}^{\infty} \kappa_n \frac{t^n}{n!}. \tag{2.16}$$

Similar to the moments the n-th cumulant is recovered from the cumulant generating function as

$$\kappa_n = \left.\frac{\partial^n K(t)}{\partial t^n}\right|_{t=0}. \tag{2.17}$$

In practice, this means that given a set of samples from some unknown probability distribution, we can construct expectation values of this distribution. This situation arises when we measure some quantity but do not understand the process generating it. For some probability densities, it is possible to recover the underlying parameters from their moments or cumulants. A normal distribution

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right), \tag{2.18}$$

is fully characterized by both its first two moments $m_1 = \mu$, $m_2 = \sigma^2 + \mu^2$ or equivalently its first two cumulants $\kappa_1 = \mu$, $\kappa_2 = \sigma^2$. While the higher moments of a normal distribution are non-zero, all higher cumulants are zero. Since the cumulant generating function cannot be a finite polynomial of degree greater than two [Lukacs, 1970] higher order cumulants give an indication of how non-Gaussian the probability distribution behaves. Weakly non-Gaussian probability distributions $\frac{\kappa_n}{\kappa_2^{\frac{n}{2}}} \ll 1$ can be reconstructed

using the Gram-Charlier series

$$p(x) \approx \frac{1}{\sqrt{2\pi\kappa_2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\kappa_2}\right)\left(1+\sum_{n=3}^{\infty}\frac{\kappa_n}{\kappa_2^{\frac{n}{2}}\,n!}H_n\left(\frac{x-\kappa_1}{\kappa_2}\right)\right). \tag{2.19}$$

Here, $H_n(x)$ are the Hermite polynomials of nth order. A multivariate form of this series is presented in [Berkowitz and Garner, 1970, Juszkiewicz et al., 1995, Giesel et al., 2021]. For large cumulants this expression may diverge and not constitute a valid probability density [Cramér, 1999]. Whenever a Gaussian approximation of the posterior is sufficient, the Fisher-formalism can be used to approximate the parameter covariance [Tegmark et al., 1997]. Another way to approximate weakly non-Gaussian distributions from its cumulants or moments is derived in the DALI expansion [Sellentin, 2015a].

2.3 Sampling and inference

The previous section gives a strategy for discovering the parameter values for a normal distribution as well as finding some hint on the shape of weakly non-Gaussian distributions. Inference tasks employ Bayes theorem (2.8) to find a probability distribution in the parameters. The model mapping the parameters θ to the data space is fixed as $y_{\text{model}}(\theta)$. For the experiments in this thesis, the likelihoods are assumed to be normally distributed in the data points. They take the form

$$\mathcal{L}(y|\theta) = \frac{1}{\sqrt{(2\pi)^n \det C}} \exp\left(-\frac{1}{2}\left(y - y_{\text{model}}(\theta)\right)^i C_{ij} \left(y - y_{\text{model}}(\theta)\right)^j\right). \tag{2.20}$$

The correlations between different data points are expressed in the inverse covariance matrix C_{ij}. For non-linear models the corresponding posterior is not a normal distribution in the parameters. While it is often possible to find the functional form of the posterior $p(\theta_1, \dots \theta_N|y)$, both its normalization and its marginals $p(\theta_i|y)$ are usually not accessible through analytic calculations. Gaining a qualitative understanding of which parameter ranges are favored given a set of data points usually requires finding lower-dimensional marginals of the probability distribution.

For some classes of distributions, including uniform distributions and normal distributions, the generation of samples is straightforward. More sophisticated methods are needed for more general functional forms. There is a wide range of methods to make these posterior probabilities tractable including Markov chain Monte Carlo (MCMC) [Metropolis et al., 1953, Hastings, 1970], variational inference [Jordan et al., 1999, David M. Blei and McAuliffe, 2017], nested sampling [Skilling, 2006] and recently simulation-based inference [Schafer and Freeman, 2012, Cameron and Pettitt, 2012, Weyant et al., 2013]. In this thesis, MCMC methods are used to generate samples from the posterior distribution. While they usually do not give access to the evidence, they allow for robust estimates of the marginal probability distributions. MCMC methods are a central tool in cosmology since their introduction in [Lewis and Bridle, 2002]. They outperform grid-based evaluation methods, especially in high dimensions. This is due to the method's ability to generate more samples in parameter regions with higher likelihood.

Markov chains

A Markov chain is a set of random variables $\{\theta_1, \theta_2, ...\}$, where the probability of drawing the next number θ_{n+1} depends only on the previous number $T(\theta_{n+1}|\theta_1, \theta_2, \ldots, \theta_n) = T(\theta_{n+1}|\theta_n)$. When there is a unique stationary distribution such that

$$\int_{-\infty}^{\infty} p(\theta_i)T(\theta_i|\theta_j)d\theta_i = p(\theta_j), \quad \forall\theta_j, \tag{2.21}$$

an infinitely long ergodic Markov chain can be understood as samples from the stationary distribution. A distribution is stationary if it fulfills the detailed balance condition

$$p(\theta_j)T(\theta_i|\theta_j) = p(\theta_i)T(\theta_j|\theta_i). \tag{2.22}$$

This allows to generate samples from the probability distribution $p(\theta)$ by designing a transition probability $T(\theta_i|\theta_j)$ such that detailed balance is fulfilled.

Metropolis-Hastings

One of the earliest algorithms still in use for this problem is the Metropolis-Hastings algorithm [Hastings, 1970]. It is used to generate samples $\{\theta_1, \theta_2, \ldots, \theta_N\}$ from the distribution $p(\theta)$. It uses a proposal distribution $q(\theta_i|\theta_j)$ to generate a proposal for the next step. This proposal is then accepted with the probability

$$\alpha(\theta_i, \theta_j) = \min\left\{1, \frac{p(\theta_i)q(\theta_j|\theta_i)}{p(\theta_j)q(\theta_i|\theta_j)}\right\}. \tag{2.23}$$

This defines the transition probability as $T(\theta_i|\theta_j) = \alpha(\theta_i, \theta_j)q(\theta_i|\theta_j)$, ensuring that detailed balance holds for the desired stationary distribution

$$p(\theta_j)T(\theta_i|\theta_j) = \min\{p(\theta_j)q(\theta_i|\theta_j), p(\theta_i)q(\theta_j|\theta_i)\} = p(\theta_i)T(\theta_j|\theta_i). \tag{2.24}$$

The desired set of samples is generated through the following steps:

1. Choose an initial θ_i and compute probability $p(\theta_i)$
2. Generate a candidate $\theta_j \sim q(\theta_j|\theta_i)$ and compute $p(\theta_j)$
3. Accept or reject candidate according to $\alpha(\theta_i, \theta_j)$
4. If the candidate is rejected repeat the previous point in the chain
5. Repeat steps 2 to 4 until the desired chain length is reached.

The resulting set of samples depends on the choice of initial point θ_i, especially when it is far away from the mass of the probability distribution sampled. The movement of the chain towards the mass of the distribution is referred to as burn-in. Usually, the burn-in points are dropped when analyzing the samples. The performance of this algorithm depends on how well the proposal distribution is tailored towards the stationary distribution.

Hamilton Monte Carlo

There are several ways to improve on this algorithm, section 3.3 focuses on Hybrid (Hamilton) Monte Carlo (HMC). These modifications of the Metropolis-Hastings algorithm were proposed by [Duane et al., 1987, Betancourt, 2017], and their working principles are reviewed in [Neal, 2012, Jasche and Kitaura, 2010]. In HMC, the original probability distribution $P(\theta) = \frac{1}{N} \exp\left(-\frac{1}{2}\chi^2(\theta)\right)$ is modified to include a set of momentum directions

$$\begin{aligned} P(\theta, p) &= \frac{1}{Z} \exp\left(-\mathcal{H}(\theta, p)\right) \\ &= \frac{1}{Z} \exp\left(-\frac{1}{2}\chi^2(\theta) - \frac{1}{2}K(p)\right) \\ &= \frac{1}{Z} \exp\left(-\frac{1}{2}\chi^2(\theta)\right) \exp\left(-\frac{1}{2}K(p)\right). \end{aligned} \tag{2.25}$$

Usually $K(p) = p^2$. This effectively doubles the number of parameters while the distribution in the momenta is a multivariate normal with a diagonal covariance matrix. The Hamiltonian equations of motion

$$\begin{aligned} \frac{d\theta^\alpha}{dt} &= \frac{\partial \mathcal{H}}{\partial p_\alpha} = \frac{1}{2}\frac{\partial K(p)}{\partial p_\alpha} \\ \frac{dp_\alpha}{dt} &= -\frac{\partial \mathcal{H}}{\partial \theta^\alpha} = -\frac{1}{2}\frac{\partial \chi(\theta)}{\partial \theta^\alpha}, \end{aligned} \tag{2.26}$$

allow us to move in parameter and momentum space without changing the overall Hamiltonian $\mathcal{H}(\theta, p)$. Greek indices denote entries in the parameter vectors. It is worth noting that these equations incorporate information on the gradients of χ. Trajectories fulfilling the Hamilton equations of motion do not change the joint momentum and parameter distribution (2.25). HMC exploits this, in addition to the fact that the momentum and parameter distribution are independent, to generate proposals in a Markov chain.

At each step in the algorithm, a momentum p is drawn from the normal momentum probability $P(p) \propto \exp\left(-\frac{1}{2}K(p)\right)$. The algorithm effectively moves to a different energy shell. The previous parameters θ_{old} are evolved according to the Hamilton equations (2.26). The resulting combination (θ_c, p_c) is treated as the candidate. Similar to the Metropolis-Hastings algorithm an acceptance probability $\alpha((\theta_c, p_c), (\theta_{old}, p))$ is computed. Since the energy shell has not changed between these two combinations the proposal is nearly always accepted.

While this algorithm samples from the joint probability distribution in momentum and parameter space, the Hamiltonian is designed such that the distributions are independent of each other. The probability distribution in parameter space can be recovered by marginalizing over momentum space

$$\begin{aligned} \int d^n p P(\theta, p) &= \frac{N}{NZ} \exp\left(-\frac{1}{2}\chi^2(\theta)\right) \int d^n p \exp\left(-\frac{1}{2}K(p)\right) \\ &= \frac{1}{N} \exp\left(-\frac{1}{2}\chi^2(\theta)\right) \\ &= P(\theta). \end{aligned} \tag{2.27}$$

When computing expectation values this marginalization is performed by summing over

all possible momenta.

Compared to the Metropolis-Hastings algorithm, HMC converges to the desired stationary distribution quicker for curved and narrow probability distributions. Additionally, HMC performs better for higher dimensional parameter spaces. For an application in cosmology see [Jasche and Kitaura, 2010, Kitaura and Ensslin, 2008].

Other commonly used methods in cosmology include ensemble samplers such as EMCEE [Foreman-Mackey et al., 2013]. Where multiple interacting Markov chains are used to map the parameter space efficiently. While EMCEE uses the distance to other Markov chains to propose candidate steps [Herzog et al., 2024] allows for a variable number of active chains, mirroring a macrocanonical ensemble in statistical physics.

Convergence

The algorithms discussed previously use the properties of Markov chains to generate a set of samples $\{\theta_0, \theta_1, \ldots, \theta_n\}$. However, detailed balance with a unique stationary distribution only guarantees the convergence to this distribution for infinitely long Markov chains. One of the most commonly used methods to quantify the convergence of a set of Markov chains is the Gelman-Rubin criterion [Gelman and Rubin, 1992, Brook and Gelman, 1997, Roberts et al., 1997]. It compares the variance of samples within a single chain with the variance of samples between m different chains. The parameter mean from a single chain i is denoted as $\overline{\theta}_i$ with the overall mean $\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m}\overline{\theta}_i$. The samples of the ith chain have the variance $s_i^2 = \frac{1}{n-1}\sum_{t=1}^{n}\left(\theta_{i,t} - \overline{\theta_i}\right)^2$. The first estimator for the variance is constructed as the average of the different chain variances

$$s^2 = \frac{1}{m}\sum_{i=1}^{m} s_i^2. \tag{2.28}$$

Due to positive correlation in the Markov chain, s^2 underestimates the target variance. A second estimate for the true variance of the distribution can be found with the help of the variance of the means between chains

$$\frac{B}{n} = \frac{1}{m-1}\sum_{i=1}^{m}(\overline{\theta}_i - \hat{\mu})^2. \tag{2.29}$$

The second estimator is then defined as

$$\hat{\sigma}^2 = \frac{n-1}{n}s^2 + \frac{B}{n}. \tag{2.30}$$

This estimator is designed to overestimate the true variance. The ratio of the two estimators converges to one as the chains converge. This defines the Gelman-Rubin R

$$R = \sqrt{\frac{\hat{\sigma}^2}{s^2}}. \tag{2.31}$$

A more thorough derivation, as well as a multivariate version can be found in [Brook and Gelman, 1997]. Additionally, [Vats and Knudson, 2021] includes variants applicable to singular Markov chains. Other methods of assessing the convergence are based on the effective sample size of a Markov chain [Gong and Flegal, 2014, Vats et al., 2019]. Additionally, [Jones et al., 2006] propose computing the expected error for the estimated

quantities from different batches of a single chain. Section 3.3 proposes convergence criteria based on a partition function approach to Markov chains.

2.4 Partition functions in probability theory

This section gives an introduction to the partition function formalism used in [Röver et al., 2023a,b]. It contains parts of the introductions and calculations first published there. Partition functions can be used as an approach to understand some of the tools discussed in the previous sections. This approach is based on [Jaynes, 1957]. Here, thermodynamics is cast as a theory of information. Our approach establishes methods similar to partition functions from thermodynamics as a tool to gain insight into the relations between physical models and the shape of the likelihood. In particular, when likelihood, prior and posterior belong to the exponential family they can be expressed as

$$\begin{aligned} \mathcal{L}(x|\theta) &\propto \exp\left(-\frac{1}{2}\chi^2(y|\theta)\right), \\ \pi(\theta) &\propto \exp\left(-\phi(\theta)\right), \\ p(\theta|y) &\propto \exp\left(-\left(\frac{1}{2}\chi^2(y|\theta)+\phi(\theta)\right)\right). \end{aligned} \tag{2.32}$$

This exponential structure is reminiscent of the probability of finding a microstate with energy $\Delta\Phi$ in statistical physics. A thermal bath can provide this energy at the Boltzmann probability $p = \exp(-\beta\Delta\Phi)$. By defining the potential as

$$\Phi(\theta) = \chi^2(y|\theta)/2 + \phi(\theta) \tag{2.33}$$

we can understand the parameters θ^μ in a Markov chain as a time series of positions in parameter space obtained by performing a thermal random walk. The properties of the time series are determined by the sampling algorithm. The probability of finding a particle at a position θ^μ is determined by the posterior distribution.

2.4.1 Metropolis-Hastings partitions

The construction of the partition function is based on the Bayesian evidence

$$p(y) = \frac{1}{N_\mathcal{L} N_\pi}\int \mathrm{d}^n\theta\, \exp\left(-\left[\frac{1}{2}\chi^2(y|\theta)+\phi(\theta)\right]\right). \tag{2.34}$$

The likelihood and the prior are assumed to belong to the exponential family. In analogy to a canonical partition function, we introduce sources J_α and an inverse temperature β. The partition function can then be defined as

$$Z[\beta, J_\alpha] = \frac{1}{N_\mathcal{L} N_\pi}\int \mathrm{d}^n\theta\, \exp\left(-\beta\left[\frac{1}{2}\chi^2(y|\theta)+\phi(\theta)\right]\right)\exp(\beta J_\alpha\theta^\alpha) \tag{2.35}$$

where the inverse temperature β and the sources J_α play the role of state variables. The evidence is recovered for $\beta = 1$ and $J_\alpha = 0$. This construction closely resembles the definition of the moment-generating function of the posterior (2.14) for $\beta = 1$. The difference is a normalization factor given by the evidence. This difference is intentional

since the evidence is difficult to access for most inference problems. The temperature scaled logarithm of the partition function can be understood as the Helmholtz energy

$$F\left[\beta, J_\alpha\right] = \frac{1}{\beta} \ln Z[\beta, J_\alpha]. \tag{2.36}$$

This corresponds to the cumulant generating functional (2.16), offset by the evidence. The evidence is constant in the parameters and does not affect the derivatives of the Helmholtz energy with respect to the sources.

In this system, the energy of a particle is entirely determined by its position in parameter space through its potential

$$\Phi(\theta) = \frac{1}{2}\chi^2(y|\theta) + \phi(\theta). \tag{2.37}$$

A thermal random walk with respect to this potential can be generated using the Metropolis-Hastings algorithm.

2.4.2 Hamilton Monte-Carlo partitions

To fully accommodate Hamilton Monte Carlo sampling the partition function formalism needs to be extended by a kinetic term $T(p)$. For a given position θ^μ and conjugate momentum p_μ the microscopic energy can be described using the Hamiltonian function $\mathcal{H}(p, \theta)$ [Liu, 2004, Betancourt, 2017]

$$\mathcal{H}(p, \theta) = T(p) + \Phi(\theta). \tag{2.38}$$

This Hamiltonian function can then be used to define a new partition function of the form

$$Z[\beta, J_\alpha, K^\alpha] = \frac{1}{N} \int \mathrm{d}^n p \int \mathrm{d}^n \theta \, \exp\left(-\beta \mathcal{H}(p, \theta)\right) \exp\left(\beta J_\alpha \theta^\alpha\right) \exp\left(\beta K^\alpha p_\alpha\right) \tag{2.39}$$

with analogous sources K^α for the canonical momenta p_α. Here, N incorporates the normalization factors of the likelihood, the prior and the kinetic terms. By choosing to represent the parameter tuples θ^μ as vectors, the conjugate momenta are assigned covariant indices to be consistent with the Hamilton equations of motion (2.26). This also determines that J_μ are linear forms and K^μ are vectors.

As the energies $\mathcal{H}(p, \theta)$ are constructed additively from the kinetic term $T(p)$ and the potential term $\Phi(\theta) = \chi^2/2 + \phi$, the partition function separates

$$\begin{aligned} Z[\beta, J_\alpha, K^\alpha] &= \int \mathrm{d}^n \theta \, \exp\left(-\beta \Phi(\theta)\right) \exp\left(\beta J_\alpha \theta^\alpha\right) \times \int \mathrm{d}^n p \, \exp\left(-\beta T(p)\right) \exp\left(\beta K^\alpha p_\alpha\right) \\ &= Z_\Phi[\beta, J_\alpha] \times Z_T[\beta, K_\alpha] \end{aligned} \tag{2.40}$$

and its logarithm

$$\ln Z[\beta, J_\alpha, K^\alpha] = \ln Z_\Phi[\beta, J_\alpha] + \ln Z_T[\beta, K^\alpha] \tag{2.41}$$

can be used as a generating function for both the cumulants of the posterior distribution $p(\theta|y)$ in configuration space and the cumulants of the posterior distribution in momentum space $p(p|y)$. This factorization of the partition function mirrors the independence of

the two posterior distributions. The posterior distribution in parameter space can be recovered from the joint distribution through marginalization.

Physically, this canonical partition $Z[\beta, J_\alpha, K^\alpha] = Z_T[\beta, K^\alpha] \times Z_\Phi[\beta, J_\alpha]$ would correspond to a classical, non-relativistic, ideal gas in thermal equilibrium inside a potential Φ.

While the kinetic term is often chosen as $T(p) = p^2/2$ generalizations to a positive symmetric quadratic form can yield numerical advantages [Betancourt, 2017]. A parabolic likelihood $\chi^2 = F_{\alpha\beta}\theta^\alpha\theta^\beta$ with a Fisher matrix $F_{\alpha\beta}$ would lead to a canonical partition

$$\begin{aligned} Z[\beta, J_\alpha, K^\alpha] = \int \mathrm{d}^n\theta \int \mathrm{d}^n p \, \exp\left(-\frac{\beta}{2} M^{\alpha\beta} p_\alpha p_\beta\right) \exp\left(-\frac{\beta}{2} F_{\alpha\beta}\theta^\alpha\theta^\beta\right) \cdot \\ \exp\left(\beta J_\alpha \theta^\alpha\right) \exp\left(\beta K^\alpha p_\beta\right). \end{aligned} \tag{2.42}$$

A choice of $M^{\alpha\beta}$ proportional to the inverse Fisher matrix $F^{\alpha\beta}$ is convenient, as the inverse Fisher matrix encodes the covariance of the distribution. Then, $M^{\alpha\beta}$ assigns low inertia to motion in the directions in which the distribution is broad. This introduces an anisotropy in the proposal distribution which could make sampling more efficient, similar to affine-invariant sampling [Foreman-Mackey et al., 2013, Hou et al., 2012]. Strong statistical degeneracies of the likelihood might even suggest a prior in momentum space. Such a prior $\pi(p)$ does not change the posterior distribution $p(\theta|y)$ but could be set up to make sampling more efficient by covering the degeneracies in parameter space more efficiently with samples compared to random, diffusive motion.

2.4.3 Ensemble Monte-Carlo partitions

Having N instead of a single Markov chain bridges towards ensemble methods [Foreman-Mackey et al., 2013]. If the chains are non-interacting, this amounts to a factorizing N-particle partition $Z[\beta, J_\alpha, N]$

$$Z[\beta, J_\alpha, N] = Z[\beta, J_\alpha]^N. \tag{2.43}$$

A variable number of chains is explored in [Herzog et al., 2024] by introducing a chemical potential μ promoting this to a macrocanonical partition function

$$\begin{aligned} \Xi[\beta, J_\alpha, \mu] &= \sum_N \frac{1}{N!} Z[\beta, J_\alpha, N] \exp(\beta\mu N) \\ &= \sum_N \frac{1}{N!} \left(Z[\beta, J_\alpha] \exp(\beta\mu)\right)^N \\ &= \exp\left(z Z[\beta, J_\alpha]\right) \end{aligned} \tag{2.44}$$

with a fugacity $z = \exp(\beta, \mu)$.

3 Partition Functions for Weakly non-Gaussian Likelihoods

This chapter is based on both [Röver et al., 2023a] and [Röver et al., 2023b] it contains the calculations and results first published there. The chapter is split into three sections, the theoretical results derived from the partition function approach in 3.1, the computation of cumulants for weakly non-Gaussian likelihoods in section 3.2 and the construction of convergence measures in section 3.3.

3.1 Theoretical results

3.1.1 Entropy measures

As mentioned in the introduction 2.4 the partition function $Z[\beta, J_\alpha]$ evaluated at unit temperature $\beta = 1$ and for vanishing sources $J_\alpha = 0$ falls back onto the Bayesian evidence as the normalizing factor for the posterior distribution

$$Z[\beta, J_\alpha = 0] = \frac{1}{N_{\mathcal{L}} N_\pi} \int \mathrm{d}^n\theta \, \exp\left(-\beta \left[\frac{1}{2}\chi^2(y|\theta) + \phi(\theta)\right]\right) \quad \text{such that} \quad Z[\beta = 1, J_\alpha = 0] = p(y). \tag{3.1}$$

Here posterior, likelihood and prior are taken from the exponential family as described in eqn. (2.32). In analogy to statistical physics the entropy $S(\beta)$ can be derived from the Helmholtz free energy $F(\beta, J_\alpha)$ through differentiation with the inverse temperature β as

$$S(\beta) = \beta^2 \frac{\partial F(\beta)}{\partial \beta} = -\beta^2 \frac{\partial}{\partial \beta}\left(\frac{1}{\beta} \ln Z[\beta]\right) = \ln Z[\beta] - \beta \frac{\partial}{\partial \beta} \ln Z[\beta]. \tag{3.2}$$

Evaluated at the temperature $\beta = 1$ the first term corresponds to the logarithm of the evidence $\ln p(y)$. The second term reads as

$$\begin{aligned} \beta \frac{\partial}{\partial \beta} \ln Z[\beta]\bigg|_{\beta=1} &= -\frac{1}{p(y)} \int \mathrm{d}^n\theta \, \left(\frac{1}{2}\chi^2(y|\theta) + \phi(\theta)\right) \exp\left(-\left[\frac{1}{2}\chi^2(y|\theta) + \phi(\theta)\right]\right) \\ &= -\int \mathrm{d}^n\theta \left(\frac{1}{2}\chi^2(y|\theta) + \phi(\theta)\right) p(\theta|y). \end{aligned} \tag{3.3}$$

With the logarithmic Bayes theorem

$$\ln p(\theta|y) = -\left(\frac{1}{2}\chi^2(y|\theta) + \phi(\theta)\right) - \ln p(y) \tag{3.4}$$

the second term simplifies to

$$\beta \frac{\partial}{\partial \beta} \ln Z[\beta] \bigg|_{\beta=1} = \int \mathrm{d}^n\theta \, p(\theta|y) \ln p(\theta|y) + \ln p(y). \tag{3.5}$$

This relates the entropy S directly to Shannon's measure of information entropy

$$S(\beta = 1) = -\beta^2 \frac{\partial}{\partial \beta} \left(\frac{1}{\beta} \ln Z[\beta] \right) \bigg|_{\beta=1} = -\int \mathrm{d}^n\theta \, p(\theta|y) \ln p(\theta|y) \tag{3.6}$$

This is another way to show the compatibility of Shannon's entropy over the wider class of Rényi-entropies with Bayes' law [Van Erven and Harremos, 2014, Baez and Fritz, 2014]. Applications of the Shannon entropy in cosmology can be found in [Carron et al., 2011, Grandis et al., 2016, Pinho et al., 2021, Nicola et al., 2019].

3.1.2 Cumulants

The partition function defined in eqn. (2.35) can be understood as a cumulant-generating function similar to the expression in (2.16). Cumulants of the posterior distribution of order n follow from n-fold differentiation with respect to the sources J_α, i.e. the first cumulant coincides with the expectation value as

$$\kappa^\mu = \langle \theta^\mu \rangle = \frac{\partial}{\partial J_\mu} \left(\frac{1}{\beta} \ln Z[\beta, J_\alpha] \right) \bigg|_{J=0,\beta=1}, \tag{3.7}$$

evaluated at $J_\mu = 0$ for all μ. Additionally, the second cumulant coincides with the covariance and is given by

$$\kappa^{\mu,\nu} = \langle \theta^\mu \theta^\nu \rangle - \langle \theta^\mu \rangle \langle \theta^\nu \rangle = \frac{\partial^2}{\partial J_\mu \partial J_\nu} \left(\frac{1}{\beta} \ln Z[\beta, J_\alpha] \right) \bigg|_{J=0,\beta=1}. \tag{3.8}$$

The higher-order derivatives correspond to skewness

$$\begin{aligned} \kappa^{\mu,\nu,\rho} &= \langle \theta^\mu \theta^\nu \theta^\rho \rangle - \langle \theta^\mu \rangle \langle \theta^\nu \theta^\rho \rangle - \langle \theta^\nu \rangle \langle \theta^\mu \theta^\rho \rangle - \langle \theta^\rho \rangle \langle \theta^\mu \theta^\nu \rangle + 2 \langle \theta^\mu \rangle \langle \theta^\nu \rangle \langle \theta^\rho \rangle \\ &= \frac{\partial^3}{\partial J_\mu \partial J_\nu \partial J_\rho} \left(\frac{1}{\beta} \ln Z[\beta, J_\alpha] \right) \bigg|_{J=0,\beta=1}, \end{aligned} \tag{3.9}$$

and a non-Gaussian kurtosis

$$\begin{aligned} \kappa^{\mu,\nu,\rho,\sigma} &= \langle \theta^\mu \theta^\nu \theta^\rho \theta^\sigma \rangle - \langle \theta^\mu \theta^\nu \rangle \langle \theta^\rho \theta^\sigma \rangle - \langle \theta^\mu \theta^\rho \rangle \langle \theta^\nu \theta^\sigma \rangle - \langle \theta^\mu \theta^\sigma \rangle \langle \theta^\nu \theta^\rho \rangle \\ &= \frac{\partial^4}{\partial J_\mu \partial J_\nu \partial J_\rho \partial J_\sigma} \left(\frac{1}{\beta} \ln Z[\beta, J_\alpha] \right) \bigg|_{J=0,\beta=1}, \end{aligned} \tag{3.10}$$

all taken at $J_\mu = 0$ and $\beta = 1$ after differentiation. Note that differentiation of the Helmholtz free energy yields the expectation values of the parameters $\langle \theta^\mu \rangle$ and not the best-fit values. This is due to the definition of the partition function as an integrated quantity.

For Hamilton Monte Carlo partitions the separation of the partition function into its

potential and kinetic part (2.40) allows for separate computation of the cumulants of the joint probability distribution. The cumulants of the posterior are computed as

$$\begin{aligned}\kappa_{\Phi}^{\mu_1,\ldots,\mu_n} &= \frac{\partial^n}{\partial J_{\mu_1}\ldots\partial J_{\mu_n}}\left(\frac{1}{\beta}\ln Z[\beta, J_\alpha, K^\alpha]\right)\Big|_{J=0,K=0,\beta=1} \\ &= \frac{\partial^n}{\partial J_{\mu_1}\ldots\partial J_{\mu_n}}\left(\frac{1}{\beta}\ln Z_\Phi[\beta, J_\alpha]\right)\Big|_{J=0,\beta=1}.\end{aligned} \tag{3.11}$$

The cumulants of the momentum distribution can be computed with a similar prescription using

$$\begin{aligned}\kappa_{T}^{\mu_1,\ldots,\mu_n} &= \frac{\partial^n}{\partial K^{\mu_1}\ldots\partial K^{\mu_n}}\left(\frac{1}{\beta}\ln Z[\beta, J_\alpha, K^\alpha]\right)\Big|_{J=0,K=0,\beta=1} \\ &= \frac{\partial^n}{\partial K^{\mu_1}\ldots\partial K^{\mu_n}}\left(\frac{1}{\beta}\ln Z_T[\beta, K^\alpha]\right)\Big|_{K=0,\beta=1}.\end{aligned} \tag{3.12}$$

Computing cumulants for a given non-Gaussian probability distribution commonly involves generating samples using Markov chain Monte Carlo. These samples are then used to compute moments and equivalently cumulants, as the two are related by Faà di Bruno's formula [Johnson, 2002]. Whenever a numerical approximation of the logarithmic partition function is viable this approach yields an analytical approximation of the cumulants. In the case of weakly non-Gaussian likelihoods this yields an analytical approximation to the posterior distribution, see section 3.2.

3.1.3 Linear models

While the partition function formalism allows for general prescriptions to find cumulants and entropies the analytical calculation is challenging for non-Gaussian likelihoods. In this section we assume that the data y^i follows a Gaussian error process. The data covariance is obtained as $C^{ij} = \langle y^i y^j\rangle - \langle y^i\rangle\langle y^j\rangle$. Additionally, the physical model is assumed to be linear in the parameters $y^i_{\text{model}} = A^i{}_\alpha\theta^\alpha$. This determines the χ^2-functional as

$$\begin{aligned}\chi^2 &= \left(y^i - A^i{}_\alpha\theta^\alpha\right)C_{ij}\left(y^j - A^j{}_\beta\theta^\beta\right) \\ &= \underbrace{y^i C_{ij} y^j}_{c} - 2\underbrace{y^j C_{ij} A^i{}_\alpha}_{Q_\alpha}\theta^\alpha + \underbrace{A^i{}_\alpha C_{ij} A^j{}_\beta}_{F_{\alpha\beta}}\theta^\alpha\theta^\beta \\ &= c - 2Q_\alpha\theta^\alpha + F_{\alpha\beta}\theta^\alpha\theta^\beta.\end{aligned} \tag{3.13}$$

The Jacobian $A^i{}_\alpha = \partial y^i/\partial\theta^\alpha$ transforms between parameter and data space. For a likelihood $\mathcal{L}(y|\theta) \propto \exp\left(-\frac{1}{2}\chi^2(y|\theta)\right)$, we identify the term $A^i{}_\alpha C_{ij} A^j{}_\beta$ as the Fisher matrix $F_{\alpha\beta}$, since

$$\begin{aligned}F_{\alpha\beta} &= \left\langle \frac{\partial\ln\mathcal{L}}{\partial\theta^\alpha}\frac{\partial\ln\mathcal{L}}{\partial\theta^\beta}\right\rangle_{y\sim\mathcal{L}} \\ &= \frac{\partial y^i_{\text{model}}}{\partial\theta^\alpha}\left\langle C_{ik}(y^k - y^k_{\text{model}})C_{j\ell}(y^\ell - y^\ell_{\text{model}})\right\rangle_{y\sim\mathcal{L}}\frac{\partial y^j_{\text{model}}}{\partial\theta^\beta} \\ &= A^i_\alpha C_{ik}C^{k\ell}C_{j\ell}A^j_\beta = A^i{}_\alpha C_{ij}A^j{}_\beta.\end{aligned} \tag{3.14}$$

The best-fit parameter tuple $\bar{\theta}^\mu$ is computed by minimizing χ^2 as a function of the

parameters θ^μ as

$$\begin{aligned}\frac{\partial}{\partial\theta^\mu}\chi^2 &= -2Q_\alpha\frac{\partial\theta^\alpha}{\partial\theta^\mu} + F_{\alpha\beta}\frac{\partial}{\partial\theta^\mu}\theta^\alpha\theta^\beta \\ &= -2Q_\alpha\delta^\alpha_\mu + F_{\alpha\beta}\left(\delta^\alpha_\mu\theta^\beta + \theta^\alpha\delta^\beta_\mu\right) \\ &= -2Q_\mu + F_{\mu\beta}\theta^\beta + F_{\alpha\mu}\theta^\alpha \\ \rightarrow \quad \bar{\theta}^\mu &= F^{\mu\alpha}Q_\alpha = A^\mu{}_i y^i.\end{aligned} \tag{3.15}$$

This expression uses the (pseudo) inverse Jacobian $A^\alpha{}_i = \partial\theta^\alpha/\partial y^i$. For linear models this estimate is unbiased $\langle\bar{\theta}^\mu\rangle_{y\sim\mathcal{L}} = A^\mu{}_i\langle y^i\rangle_{y\sim\mathcal{L}} = A^\mu{}_i A^i{}_\beta\theta^\beta = \theta^\mu$.

Constructing the partition function for such a linear model amounts to

$$Z[\beta, J_\alpha] = \frac{1}{N}\int \mathrm{d}^n\theta\, \exp\left(-\frac{\beta}{2}F_{\alpha\beta}\theta^\alpha\theta^\beta + \beta Q_\alpha\theta^\alpha\right)\exp\left(\beta J_\alpha\theta^\alpha\right). \tag{3.16}$$

The Gaussian integrals can be carried out to yield

$$Z[\beta, J_\alpha] = \frac{1}{N}\sqrt{\left(\frac{2\pi}{\beta}\right)^n\frac{1}{\det(F)}}\exp\left(\frac{\beta}{2}F^{\alpha\beta}(J_\alpha + Q_\alpha)(J_\beta + Q_\beta)\right), \tag{3.17}$$

with the inverse Fisher matrix $F^{\alpha\beta}$. We absorb the constant c from eqn. (3.13) into the normalization N and disregard the prior $\pi(\theta)$ for simplicity.

The expectation value of the posterior distribution follows directly from differentiation of $\ln Z[\beta, J_\alpha]/\beta$, evaluated at $J_\mu = 0$

$$\begin{aligned}\kappa^\mu = \langle\theta^\mu\rangle_{\theta\sim p(\theta|y)} &= \frac{\partial}{\partial J_\mu}\left(\frac{1}{\beta}\ln Z[\beta, J_\alpha]\right)\Bigg|_{J=0} \\ &= \frac{F^{\alpha\beta}}{2}\left(\frac{\partial J_\alpha}{\partial J_\mu}(J_\beta + Q_\beta) + (J_\alpha + Q_\alpha)\frac{\partial J_\beta}{\partial J_\mu}\right)\Bigg|_{J=0} \\ &= \frac{F^{\alpha\beta}}{2}\left(\delta^\mu_\alpha(J_\beta + Q_\beta) + (J_\alpha + Q_\alpha)\delta^\mu_\beta\right)\Bigg|_{J=0} \\ &= F^{\mu\alpha}Q_\alpha.\end{aligned} \tag{3.18}$$

The last step uses the symmetry of the Fisher matrix to recover the result from the direct calculation in eqn. (3.15). For a symmetric distribution the most likely value and expectation value coincide with the true parameter value $\kappa^\mu = \langle\theta^\mu\rangle = \hat{\theta}^\mu$, as a reflection of the Gauss-Markov theorem in this formalism.

The second cumulant $\kappa^{\mu,\nu}$, corresponding to the parameter covariance, is computed as

$$\kappa^{\mu,\nu} = \frac{\partial^2}{\partial J_\mu\partial J_\nu}\left(\frac{1}{\beta}\ln Z[\beta, J_\alpha]\right)\Bigg|_{J=0} = \frac{F^{\alpha\beta}}{2}\left(\delta^\mu_\alpha\delta^\nu_\beta + \delta^\nu_\alpha\delta^\mu_\beta\right)\Bigg|_{J=0} = F^{\mu\nu} \tag{3.19}$$

Since any higher-order cumulants are zero, the posterior distribution $p(\theta|y)$ is Gaussian. The inverse parameter covariance is determined by the Fisher matrix and the first cumulant determines the mean. This allows to recover the full posterior as

$$p(\theta|y) = \sqrt{\frac{1}{(2\pi)^n\det(F)}}\exp\left(-\frac{1}{2}F_{\mu\nu}(\theta^\mu - \kappa^\mu)(\theta^\nu - \kappa^\nu)\right). \tag{3.20}$$

3.2 Cumulants for weakly non-Gaussian posteriors

Non-Gaussian posteriors result from nonlinear models. In this case the function $y^i(\theta^\alpha)$ can not be written as $y^i = A^i{}_\alpha \theta^\alpha$ with a constant $A^i{}_\alpha$. The Helmholtz free energy does not truncate after second order in the sources J_α. The posterior distribution becomes genuinely non-Gaussian. However, cumulants of order n with $n \geq 3$ remain numerically computable from the partition function. In this sense, the partition function formalism provides an approximation for non-Gaussian posterior distributions at a given order. In this section the models are rescaled such that the fiducial parameter values are zero, $\langle \theta^\mu \rangle = \kappa^\mu = 0$ to simplify the notation.

3.2.1 Weakly non-Gaussian posteriors

When introducing a weak non-Gaussianity, e.g. through introducing a model where the parameters are not quite linearly linked to the data, the partition function factorizes into a Gaussian and a non-Gaussian part. Non-Gaussianity is introduced in the χ^2-functional as

$$\frac{\chi^2}{2} = \frac{1}{2} F_{\alpha\beta}\theta^\alpha\theta^\beta - \sum_{k=3}^{N} \frac{1}{k!} C_{\mu_1\dots\mu_k}\theta^{\mu_1}\dots\theta^{\mu_k}, \tag{3.21}$$

i.e. with a Taylor-expansion of χ^2 beyond quadratic order. Coefficients in the Taylor expansion are assumed to be small compared to the entries of the covariance matrix $F_{\alpha\beta}$. The minus sign of the nonlinear term $C_{\mu_1\dots\mu_k}$ is chosen out of convenience. The separation of the partition function then follows as

$$\begin{aligned} Z[\beta, J_\alpha] &= \int \frac{\mathrm{d}^n\theta}{N} \exp\left(-\frac{\beta}{2} F_{\alpha\beta}\theta^\alpha\theta^\beta + \beta \sum_{k=3}^{N} \frac{1}{k!} C_{\mu_1\dots\mu_k}\theta^{\mu_1}\dots\theta^{\mu_k} + \beta J_\alpha\theta^\alpha\right) \\ &\approx \int \frac{\mathrm{d}^n\theta}{N} \exp\left(-\frac{\beta}{2} F_{\alpha\beta}\theta^\alpha\theta^\beta + \beta J_\alpha\theta^\alpha\right)\left(1 + \beta \sum_{k=3}^{N} \frac{1}{k!} C_{\mu_1\dots\mu_k}\theta^{\mu_1}\dots\theta^{\mu_k}\right). \end{aligned} \tag{3.22}$$

Here we assumed that the χ^2-functional contains the prior dependence on the parameters θ. The normalization N denotes the normalizations of both the likelihood and the prior.

Next, the computation of the moments is replaced with a differentiation with respect to the sources J_α. The partition function can then be expressed as

$$\begin{aligned} Z[\beta, J_\alpha] &\approx \frac{1}{N}\left(1 + \beta \sum_{k=3}^{N} \frac{1}{k!} C_{\mu_1\dots\mu_k} \frac{\partial}{\partial J_{\mu_1}} \dots \frac{\partial}{\partial J_{\mu_k}}\right) \int \mathrm{d}^n\theta \exp\left(-\frac{\beta}{2} F_{\alpha\beta}\theta^\alpha\theta^\beta + \beta J_\alpha\theta^\alpha\right) \\ &= \frac{1}{N}\sqrt{\left(\frac{2\pi}{\beta}\right)^n \frac{1}{\det(F)}}\left(1 + \beta \sum_{k=3}^{N} \frac{1}{k!} C_{\mu_1\dots\mu_k} \frac{\partial}{\partial J_{\mu_1}} \dots \frac{\partial}{\partial J_{\mu_k}}\right) \exp\left(\frac{\beta}{2} F^{\alpha\beta} J_\alpha J_\beta\right) \\ &= \frac{1}{N}\sqrt{\left(\frac{2\pi}{\beta}\right)^n \frac{1}{\det(F)}} \exp\left(\frac{\beta}{2} F^{\alpha\beta} J_\alpha J_\beta\right) \cdot \\ &\quad \left(1 + \beta \sum_{k=3}^{N} \frac{1}{k!} C_{\mu_1\dots\mu_k} \exp\left(-\frac{\beta}{2} F^{\alpha\beta} J_\alpha J_\beta\right) \frac{\partial}{\partial J_{\mu_1}} \dots \frac{\partial}{\partial J_{\mu_k}} \exp\left(\frac{\beta}{2} F^{\alpha\beta} J_\alpha J_\beta\right)\right). \end{aligned} \tag{3.23}$$

The first factor in this equation is equivalent to the partition function for a linear model as given in eqn. (3.17). Here the mean values are set to zero. The second factor in the

equation is the contribution of non-Gaussianities to the partition function. With the definitions

$$Z_G[\beta, J_\alpha] = \frac{1}{N}\sqrt{\left(\frac{2\pi}{\beta}\right)^n \frac{1}{\det(F)}}\exp\left(\frac{\beta}{2}F^{\alpha\beta}J_\alpha J_\beta\right) \tag{3.24}$$

and

$$Z_{NG}[\beta, J_\alpha] = 1 + \beta\sum_{k=3}^{N}\frac{1}{k!}C_{\mu_1\dots\mu_k}\exp\left(-\frac{\beta}{2}F^{\alpha\beta}J_\alpha J_\beta\right)\frac{\partial^k}{\partial J_{\mu_1}\dots\partial J_{\mu_k}}\exp\left(\frac{\beta}{2}F^{\alpha\beta}J_\alpha J_\beta\right) \tag{3.25}$$

a factorization of the partition function into a contribution due to its Gaussian part and the influence of the non-Gaussianities is observed,

$$Z[\beta, J_\alpha]) = Z_G[\beta, J_\alpha]Z_{NG}[\beta, J_\alpha]. \tag{3.26}$$

Using the fact that the inverse Fisher matrix can be Cholesky decomposed as $F^{\alpha\beta} = L^{\gamma\beta}L_\gamma{}^\alpha$ the non-Gaussian part is expressed in terms of multivariate Hermite polynomials of the form

$$H^{(\nu_1\dots\nu_\ell)}(J_\alpha) = \exp\left(\frac{1}{2}J_\alpha\delta^{\alpha\beta}J_\beta\right)(-1)^\ell\frac{\partial^\ell}{\partial J_{\nu_1}\dots\partial J_{\nu_\ell}}\exp\left(-\frac{1}{2}J_\alpha\delta^{\alpha\beta}J_\beta\right) \tag{3.27}$$

as

$$Z_{NG}[\beta, J_\alpha] = \left(1 + \beta\sum_{k=3}^{N}\frac{(-\mathrm{i}^k)}{k!}C_{\mu_1\dots\mu_k}L^{\mu_1}_{\nu_1}\dots L^{\mu_k}_{\nu_k}H^{(\nu_1\dots\nu_k)}(\mathrm{i}L^\mu_\alpha J_\mu)\right). \tag{3.28}$$

Cumulants are computed from the Helmholtz free energy and the contributions of the Gaussian and non-Gaussian parts can be expressed as a sum. Up to first order in the non-Gaussianities the expression is

$$\begin{aligned}\frac{1}{\beta}\ln Z[\beta, J_\alpha] &= \frac{1}{\beta}\ln Z_G[\beta, J_\alpha] + \frac{1}{\beta}\ln Z_{NG}[\beta, J_\alpha] \\ &\approx \frac{1}{2}F^{\alpha\beta}J_\alpha J_\beta + \sum_{k=3}^{N}\frac{(-\mathrm{i}^k)}{k!}C_{\mu_1\dots\mu_k}L^{\mu_1}_{\nu_1}\dots L^{\mu_k}_{\nu_k}H^{(\nu_1\dots\nu_k)}(\mathrm{i}L^\mu_\alpha J_\mu) + \text{const.}\end{aligned} \tag{3.29}$$

Note that in eqn. (3.22) the approximation can be performed to higher order in the non-Gaussianities for the cost of including higher-order Hermite polynomials in the result. The factorization itself can still be performed.

For non-vanishing expectation values such that

$$\frac{\chi^2}{2} = \frac{1}{2}F_{\alpha\beta}\theta^\alpha\theta^\beta + \kappa_\alpha\theta^\alpha - \sum_{k=3}^{N}\frac{1}{k!}C_{\mu_1\dots\mu_k}\theta^{\mu_1}\dots\theta^{\mu_k} \tag{3.30}$$

the Gaussian term is modified by a term linear in the sources

$$\frac{1}{\beta}\ln Z_G[\beta, J_\alpha] = \frac{1}{2}F^{\alpha\beta}J_\alpha J_\beta - \kappa^\alpha J_\alpha, \tag{3.31}$$

and consequently, the non-Gaussian term in the partition function remains unchanged.

Higher-order cumulants of the posterior can be computed as

$$\begin{aligned}\kappa^{\mu_1,\dots,\mu_n} &= \frac{\partial^n}{\partial J_{\mu_1}\dots\partial J_{\mu_n}}\left(\frac{1}{\beta}\ln Z[\beta,J_\alpha]\right)\Bigg|_{J=0} \\ &= \frac{\partial^n}{\partial J_{\mu_1}\dots\partial J_{\mu_n}}\left(\frac{1}{2}F^{\alpha\beta}J_\alpha J_\beta - \sum_{k=3}^{N}\frac{(-\mathrm{i}^k)}{k!}C_{\mu_1\dots\mu_k}L^{\mu_1}_{\nu_1}\dots L^{\mu_k}_{\nu_k}H^{(\nu_1\dots\nu_k)}(\mathrm{i}L^\mu_\alpha J_\mu)\right)\Bigg|_{J=0}.\end{aligned} \tag{3.32}$$

The first and second cumulants contain contributions from the odd and even Hermite polynomials in the non-Gaussian term respectively. The higher cumulants are completely determined by the non-Gaussian term.

Whether non-Gaussianities are genuine or an artifact of an unfortunate choice of random variables and can be removed by a suitable coordinate transform can be traced to the existence of curvature on the manifold whose metric is given by the Fisher matrix $F_{\mu\nu}$ [see the foundational work by Amari, 2016]. Applications in cosmology are discussed in [Giesel et al., 2021], and variations of the Fisher matrix over the parameter manifold and the resulting non-Gaussianities in [Schäfer and Reischke, 2016] and [Reischke et al., 2017].

3.2.2 Gram-Charlier series

The cumulants of a weakly non-Gaussian posterior allow us to reconstruct it using the multivariate Gram-Charlier series. It can be written as [Berkowitz and Garner, 1970, Juszkiewicz et al., 1995, Giesel et al., 2021]

$$\begin{aligned}p(\theta|y) =& \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det(\kappa^{(2)})}}\exp\left(-\frac{1}{2}\kappa_{\alpha,\beta}(\theta^\alpha-\kappa^\alpha)(\theta^\beta-\kappa^\beta)\right)\cdot \\ & \left(1+\sum_{\ell=3}^{\infty}\frac{\kappa^{\alpha_1,\dots,\alpha_\ell}}{\ell!}L^{\beta_1}_{\alpha_1}\dots L^{\beta_\ell}_{\alpha_\ell}H_{(\beta_1\dots\beta_\ell)}(L^\alpha_\beta(\theta-\mu)^\beta)\right).\end{aligned} \tag{3.33}$$

The mean value of the distribution is chosen as the first cumulant of the posterior while the covariance matrix is the inverse of the second cumulant. The Hermite polynomials in this expression again follow the definition in eqn. (3.27). In this definition the multidimensional Hermite polynomials factorize into one-dimensional ones. The variable transformation necessary to achieve this form is in the argument of the Hermite polynomials. This accounts for the Cholesky decomposed second cumulants in their prefactors.

In the following, the Gram-Charlier series is used to reconstruct a probability distribution of the form

$$p(\theta|y) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\det(\Sigma)}}\exp\left(-\frac{1}{2}\Sigma_{\alpha\beta}\theta^\alpha\theta^\beta\right)\left(1+\frac{1}{3!}C_{\alpha\beta\gamma}\theta^\alpha\theta^\beta\theta^\gamma\right). \tag{3.34}$$

This is a first-order approximation in the coefficients $C_{\alpha\beta\gamma}$ to a posterior with a third-order non-Gaussianity. For small coefficients, it is possible to compute the cumulants of this posterior according to eqn. (3.32). They can then be inserted into expression eqn. (3.33) to recover an approximation of the posterior. This reconstruction is depicted in Fig. 3.1.

As demonstrated in the above example it is possible to approximately construct cumulants for a given expansion coefficient in the DALI expansion. However, it is worth noting that

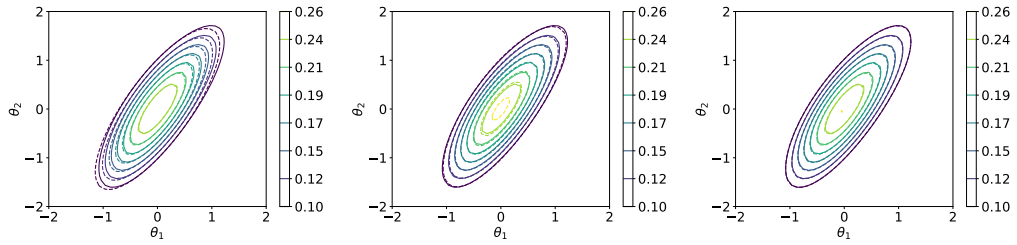

Figure 3.1: Reconstruction of non-Gaussianity of second order using the Gram-Charlier series. The left plot shows the posterior constructed according to eqn. (3.34) and a Gaussian (dashed) with the same width in comparison. The middle (dashed) is the Gram-Charlier series to first order in the non-Gaussianities. The dashed lines on the right depict the Gram-Charlier series to second order in the non-Gaussianities

even the first order in the DALI expansion does not correspond to any specific order in the Gram-Charlier expansion. This can be seen by considering

$$\exp\left(-\frac{1}{2}F_{\alpha\beta}\theta^\alpha\theta^\beta + \frac{1}{3!}S_{\alpha\beta\gamma}\theta^\alpha\theta^\beta\theta^\gamma\right) = \exp\left(-\frac{1}{2}F_{\alpha\beta}\theta^\alpha\theta^\beta\right)\sum_{n=0}^{\infty}\frac{1}{n!}(S_{\alpha\beta\gamma}\theta^\alpha\theta^\beta\theta^\gamma)^n. \quad (3.35)$$

In this expansion terms of order 3^n in the parameters θ can only be obtained by including an ever-increasing order expansion in the Gram-Charlier series. In this sense, the Gram-Charlier series seems to be incompatible with a straightforward expansion of the logarithmic likelihood in terms of a polynomial as done in DALI [Sellentin et al., 2014, Sellentin, 2015b]. A large number of terms is needed to describe even moderately non-Gaussian distributions, and set aside issues with non-positive definite probability densities, as pointed out in [Cramér, 1999].

3.2.3 Application to supernova data

As an example for a non-Gaussian likelihood from cosmology, we consider constraints on Ω_m and w from the distance redshift relation of supernovae [Riess et al., 1998, Goobar and Leibundgut, 2011]. We focus on spatially flat FLRW-cosmologies with a constant dark energy equation of state, and derive constraints on Ω_m and w for the Union2.1-data set [Suzuki et al., 2012, Amanullah et al., 2010, Kowalski et al., 2008]. For these cases, the distance modulus $y(z)$ as a function of redshift z is given by

$$y(z,\Omega_m,w) = 10 + 5\log\left((1+z)\,\chi_H\int_0^z \mathrm{d}z'\,\frac{1}{\sqrt{\Omega_m(1+z')^3 + (1-\Omega_m)(1+z')^{3(1+w)}}}\right). \quad (3.36)$$

For this type of cosmology the relevant integral can be expressed in terms of a hypergeometric function ${}_2F_1$ [Arutjunjan et al., 2022]

$$\int \mathrm{d}u\,\frac{1}{\sqrt{A\,u^3 + B\,u^c}} = -\frac{2u\sqrt{\frac{A\,u^{3-c}}{B}+1}\,{}_2F_1\left(\frac{1}{2},\frac{c-2}{2c-6};\frac{3c-8}{2c-6};-\frac{A\,u^{3-c}}{B}\right)}{(c-2)\sqrt{A\,u^3 + B\,u^c}} + \text{const.} \quad (3.37)$$

Expressing the likelihood for the two parameters Ω_m and w for Gaussian errors σ_i in the distance moduli y_i yields a simplified expression, where we neglect correlations between

the data points,

$$\begin{aligned} \mathcal{L}(y|\Omega_m, w) &\propto \exp\left(-\frac{1}{2}\chi^2(y|\Omega_m, w)\right) \\ \text{with} \quad \chi^2(y|\Omega_m, w) &= \sum_i \left(\frac{y_i - y(z_i, \Omega_m, w)}{\sigma_i}\right)^2 . \end{aligned} \tag{3.38}$$

We use this formulation to construct partition functions $Z[\beta, J_\alpha]$ and implementation in a Monte Carlo Markov chain. For simplicity, we employ a flat prior $\pi(\Omega_m, w)$ on the two cosmological parameters. The model for the distance modulus as a function of the model parameters Ω_m and w is nonlinear, giving rise to a non-Gaussian likelihood, on which we demonstrate a Gram-Charlier expansion.

In this case the partition function is constructed from a Monte Carlo Markov chain sampling from the posterior distribution. Up to a constant normalization the partition function is the expectation value of $\exp(J_\alpha \theta^\alpha)$

$$Z[\beta = 1, J_\alpha] \propto \langle \exp(J_\alpha \theta^\alpha) \rangle \approx \frac{1}{N} \sum_{i=1}^{N} \exp(J_\alpha (\theta^\alpha)_i) . \tag{3.39}$$

Here, i denotes the i-th sample in the Markov chain. This allows us to construct the Helmholtz free energy for a given J_α and to find the cumulants following eqn. (3.11). The cumulants are computed through finite differencing with respect to all J_α.

The numerical precision is verified by computing moments of order $a + b$ of the posterior using the samples drawn using Markov chain Monte Carlo. The moments are computed as

$$\left\langle {\Omega_m}^a w^b \right\rangle = \int \mathrm{d}\Omega_m \mathrm{d}w \, p(\Omega_m, w|y) \, {\Omega_m}^a w^b \approx \frac{1}{N} \sum_{i=1}^{N} ({\Omega_m}^a w^b)_i . \tag{3.40}$$

They are converted to cumulants with Faà di Bruno's formula.

At Gaussian and lowest non-Gaussian order, posterior distribution $p(\Omega_m, w|y)$ is depicted in Fig. 3.2 along with the samples generated by the Monte Carlo algorithm. The Gaussian isoprobability contours correspond exactly to the Fisher matrix and the lowest non-Gaussian approximation to a Gram-Charlier expansion including skewness. Driving the Gram-Charlier expansion to higher-order shows the known deficiency in reproducing distributions with strong non-Gaussianities. This causes the Gram-Charlier expansion to lose positive definiteness. Those are cases where DALI plays its unique strength [Sellentin, 2015b, Sellentin et al., 2014]. We would like to emphasize that the computation of the cumulants from $\ln Z$ is numerically sound.

The computations of cumulants of the posterior distribution from the two ways considered in this work give very similar results. Table 3.1 collects all cumulants up to fourth-order from the posterior distribution $p(\Omega_m, w|y)$ of the supernova example. These cumulants follow either through estimation of the moments from MCMC samples and successive conversion into cumulants using Faà di Bruno's formula, or by finite differencing of the numerically evaluated partition function $\ln Z$. Note however that both of these results are entirely dependent on the same part of a Markov chain. Comparing the results in Table 3.1 to cumulants obtained from three different parts of the same Markov chain, each containing 10^6 elements leads to differences in the cumulant values of about 0.07%, 0.4%, 20%, 20%, for the first, second, third and fourth cumulant respectively. Given the

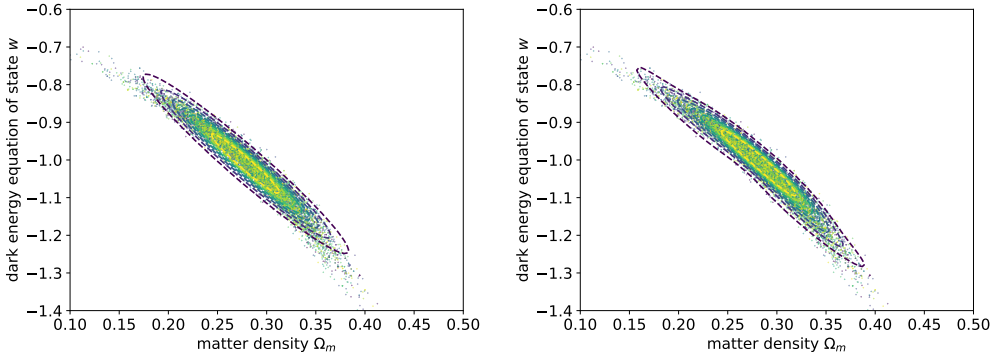

Figure 3.2: The scatter plot depicts 10^4 points from a Monte-Carlo Markov chain with a likelihood as described in eqn. (3.38). The superimposed contours are the 0th (left) and 1st (right) order approximations to this posterior distribution using the Fisher matrix (left) and Gram-Charlier series with nonzero skewness (right).

large number of samples needed, we replaced the Metropolis-Hastings algorithm with the affine-invariant sampler emcee [Foreman-Mackey et al., 2013] for better efficiency.

3.3 Measuring convergence

The previous sections dealt with approximations of posterior properties using the partition function formalism. This section focuses on the analogy between Markov chain Monte Carlo and statistical physics. We identify virialization, equipartition and thermalization as quantifiers of convergence when treating the Markov chain as the time evolution of the state of a physical system. In the following, kinetic energy is introduced as described in section 2.4.2. The Hamiltonian is defined as the sum of the kinetic and potential energy, $\mathcal{H}(p, \theta) = T(p) + \Phi(\theta)$. The potential is determined by the underlying likelihood and prior, $\Phi(\theta) = \frac{1}{2}\chi^2(y|\theta) + \phi(\theta)$.

3.3.1 Virialization

Bounded motion inside a potential exhibits the virial relation

$$\left\langle \theta^\mu \frac{\partial \mathcal{H}}{\partial \theta^\mu} \right\rangle = \left\langle p_\mu \frac{\partial \mathcal{H}}{\partial p_\mu} \right\rangle \tag{3.41}$$

which translates to the relation $2\langle T \rangle = k\langle \Phi \rangle$ between the average kinetic and potential energies for Hamiltonian functions that are homogeneous of order 2 in p and of order k in θ. For ergodic systems it does not matter whether the averages are taken over time or a statistical ensemble. As a Markov chain starts exploring the potential Φ the virial relation does not hold straight away. Rather, it can only be expected to hold over a few dynamical timescales of the system.

For equilibrated Markov chains which are a proper realization of the canonical partition

cumulants	MCMC sampling	partition function
κ^{Ω_m}	0.27881995	0.27882000
κ^{w}	−1.01051052	−1.01051010
$\kappa^{\Omega_m,\Omega_m}$	0.0021031685	0.0021031674
$\kappa^{\Omega_m,w}$	−0.004649421	−0.004649417
$\kappa^{w,w}$	0.010934010	0.010934001
$\kappa^{\Omega_m,\Omega_m,\Omega_m}$	$-3.565244 \cdot 10^{-5}$	$-3.565206 \cdot 10^{-5}$
$\kappa^{\Omega_m,\Omega_m,w}$	$3.7907932 \cdot 10^{-5}$	$3.7907907 \cdot 10^{-5}$
$\kappa^{\Omega_m,w,w}$	$1.082203 \cdot 10^{-5}$	$1.082211 \cdot 10^{-5}$
$\kappa^{w,w,w}$	−0.000250971	−0.000250968
$\kappa^{\Omega_m,\Omega_m,\Omega_m,\Omega_m}$	$1.2953 \cdot 10^{-6}$	$1.2961 \cdot 10^{-6}$
$\kappa^{\Omega_m,\Omega_m,\Omega_m,w}$	$-1.23547 \cdot 10^{-6}$	$-1.23531 \cdot 10^{-6}$
$\kappa^{\Omega_m,\Omega_m,w,w}$	$7.911 \cdot 10^{-7}$	$7.906 \cdot 10^{-7}$
$\kappa^{\Omega_m,w,w,w}$	$-1.27243 \cdot 10^{-6}$	$-1.27230 \cdot 10^{-6}$
$\kappa^{w,w,w,w}$	$1.1046275 \cdot 10^{-5}$	$1.1046293 \cdot 10^{-5}$

Table 3.1: Comparison of the cumulants κ of order 1, 2, 3 and 4 of the supernova posterior distribution $p(\Omega_m, w|y)$, evaluated by MCMC sampling and by finite differencing of the logarithmic partition function $\ln Z$. The cumulants are computed from a Markov chain with 10^6 elements.

function $Z[\beta, J_\alpha, K^\alpha]$, the expectation values in the virial theorem are computed as

$$\begin{aligned}\left\langle \theta^\mu \frac{\partial \mathcal{H}}{\partial \theta^\mu} \right\rangle &= \frac{1}{Z} \int \mathrm{d}^n\theta \int \mathrm{d}^n p \, \exp(-\beta \mathcal{H}(\theta, p)) \theta^\mu \frac{\partial \mathcal{H}}{\partial \theta^\mu} \\ &= -\frac{1}{\beta Z} \int \mathrm{d}^n\theta \int \mathrm{d}^n p \, \theta^\mu \frac{\partial}{\partial \theta^\mu} \exp(-\beta \mathcal{H}(\theta, p)) \\ &= \frac{n}{\beta}.\end{aligned} \tag{3.42}$$

This calculation uses integration by parts and the fact that the trace $\partial\theta^\mu/\partial\theta^\mu = \delta^\mu_\mu = n$ gives the dimensionality n of the parameter space. Analogously,

$$\begin{aligned}\left\langle p_\mu \frac{\partial \mathcal{H}}{\partial p_\mu} \right\rangle &= \frac{1}{Z} \int \mathrm{d}^n\theta \int \mathrm{d}^n p \, \exp(-\beta \mathcal{H}(\theta, p)) p_\mu \frac{\partial \mathcal{H}}{\partial p_\mu} \\ &= -\frac{1}{\beta Z} \int \mathrm{d}^n\theta \int \mathrm{d}^n p \, p_\mu \frac{\partial}{\partial p_\mu} \exp(-\beta \mathcal{H}(\theta, p)) \\ &= \frac{n}{\beta}\end{aligned} \tag{3.43}$$

again with the trace $\partial p_\mu/\partial p_\mu = \delta^\mu_\mu = n$. Both results apply regardless of the shape of the potential Φ. We argue that the virial relation might serve as a convergence criterion for Markov chains, with a well-defined value of n for $\beta = 1$. This value is reached after equilibration or burn-in. Naturally, derivatives of the Hamilton function $\mathcal{H}(\theta, p)$ with respect to the canonical momentum are trivial, with $T(p) = \delta^{\alpha\beta} p_\alpha p_\beta/2$ implying for the derivative $\partial T/\partial p_\mu = \delta^{\alpha\beta}(\delta^\mu_\alpha p_\beta + p_\alpha \delta^\mu_\beta)/2 = p^\mu$, such that the virial expression for the momenta becomes $\langle p_\mu p^\mu \rangle = 2\langle T \rangle$. It should be noted, however, that the validity

of the virialization condition does not require a kinetic energy that is quadratic in the momenta. The natural value of $\beta = 1$ for the inverse temperature suggests that both virial expressions should equal the dimensionality in thermal equilibrium.

3.3.2 Stationarity

The thermal ensemble is stationary, since the posterior distribution $p(\theta, p|y)$ that the Markov chain samples from does not evolve with time. As demonstrated in the previous sections cumulants κ^m of the posterior $p(\theta|y)$ can be computed as

$$\kappa_\Phi^m = \frac{\partial^m}{\partial J^m} \frac{1}{\beta} \ln Z[\beta, J_\alpha, K^\alpha]\Big|_{K=0=J}. \tag{3.44}$$

The time derivative of the cumulant is given by

$$\frac{\partial}{\partial t}\kappa_\Phi^m = \frac{\partial^m}{\partial J^m} \frac{1}{N} \int \mathrm{d}^n\theta \int \mathrm{d}^n p \, \exp(-\beta\mathcal{H}(\theta, p)) \exp(\beta J_\alpha \theta^\alpha) \, J_\gamma \dot{\theta^\gamma}\Big|_{J=0} \tag{3.45}$$

as partial differentiations interchange and the Hamiltonian is constant in time. Here, we already discard the non-contributing terms involving K^α. The normalization N contains the normalizations of the likelihood, the prior and the kinetic term. Time derivatives of the parameters can be rewritten with the Hamilton equation of motion,

$$\dot{\theta}^\gamma = +\frac{\partial \mathcal{H}}{\partial p_\gamma} \tag{3.46}$$

leading to

$$\frac{\partial}{\partial t}\kappa^m = -\frac{\partial^m}{\partial J^m} \frac{1}{\beta N} \int \mathrm{d}^n\theta \int \mathrm{d}^n p \, \exp(\beta[J_\alpha \theta^\alpha]) \left[J_\gamma \frac{\partial}{\partial p_\gamma} \exp(-\beta\mathcal{H}(\theta, p)) \right] \Big|_{J=0}. \tag{3.47}$$

Integration by parts then yields a vanishing integral,

$$\frac{\partial}{\partial t}\kappa^m = \frac{\partial^m}{\partial J^m} \frac{1}{\beta N} \int \mathrm{d}^n\theta \int \mathrm{d}^n p \, \exp(-\beta\mathcal{H}(\theta, p)) \, J_\gamma \frac{\partial}{\partial p_\gamma} \exp(\beta J_\alpha \theta^\alpha) \Big|_{J=0} = 0, \tag{3.48}$$

since $\exp(\beta J_\alpha \theta^\alpha)$ does not depend on p. This implies that there is no time evolution of the configuration space cumulants.

Likewise, the momentum space cumulants are given by

$$\kappa_T^m = \frac{\partial^m}{\partial K^m} \frac{1}{\beta} \ln Z[\beta, J_\alpha, K^\alpha]\Big|_{K=0=J}. \tag{3.49}$$

Their time derivative follows analogously

$$\frac{\partial}{\partial t}\kappa_T^m = \frac{\partial^m}{\partial K^m} \frac{1}{N} \int \mathrm{d}^n\theta \int \mathrm{d}^n p \, \exp(-\beta\mathcal{H}(\theta, p)) \exp(\beta K^\alpha p_\alpha) \, K^\gamma \dot{p_\gamma}\Big|_{K=0} \tag{3.50}$$

using the other Hamilton equation of motion at this point

$$\dot{p}_\gamma = -\frac{\partial \mathcal{H}}{\partial \theta^\gamma} \tag{3.51}$$

implying

$$\frac{\partial}{\partial t}\kappa_T^m = \frac{\partial^m}{\partial K^m}\frac{1}{\beta N}\int \mathrm{d}^n\theta \int \mathrm{d}^n p\, \exp(\beta[K^\alpha p_\alpha]) \left[K^\gamma \frac{\partial}{\partial \theta^\gamma}\exp(-\beta\mathcal{H}(\theta,p))\right]\Big|_{K=0}. \quad (3.52)$$

Again, integration by parts then yields a vanishing integral,

$$\frac{\partial}{\partial t}\kappa_T^m = -\frac{\partial^m}{\partial K^m}\frac{1}{\beta N}\int \mathrm{d}^n\theta \int \mathrm{d}^n p\, \exp(-\beta\mathcal{H}(\theta,p))\, K^\gamma \frac{\partial}{\partial \theta^\gamma}\exp(\beta K^\alpha p_\alpha)\Big|_{K=0} = 0, \quad (3.53)$$

such that the cumulants become stationary. After equilibration, the Markov chain samples from a stationary posterior distribution and that the cumulants do not evolve.

3.3.3 Equipartition

In contrast to virialization, equipartition is a characteristic of thermalized systems. Virialization does not make assumptions about thermodynamic equilibrium. Calculating the expectation values of the quantities $\theta^\mu \partial_\nu \Phi$

$$\begin{aligned}\left\langle \theta^\mu \frac{\partial \Phi}{\partial \theta^\nu}\right\rangle &= \frac{1}{Z}\int \mathrm{d}^n\theta \int \mathrm{d}^n p\, \exp(-\beta\mathcal{H})\, \theta^\mu \frac{\partial \Phi}{\partial \theta^\nu} \\ &= -\frac{1}{\beta Z}\int \mathrm{d}^n\theta \int \mathrm{d}^n p\, \theta^\mu \frac{\partial}{\partial \theta^\nu}\exp(-\beta\mathcal{H}) \\ &= \frac{1}{\beta Z}\int \mathrm{d}^n\theta \int \mathrm{d}^n p\, \frac{\partial \theta^\mu}{\partial \theta^\nu}\, \exp(-\beta\mathcal{H}) \\ &= \frac{\delta^\mu_\nu}{\beta}\end{aligned} \quad (3.54)$$

and $p_\mu \partial^\nu T$

$$\begin{aligned}\left\langle p_\mu \frac{\partial T}{\partial p_\nu}\right\rangle &= \frac{1}{Z}\int \mathrm{d}^n\theta \int \mathrm{d}^n p\, \exp(-\beta\mathcal{H})\, p_\mu \frac{\partial T}{\partial p_\nu} \\ &= -\frac{1}{\beta Z}\int \mathrm{d}^n\theta \int \mathrm{d}^n p\, p_\mu \frac{\partial}{\partial p_\nu}\exp(-\beta\mathcal{H}) \\ &= \frac{1}{\beta Z}\int \mathrm{d}^n\theta \int \mathrm{d}^n p\, \frac{\partial p_\mu}{\partial p_\nu}\, \exp(-\beta\mathcal{H}) \\ &= \frac{\delta^\nu_\mu}{\beta}\end{aligned} \quad (3.55)$$

suggests that the degrees of freedom become independent of each other. Furthermore, the expectation values are equal and proportional to temperature in equilibrium. From this we define a further convergence criterion for Markov chains, for the specific value of $\beta = 1$.

Equipartition is a much stronger condition than virialization. While virialization sums over all degrees of freedom, equipartition makes a statement about the individual degrees of freedom of the system. The virialization condition follows from equipartition by summing over different degrees of freedom since

$$\begin{aligned}\left\langle \theta^\mu \frac{\partial \Phi}{\partial \theta^\mu}\right\rangle &= \sum_{\mu\nu}\left\langle \theta^\mu \frac{\partial \Phi}{\partial \theta^\nu}\right\rangle = \sum_{\mu\nu}\frac{\delta^\mu_\nu}{\beta} = \frac{n}{\beta} \\ \text{and}\quad \left\langle p_\mu \frac{\partial T}{\partial p_\mu}\right\rangle &= \sum_{\mu\nu}\left\langle p_\mu \frac{\partial T}{\partial p_\nu}\right\rangle = \sum_{\mu\nu}\frac{\delta^\nu_\mu}{\beta} = \frac{n}{\beta}.\end{aligned} \quad (3.56)$$

In addition, as the virialization condition is built as an average over the equipartition conditions, fluctuations are suppressed according to the law of large numbers and the expectation value n/β is reached faster, again indicating that virialization is the weaker criterion.

Mixed expectation values are zero as coordinates and momenta are independent in Hamiltonian mechanics,

$$\begin{aligned}\left\langle \theta^\mu \frac{\partial T}{\partial p_\nu} \right\rangle &= \frac{1}{Z} \int \mathrm{d}^n\theta \int \mathrm{d}^n p \, \exp(-\beta\mathcal{H}) \, \theta^\mu \frac{\partial T}{\partial p_\nu} \\ &= -\frac{1}{\beta Z} \int \mathrm{d}^n\theta \int \mathrm{d}^n p \, \theta^\mu \frac{\partial}{\partial p_\nu} \exp(-\beta\mathcal{H}) \\ &= \frac{1}{\beta Z} \int \mathrm{d}^n\theta \int \mathrm{d}^n p \, \frac{\partial \theta^\mu}{\partial p_\nu} \, \exp(-\beta\mathcal{H}) = 0\end{aligned} \tag{3.57}$$

and similarly,

$$\begin{aligned}\left\langle p_\mu \frac{\partial \Phi}{\partial \theta^\nu} \right\rangle &= \frac{1}{Z} \int \mathrm{d}^n\theta \int \mathrm{d}^n p \, \exp(-\beta\mathcal{H}) \, p_\mu \frac{\partial \Phi}{\partial \theta^\nu} \\ &= -\frac{1}{\beta Z} \int \mathrm{d}^n\theta \int \mathrm{d}^n p \, p^\mu \frac{\partial}{\partial \theta^\nu} \exp(-\beta\mathcal{H}) \\ &= \frac{1}{\beta Z} \int \mathrm{d}^n\theta \int \mathrm{d}^n p \, \frac{\partial p_\mu}{\partial \theta^\nu} \, \exp(-\beta\mathcal{H}) = 0.\end{aligned} \tag{3.58}$$

This illustrates that the sampling in parameter space and momentum space is independent. Again, this characteristic of thermal equilibrium can be investigated in the burn-in of Markov chains.

Gelman-Rubin criterion as a particular case

The Gelman-Rubin criterion [Gelman and Rubin, 1992, Brook and Gelman, 1997, Roberts et al., 1997] quantifies convergence in Markov chain Monte Carlo by comparing the (co)-variance generated by a single chain in its evolution with the (co)-variance of an ensemble of chains at the same instant, see section 2.3. In ergodic cases, the two averages should coincide, and if properly equilibrated, the variance does not evolve anymore.

While the Gelman-Rubin criterion [for reviews, see Brooks et al., 2011, Vats and Knudson, 2021] quantifies stationarity, it is remarkable that a criterion based on (co)-variance alone is sufficient to ensure that the sampling is representative of the posterior distribution. The physical interpretation of the Gelman-Rubin criterion, however, seems to be identical to equipartition for Gaussian distributions. Choosing a parabolic potential

$$\begin{aligned}&\Phi(\theta) = \frac{1}{2} F_{\alpha\beta} \theta^\alpha \theta^\beta \\ \rightarrow \quad &\frac{\partial \Phi}{\partial \theta^\nu} = \frac{F_{\alpha\beta}}{2} \left(\delta^\alpha_\nu \theta^\beta + \theta^\alpha \delta^\beta_\nu \right) = F_{\alpha\nu} \theta^\alpha\end{aligned} \tag{3.59}$$

allows to rewrite the covariance as

$$\begin{aligned}&F_{\alpha\nu} \langle \theta^\mu \theta^\alpha \rangle = \left\langle \theta^\mu \frac{\partial \Phi}{\partial \theta^\nu} \right\rangle = \frac{\delta^\mu_\nu}{\beta} \\ \rightarrow \quad &\langle \theta^\mu \theta^\nu \rangle = F^{\mu\nu}.\end{aligned} \tag{3.60}$$

In an equilibrated Markov chain at unit β the covariance is the inverse Fisher matrix. Monitoring the covariance using the Gelman-Rubin criterion or the equipartition condition for the corresponding degree of freedom is equivalent. The Gelman-Rubin criterion compares two variances and is formulated as a statistical test for their equality. In contrast, virialization, equipartition and thermalization make statements about an expectation value with a physically defined target value in thermal equilibrium.

While the Gelman-Rubin criterion requires a comparison between in-chain variances and a variance between chains, the virialization, equipartition and thermalization conditions can naturally be applied to a single chain. Additionally, virialization, equipartition and thermalization can be sensible even in the case of distributions that do not have a finite second moment such as the Cauchy distribution. In the case of multiple parameters, the computation of the Gelman-Rubin R requires the inversion of the covariance matrix. This slows down the computation of the convergence criterion.

3.3.4 Thermalization

Driven by physical intuition one might keep a record of the thermal energy transferred to and dissipated from the system in the sampling process. Equilibration is characterized by no net exchange of energy with the heat bath. Initializing the Markov chain close to the minimum position of the potential requires an investment of energy for equilibration. Initialization far away from the minimum results in dissipation of energy until equilibrium is reached. For an equilibrated Markov chain the average energy is determined by the evidence and the entropy of the posterior distribution. The average energy is computed as

$$\begin{aligned}\langle \mathcal{H} \rangle &= \langle T(p) \rangle_p + \frac{1}{Z(\beta)} \int \mathrm{d}^n\theta \exp\left(-\beta\Phi(\theta)\right) \Phi(\theta) \\ &= \langle T(p) \rangle_p - \frac{1}{Z(\beta)} \frac{\partial}{\partial\beta} \int \mathrm{d}^n\theta \exp\left(-\beta\Phi(\theta)\right) \\ &= \langle T(p) \rangle_p - \frac{1}{Z(\beta)} \frac{\partial}{\partial\beta} Z(\beta) \\ &= \langle T(p) \rangle_p - \frac{\partial}{\partial\beta} \ln Z(\beta).\end{aligned} \tag{3.61}$$

Comparing to the result of section 3.1.1 yields

$$\langle \mathcal{H} \rangle = \langle T(p) \rangle_p + \frac{1}{\beta} \ln Z_\Phi \left[\beta, J_\alpha = 0\right] - \frac{1}{\beta} S_\Phi(\beta). \tag{3.62}$$

The difference in normalization is lost due to the differentiation. At unit inverse temperature, the expectation value of the potential energy is given by the difference between the entropy of the posterior and the logarithm of the evidence

$$\langle \Phi \rangle = \ln p(y) - S(p(\theta|y)). \tag{3.63}$$

The kinetic term $T(p)$ is usually constructed without connection to the data. The momenta are sampled from

$$p \sim \frac{1}{N_T} \exp\left(-T(p)\right). \tag{3.64}$$

The expectation value of the kinetic term then consists of the entropy of the momenta S_T and the normalization $\langle T(p) \rangle_p = \ln N_T - S_T$. This is analogous to the structure derived

for the parameters. The expectation value of the full Hamiltonian then reads

$$\langle \mathcal{H} \rangle = \ln p(y) - S(p(\theta|y)) + \ln N_T - S_T. \tag{3.65}$$

In equilibrium this expectation value is a fixed, if often a priori inaccessible number. Since the average energy of an equilibrated Markov chain is constant the energy exchange with the heat bath fluctuates around an expectation value of zero.

It is important to notice that the exchange of thermal energy in burn-in takes place outside thermal equilibrium. This criterion is not directly linked to the thermodynamic entropy $\mathrm{d}S$. The differential entropy is defined as the reversibly exchanged heat normalized by the equilibrium temperature. However, there is no notion of temperature outside equilibrium. The criterion is attractive from a technical point of view, since keeping track of the energy while sampling is a straightforward addition to a Markov chain implementation. It also allows the definition of a convergence criterion without calculating the derivative of the potential. Consequently, the energy exchange with the heat bath can measure the convergence of conventional Metropolis-Hastings algorithms. Here the energy exchanged is measured by the change in the potential energy $\Phi(\theta) = \chi^2(y|\theta)/2 + \phi(\theta)$, equivalent to $\Delta\chi^2/2$ if the prior is neglected.

3.3.5 Numerical results

We investigate physically motivated convergence criteria for Markov chains with a Hamilton Monte Carlo algorithm. It efficiently samples microstates (p_μ, θ^ν) from the canonical partition function

$$Z[\beta, J_\alpha, K^\alpha] = \frac{1}{N_{\mathcal{H}}} \int \mathrm{d}^n\theta \int \mathrm{d}^n p \, \exp\left(-\beta \left[T(p) + \Phi(\theta)\right]\right) \exp\left(\beta \left[J_\alpha \theta^\alpha + K^\alpha p_\alpha\right]\right). \tag{3.66}$$

Here $N_{\mathcal{H}}$ combines the normalizations of the prior, the likelihood and the kinetic part. The Hamiltonian function $\mathcal{H}(p, \theta) = T(p) + \Phi(\theta)$ separates into a conventional quadratic kinetic part and a potential,

$$T(p) = \frac{1}{2m} \delta^{\mu\nu} p_\mu p_\nu \quad \text{as well as} \quad \Phi(\theta) = \frac{\chi^2(y|\theta)}{2} + \phi(\theta). \tag{3.67}$$

Expectation values of any phase space function $g(p, \theta)$ can be estimated from the samples $\left(p_\mu^{(i)}, \theta^{\nu,(i)}\right)_{i=1\ldots N}$ provided by the Markov chain

$$\begin{aligned} \langle g(p,\theta) \rangle &= \int \mathrm{d}^n\theta \int \mathrm{d}^n p \, p(\theta, p|y) g(p,\theta) \\ &= \frac{1}{Z} \int \mathrm{d}^n\theta \int \mathrm{d}^n p \, \exp\left(-\beta \mathcal{H}(p,\theta)\right) g(p,\theta) \\ &\approx \frac{1}{N} \sum_{i=1}^{N} g\left(p^{(i)}, \theta^{(i)}\right). \end{aligned} \tag{3.68}$$

For instance, equipartition conditions in the previous section would be computed as

$$\left\langle \theta^\mu \frac{\partial \mathcal{H}}{\partial \theta^\nu} \right\rangle \approx \frac{1}{N} \sum_{i=1}^{N} \theta^{\mu,(i)} \frac{\partial \Phi}{\partial \theta^\nu} \left(\theta^{(i)}\right) \tag{3.69}$$

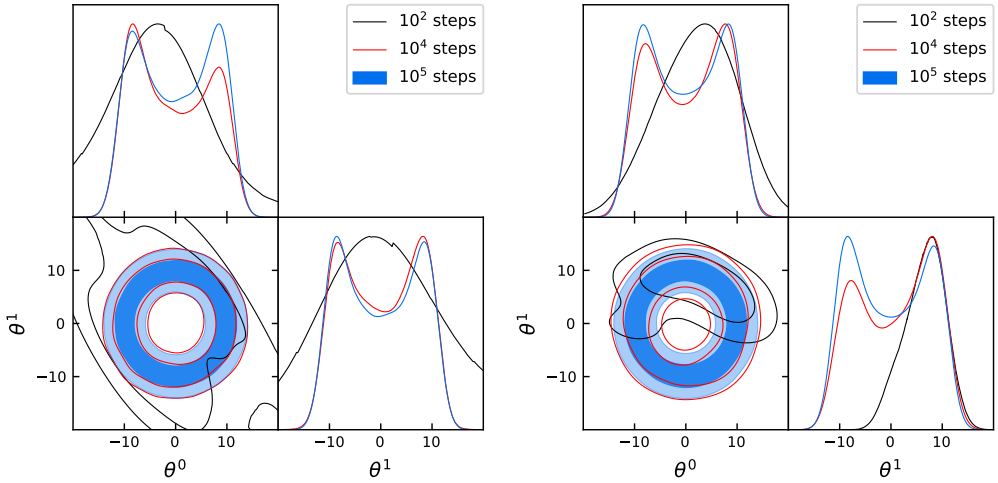

Figure 3.3: Kernel density estimates performed on the first 10^2, 10^4 and 10^5 points of an HMC chain for the toy example. For the plot on the left initial conditions for the HMC are chosen away from the maximum posterior region, while for the plot on the right one of the most probable points was chosen as the initial condition.

where the gradient $\partial\Phi/\partial\theta^\nu$ at the position $\theta^{(i)}$ can be evaluated by finite differencing. We work with an analytical expression of the gradients of Φ in the example Sect. 3.3.5 and use autodifferentiability of the physics-informed neural network implementation in Sect. 3.3.5.

First numerical experiments

To demonstrate that the derived convergence criteria perform well in practice they are applied to a two-dimensional toy example with non-Gaussian shape and a strong degeneracy. The positions, associated momenta and derivatives of the potential are obtained using a basic Hamilton Monte Carlo algorithm as described in [Neal, 2012]. The likelihood is chosen as

$$\mathcal{L}\left(\theta|R,\sigma\right) \propto \exp\left(-\frac{\left(\sqrt{\theta_\nu\theta^\nu}-R\right)^2}{2\sigma^2}\right), \tag{3.70}$$

with the analytic derivative

$$\frac{\partial}{\partial\theta^\mu}\left(-\ln\mathcal{L}\left(\theta|R,\sigma\right)\right) = \frac{\sqrt{\theta_\nu\theta^\nu}-R}{\sqrt{\theta_\rho\theta^\rho}\sigma^2}\theta_\mu. \tag{3.71}$$

The Hamilton Monte Carlo algorithm uses the derivatives of the potential to find trajectories on which new points are proposed. Estimates of the convergence criteria, eqn. (3.69) are computed on these points. Fig. 3.3 shows kernel density estimates, performed with GetDist [Lewis, 2019] on the first 10^2, 10^4 and 10^5 points of the Markov chain. This gives some intuition of how well the chain reproduces the actual posterior after accumulating a certain number of samples.

The cumulative values of the convergence criteria up to a specific step along the Markov chain are shown in Fig. 3.4. For the left column of the Figure, the initial conditions of

the Markov chain are chosen far away from the minimum of the potential, whereas the initial conditions of the right column are at the (degenerate) minimum of the potential. The top row of plots illustrates the evolution of the mixed expectation value terms. They converge to zero as more samples are drawn. This is realized surprisingly early in the evolution of the Markov chain, even before 10^2 steps are performed. In the center row partition into different degrees of freedom is illustrated. The quantities $\langle \theta^\mu \partial_\nu \Phi \rangle$ and $\langle p_\mu \partial^\nu T \rangle$, for $\mu \neq \nu$, tend towards zero as a larger amount of samples is accumulated. In these plots it is worth noting that the partition is significantly faster in the momentum degrees of freedom. This can be easily understood by recalling that the underlying distribution of the momenta is an uncorrelated normal distribution, which is sampled from directly in the HMC algorithm. The lower row shows that the virial relations, i.e. the expectation values for $\mu = \nu$, tend towards one, after a similar number of steps. While the left column illustrates the effect a long burn-in phase has on the different convergence criteria, the right column shows the effect of thermal fluctuations when the chain is started at a potential minimum. Even though we compute all expectation values cumulatively over all samples including those in the burn-in phase, a clear trend towards the thermal expectation values is seen, which can help to quantify convergence.

Lastly, Fig. 3.5 illustrates that the convergence of the Gelman-Rubin R is commensurate with the virialization conditions, in both cases of a well and badly chosen initial condition. Here we would like to emphasize that R is a test statistic akin to a t-test and helps to decide between the hypothesis that the variances along a single Markov chain and between an ensemble of independent Markov chains are identical versus the hypothesis that this is not true, at a selected confidence level. Similarly, one would quantify equality of the virialization or equipartition conditions with the thermal expectation value by formulating a similar statistical test, in this case an F-test.

Application to supernova data

As a straightforward and relevant example for non-Gaussian likelihoods, we consider constraints on the matter density Ω_m and the dark energy equation of state w from the distance redshift relation of the type Ia supernovae [Riess et al., 1998, Goobar and Leibundgut, 2011]. We impose a prior on spatial flatness and assume the equation of state to be constant in time. Constraints are derived from the Union2.1-data set [Suzuki et al., 2012, Amanullah et al., 2010, Kowalski et al., 2008]. The FLRW-distance modulus $y(z)$ as a function of redshift z is given by

$$y(z|\Omega_m, w) = 10 + 5 \log \left((1+z)\, \chi_H \int_0^z \mathrm{d}z' \, \frac{1}{\sqrt{\Omega_m (1+z')^3 + (1-\Omega_m)(1+z')^{3(1+w)}}} \right). \tag{3.72}$$

Constructing the likelihood for the two parameters Ω_m and w for Gaussian errors σ_i in the distance moduli y_i yields the same simplified expression as in eqn. 3.38. This likelihood is implemented in a Hamilton Monte Carlo sampler, with a uniform prior $\pi(\theta)$ for simplicity. To speed up the computations, we employ physics-informed neural networks (PINN) [Raissi et al., 2017]. Details on the emulation of distance moduli using a PINN are discussed in section 5.2. For this work we used a dense neural network with three hidden layers and a width of 50 neurons. As the model prediction $y_i(z_i|\Omega_m, w)$ is given as an explicit function, we use its automatic differentiation functionality to derive the gradients of $\chi^2(y|\Omega_m, w)$ needed in Hamilton Monte Carlo sampling.

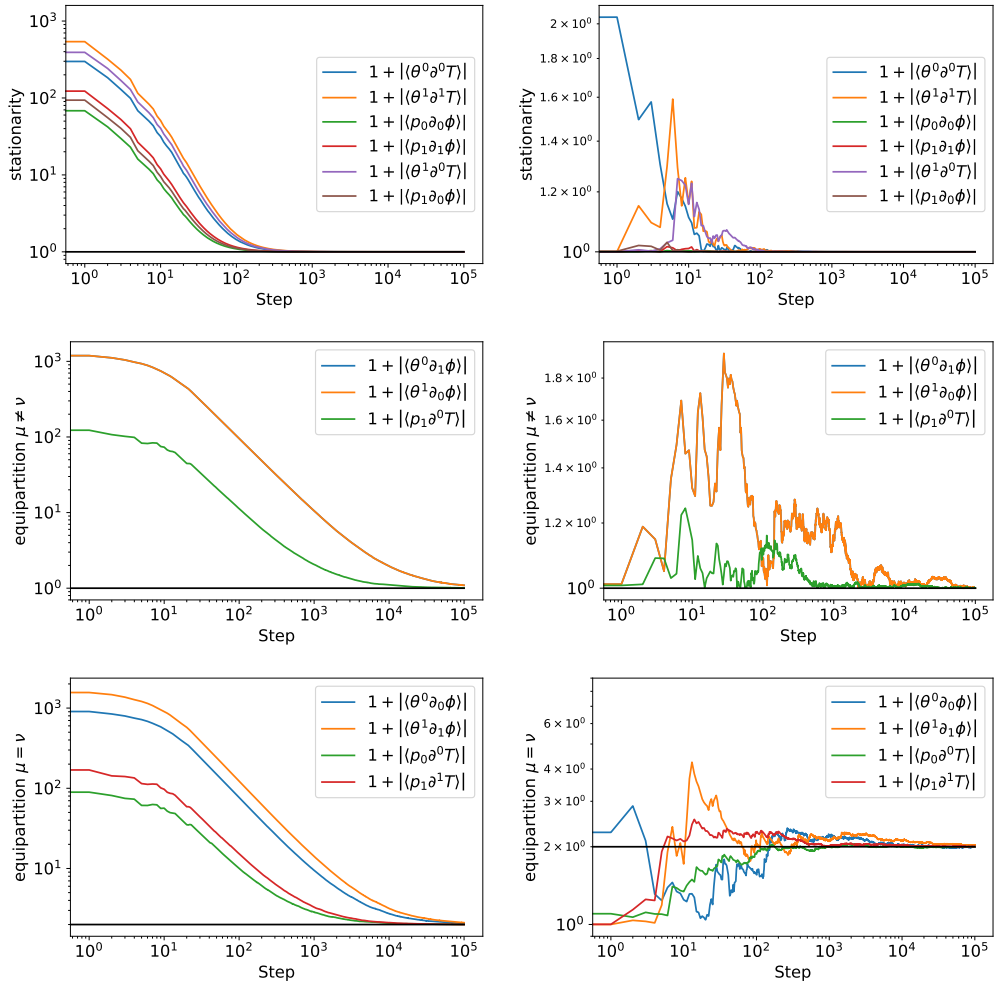

Figure 3.4: Progression plots of the stationarity condition and the equipartition of the different degrees of freedom in an HMC chain sampling the toy example. For the plots on the left initial conditions for the HMC chain are chosen away from the maximum posterior region, while for the right column the maximum of the posterior was chosen as the initial condition.

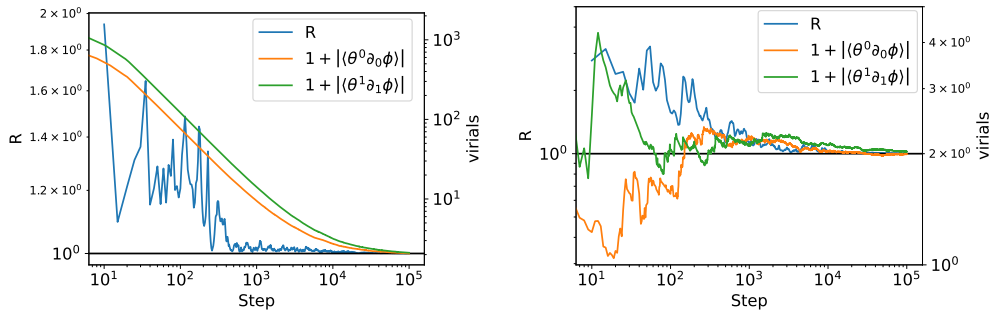

Figure 3.5: Comparison between the virialization conditions and the Gelman-Rubin criterion R. For the ensemble averaging in the determination of the Gelman-Rubin criterion the Markov chain was split into 10 batches. For the plot on the left initial conditions for the HMC are chosen away from the maximum posterior region, while for the plot on the right one of the most probable points was chosen as the initial condition.

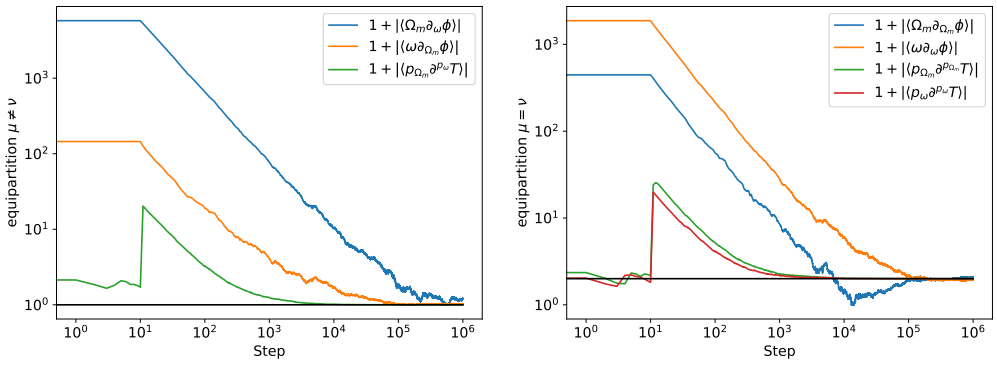

Figure 3.6: Application of the equipartition criterion to Hamilton Markov chain Monte Carlo sampling the supernova likelihood. The left plot shows the partition into different degrees of freedom while the plot on the right shows that they are equipartitioned.

The convergence criteria discussed in the previous sections are applied to the PINN-enhanced supernova likelihoods in Fig. 3.6: There is a clear trend towards the values expected for thermal equilibrium, with a scaling $\propto \text{step}^{-1}$ for the cumulatively computed values.

Fig. 3.7 shows the average energy $\mathcal{H}(\theta, p)$ in the HMC system. For the left plot, energies are averaged over 10^2 steps each and then, all 10^4 batches are plotted successively. This allows to see the thermal fluctuations of these batch averages around the overall average value defined by the entire chain. The right plot depicts the cumulative average energy of the Markov chain. For an equilibrated Markov chain the average of the energy can be approximated under the assumption that the data points are uncorrelated and that the model fits the data within the data variance σ^2 as

$$\begin{aligned}\langle \mathcal{H} \rangle &= \langle T(p) \rangle_p + \langle \Phi(\theta) \rangle_\theta \\ &= 1 + \left\langle \frac{1}{2} \sum_{j=1}^{D} \left(\frac{y_j - y(z_j|\Omega_m, w)}{\sigma_j} \right)^2 \right\rangle_{\Omega_m, w} \\ &\approx 1 + \frac{D}{2}.\end{aligned} \tag{3.73}$$

Where D is the number of data points. The average potential energy approximately equals half the number of data points. This approximation disregards the correlations between the data points, which explains why the numerical value of $\langle \mathcal{H} \rangle$ falls short of half the number of actual data points, which is 290.

3.4 Summary and discussion

Partition functions are constructed from the likelihood and prior by introducing an inverse temperature β and carrying out a Laplace transform from the model parameters θ^α to the sources J_α. In the case of HMC the momenta p_β are also transformed to sources K^γ. Evaluating at $J_\alpha = 0$, $K^\gamma = 0$ and $\beta = 1$ recovers the Bayesian evidence.

Most of the analytical calculations are based on the Helmholtz free energy $F = -\frac{1}{\beta} \ln Z$, which is introduced based on the partition function Z. This definition allows to recover

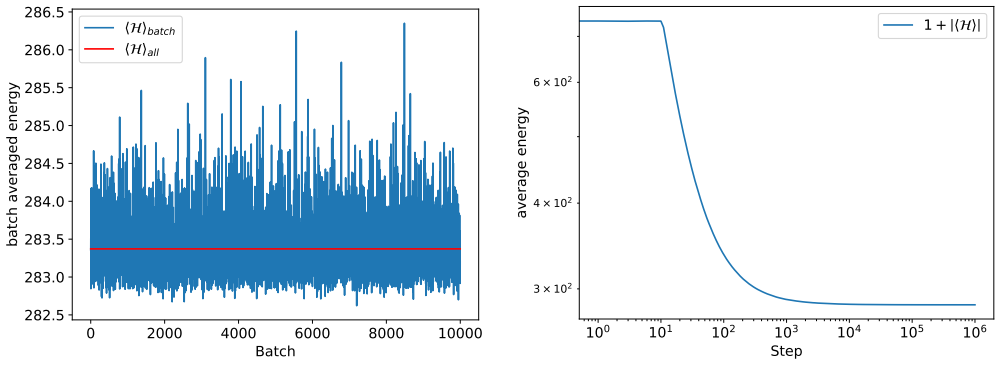

Figure 3.7: Application of the energy transfer convergence criteria to Hamilton Markov chain Monte Carlo sampling the supernova likelihood. The plot on the left shows the averaged energy over 100 steps each, while the plot on the right shows the cumulative average at each step.

the Shannon entropy in analogy to statistical physics as $S(\beta) = \beta^2 \frac{\partial F}{\partial \beta}$. It can also be used as a cumulant-generating function by taking derivatives with respect to the sources and evaluating at $J_\alpha = 0$, $K^\gamma = 0$ and $\beta = 1$.

For linear models, the integrand of the partition function is of Gaussian shape in the parameters and has an analytical solution. The sequence of cumulants truncates at second order. Thus, the posterior is a Gaussian distribution. While the first cumulant becomes equal to the true model parameters due to the Gauß-Markov theorem, the second cumulant recovers the inverse Fisher matrix. For non-linear models, the partition function factorizes into a Gaussian and a non-Gaussian term. In the case of weak non-Gaussianities, the corresponding term can be Taylor expanded leading to an analytical expression based on the multivariate Hermite polynomials. This analytical expression is successfully applied to a toy model. The cumulants are computed numerically for supernova data up to fourth order via finite differencing.

A Markov chain is considered converged when the samples θ allow the computation of any expectation value of a function $g(\theta)$ through

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} g(\theta^{(i)}) = \int \mathrm{d}\theta \, p(\theta|y) g(\theta). \tag{3.74}$$

The Gelman-Rubin criterion probes this relation for the choice $g(\theta) = \theta^2$. The canonical partition function in combination with the analogy to statistical physics provides quantitative methods to characterize thermal equilibrium.

Any bounded mechanical system satisfies the virialization condition $\langle \theta^\mu \partial\Phi/\partial\theta^\mu \rangle = \langle p_\mu \partial T/\partial p_\mu \rangle$. For an HMC chain in thermal equilibrium, the expectation values can be evaluated as n/β, giving a clearly defined value that can be checked against expectation values computed from the Markov chain. In contrast to this parameter averaged criterion the equipartition criterion makes a statement about the individual degrees of freedom $\langle \theta^\mu \partial\Phi/\partial\theta^\nu \rangle \propto \delta^\mu_\nu$ and $\langle p_\mu \partial T/\partial p_\nu \rangle \propto \delta^\nu_\mu$. Mixed derivatives $\langle p_\mu \partial\Phi/\partial\theta^\nu \rangle$ and $\langle p_\mu \partial T/\partial\theta^\nu \rangle$ evaluate to zero in thermal equilibrium.

The criteria in the previous paragraph require a notion of the gradient of the potential Φ or the kinetic energy. Thermal equilibrium can also be characterized as no net energy exchange with the heat bath. This can be evaluated from the average energy

difference $\langle \Delta \mathcal{H} \rangle$ between Markov chain steps. This suggests that the average energy of an equilibrated Markov chain $\langle \mathcal{H} \rangle = \langle T \rangle + \langle \Phi \rangle$ is constant. From the partition function we obtain $\langle \Phi \rangle = \ln p(y) - S(p(\theta|y))$ while the average kinetic energy is determined by the choice of the kinetic term in an HMC algorithm.

We demonstrate the viability of the virialization, equipartition and thermalization conditions as convergence criteria using a two-dimensional toy example. The numerical approximations of the thermodynamic criteria approach their target values as the Markov chain converges towards the true distribution. During burn-in the evolution of the virials shows similar properties to the Gelman-Rubin criterion. As a physical application the convergence criteria are applied to the inference of the matter density Ω_m and dark-energy equation of state parameter w from the magnitude redshift relation of the type Ia supernovae.

Part II

Learning Cosmological Functions

4 A Short History of the Universe

This introduction to the Geometry and dynamics of the Universe and the introduction to inflation are strongly influenced by [Baumann, 2012]. In all calculations in this chapter we set the speed of light $c = 1$ as well as the gravitational constant $8\pi G = 1$. Unless otherwise specified we work under the assumption of ΛCDM.

This chapter gives a brief introduction to the geometry and dynamics governing the evolution of the Universe in section 4.1. Section 4.2 introduces inflation and the differential equations governing both the evolution of the inflaton field and the evolution of its perturbations. The primordial power spectrum of these perturbations at the end of inflation is linked to observations in section 4.3. Section 4.4 gives a brief introduction to the dark energy driven expansion of the current Universe and the supernovae observations used to measure it. Finally, section 4.5 introduces the aspects of machine learning needed in the later chapters.

4.1 Geometry and dynamics

On large scales the universe can be described using a Friedmann-Lemaitre-Robertson-Walker (FLRW) metric

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + a^2(t)\left[\frac{\mathrm{d}r^2}{1-kr^2} + r^2\mathrm{d}\Omega^2\right]. \tag{4.1}$$

Here, a is the scale factor and $k \in \{-1, 0, 1\}$ characterizes the curvature. The rescaling symmetry

$$a \to \lambda a, \quad r \to \frac{r}{\lambda}, \quad k \to \lambda^2 k \tag{4.2}$$

is used to fix the value of the scale factor today to $a(t_0) = 1$. By defining comoving coordinates χ through

$$r^2 = \Phi_k(\chi^2) = \begin{cases} \sinh^2\chi & k = -1 \\ \chi^2 & k = 0 \\ \sin^2\chi & k = +1 \end{cases} \tag{4.3}$$

and conformal time η such that $\mathrm{d}t = a(\eta)\mathrm{d}\eta$ the FLRW metric is rewritten as

$$\mathrm{d}s^2 = a(\eta)^2\left(-\mathrm{d}\eta^2 + \mathrm{d}\chi^2 + \Phi_k(\chi^2)\right). \tag{4.4}$$

This defines the underlying geometry for the physics problems considered in this work.

The dynamics of the scale factor as well as any cosmological fields are governed by the Einstein equation

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = 8\pi G T_{\mu\nu}. \tag{4.5}$$

Here, Λ represents the cosmological constant. The Einstein tensor $G_{\mu\nu}$ is governed by spacetime curvature while the stress-energy tensor $T_{\mu\nu}$ is determined by the matter content of the Universe. The stress-energy tensor is conserved in the sense $\nabla_\mu T^\mu_\nu = 0$. Under the assumption that the matter content behaves like an ideal fluid the stress-energy tensor can be completely described with its energy density ρ and pressure P in the fluid rest frame. The FLRW metric is constructed under the assumption of isotropy and homogeneity. Combining this assumption with both the continuity equation and Einstein equation for a FLRW metric yields the Friedmann equations

$$\begin{aligned} H^2 &= \frac{\rho}{3} - \frac{k}{a^2} + \frac{\Lambda}{3} \\ \frac{\ddot{a}}{a} &= -\frac{1}{6}(\rho + 3P) + \frac{\Lambda}{3}. \end{aligned} \tag{4.6}$$

Here, the Hubble function is defined as $H = \frac{\dot{a}}{a}$.

This set of equations governs the history of cosmic expansion. In the inflationary paradigm the universe undergoes a period of rapid expansion immediately after the Big Bang. This period of inflation is described in more detail in section 4.2. After the end of inflation matter in the Universe consists of Standard Model particles. The scaling behavior of their energy densities can be derived from the Friedmann equations as

$$\rho \propto a^{-3(1+w)}. \tag{4.7}$$

Here, we assumed a constant equation of state $w = \frac{P}{\rho}$. Cosmology usually distinguishes between radiation with an equation of state is $w = 1/3$, matter with $w = 0$ and dark energy. In ΛCDM the cosmological constant Λ plays the role of dark energy. It can be interpreted as a fluid with $\rho = \Lambda$, $P = -\Lambda$. This leads to an equation of state of $w = -1$. Modifications of ΛCDM often propose different models for dark energy [Wetterich, 1988, Ratra and Peebles, 1988, Linder, 2008, Tsujikawa, 2013, Mortonson et al., 2013]. This can lead to a scale factor dependent dark energy equation of state.

The scaling behavior of the different cosmological fluids can be used to rewrite the first Friedmann equation (4.6) as

$$\frac{H^2}{H_0^2} = \Omega_\gamma a^{-4} + \Omega_m a^{-3} + \Omega_k a^{-2} + \Omega_\Lambda. \tag{4.8}$$

The index 0 denotes quantities at present time. The energy densities Ω are computed relative to the value of the Hubble function today as $\Omega_i = \frac{\rho_{i,0}}{3H_0^2}$, $\Omega_k = -\frac{k}{H_0^2}$. Observations of the CMB [Aghanim et al., 2020b] suggest that the Universe is flat $\Omega_k = 0.001 \pm 0.002$. The combination of the first Friedmann equation (4.6) with a positive H_0 suggests that the scale factor is always increasing. The formulation in eqn. (4.8) allows us to identify periods in which different components of the cosmological fluid dominate the dynamics.

In the early universe the scale factor is small and the evolution of the Hubble function is radiation-dominated. The evolution of the Hubble function is determined by the first term in equation (4.8). During this period gravitational attraction and pressure compete to give rise to acoustic oscillations in the primordial plasma. As the Universe continues to grow the second term in eqn. (4.8) starts to dominate the evolution and the Universe enters matter domination. About 370000 years after the Big Bang protons and electrons combine into Hydrogen and Helium (and some Lithium), in a process called recombination. The Universe transitions from ionized to neutral. Photons emitted

during recombination do not scatter off charged particles anymore and can be observed today as the CMB. As the Universe continues to expand the scale factor approaches its current value $a_0 = 1$ at these values and especially at $a > 1$ in the future the last term in eqn. (4.8) begins to dominate. The expansion becomes cosmological constant, or dark energy dominated. During this period the Friedmann equations (4.6) suggest that there will be an accelerated expansion in contrast to the decelerating expansion in the previous two periods.

4.2 Inflation

Cosmic inflation was introduced in [Sato, 1981, Starobinsky, 1980, Guth, 1981] to cure the standard cosmological model (ΛCDM) of the horizon and flatness problems.

4.2.1 Horizon problem

In a universe starting with the Big Bang, governed by the geometry and dynamics introduced in the previous section there is a maximum comoving distance that light can propagate. This distance is called the particle horizon and can be defined as

$$\chi_p(\eta) = \eta - \eta_i = \int_{t_i}^{t} \frac{\mathrm{d}t}{a(t)}. \tag{4.9}$$

Here the subscript i denotes a quantity at the beginning of the universe $a(t_i) = a_i = 0$. In terms of the comoving Hubble radius $(aH)^{-1}$ the particle horizon can be computed as

$$\chi_p = \int_{a_i}^{a} \frac{1}{\dot{a}} \frac{\mathrm{d}a}{a} = \int_{\ln a_i}^{\ln a} (aH)^{-1} \,\mathrm{d}\ln a. \tag{4.10}$$

For a universe dominated by a fluid with a constant equation of state w, the first Friedmann equation (4.6) implies that the comoving horizon can be expressed as

$$(aH)^{-1} = H_0^{-1} a^{\frac{1}{2}(1+3w)}. \tag{4.11}$$

The particle horizon is then determined by

$$\begin{aligned} \chi_p &= \frac{2H_0^{-1}}{1+3w} \left(a^{\frac{1}{2}(1+3w)} - a_i^{\frac{1}{2}(1+3w)} \right) \\ &= \frac{2}{1+3w} \left((aH)^{-1} - (aH)_i^{-1} \right). \end{aligned} \tag{4.12}$$

For radiation domination, with $w = \frac{1}{3}$, this falls back to $\chi_p = aH_0^{-1} = (aH)^{-1}$. In the standard Big Bang picture this means that two points at the end of radiation domination can not have communicated if their comoving distance from each other is greater than the comoving Hubble radius.

The horizon problem arises from the observation that the Cosmic Microwave Background (CMB) is almost perfectly isotropic. Points all across the sky are at almost exactly the same temperature. This implies that the observable points in the CMB have been in causal contact. However, the particle horizon around each of these points only covers a fraction of the sky.

In the inflationary paradigm the horizon problem is solved by introducing a period of decreasing Hubble radius $\frac{\mathrm{d}}{\mathrm{d}t}(aH)^{-1} < 0$ in the early universe. According to the calculation in eqn. (4.12) this translates to a fluid with $(1+3w) < 0$. The particle horizon is then dominated by the early universe where $a_i \to 0$. This period of shrinking Hubble radius, or equivalently exponential expansion, must last at least until all of the observable universe was in causal contact at early times.

4.2.2 Background evolution of the inflaton

For this thesis we restrict the investigations to single-field inflation. The inflaton φ is defined as a scalar field that only couples to gravity. Assuming a minimal coupling, the inflaton action can be formulated as

$$S_\varphi = \int \mathrm{d}^4x\sqrt{-g}\left(\frac{1}{2}g^{\mu\nu}\nabla_\mu\varphi\nabla_\nu\varphi - V(\varphi)\right). \tag{4.13}$$

Here, $g_{\mu\nu}$ denotes the metric during inflation. The inflaton is described with canonical kinetic term and a potential $V(\varphi)$ determining its self-interaction. Under the assumptions of the FLRW metric, eqn. (4.4) and a homogeneous inflaton field, the density and pressure of the inflaton fluid are found as [Baumann, 2012]

$$\begin{aligned}\rho_\varphi &= \frac{1}{2}\dot{\varphi}^2 + V(\varphi)\\ P_\varphi &= \frac{1}{2}\dot{\varphi}^2 - V(\varphi).\end{aligned} \tag{4.14}$$

To induce a shrinking Hubble radius the equation of state parameter

$$w_\varphi = \frac{\frac{1}{2}\dot{\varphi}^2 - V(\varphi)}{\frac{1}{2}\dot{\varphi}^2 + V(\varphi)}, \tag{4.15}$$

must fulfill $w_\varphi < -\frac{1}{3}$.

The background equations of motion are obtained through varying the inflaton field in the action. In proper time they are described by

$$\ddot{\varphi} + 3H\dot{\varphi} = -\frac{\mathrm{d}V}{\mathrm{d}\varphi}. \tag{4.16}$$

The dynamics of the Hubble function is described by the Friedmann equation as

$$H^2 = \frac{\rho_\varphi}{3} = \frac{1}{3}\left(\frac{1}{2}\dot{\varphi}^2 + V(\varphi)\right). \tag{4.17}$$

This description determines the dynamics of the inflaton field in the absence of other matter and under the assumptions of homogeneity and isotropy in both the metric and the inflaton field.

Slow-Roll inflation

The equation of state parameter, eqn. (4.15) together with the restriction $w_\varphi < -\frac{1}{3}$ suggests that the kinetic energy $\frac{1}{2}\dot{\varphi}^2$ must be at least twice as large as the potential $V(\varphi)$ to sustain inflation. To shrink the Hubble radius enough to explain the homogeneity

of the CMB inflation needs to be sustained for some time. This admits to the intuitive picture of the inflaton slowly rolling down a potential until its kinetic energy exceeds the potential energy and inflation ends. The ratio of kinetic and potential energy can be quantified with the first slow-roll parameter

$$\epsilon = -\frac{\dot{H}}{H^2} = 3\frac{\frac{1}{2}\dot{\varphi}^2}{\frac{1}{2}\dot{\varphi}^2 + V(\varphi)}. \tag{4.18}$$

During inflation $\epsilon < 1$, inflation ends when this condition is not fulfilled. To sustain inflation the acceleration of the scalar field

$$\delta = -\frac{\ddot{\varphi}}{H\dot{\varphi}} \tag{4.19}$$

has to remain small as well.

Slow-roll inflation is realized whenever $\epsilon, \delta \ll 1$. The equations of motion simplify since the potential energy dominates over the kinetic energy $V(\varphi) \gg \frac{1}{2}\dot{\varphi}^2$ and the second condition implies $|\ddot{\varphi}| \ll |H\dot{\varphi}|$. The simplified background equations of motion read

$$\begin{aligned} 3H\dot{\varphi} &= -\frac{\mathrm{d}V}{\mathrm{d}\varphi} \\ H^2 &= \frac{V}{3}. \end{aligned} \tag{4.20}$$

Reheating

For the standard cosmic history to resume inflation needs to end. In the numerical calculations we remain agnostic to this process. In the picture of single-field inflation the energy density in the inflaton field must be transformed into standard model degrees of freedom. This process is referred to as reheating. After reheating the universe is radiation-dominated and highly ionized. While the physics during reheating is largely unknown more detailed descriptions can be found in [Baumann, 2012] and [Riotto, 2003].

4.2.3 Perturbations

The previous section describes the evolution of the inflaton field in a perfectly homogeneous and isotropic universe. This explains the homogeneity of the CMB. However, the CMB is only homogeneous up to fluctuations at all length scales of order 10^{-5} [Bennett et al., 1992]. These can be seeded in the inflaton and the metric during inflation by allowing for a small deviation from homogeneity and isotropy. The fluctuations can be characterized in a gauge invariant fashion through the comoving curvature perturbations $\mathcal{R}$. In Fourier space the modes $\mathcal{R}_k$ of the fluctuations are associated to a comoving wavenumber k. For adiabatic expansion the curvature perturbation modes are constant on superhorizon scales $k \ll aH$. While fluctuations are created at all length scales k only modes inside the horizon at their creation evolve during inflation. All relevant modes are created on subhorizon scales $k \gg aH$. While their comoving wavenumber stays constant the comoving Hubble radius $(aH)^{-1}$ decreases during inflation. The curvature modes evolve until they exit the horizon $k < aH$ and freeze.

Immediately after inflation ends the observable modes are outside the horizon and remain unchanged by reheating. As the universe returns to the standard Big Bang scenario

the comoving horizon begins to grow until the observable modes reenter the horizon when $k > aH$. They continue evolving at this point. This allows to probe the physics of inflation from observables such as the CMB or Galaxy surveys without knowledge about the physics of reheating.

Similar to the differential equations for the background evolution the differential equations governing the evolution of the curvature modes can be derived starting from an action defined on the background and the perturbations [Mukhanov et al., 1992, Maldacena, 2003]. The differential equation governing the conformal time evolution of the Mukhanov-Sasaki potentials

$$u_k = z\mathcal{R}_k,\ z = a\frac{\dot{\varphi}}{H}, \tag{4.21}$$

[Baumann, 2012] is given by

$$\partial_\eta^2 u_k + \left[k^2 - \frac{\partial_\eta^2 z}{z}\right] u_k = 0,\ z = \frac{\partial_\eta \varphi}{H}. \tag{4.22}$$

An analogous derivation is performed for the tensor perturbations

$$h_{ij} = \int \frac{\mathrm{d}^3 k}{(2\pi)^3} \sum_{s=+,\times} \epsilon_{ij}^s h_k^s(\eta) e^{ikx}. \tag{4.23}$$

Here $+, \times$ represent two different polarizations, while the coefficients fulfill $\epsilon_{ii} = k^i \epsilon_{ij} = 0$ and $\epsilon_{ij}^s \epsilon_{ij}^{s'} = 2\delta^{ss'}$. Similar to the scalar perturbations a differential equation for the Mukhanov-Sasaki potentials $v_k^s = \frac{a}{2} h_k^s$ is derived as

$$\partial_\eta^2 v_k + \left[k^2 - \frac{\partial_\eta^2 a}{a}\right] v_k = 0. \tag{4.24}$$

In this formulation polarization indices are omitted.

The Mukhanov-Sasaki potentials are initialized at $\eta \to -\infty$ where $k \gg \frac{\partial_\eta^2 a}{a}, \frac{\partial_\eta^2 z}{z}$. In this limit the differential equations simplify to a harmonic oscillator with the solution

$$\lim_{\eta \to -\infty} u_k(\eta) = \frac{e^{-ik\eta}}{\sqrt{2k}}. \tag{4.25}$$

Note that the differential equations are independent under constant phase shifts in u_k or v_k. This freedom can be used to fix the phase of the variables at the initial time to 0. In numerical calculations conformal time translation invariance is used to fix the initial conditions using the Bunch-Davies vacuum [Chernikov and Tagirov, 1968, Bunch and Davies, 1978]

$$\begin{aligned} \mathrm{Re}(u_k) = \mathrm{Re}(v_k) &= \frac{1}{\sqrt{2k}} \\ \mathrm{Im}(u_k) = \mathrm{Im}(v_k) &= 0 \\ \mathrm{Re}(\partial_\eta u_k) = \mathrm{Re}(\partial_\eta v_k) &= 0 \\ \mathrm{Im}(\partial_\eta u_k) = \mathrm{Im}(\partial_\eta v_k) &= \frac{-k}{\sqrt{2k}}. \end{aligned} \tag{4.26}$$

The assumption on the differential equation is justified when $k \gg aH$. In numerical calculations this can be enforced by setting these initial conditions at $k = aH/100$.

There are tools to compute the evolution of primordial fluctuation such as CLASS [Blas

et al., 2011]. Figure 6.1 in section 6.1.1 gives some intuition on how the curvature modes u_k evolve with conformal time.

Primordial power spectrum

The power spectrum of the curvature $\mathcal{R}$ provides a statistical measure of the primordial perturbations. It is defined as the ensemble average

$$\langle \mathcal{R}_k \mathcal{R}_{k'} \rangle = (2\pi)^3 \delta(k + k') \Delta^2_{\mathcal{R}}. \tag{4.27}$$

The scalar power spectrum is defined as

$$\mathcal{P}_{\mathcal{R}} = \frac{k^3}{2\pi^2} \Delta^2_{\mathcal{R}}, \tag{4.28}$$

and can also be obtained from the Mukhanov-Sasaki potentials at the end of inflation as [Powell and Kinney, 2007]

$$\mathcal{P}_{\mathcal{R}} = \frac{k^3}{2\pi^2} \left| \frac{u_k}{z} \right|^2. \tag{4.29}$$

In ΛCDM the primordial scalar power spectrum is described through its scale dependence. The scalar spectral index is defined as

$$n_s - 1 \equiv \frac{\mathrm{d} \ln \mathcal{P}_{\mathcal{R}}}{\mathrm{d} \ln k}, \tag{4.30}$$

while its scale dependence is in turn characterized through

$$\alpha_s \equiv \frac{\mathrm{d}^2 \ln \mathcal{P}_{\mathcal{R}}}{\mathrm{d} \ln k^2}. \tag{4.31}$$

In proximity of some pivot scale k_* the logarithmic primordial scalar power spectrum is approximated by

$$\ln \mathcal{P}_{\mathcal{R}}(k) = \ln A_s(k_*) + (n_s(k_*) - 1) \ln \frac{k}{k_*} + \alpha_s(k_*) \left(\ln \frac{k}{k_*} \right)^2. \tag{4.32}$$

For a better approximation this Taylor series can be extended to higher order in logarithmic comoving wavenumber. In ΛCDM the primordial power spectrum is characterized by the amplitude of the primordial power spectrum at the pivot scale A_s and the scalar spectral index n_s.

The primordial tensor power spectrum can similarly be defined based on the ensemble average

$$\langle h_k h_{k'} \rangle = (2\pi)^3 \delta(k + k') \Delta^2_h, \tag{4.33}$$

as

$$\mathcal{P}_t = 2\mathcal{P}_h = 2 \frac{k^3}{2\pi^2} \Delta^2_h. \tag{4.34}$$

Here, the factor two comes from including both polarization modes. Again the tensor power spectrum can be computed from the Mukhanov-Sasaki variables at the end of inflation as [Powell and Kinney, 2007]

$$\mathcal{P}_t = \frac{32k^3}{\pi} \left| \frac{v_k}{a} \right|^2. \tag{4.35}$$

Its scale dependence is characterized by the tensor spectral index

$$n_t \equiv \frac{\mathrm{d}\ln \mathcal{P}_t}{\mathrm{d}\ln k}. \tag{4.36}$$

The amplitude of the tensor power spectrum is typically defined relative to the scalar power spectrum through the tensor-to-scalar ratio

$$r \equiv \frac{\mathcal{P}_t(k_*)}{\mathcal{P}_\mathcal{R}(k_*)}. \tag{4.37}$$

4.3 Cosmic Microwave Background

After inflation ends, and the subsequent reheating, the universe is radiation-dominated. During this period highly energetic photons ionize the matter content of the universe. As it expands the number density of photons with energies above the hydrogen ionization energy drops below the baryon density of the universe. Electrons and protons combine for the first time. This process is known as recombination. The free electron density drops sharply and the photon mean free path grows large. At redshift $z_{dec} \sim 1100$ photons decouple from the electrons and the universe becomes transparent [Durrer, 2020]. They travel until redshift $z = 0$ and constitute the Cosmic Microwave Background observed today. This is one of the earliest observables with a signature of the primordial power spectrum.

The CMB has a thermal black body spectrum at a temperature of $T_0 \approx 2.73\mathrm{K}$ [Fixsen, 2009] and is almost perfectly isotropic. The anisotropies are of order $\Delta T/T \sim 10^{-5}$ and directly result from the curvature perturbations at the time of photon decoupling. The power spectrum of the anisotropies is shaped by the physics in the early universe as described in the previous section. Additionally, modes that reenter the horizon before recombination undergo an evolution in the radiation-dominated universe. A prominent effect of this on both the angular power spectra of the CMB and the matter power spectrum at later times are Baryon Acoustic Oscillations (BAOs).

Curvature fluctuations sourced during the primordial universe lead to anisotropies in the radiation-dominated universe. While the dark matter in the universe follows the curvature fluctuations the equation of state parameter of radiation suggests a non-vanishing effect from photon pressure. During radiation domination baryonic matter is coupled to photons leading to sound waves of photons and baryons traveling away from the dark matter overdensities. At recombination, the origin of the CMB, photons decouple. They leave a shell of baryonic matter at a comoving distance of about 150 Mpc away from the dark matter overdensities. This overdense shell in real space is imprinted on the angular power spectra of the CMB in the structure of its peaks [Aghanim et al., 2020a]. It can also be observed in the matter power spectrum reconstructed from Galaxy surveys [Paillas et al., 2024].

Angular power spectra

Light from the CMB reaches us from all directions $\hat{n}$ in the sky from an approximately isotropic sphere around us. The CMB photons are released at similar redshift giving the sphere a small but finite thickness. The anisotropies in this sphere can be analyzed

through a harmonic expansion. For the temperature anisotropies $\Theta(\hat{n}) = \frac{\Delta T(\hat{n})}{T_0}$ the expansion reads

$$\Theta(\hat{n}) = \sum_{\ell m} a_{\ell m} Y_{\ell m}(\hat{n}). \tag{4.38}$$

Here $Y_{\ell m}$ are the spherical harmonics. The angular power spectrum of the temperature anisotropies can be reconstructed from the expansion coefficients

$$C_\ell^{TT} = \frac{1}{2\ell+1} \sum_m \langle a_{\ell m}^* a_{\ell m} \rangle. \tag{4.39}$$

In addition to the anisotropies in temperature, the CMB photons also exhibit an anisotropy in their polarizations. For an introduction to CMB polarizations see [Hu and White, 1997, Baumann et al., 2009]. The polarizations arise during recombination from electron-photon scattering. Unpolarized photons lead to a linear polarization of the scattered photons in the plane orthogonal to the line of sight. When the radiation field influencing the electron is isotropic the polarizations cancel out. However, the incoming radiation is dependent on the temperature anisotropies and can have a quadrupole component leading to a linear polarization. The polarization anisotropies are characterized by the curl-free modes E and the divergence-free modes B defined as

$$E(\hat{n}) = \sum_{\ell m} a_{E,\ell m} Y_{\ell m}(\hat{n}), \quad B(\hat{n}) = \sum_{\ell m} a_{B,\ell m} Y_{\ell m}(\hat{n}). \tag{4.40}$$

The coefficients follow from a decomposition of the intensity tensor into Stokes parameters and subsequent expansion into tensor spherical harmonics [Baumann et al., 2009]. The expansion of the desired modes is then found from their properties. The angular power spectra of the polarization anisotropies and the correlations between different anisotropies are defined as

$$C_\ell^{XY} = \frac{1}{2\ell+1} \sum_m \langle a_{X,\ell m}^* a_{Y,\ell m} \rangle, \quad X, Y \in \{T, E, B\}. \tag{4.41}$$

Modern measurements of the angular power spectra can be found in [Akrami et al., 2020]. Figure 4.1 depicts the angular CMB spectra of the CMB as measured by the Planck satellite. In this plot the TT and TE spectra are transformed to $D_\ell^{TX} = \frac{\ell(\ell+1)}{2\pi} C_\ell^{TX}$ with $X \in T, E$ to make the BAO peaks more visible.

To find constraints on the primordial power spectrum $\mathcal{P}(k)$ the angular power spectra are simulated using Boltzmann codes such as CLASS and CAMB [Blas et al., 2011, Lewis et al., 2000]. The evolution of the curvature and tensor modes after horizon reentry is modeled by transfer functions $\Delta_{X,\ell}(k)$ for each of the different observable modes $X \in \{T, E, B\}$. The angular power spectra are obtained from the primordial power spectra through the integrations

$$\begin{aligned} C_\ell^{TT} &= \frac{2}{\pi} \int k^2 \mathrm{d}k \mathcal{P}_\mathcal{R}(k) \Delta_{T,\ell}^2(k) \\ C_\ell^{XY} &\approx (4\pi)^2 \int k^2 \mathrm{d}k \mathcal{P}_\mathcal{R}(k) \Delta_{X,\ell}(k) \Delta_{Y,\ell}(k), \; XY \in \{TE, EE\} \\ C_\ell^{BB} &= (4\pi)^2 \int k^2 \mathrm{d}k \mathcal{P}_t(k) \Delta_{B,\ell}^2(k), \end{aligned} \tag{4.42}$$

while the EB and TB correlations are zero [Baumann, 2012]. Note that the TT and BB modes are determined by the scalar power spectrum and the tensor power spectrum

respectively. The TE and EE modes are written as an approximation because they contain a small contribution from the tensor power spectrum not reflected in the formulae above.

A central prediction of inflation are primordial gravitational waves, sourced by tensor perturbations and governed by eqn. (4.24). Following eqn. (4.42) they lead to B-modes in the CMB polarization. In current experiments, such as the Planck satellite, the angular power spectrum of these modes is consistent with zero [Akrami et al., 2020]. Future experiments such as CMB-S4 [Abazajian et al., 2022] and LiteBIRD [Allys et al., 2023] are designed to probe them with higher sensitivity.

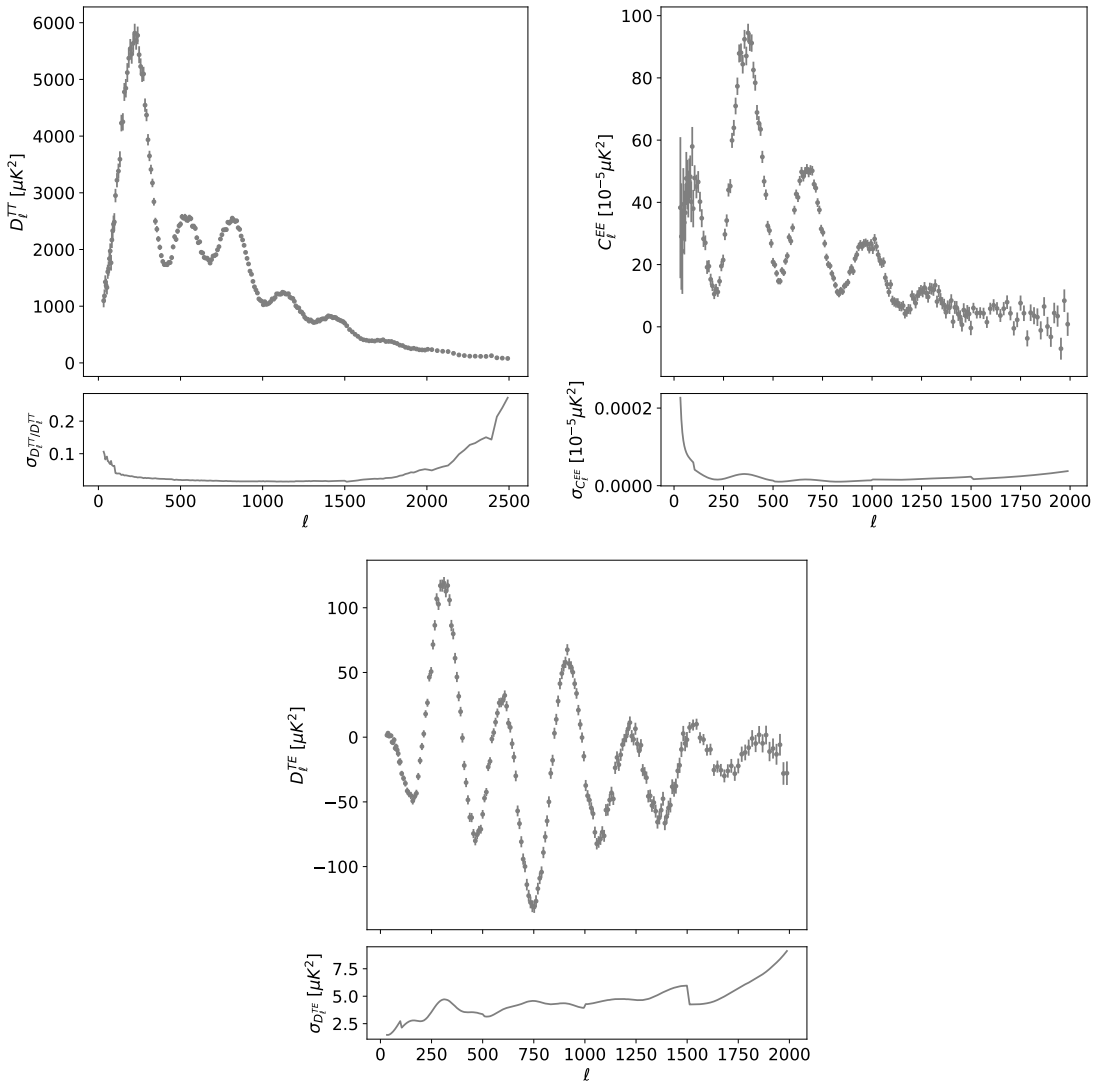

Figure 4.1: Angular power spectra of the CMB as reported in [Aghanim et al., 2020c]. This shows the binned data used in the plik_lite likelihoods. The data is extracted using the tool published together with [Prince and Dunkley, 2019].

Matter power spectrum

In addition to the CMB the primordial power spectrum can also be investigated using the density fluctuations of matter. If the dynamics of the universe after horizon reentry of the curvature modes is understood well enough the power spectrum of the matter density fluctuations can be predicted from the primordial scalar power spectrum. Schematically, when the matter dynamics after horizon reentry is encoded in the dark matter transport function $T_\delta^2(k,\eta)$ the density contrast power spectrum is calculated as [Baumann, 2012]

$$P_\delta(k,\eta) = \frac{4}{25}\left(\frac{k}{aH}\right)^4 T_\delta(k,\eta)\mathcal{P}_\mathcal{R}(k). \tag{4.43}$$

Similar to the CMB, the transfer function is computed numerically using the Boltzmann solver CLASS [Blas et al., 2011] throughout this thesis. An analytical approximation can be found in [Eisenstein and Hu, 1998].

There are different ways to probe the matter power spectrum, such as measuring the fluctuations of galaxy populations and weak lensing [Adame et al., 2024, Abbott et al., 2022, Almeida et al., 2023]. More recently 21cm intensity mapping experiments such as the SKA [Bacon et al., 2020] have been brought up as a future probe of the density power spectrum. This would make the matter power spectrum accessible at higher redshifts.

4.4 Type Ia supernovae

Type Ia supernovae occur when a white dwarf in a binary system exceeds the Chandrasekhar limit by accreting mass from the companion star [Mazzali et al., 2007]. The emitted radiation is measured as the light curve of the supernova. These light curves can be standardized to yield simultaneous measurements of the distance modulus μ, in magnitudes, and redshift z [Brout et al., 2022]. The distance modulus can be related to the Hubble function $H(z)$ through the luminosity distance d_L as

$$\mu(z) = 5\log_{10} d_L(z) + 10, \quad d_L(z) = (1+z)c\int_0^z \mathrm{d}z' \frac{1}{H(z')}. \tag{4.44}$$

Essentially, this measurement provides a way to measure the Hubble function at late times. Supernovae measurements have been used to provide the first evidence for the accelerated expansion of the universe [Riess et al., 1998, Perlmutter and et al., 2003, Perlmutter et al., 1999].

In the framework of ΛCDM the Hubble function is determined by eqn. (4.8). The supernovae type Ia measurements can be used to find constraints on the energy densities of the cosmological constant Ω_Λ and matter Ω_m under the assumption of a flat universe [Brout et al., 2022, Rubin et al., 2023]. Combining the supernova measurement with local distance indicators of their host galaxies allows to measure the Hubble constant H_0 [Riess et al., 2022]. The value obtained through this late time measurement is in tension with the value obtained from CMB measurements [Aghanim et al., 2020b].

Common extensions of ΛCDM postulate dark energy as a cosmological fluid driving the current expansion of the universe [Wetterich, 1988, Ratra and Peebles, 1988, Linder, 2008, Tsujikawa, 2013, Mortonson et al., 2013]. The dark energy equation of state can deviate from the constant value of -1 implied by ΛCDM. It is often restricted to be constant or linearly evolving [Chevallier and Polarski, 2001, Linder, 2003]. These types

of dark energy models are investigated in [Brout et al., 2022, Rubin et al., 2023]. For a general dark energy equation of state the Hubble function is expressed as [Takada and Jain, 2004]

$$\frac{H^2(a)}{H_0^2} = \frac{\Omega_m}{a^3} + (1-\Omega_m)\ \exp\left[-3\int_1^a da' \frac{1+w(a')}{a'}\right]\ . \tag{4.45}$$

This equation assumes a flat universe and an insignificant contribution from the energy density of radiation.

Throughout this work we use the Union2.1 [Suzuki et al., 2012, Amanullah et al., 2010, Kowalski et al., 2008] and Pantheon+ [Scolnic et al., 2022] datasets. The redshift distribution and distance modulus values for these surveys are depicted in Fig. 4.2. The more recent Pantheon+ data set includes a larger number of supernovae and can probe a wider redshift range.

4.5 Neural networks

This introduction to neural networks is by no means exhaustive. It covers the concepts and ideas needed to arrive at the models used in [Röver et al., 2024]. It is roughly based on [Plehn et al., 2022].

4.5.1 Function approximation

In this thesis neural networks are used to approximate some function f mapping an input vector $x \in \mathbb{R}^d$ onto the output space $\mathbb{R}^n$. The neural network is described by the map

$$x \to f_\theta(x),\ x \in \mathbb{R}^d,\ f_\theta(x) \in \mathbb{R}^n, \tag{4.46}$$

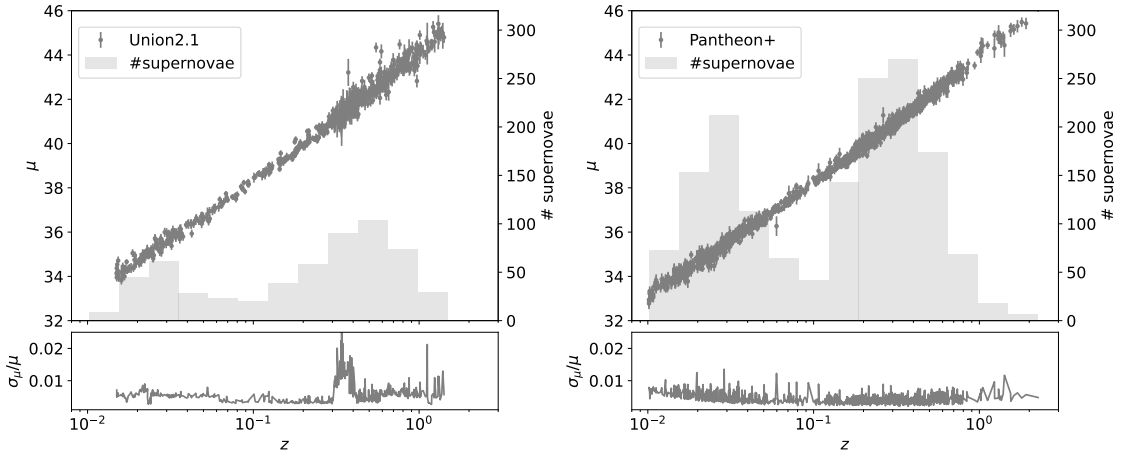

Figure 4.2: Redshift dependence of the distance modulus of the Union2.1 (left) and Pantheon+ data (right). The histograms capture the distribution of the supernovae in redshift. The lower sub-panels show the relative error bars on the distance moduli.

where θ are the network parameters. The loss function $\mathcal{L}$ measures how well the network approximates the target function f. It compares the output of the target function to the network output using some distance measure. A common loss function is the mean-squared error loss which compares the squared ℓ^2-norm of the two functions $\mathcal{L}(\theta) = \|f_\theta(x) - f(x)\|_2^2$ for an input vector x. For a batch of B input vectors the mean-squared error loss is computed as

$$\langle \mathcal{L}(\theta) \rangle_{batch} = \frac{1}{B} \sum_{i=1}^{B} \|f_\theta(x_i) - f(x_i)\|_2^2 . \tag{4.47}$$

The network parameters θ are optimized on a training data set $\{(x_i, f(x_i))\}_{i\in\{1,\dots,N\}}$ by minimizing the loss functions. In many applications the training data set is large such that the computation of the loss for all input vectors at once is slow. In these cases the training data is divided into batches of fixed size. The optimization of the loss function is usually based on gradient descent, where the parameters are updated using the gradient of the loss function

$$\theta^{t+1} = \theta^t - \alpha \left\langle \nabla_\theta \mathcal{L}^t \right\rangle_{batch} , \tag{4.48}$$

with a step size α. Throughout this work network parameters are optimized using the stochastic gradient descent algorithm Adam [Kingma and Ba, 2014].

In a multilayer perceptron (MLP) or fully connected neural network the neural network f_θ is composed of multiple layers of affine transformations. Each layer $\tilde{g}_n$ transforms an input vector $x^{(n)} \in \mathbb{R}^{d_n}$ into an output vector $y^{(n)} \in \mathbb{R}^{o_n}$ using

$$y^{(n)} = \tilde{g}_n(x^{(n)}) = W^{(n)} x^{(n)} + b^{(n)}. \tag{4.49}$$

Note that the matrices $W^{(n)}$ are not necessarily square matrices. Together with the biases $b^{(n)}$ they constitute the network parameters. To approximate non-linear functions an activation function σ is introduced. It acts on every node, i.e. every entry of each layer of the network. Including the non-linearity each layer can be written as

$$g_n(x^{(n)}) = \sigma\left(W^{(n)} x^{(n)} + b^{(n)}\right). \tag{4.50}$$

The fully connected neural network is a composition of these layers

$$f_\theta(x) = g_L \circ g_{L-1} \cdots \circ g_1(x), \tag{4.51}$$

where we have assumed a network with L layers. For sufficiently many network parameters this architecture is a universal approximator [Hornik et al., 1989].

Optimizing a loss function using a fully connected network requires computing derivatives of the loss function with respect to the network parameters. If the non-linearity σ is chosen such that its analytical derivative σ' is known this can be expressed using the chain rule

$$\begin{aligned}
\partial_{W_{ij}^{(n)}} \mathcal{L} &= \frac{\partial f_{\theta,k}}{\partial W_{ij}^{(n)}} \partial_{f_{\theta,k}} \mathcal{L} \\
&= \frac{\partial g_n(g_{n-1})_{l_n}}{\partial W_{ij}^{(n)}} \frac{\partial (g_L \circ g_{L-1} \cdots \circ g_n)_k}{\partial (g_n)_{l_n}} \partial_{f_{\theta,k}} \mathcal{L} \\
&= \sigma' \delta_{i l_n} (g_{n-1})_j \sigma' W_{k l_{L-1}}^{(L)} \dots \sigma' W_{l_{n+1} l_n}^{(n+1)} \partial_{f_{\theta,k}} \mathcal{L} \\
&= \sigma' W_{k l_{L-1}}^{(L)} \dots \sigma' W_{l_{n+1} i}^{(n+1)} \sigma' (g_{n-1})_j \partial_{f_{\theta,k}} \mathcal{L}.
\end{aligned} \tag{4.52}$$

In numerical calculations this back-propagation is done by saving a graph of the operations performed on a tensor. This composition of operations is then used to find the analytical derivatives similar to eqn. (4.52), in a process called automatic differentiation [Griewank and Walther, 2008, Paszke et al., 2017].

4.5.2 Uncertainty estimation

While there is a lot of variability in the choice of network architecture the loss function determines what the network output approximates. Using the loss described in eqn. (4.47) ensures that the Euclidean distance between the network approximation and the data points used in the network is minimized. Alternatively, the problem can be approached from an inference perspective. The training data is distributed according to the data distribution $p(x)$. We can understand the network, with parameters θ, as a model used to explain the data distribution and construct a likelihood $p(x|\theta)$. The posterior distribution of the network parameters $p(\theta|x)$ is then constructed using Bayes theorem as

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}, \tag{4.53}$$

where $p(\theta)$ is a prior on the network parameters. By constructing a loss as the negative logarithm of the posterior the training objective becomes finding the most probable parameter configuration that describes the data distribution.

In practice the loss can be defined as

$$\mathcal{L} = -\log p(x|\theta) - \log p(\theta) \tag{4.54}$$

since the data distribution does not contribute to derivatives with respect to the parameters and does not change the shape of the loss landscape. While this loss is defined for a single data point x, gradients in parameter space are usually determined as averages over the batch entries $\langle \nabla_\theta \mathcal{L} \rangle_{x \in batch}$. The prior can be chosen to implement additional constraints on the network parameters the likelihood determines the effect of the training data on the network. For an n-dimensional problem $f(x) \in \mathbb{R}^n$ a Gaussian likelihood leads to a loss defined as

$$\mathcal{L}_G(\theta) = \frac{1}{2}(f(x) - f_\theta(x))^T \Sigma_\theta^{-1}(x)(f(x) - f_\theta(x)) + \frac{1}{2}\log\det\Sigma_\theta(x) + \log p(\theta). \tag{4.55}$$

In this description the network approximates both the mean of the likelihood $f_\theta(x)$ and the covariance $\Sigma_\theta(x)$ at each data point. Since the problem is n-dimensional the neural network needs to be expressive enough to find all $n^2 + n$ entries of the mean and the covariance for each data point. This expression can be simplified by assuming an uncorrelated Gaussian likelihood to arrive at the heteroscedastic loss [Le et al., 2005, Gal, 2016]

$$\mathcal{L}_{het} = \frac{|f(x) - f_\theta(x)|^2}{2\sigma_\theta(x)^2} + \log\sigma_\theta(x) + \log p(\theta). \tag{4.56}$$

Here, the variance of the network is allowed to vary with the data and the network only needs to be able to express a $2n$ dimensional function. This loss can be further simplified to the homoscedastic loss by enforcing the same variance for each data point

$$\mathcal{L}_{hom} = \frac{|f(x) - f_\theta(x)|^2}{2\sigma_\theta^2} + \log\sigma_\theta + \log p(\theta). \tag{4.57}$$

In this case there is only one parameter σ_θ added to the network when compared to approximating the mean value using a mean-squared error loss.

The variance and covariance in these loss functions capture the width of the likelihood providing a quantification for the variability of the data labels $f(x)$ around the mean predicted by the network $f_\theta(x)$. The likelihood is linked to the data distribution $p(x)$ through Bayes theorem. If the mean predicted by the network is close to the physical truth underlying the measurement producing $p(x)$ the covariance captures the variability of the measurements around this truth. For a well-trained model the aleatoric uncertainty described in this section makes a statement about the uncertainty in the data distribution.

Epistemic uncertainty

The average loss as a function of the parameters $\langle \mathcal{L}(\theta) \rangle_x$ is rarely unimodal. Loss minimization usually introduces a stochastic element to avoid getting stuck in suboptimal minima. However, this stochasticity, as well as many, often equally good, minima can lead to different results when training similar networks on the same data set. This variability is often an indication of little or no data in the regions where differences between the network realizations occur.

While training an ensemble of networks [Lakshminarayanan et al., 2016] provides a measure of the epistemic uncertainty it does not come with a guarantee on the diversity of the trained networks. This diversity can be enforced by introducing a repulsive term as demonstrated in [D'Angelo and Fortuin, 2021]. Section 4.5.3 provides an introduction to this method of uncertainty quantification. Another approach to epistemic as well as aleatoric uncertainties is provided by Bayesian neural networks [Gal, 2016, Bollweg et al., 2020, Kasieczka et al., 2020, Bellagente et al., 2022, Butter et al., 2023]. In this approach each of the network parameters is promoted to a random variable drawn from a probability distribution. A principled if computationally challenging approach to this is to perform HMC parameter inference for the whole network [Izmailov et al., 2021]. However, there are also faster to compute approaches parametrizing the probability distribution for each network parameter in a traditional network with a few parameters [Graves, 2011].

4.5.3 Repulsive ensembles

This section is based on [Röver et al., 2024] and reproduces the derivation of repulsive ensembles described in [D'Angelo and Fortuin, 2021, Plehn et al., 2022]. This derivation is the work of Tilman Plehn and Theo Heimel.

For a repulsive ensemble the update rule to minimize the log-probability $p(\theta^t|x)$ is extended to an ensemble of networks, while its coverage of the parameter space is improved by a repulsive interaction. This interaction is based on the proximity of the ensemble member θ to all other members. The kernel $k(\theta, \theta_j)$ describes the interaction with a second ensemble member θ_j. Adding up the interactions with all other weight configurations yields

$$\theta^{t+1} = \theta^t + \alpha \nabla_{\theta^t} \left[\log p(\theta^t|x) - \sum_j k(\theta^t, \theta_j^t) \right] . \tag{4.58}$$

The kernel is chosen such that after training each ensemble member is a sample from the weight probability, $\theta \sim p(\theta|x)$.

Weight-space density

To ensure this sampling property the discretized t-dependence of the network parameters is related to a time-dependent probability density $\rho(\theta, t)$. The time evolution of the parameters can either be described through an ODE or a continuity equation,

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = v(\theta, t) \qquad \text{or} \qquad \frac{\partial \rho(\theta, t)}{\partial t} = -\nabla_\theta \left[v(\theta, t) \rho(\theta, t) \right] \ . \tag{4.59}$$

For a given velocity field $v(\theta, t)$ the individual paths $\theta(t)$ describe the evolving density $\rho(\theta, t)$ and the two descriptions are equivalent. If we choose the velocity field as

$$v(\theta, t) = -\nabla_\theta \log \frac{\rho(\theta, t)}{\pi(\theta)} \ , \tag{4.60}$$

the two descriptions read

$$\begin{aligned} \frac{\mathrm{d}\theta}{\mathrm{d}t} &= -\nabla_\theta \log \frac{\rho(\theta, t)}{\pi(\theta)} \\ \frac{\partial \rho(\theta, t)}{\partial t} &= -\nabla_\theta \left[\rho(\theta, t) \nabla_\theta \log \pi(\theta) \right] + \nabla_\theta^2 \log \rho(\theta, t) \ . \end{aligned} \tag{4.61}$$

The continuity equation becomes the Fokker-Planck equation, for which $\rho(\theta, t) \to \pi(\theta)$ is the unique stationary probability distribution.

Based on this ODE description we can construct an update rule similar to eqn. (4.58). The discretized version of the ODE is

$$\frac{\theta^{t+1} - \theta^t}{\alpha} = -\nabla_{\theta^t} \log \frac{\rho(\theta^t)}{\pi(\theta^t)} \ . \tag{4.62}$$

An unknown density $\rho(\theta^t)$, can be approximated as a superposition of kernels,

$$\rho(\theta^t) \approx \frac{1}{n} \sum_{i=1}^{n} k(\theta^t, \theta_i^t) \qquad \text{with} \qquad \int \mathrm{d}\theta^t \rho(\theta^t) = 1 \ . \tag{4.63}$$

Inserting this approximation into the discretized ODE yields

$$\frac{\theta^{t+1} - \theta^t}{\alpha} = \nabla_{\theta^t} \log \pi(\theta^t) - \frac{\nabla_{\theta^t} \sum_i k(\theta^t, \theta_i^t)}{\sum_i k(\theta^t, \theta_i^t)} . \tag{4.64}$$

This form can be related to eqn. (4.58) by setting $\pi(\theta) \equiv p(\theta|x)$. Following this description we add the normalization term of eqn. (4.64) to our original kernel in eqn. (4.58),

$$\nabla_{\theta^t} \sum_i k(\theta^t, \theta_i^t) \ \to \ \frac{\nabla_{\theta^t} \sum_i k(\theta^t, \theta_i^t)}{\sum_i k(\theta^t, \theta_i^t)} \ , \tag{4.65}$$

to ensure that the update rule with an appropriate kernel leads to the correct density.

Function-space density

The derivation in the previous section holds for ensembles with a repulsive force in weight space. However, we are interested in the function the network encodes and not the weight representation. Two networks encoding the same function could be constructed by permuting weights of the hidden layers. The resulting network configurations can be well separated in weight space while exhibiting very similar network outputs, unaffected by a repulsive force in weight space. To properly approximate epistemic uncertainty the repulsion should take place in the space of the network outputs $f_\theta(x)$.

In this space we can symbolically write the update rule from eqn. (4.58) with the normalization of eqn. (4.65) as

$$\frac{f^{t+1} - f^t}{\alpha} = \nabla_{f^t} \log p(f|x) - \frac{\sum_j \nabla_{f^t} k(f, f_j)}{\sum_j k(f, f_j)} \,. \tag{4.66}$$

Since the network training is defined in weight space, we have to translate the function-space update rule into weight space using the appropriate Jacobian

$$\frac{\theta^{t+1} - \theta^t}{\alpha} = \nabla_{\theta^t} \log p(\theta^t|x) - \frac{\partial f^t}{\partial \theta^t} \frac{\sum_j \nabla_f k(f_{\theta^t}, f_{\theta_j^t})}{\sum_j k(f_{\theta^t}, f_{\theta_j^t})} \,. \tag{4.67}$$

Since the kernel cannot be evaluated in function space we have to evaluate the function for a finite batch of points x,

$$\frac{\theta^{t+1} - \theta^t}{\alpha} \approx \nabla_{\theta^t} \log p(\theta^t|x) - \frac{\sum_j \nabla_{\theta^t} k(f_{\theta^t}(x), f_{\theta_j^t}(x))}{\sum_j k(f_{\theta^t}(x), f_{\theta_j^t}(x))} \,. \tag{4.68}$$

Loss function

The update rule derived in eqn. (4.68) can be used to define a loss function for the repulsive ensemble training. To that end we transform the posterior into a tractable likelihood loss with a Gaussian prior,

$$\log p(\theta|x) = \log p(x|\theta) \; - \frac{|\theta|^2}{2\sigma^2} + \text{const} \,. \tag{4.69}$$

Given a training dataset of size N, we evaluate the likelihood on batches of size B, so eqn. (4.66) becomes

$$\frac{\theta^{t+1} - \theta^t}{\alpha} \approx \nabla_{\theta^t} \frac{N}{B} \sum_{b=1}^{B} \log p(x_b|\theta) - \frac{\sum_j \nabla_{\theta^t} k(f_{\theta^t}(x), f_{\theta_j^t}(x))}{\sum_j k(f_{\theta^t}(x), f_{\theta_j^t}(x))} - \nabla_{\theta^t} \frac{|\theta|^2}{2\sigma^2} \,. \tag{4.70}$$

Here, $f_{\theta^t}(x)$ is to be understood as evaluating the function for all samples $x_1, \dots, x_B$ in the batch.

The loss function is obtained from the update rule by dividing by N to remove the scaling with the size of the training dataset and summing over all members of the ensemble. Since the gradients of the loss function are computed with respect to the parameters of all networks in the ensemble, we need to ensure the correct gradients of the repulsive

term using a stop-gradient operation, denoted with an overline $\overline{f_{\theta_j}(x)}$. The loss function for repulsive ensembles then reads

$$\mathcal{L} = \sum_{i=1}^{n} \left[-\frac{1}{B} \sum_{b=1}^{B} \log p(x_b|\theta_i) + \frac{1}{N} \frac{\sum_{j=1}^{n} k(f_{\theta_i}(x), \overline{f_{\theta_j}(x)})}{\sum_{j=1}^{n} k(\overline{f_{\theta_i}(x)}, \overline{f_{\theta_j}(x)})} + \frac{|\theta_i|^2}{2N\sigma^2} \right] . \tag{4.71}$$

The prior is used to enforce an L2-regularization with prefactor $1/(2N\sigma^2)$.

Kernel in function space

A typical choice for the kernel introduced in eqn. (4.63) is a normal distribution. In eqn. (4.71) this is a Gaussian in the multidimensional function space, evaluated over a sample,

$$k(f_{\theta_i}(x), f_{\theta_j}(x)) = \prod_{b=1}^{B} \exp\left(-\frac{|f_{\theta_i}(x_b) - f_{\theta_j}(x_b)|^2}{h} \right) . \tag{4.72}$$

The width h should be chosen such that the width of the distribution is not overestimated while still ensuring that it is sufficiently smooth. This can be achieved with the median heuristic [Liu and Wang, 2016],

$$h = \frac{\text{median}_{ij} \left(\sum_b |f_{\theta_i}(x_b) - f_{\theta_j}(x_b)|^2 \right)}{2 \log(n+1)} , \tag{4.73}$$

with the number of ensemble members n.

4.5.4 Physics-informed neural networks

This section is based on the introduction to physics-informed neural networks (PINNs) in [Röver et al., 2024].

Physics-informed neural networks [Raissi et al., 2017, Piscopo et al., 2019, Araz et al., 2021, Li et al., 2021, Cuomo et al., 2022, Hao et al., 2022] together with neural differential equations and neural operators form a group of machine learning methods relating neural networks to solutions of differential equations. PINNs learn a prediction for a given parameter choice without really solving an ODE at the stage of evaluation. Neural ODEs [Chen et al., 2018] use neural networks as part of a system of differential equations that is solved with conventional methods. Neural operators [Patel and Desjardins, 2018] provide a parametrized mapping of e.g. initial conditions to a state at a given time but can be used in a more general context.

The PINN setup requires training data that can be understood as the solution to a differential equation. The connection to the differential equation is encoded in the loss function. For an ODE,

$$\dot{u}(t) = F(u,t) \qquad \text{with initial conditions} \qquad u(t=0) = u_0 , \tag{4.74}$$

the MSE loss for a PINN consists of two terms,

$$\begin{aligned} \mathcal{L} &= (1-\beta)\mathcal{L}_{\text{IC}} + \beta\mathcal{L}_{\text{ODE}} \\ \text{with} \qquad \mathcal{L}_{\text{IC}} &= \left[u_\theta(t=0) - u_0\right]^2 \\ \mathcal{L}_{\text{ODE}} &= \left[\dot{u}_\theta(t) - F(u_\theta, t)\right]^2 \ . \end{aligned} \tag{4.75}$$

The first term drives the PINN to fulfill the initial conditions, and can be used without any additional training data. The second term ensures that the network approximates a solution to the differential equation. The parameter β balances the two contributions.

Training through the ODE loss uses two kinds of data. First, unlabeled or residual data points consist of points in time, where the differential equation is evaluated during the training [Raissi et al., 2017]. Second, labeled time points can include other information, in our case the corresponding true values for $u(t)$ and $\dot{u}(t)$. This can either be only the initial condition or a larger number of points used to encourage the PINN towards the correct differential equation solution.

In addition to the MSE loss we can define an uncertainty-aware heteroscedastic loss function similar to eqn. (4.56). For a d-dimensional function, defined through eqn. (4.74), evaluated on N residual points the heteroscedastic PINN loss reads

$$\begin{aligned} \mathcal{L}_{\text{IC,het}} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{d}\left[\frac{|u_{\theta,k}(t_i=0) - u_{0,k}|^2}{2\sigma_{\theta,k}(t_i=0)^2} + \log\sigma_{\theta,k}(t_i=0)\right] \\ \mathcal{L}_{\text{ODE,het}} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{d}\left[\frac{|\dot{u}_{\theta,k}(t_i) - F_k(u_\theta(t_i))|^2}{2\sigma_{\theta,k}(t_i)^2} + \log\sigma_{\theta,k}(t_i)\right] \ . \end{aligned} \tag{4.76}$$

The heteroscedastic loss is based on the likelihood that the network parameters describe a solution to the differential equation. When training with residual points that fulfill the differential equation exactly the widths $\sigma_{\theta,k}$ should approach zero. Regions with a large width, after training only with residual points, can indicate a large difference between the network approximation of the mean and the residual points.

We can also construct a heteroscedastic loss for the labeled data points following eqn. (4.56). In all numerical applications in this thesis heteroscedastic uncertainty is implemented by doubling the number of output parameters of the network. Half of them are used to approximate the mean, while the other half describe the uncertainty.

As an alternative to repulsive ensembles Bayesian neural networks offer a way to include epistemic uncertainty on top of the approach to aleatoric uncertainty described in this section. For PINNs this has been investigated in [Yang et al., 2021].

5 Inferring the Hubble Function with Uncertainties

This chapter is based on [Röver et al., 2024]. It explores the supernova data introduced in section 4.4 by constructing an uncertainty-aware emulator as well as inferring the Hubble function in a model-independent way. These tasks are performed using physics-informed neural networks, described in section 4.5.4. Section 5.1 describes an uncertainty-aware emulator setup using a toy example. This approach is then applied to the supernova data in section 5.2 and subsequently expanded to find a network reconstruction of the Hubble function in 5.3.

5.1 PINNcertainties

5.1.1 Toy example

This section demonstrates some properties of PINNs and the uncertainties we introduce. We explore the influence of the number of residual points as well as the effect of introducing labeled data points, representing a noisy measurement of the truth. We concentrate on a toy model, defined by the two-dimensional differential equation,

$$\ddot{u} + \frac{u}{2} = 0 \qquad \text{with} \qquad u(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad \dot{u}(0) = \begin{pmatrix} 0 \\ 1 \end{pmatrix} . \tag{5.1}$$

Numerically the PINNs are trained on the four-dimensional, first-order ODE describing the evolution of $(u, \dot{u})$. Apart from the differential equations we do not enforce an additional relation between u and $\dot{u}$. This has the advantage of increasing training speed. However, while the loss drives $\dot{u}$ towards the derivative of u this relation is not exact. For all results, we show one of the two components $u_{1,2}(t)$.

Note that the harmonic oscillator has a trivial solution $u(t) = 0$. For a PINN loss constructed as in eqn. (4.75) $\mathcal{L}_{\text{IC}}$ is not minimal, but $\mathcal{L}_{\text{ODE}}$ does not lead to any gradient. Non-trivial approximations to the ODE solution are constructed by including both loss terms in the training. However, for times far away from the initial condition the influence of the initial conditions weakens, and the network predicts an oscillation with a decreasing amplitude over time.

Unlabeled or residual data

This section establishes the effect of the number of residual points on the network estimation of the ODE solution. Using the MSE loss described in eqn. (4.75) we obtain uncertainties through training an ensemble of ten networks. Our basic architecture consists of five layers with 200 nodes per hidden layer. All our networks are written in

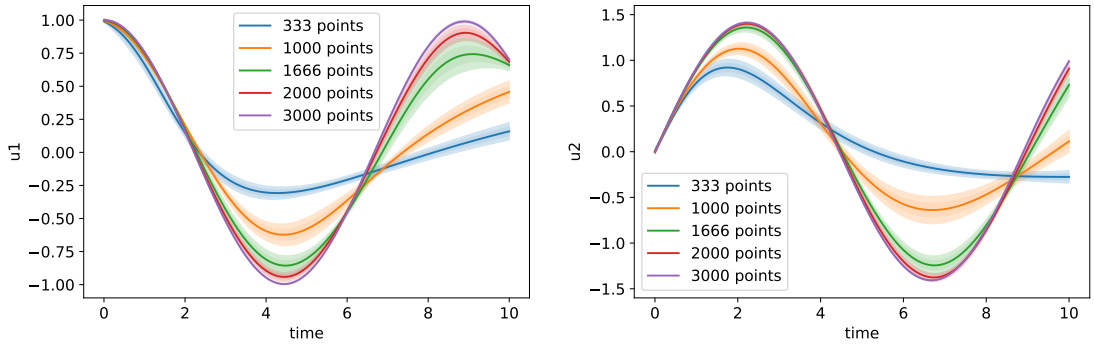

Figure 5.1: Learned harmonic oscillator, $u(t)$ on the left and $\dot{u}(t)$ on the right, for a varying number of uniformly distributed residual points. For the ensemble spread we train 10 independent models trained on different data points.

PyTorch [Paszke et al., 2019]. The training uses the ADAM optimizer [Kingma and Ba, 2014] in a batch learning setup. For the loss, we choose equal contributions, $\beta = 1/2$. We train the networks on 333, 1000, 1666, 2000, and 3000 uniformly distributed residual points in $t \in [0, 10]$. The means and standard deviations of this ensemble are shown in Figure 5.1.

As expected, the approximation improves when the number of residual points increases. While the initial condition and the evolution shortly after are learned from even a few residual points, good predictions at later times require more training data. Since the networks are designed as continuous functions their value at early times is close to the initial conditions and thus approximates the true solution. However, the network is trained for all times simultaneously and is initialized close to the trivial solution $u(t) = 0$. The network appears to get stuck in a local minimum where later times still fulfill the differential equation approximately, while also finding a continuous connection to the initial conditions. For more residual points, the agreement with the true solution improves quantitatively at early times and qualitatively at late times.

While the uncertainty estimate from the network ensemble appears to decrease as the networks approach the true solution they do not capture the poor agreement with the true solution. The different networks appear to be drawn to the same local minimum in the loss function even for different sets of residual points.

Labeled data

In many physics problems we have measurements of the desired quantity on top of information on its evolution through the differential equation. To judge the impact of this data we combine 1000 residual points t_i and 6000 labeled points $(t, u, \dot{u})_i$. The additional information can be incorporated in the ODE loss of eqn. (4.75) similar to the initial condition loss. This helps anchor the network to the true solution at different times. We train the network alternatingly. In a first step we minimize the MSE between the network prediction and the labels, and a second step minimizes the PINN loss from eqn. (4.75) on the residual points. Training with labeled data points can be considered standard network training. On the other hand for the labeled data points we first generate the information for the network training by inserting the network into the differential equation. In particle

physics, efficient integration and sampling build on a very similar combination of online and buffered or sample-based training [Heimel et al., 2023, 2024].

For the harmonic oscillator uniformly distributed labeled data points are not optimal. In Figure 5.2 we show how the PINN training improves when we include labeled data in specific time windows, while the unlabeled data remains distributed uniformly.

The left panel shows that training with 6000 labeled data points close to the initial condition yields a significant improvement in the region of the labeled data points. Additionally, for a short time after leaving the labeled data region the PINNs are able to extrapolate well. This effect is comparable to the unlabeled data case where the PINNs find a good approximation close to the initial condition. The ensemble uncertainties do not cover any of the deviations from the true solution. In the right panel the labeled points are positioned at later times. Combined with the IC-loss this allows the networks to learn a good approximation over the entire time range. If we consider the initial condition as labeled data as well, this setup reduces our problem to an interpolation. The gap between the initial condition and the additional labeled points does not cover the first maximum of the oscillation, its position is however captured by the PINN loss.

5.1.2 Uncertainties

This section demonstrates the estimation of PINN uncertainties using a heteroscedastic loss from eqn. (4.56) and repulsive ensembles, see section 4.5.3. We use the harmonic oscillator toy model from section 5.1.1 while adding noisy labeled data points. To determine how well each uncertainty estimate captures the statistic uncertainty in the training data the labeled data points are distributed such that they become sparse for late times.

Sparse and stochastic data

To determine the effect of the PINN loss we first establish the effect of noisy and sparse data. For this experiment the neural networks are trained only on the labeled data points. We generate two datasets with a reduced training point density towards late times. For

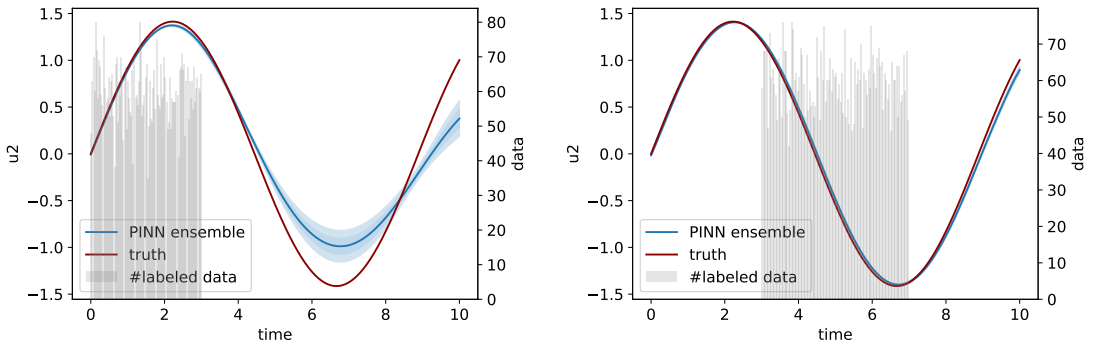

Figure 5.2: Learned harmonic oscillator adding labeled data points at small times (left) and intermediate times (right). The light histogram gives the distribution of training points. For the ensemble spread we train 10 independent models on different residual data points.

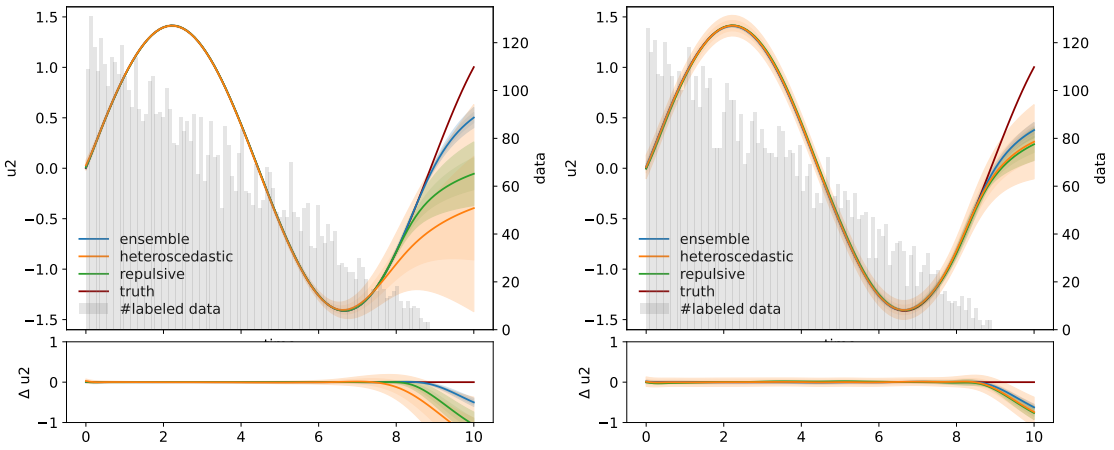

Figure 5.3: Learned harmonic oscillator with sparse training data at late times. For the training we only use labeled data points, defining a simple regression task. In the left panel the training data is exact, in the right panel it includes noise. The error bars correspond to 68% and 95% CL.

one of them we include noise while the other is exact. In this setup the labels u and $\dot{u}$ are separate, the network architecture does not include any information on the differential equation. The decreasing distribution of labeled data points is given in the background histogram of Figure 5.3, creating a smooth transition from abundant to sparse data and ultimately to an extrapolation problem.

The left panel of this Figure demonstrates the effect of increasingly sparse data without noise. The heteroscedastic uncertainty increases with time, as the density of labeled training points decreases [Seitzer et al., 2022]. Both, the repulsive ensemble and the heteroscedastic network deviate from the true solution for $t > 8$. They learn the shape of the minimum even though there is very little data beyond $t = 6$. The repulsive ensemble remains more stable than the heteroscedastic network, which can be explained by the stabilizing effect of ensembling. For both, the heteroscedastic network and the repulsive ensembles, the error bar increases fast enough to cover the deviation from the true solution up to $t = 9$. Beyond this point the error bar is not conservative in covering the uncertainty related to missing training data altogether. The classic ensemble without repulsive term approximates the solution well up to $t = 9$ but without a meaningful spread beyond that.

The right panel of Figure 5.3 focuses on noisy data. The labeled data points still encode the solution to the differential equation, but with Gaussian noise on u and $\dot{u}$ information of mean zero and width 0.1. The heteroscedastic network captures this stochasticity as an additional source of uncertainty over the entire time range. While each member of the repulsive ensemble is determined using a heteroscedastic loss we only plot the spread of their means to capture the effects of epistemic uncertainty. They approximate the truth well, without a visible spread. The ensemble trained with an MSE loss does not capture the noise in the data. At late times, the noise has a counter-intuitive effect on the extrapolation; all predictions using a heteroscedastic loss become better, and the reduced uncertainties confirm this trend. The central values and the error bars for the heteroscedastic network and the repulsive ensembles lose their reliability in the region

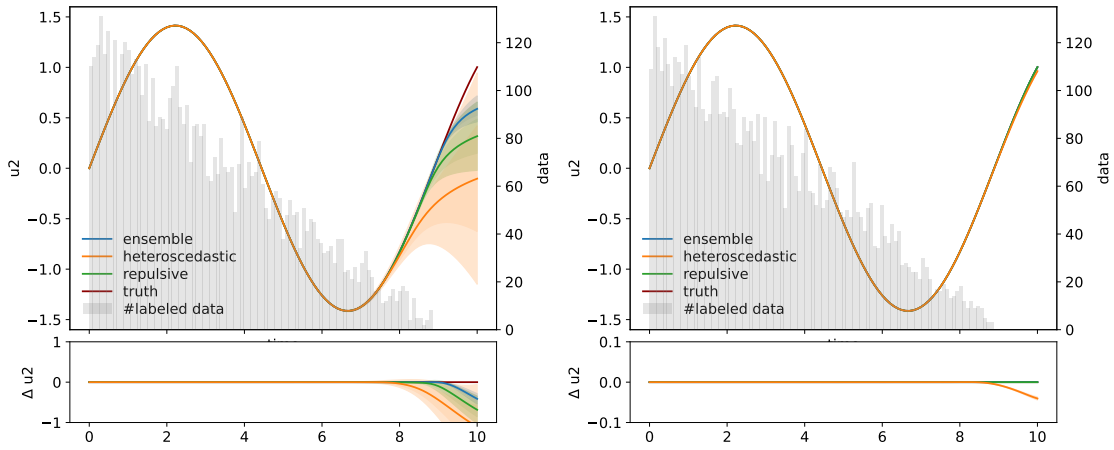

Figure 5.4: Learned harmonic oscillator adding the ODE loss enforcing the differential equation. For the left panel the additional residual points are distributed like the labeled point, for the right panels we add 10000 residual points uniformly over time. The error bars correspond to 68% and 95% CL.

without data, $t > 9$.

For both test cases regression works as long as there is some training data. However, once we enter the regime of extrapolation the networks fail to approximate the true solution. This is true for the central value as well as the learned uncertainty estimates.

ODE extrapolation

Adding the ODE loss to the network training allows them to extrapolate to regions without labeled data by using the additional residual data. At these points the network confirms that its output fulfills the differential equation. We train with the two datasets alternatingly, one epoch using the labeled data point and one epoch using residual points, both computing the loss in eqn. (4.76).

As a first experiment, we include residual data with the same time distribution as the labeled data. In practice, we strip the labeled data of the additional information and add the remaining t-values as residual points. The left panel of Figure 5.4 shows the PINNs becoming slightly more accurate at large times due to the increased total number of training points. While this is true for the case without noise, the improvement is not visible for noisy data. The learned network uncertainty confirms the behavior of the central prediction.

In a further experiment, we add 10000 residual training points uniformly distributed over time. Without noise, these models reproduce the true function extremely well, over the entire time range and with correspondingly small uncertainties from the heteroscedastic loss and the repulsive ensembles. Note, however, that this extrapolation away from the labeled data points requires residual data points in the regions where the solution is predicted.

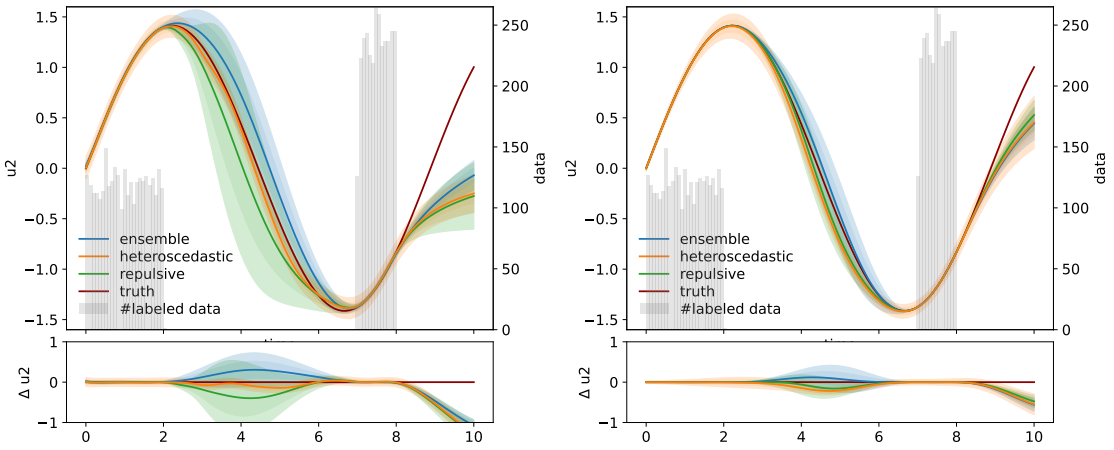

Figure 5.5: Learned harmonic oscillator with split training data and no noise. In the left panel we only use the labeled training point, in the right panel we add residual points distributed the same way as the labeled points. The error bars correspond to 68% and 95% CL.

Interpolation with a gap

As illustrated in Figure 5.3 both the heteroscedastic loss and the repulsive ensembles assign an increasing error bar towards the data-deprived region, with a conservative uncertainty estimate for as long as there is training data. This section addresses wide interpolation.

Here noisy training data is distributed uniformly in the ranges, $t = [0, 2]$, and $t = [7, 8]$ with higher density in the second range. This forces the network to interpolate over a large time window and extrapolate to late times. In the left panel of Figure 5.5 we train with the labeled data only. The wide interpolation challenges the three different types of networks, indicated by the poor agreement with the true solution. The spread of the classic ensemble barely covers the difference from the truth. The situation improves with repulsive ensembles, which provide more conservative error bars in both data-deprived regions. In the interpolation region the heteroscedastic network covers a much smaller family of functions. It does not appear to capture the aleatoric uncertainty in this region. While at late times the central value deviates from the true solution at a similar level as the repulsive ensembles, the error bar is smaller and not conservative for the extrapolation.

In the right panel of Figure 5.5, we again add residual data following the same distribution as the labeled data. The network can learn the differential equation using the ODE loss. From the left panel of Figure 5.4 we know that this has hardly any effect on regions with enough data or actual extrapolation. However, here we see that the residual data and the ODE loss have a significant effect on the uncertainty estimate for the wide interpolation.

Note that domain knowledge can guide our expectation of network behavior for wide interpolation and thus the choice of uncertainty estimate. Either we argue that the network should consider a wide interpolation an extrapolation and admit that there is not enough data to capture possible features in the sparsely probed region. In that case

the error bar should be large. Or we assume that there are no additional features, in which case a small uncertainty reflects the confidence of the network training.

5.2 Supernova PINNulator

This section moves away from the toy example and explores the computation of the distance moduli μ of the type Ia supernovae through a PINN based neural network emulator. For a known Hubble function the luminosity distances are computed through the integration

$$\mu = 5\log_{10} d_L(z,\lambda) + 10 \quad \text{with} \qquad d_L(z,\lambda) = (1+z)\,c\int_0^z \mathrm{d}z' \frac{1}{H(z',\lambda)}\,. \tag{5.2}$$

The functional form of the Hubble function is dependent on our assumptions of the underlying cosmology. The argument λ symbolizes the dependence on cosmological parameters, and is carried through all derivations. In this section we focus on a flat two-fluid universe including matter and dark energy, wCDM, assuming a constant $w(z) < -1/3$ to ensure accelerated expansion. This model contains ΛCDM as a particular choice of the equation of state parameter $w = -1$. If we only assume the FLRW symmetries, the Hubble function $H(z)$ can take any form allowed by the data. We come back to this second option in section 5.3.

Luminosity-distance PINN

PINNs can learn luminosity distances as a solution to a differential equation based on eqn. (5.2). The resulting emulator can be used to speed up inference in classical MCMC inference as used in section 3.3.5.

Based on the integral expression luminosity distances are governed by the ODE

$$\frac{\mathrm{d}\tilde{d}_L(z,\lambda)}{\mathrm{d}z} - \frac{\tilde{d}_L(z,\lambda)}{1+z} - \frac{1+z}{\tilde{H}(z,\lambda)} = 0 \qquad \text{with} \qquad d_L(0,\lambda) = 0\,. \tag{5.3}$$

Here, $\tilde{d}_L = d_L H_0/c$ and $\tilde{H}(z,\lambda) = H(z,\lambda)/H_0$ are dimensionless and ensure solutions of order unity. This makes PINN training more stable [Wang et al., 2023]. To learn the solution to eqn. (5.3), we choose the cosmological parameters and the functional form for the Hubble function similar to [Chantada et al., 2023],

$$\frac{H(z,\lambda)}{H_0} = \sqrt{\Omega_m(1+z)^3 + (1-\Omega_m)(1+z)^{3(1+w)}}\,. \tag{5.4}$$

As cosmological input parameters we concentrate on the redshift z, the energy density of matter Ω_m and the dark energy equation of state parameter w. In this subsection, we fix the Hubble parameter to 70 km/s/Mpc.

The two relevant losses defined in eqn. (4.75) can be read off eqn. (5.3) as

$$\begin{aligned}\mathcal{L}_{\text{IC}} &= \frac{1}{N}\sum_{i=0}^{N}\left[d_{L,\theta}(0,\lambda_i)\right]^2 \\ \mathcal{L}_{\text{ODE}} &= \frac{1}{N}\sum_{i=0}^{N}\left[\frac{\mathrm{d}d_{L,\theta}(z_i,\lambda_i)}{\mathrm{d}z} - \frac{d_{L,\theta}(z_i,\lambda_i)}{1+z_i} - \frac{1+z_i}{H(z_i,\lambda_i)}\right]^2 .\end{aligned} \tag{5.5}$$

The index i counts N elements $(z,\lambda)_i$, generated uniformly over the relevant parameter ranges.

As in the toy example, we construct heteroscedastic versions of the MSE losses to learn the uncertainties from the training data,

$$\begin{aligned}\mathcal{L}_{\text{IC,het}} &= \frac{1}{N}\sum_{i=0}^{N}\left[\frac{d_{L,\theta}(0,\lambda_i)^2}{2\sigma_\theta(0,\lambda_i)^2} + \log\sigma_\theta(0,\lambda_i)\right] \\ \mathcal{L}_{\text{ODE,het}} &= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\left(\frac{\mathrm{d}d_{L,\theta}(z_i,\lambda_i)}{\mathrm{d}z} - \frac{d_{L,\theta}(z_i,\lambda_i)}{1+z_i} - \frac{1+z_i}{H(z_i,\lambda_i)}\right)^2}{2\sigma_\theta(z_i,\lambda_i)^2} + \log\sigma_\theta(z_i,\lambda_i)\right] .\end{aligned} \tag{5.6}$$

Our small network uses five hidden layers with 100 nodes each, with a one-dimensional output approximating the luminosity distance. The 10^5 residual training points are generated uniformly in the ranges $z \in [0, 1.8]$, $\Omega_m \in [0, 1]$, and $w \in [-1.6, -0.5]$. Network training with only the residual points is good enough that we do not have to consider labeled data for the PNN emulator. Section 3.3.5 uses a smaller model to constrain the matter density and the equation of state parameter using the Union2.1 dataset [Suzuki et al., 2012, Amanullah et al., 2010, Kowalski et al., 2008].

Luminosity-distance emulator

Figure 5.6 demonstrates the accuracy of the PINN emulator assuming the best-fit parameters of the Union2.1 dataset. The left panel demonstrates that the spread of ten emulators trained using MSE errors and heteroscedastic errors both vary at less than an

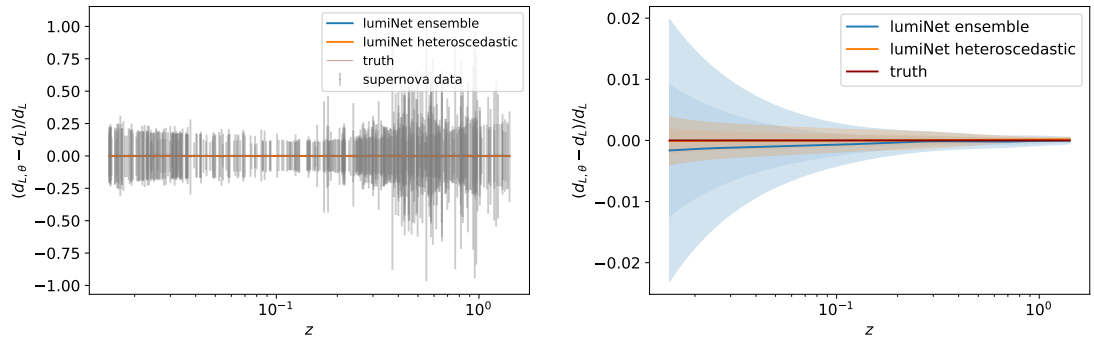

Figure 5.6: Learned luminosity distance from residual points only. The left panel compares the heteroscedastic PINN uncertainty to the experimental uncertainties in the Union2.1 dataset. The right panel shows the relative difference between the learned and true solutions. For the ensemble spread we train 10 independent models on different data points.

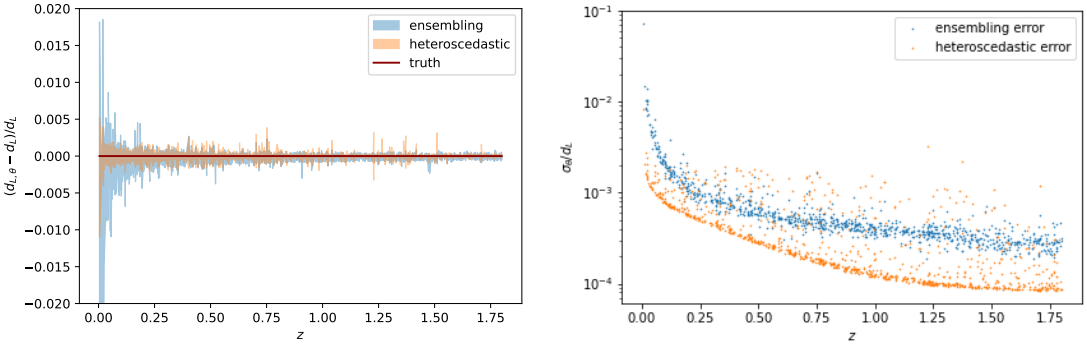

Figure 5.7: PINN accuracy for data points uniformly sampled from the same cosmological parameter ranges as the training points. The left panel shows the error bands around the true solution, the right panel the evolution of the ensemble spread and the heteroscedastic uncertainty with redshift.

order of magnitude of the experimental uncertainties. The right panel shows that the spread of ten MSE-trained PINNs is larger than the uncertainty estimation obtained when training with a heteroscedastic loss.

Since the networks are trained using only residual points the solution is probed exactly. The heteroscedastic error bars do not capture the noise in the data but the limitations of the network expressivity. The error bands computed from ten networks trained with the heteroscedastic loss lie within the heteroscedastic error bars. Rather than adjusting a network with limited expressivity to data with arbitrary precision this effect is captured in the heteroscedastic loss. This helps stabilize the training and subsequently the agreement of the trained model with the true solution.

To test the reliability of the distance modulus emulator we generate 1000 test data points from the same distribution as the training data, while computing their true luminosity distances using eqn. (5.3). The left panel of Figure 5.7 shows the deviation of the PINN prediction from the true solution. The spread of the ensemble trained with an MSE loss deviates from the truth by less than two percent. The heteroscedastic training improves this agreement to better than one percent. In the right panel of Figure 5.7 the relative uncertainties grow rapidly for small redshifts since the initial condition of the luminosity distance is also small. This requires better absolute precision.

Overall PINNs trained with either type of loss function are precise enough to use as an emulator for the Union2.1 or Pantheon+ [Scolnic et al., 2022] data, which come with experimental errors of around 10%, without resorting to labeled data training.

5.3 Supernova PINNference

The previous section demonstrates that PINNs can learn and emulate luminosity distances for a given parameterized Hubble function as a solution to a differential equation. The trained emulators can be used to infer posterior distributions of the cosmological parameters. This section instead uses the experimental data sets represented in Figure 4.2 to infer a neural network representation of the Hubble function with minimal assumptions on the underlying cosmology.

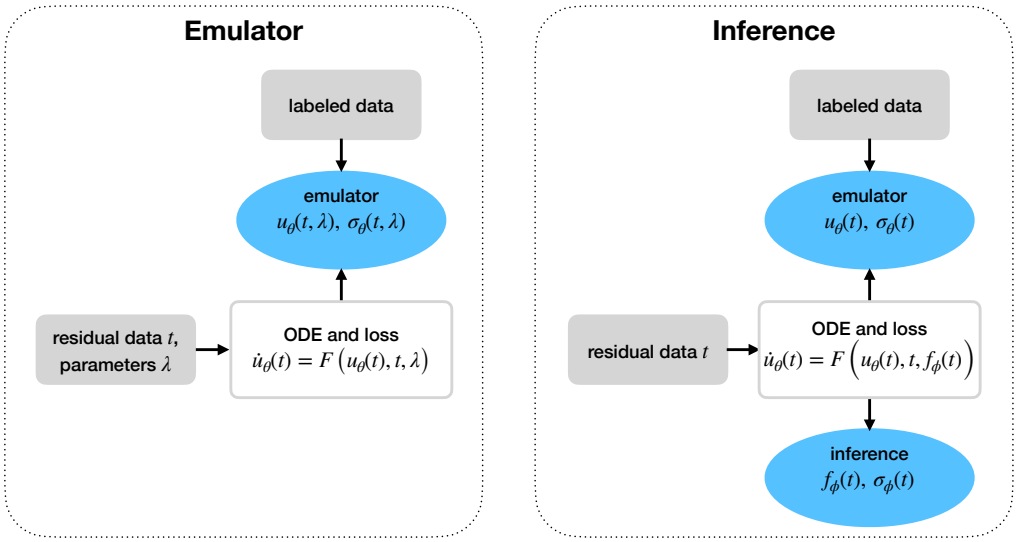

Figure 5.8: Illustration of the PINN emulation in section 5.2 and inference setups in section 5.3.

Introducing a free function $f_\phi(t) \approx f(t)$ represented by a neural network, similar to [Shukla et al., 2020], expands the structure of eqn. (4.74) to

$$\dot{u}(t) = F(u(t), t, f(t)) \quad \text{with} \quad u(0) = u_0 \,. \tag{5.7}$$

We extract information on the differential equation including $f(t)$ by training a network $u_\theta(t)$ on the labeled data. The trained network fulfills the differential equation with the true function $f(t)$. This function is approximated with a second network $f_\phi(t)$. Given N labeled data points $(t,u)_i$ and M residual points $\tilde{t}_j$ the training uses the loss functions

$$\begin{aligned} \mathcal{L}_{\text{Data}} &= \frac{1}{N}\sum_{i=1}^{N} \left[u_\theta(t_i) - u_i\right]^2 \\ \mathcal{L}_{\text{ODE}} &= \frac{1}{M}\sum_{j=1}^{M} \left[\dot{u}_\theta(\tilde{t}_j) - F(u_\theta(\tilde{t}_j), \tilde{t}_j, f_\phi(\tilde{t}_j))\right]^2 . \end{aligned} \tag{5.8}$$

In this equation the data loss plays the same role as $\mathcal{L}_{\text{IC}}$ in eqn. (4.75). It anchors the network approximation to the true solution of the differential equation and extracts the information on $f(t)$ via u_θ. The second loss term $\mathcal{L}_{\text{ODE}}$ ensures that the network $f_\phi(t)$ approximates the true $f(t)$ for all times covered by the residual points, as long as u_θ is sufficiently accurate. In all numerical experiments the losses are combined by alternating between epochs using only one of them. The network structure and training are illustrated in Figure 5.8.

As a first numerical experiment we generate 10^3 artificial, noiseless data points from a cosmological model defined by eqn. (5.4). Here w is fixed to the best-fit value of the Union2.1 dataset. In Figure 5.9 both the Hubble function and the luminosity distance are reconstructed using dense networks with five hidden layers and a width of 100 nodes. We use 10^4 residual points and compensate for the imbalance between residual and labeled data points by training ten epochs with the data loss for every epoch trained with the

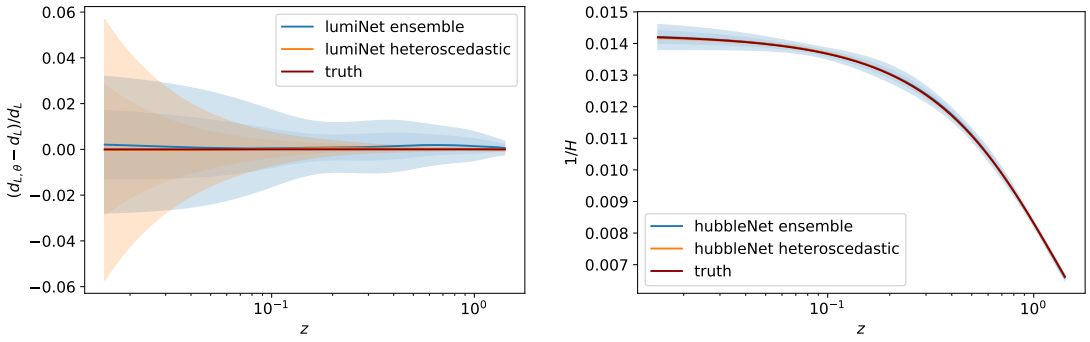

Figure 5.9: Reconstruction of the Hubble function using PINNs with an MSE loss. The left plot depicts the luminosity distance approximation compared to the true value with an ensembling error bar derived from ten models. The right-hand side depicts the corresponding Hubble reconstruction.

ODE loss. This reconstruction of the (inverse) Hubble function is performed without any input on the particular model used to generate the data. On this synthetic data set the Hubble function can be learned almost perfectly.

5.3.1 Uncertainty estimation

While the previous section reconstructs the Hubble function from noiseless labeled data this section provides a way to incorporate data uncertainties into the calculation, using a heteroscedastic loss. Additionally, we implement repulsive ensembles since they allow for a more conservative error estimate in regions with sparse data, as demonstrated in section 4.5.3.

Combining the learned luminosity distance $\tilde{d}_{L,\theta}$, with uncertainty σ_θ, and the ODE in eqn. (5.3), every luminosity distance value contributes to the reconstruction of the Hubble function as

$$\frac{1+z_i}{\tilde{H}(z_i)} \approx \frac{\mathrm{d}\tilde{d}_{L,\theta}(z_i)}{\mathrm{d}z} - \frac{\tilde{d}_{L,\theta}(z_i)}{1+z_i}. \tag{5.9}$$

In order to include the network uncertainties σ_θ in the Hubble reconstruction, both $\tilde{d}_{L,\theta}(z_i)$ and $\mathrm{d}\tilde{d}_{L,\theta}(z_i)/\mathrm{d}z$ need to be drawn from their respective probability distributions. By using a heteroscedastic loss the luminosity distance at each redshift is assumed to follow a normal distribution $\mathcal{N}(\tilde{d}_{L,\theta}(z_i), \sigma_\theta^2)$. Since samples of the luminosity distance are generated using a standard Gaussian, the width of the derivative distribution is $\mathrm{d}\sigma_\theta/\mathrm{d}z$. We can generate a distribution of Hubble function values by sampling the luminosity distance and its derivative from their distributions and inserting into eqn. (5.9).

A second network can then learn $\tilde{H}_\phi$ with an uncertainty σ_ϕ based on this distribution of Hubble function values. The uncertainty on $\tilde{H}_\phi$ is learned using the heteroscedastic loss of eqn. (5.6). This uncertainty is interpreted as the uncertainty on $(1+z)/\tilde{H}_\phi(z)$ under the assumption that for each redshift $d_{L,\theta}$ fulfills the differential equation correctly. Based on this assumption inserting eqn. (5.9) into the loss allows us to reduce it to the

expression

$$\mathcal{L}_{\text{Hubble, het}} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\left(\frac{1+z_i}{\tilde{H}(z_i)} - \frac{1+z_i}{\tilde{H}_\phi(z_i)}\right)^2}{2\left(\sigma_\phi(z_i)\right)^2} + \log\sigma_\phi(z_i)\right]. \tag{5.10}$$

The Hubble function is approximated by a normal distribution in $(1+z_i)/\tilde{H}_\phi(z_i)$ with variance $\sigma^2_\phi(z_i)$.

Note that training with the data loss only modifies the luminosity distance network $d_{L,\theta}$, while combining eqn. (5.9) and eqn. (5.10) allows us to optimize H_ϕ and $d_{L,\theta}$ simultaneously. Here both mean value and uncertainty of $d_{L,\theta}$ appear in the sampling of $d_L(z_i)$, allowing the network parameters θ to influence the loss. The ratio of labeled data epochs, where the data is extracted using $d_{L,\theta}$, to ODE epochs, where H_ϕ is inferred, is a training hyperparameter.

5.3.2 Noisy data

With the uncertainty estimation derived in the previous section we can solve the inverse problem for real observations. We consider two datasets, Union2.1 [Suzuki et al., 2012, Amanullah et al., 2010, Kowalski et al., 2008] and Pantheon+ [Scolnic et al., 2022]. Figure 4.2 depicts the measured distance moduli and their redshift distributions. We convert each of these data sets into luminosity distance following eqn. (5.2). In this process we assume that the data follows a multivariate normal distribution and generate a set of luminosity distances per redshift using the mean and the covariance matrix from the actual data.

The resulting luminosity distances and the distribution of redshifts for the ensemble of synthetic datasets are depicted in Figure 5.10. Their relative error is around 10% of the function value, and the data becomes sparse towards large redshift. The newer Pantheon+ dataset covers a larger range of redshifts and includes three times as many supernovae.

In this section the luminosity distance is learned as $d_{L,\theta}$, using five layers with 100 nodes each. The inverse Hubble function is modeled with a second network with five layers and 200 nodes wide. As suggested in [Wang et al., 2023] we impose the boundary condition of the luminosity distance network by learning $(d_L/z)_\theta$ and multiplying by z later. This ensures that $d_{L,\theta}(z=0)=0$. In addition, we find that using random Fourier features [Tancik et al., 2020] significantly reduces the required training time. For each epoch the labeled training is generated from the luminosity distance distribution shown in Figure 5.10. The resulting ensemble of luminosity distances scatters around the mean at each redshift, which can be captured by the heteroscedastic loss of the Hubble function network.

Figure 5.11 depicts the reconstruction of the Hubble function from both datasets. We show the learned luminosity distance and the reconstructed Hubble function, comparing a heteroscedastic network, an ensemble of MSE networks and a repulsive ensemble. Similar to section 5.1.2, the ensemble and the aleatoric uncertainty of the repulsive ensemble do not capture the data noise, whereas the heteroscedastic uncertainty of the luminosity distance does. The reconstructed Hubble function is consistent with a wCDM approximation of the Hubble function from a direct fit of a parameterized model.

The sharp feature in the Hubble function reconstruction from the Union2.1 dataset can be understood from eqn. (5.9). The uncertainty of the Hubble function is approximately the quadratic mean of the uncertainty of the derivative of the luminosity distance and the uncertainty of the luminosity distance itself. Fast changes in the width and scatter of the labeled data points with redshift, see Figure 5.10, leverage fast changes in the predicted error bars of the luminosity distance. The sharp increase in the uncertainty of the reconstructed Hubble function at redshift 0.3 corresponds to the change in the uncertainty in the luminosity distance leading to a maximum in the uncertainty.

The reconstruction of the Hubble function in eqn. (5.9) relies on the assumption that the network approximating the luminosity distances fulfills the differential equation exactly. The deviation from the true solution can be approximated by inserting both networks into the differential equation. In this application the deviation is small compared to the predicted uncertainties from the spread of the data.

5.3.3 Dark energy equation of state

Finally, the inferred, parameter-free Hubble function $H(a)/H_0$ can be converted to an equation of state function $w(a)$. Using the general relation [Takada and Jain, 2004],

$$\frac{H^2(a)}{H_0^2} = \frac{\Omega_m}{a^3} + (1-\Omega_m)\ \exp\left[-3\int_1^a da' \frac{1+w(a')}{a'}\right] , \tag{5.11}$$

$w(a)$ is determined through differentiation,

$$w(a) = -\frac{1}{3}\frac{\mathrm{d}}{\mathrm{d}\log a}\log\left[\frac{H^2(a)}{H_0^2} - \frac{\Omega_m}{a^3}\right] - 1 . \tag{5.12}$$

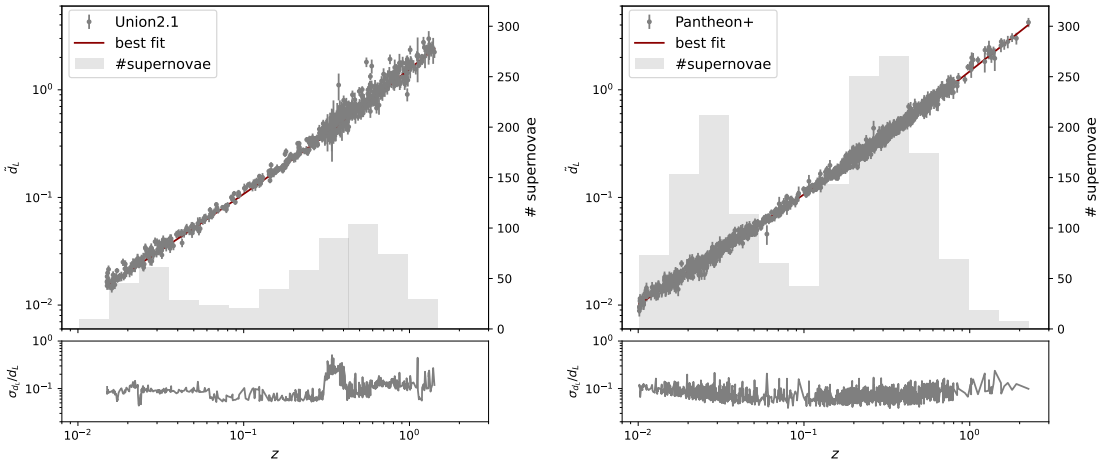

Figure 5.10: Generated redshift dependencies of the luminosity distance values of the Union2.1 (left) and Pantheon+ data (right). The histograms capture the distribution of the supernovae in redshift. The lower sub-panels show the relative error bars on the luminosity distances.

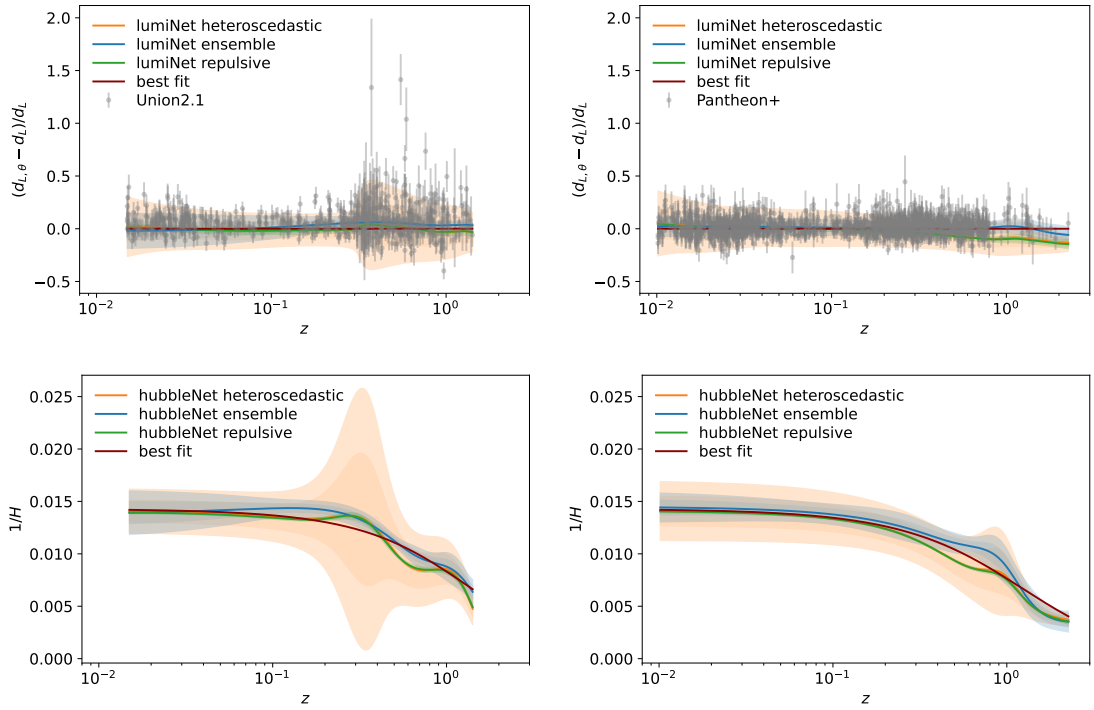

Figure 5.11: Top: PINN-learned luminosity distance from the labeled data, derived from the Union2.1 (left) and Pantheon+ (right) data. Bottom: learned inverse Hubble function from the two datasets.

We use $\Omega_m = 0.28$, as suggested by the Union2.1 dataset. Note that following this equation reintroduces more model assumptions into our setup.

The left panel of Figure 5.12 demonstrates that $w(z)$ can be reconstructed for small data uncertainties. Increasing the observational uncertainty to 5% or 10% has a large impact on the inferred uncertainty of $w(z)$ beyond $z \simeq 0.3$. This is partially caused by the increase in PINN uncertainty away from the initial conditions. More importantly, at high redshift the dark energy has a small influence on the Hubble function at high redshift, rendering $w(z)$ effectively unconstrained. Technically, by approaching $H(a)^2/H_0^2 \simeq \Omega_m/a^3$ at sufficiently high redshifts leads to a diverging logarithmic derivative in eqn. (5.12).

In the right panel of Figure 5.12 we show the reconstruction of $w(z)$ from our two datasets. The matter density for each dataset is assumed to be their respective best-fit values. At small redshifts our inference method constrains $w(z)$ well, but the uncertainties of the labeled data do not leave any sensitivity beyond $z \gtrsim 0.3$.

5.4 PINNclusions

Physics-informed neural networks are trained on the output of a parameterized system of differential equations. They can predict solutions for given parameters with a proper interpolation between parameter choices. This emulation of the space of ODE solutions provides tremendous speed-ups and therefore an excellent tool for statistical inference. The focus of our investigation was the error-awareness or uncertainty estimation of PINNs. For this purpose we have compared a heteroscedastic loss and repulsive ensembles, confirming

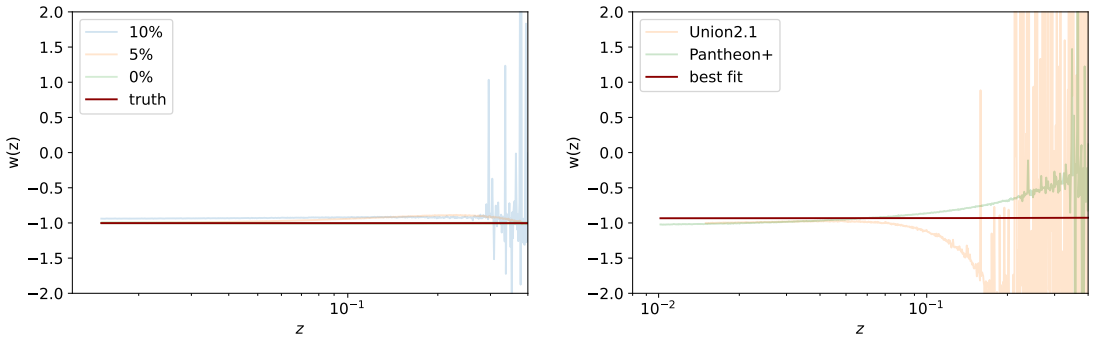

Figure 5.12: Inferred dark energy equation of state. The left panel uses simulated data with increasing assumed error bars. The right panel uses the Union2.1 and the Pantheon+ dataset, propagating the error bars estimated by the collaborations through the PINN inference.

that PINNs extrapolate into regions of sparse or low-quality data, while sensibly increasing their learned error in these regions. Testing these aspects with the harmonic oscillator as a toy example confirms the fundamental behavior of PINNs.

The functionality of PINNs as emulators was then verified with luminosity distances as functions of redshift for a conventional dark-energy dominated Friedmann-Robertson-Walker universe. PINNs correctly predict the luminosity distance for a given redshift over a wide range of dark energy equation of state parameters, without performing a numerical integration in the forward simulation.

Using PINNs for inference rather than emulation requires a statistical inversion, i.e. a mapping of the experimental uncertainty back to the parameterization. Applied to the supernova example, PINNs allow for an uncertainty-aware reconstruction of the Hubble function without any predefined parameterization. The Hubble function is reconstructed by the PINN including an error estimate. They discover peculiarities in the data, such as the sudden increase in error in the Union data set at $z \simeq 0.3$, reflecting a large uncertainty in the reconstructed Hubble function. Re-expressing the Hubble function with the dark energy equation of state function derived for a fixed matter density shows weaker constraints, as the increase in error is driven by the derivative transitioning from $H(a)$ to $w(a)$.

6 Parallelizing Madelung Modes during Inflation

This chapter introduces the parallelized inflation solver PARALLIZIS. Section 6.1 derives the Madelung transformed mode equations used in the solver, while section 6.2 constructs a parallelized approach to the mode equations using the ODE solvers implemented in TORCHDIFFEQ. Additionally, section 6.3 describes emulators to translate from primordial power spectra to observable power spectra based on fully connected neural networks.

6.1 Madelung mode equations

This section collects theoretical insights into the perturbation equations of single-field inflation. Applying the Madelung transformation from quantum mechanics to the mode equations yields a conserved quantity similar to a conserved angular momentum in the complex plane.

6.1.1 Mode equations during inflation

In single-field inflation the evolution of the perturbation modes is governed by the mode equation (4.22). Usually the evolution starts deep inside the horizon, at $100aH = k$ [Lesgourgues et al., 2008, Mortonson et al., 2011], such that the Bunch-Davies initial conditions (4.26) can be applied. This section is concerned with the evolution of the mode after that and in particular the expected freezing of the mode when $aH \ll k$. All Figures in this section depict the curvature mode u_{k_*} associated to the pivot scale $k_* = 0.05\ \mathrm{Mpc}^{-1}$.

Figure 6.1 depicts the conformal time evolution of the Mukhanov-Sasaki variable u. The absolute values of the curvature modes, shown in red, are accessible through the primordial power spectrum. The left panel shows the evolution of the mode transformed with an inverse hyperbolic sine, this allows to capture the rapid change of the curvature at late times. This happens as the mode crosses the horizon, indicated by the dashed vertical line. The transition appears fast in conformal time since the comoving horizon changes quickly at late times, see Figure 6.5. Since the curvature modes are constructed from the Mukhanov-Sasaki variables u/z the rapid increase in u_{k_*} is moderated by a similar increase in $z = \frac{\partial_\eta \varphi}{H}$. This behavior is depicted in the right panel of Figure 6.1.

Transitioning to $\ln a$ gives a better intuition into the evolution of the mode as it crosses the horizon. For this time variable the mode equations in eqn. (4.22) are transformed

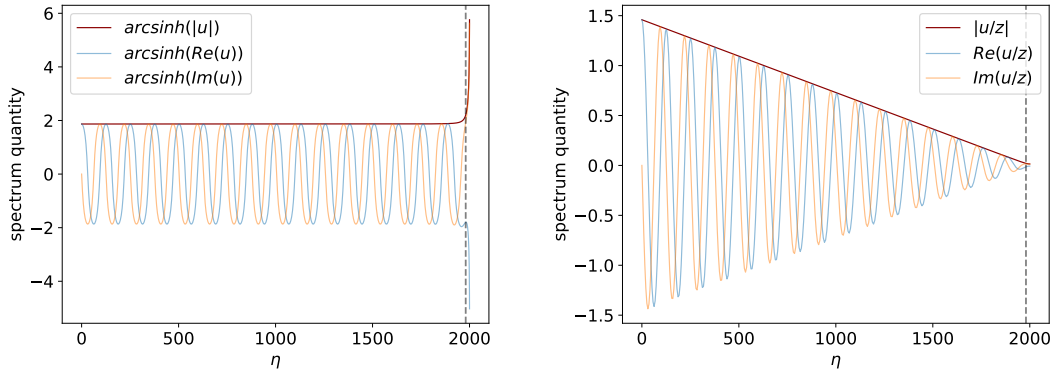

Figure 6.1: Evolution of the curvature perturbation mode u_{k_*} for $k_* = 0.05\ \mathrm{Mpc}^{-1}$ in conformal time. Blue lines depict the real part of the mode equations, while orange is the imaginary part. Red lines show their absolute value, while the gray line marks where the mode crosses the horizon. The left panel shows the evolution of the Mukhanov-Sasaki modes, while the right panel shows the evolution of the curvature modes directly.

using $\mathrm{d}\ln a = aH\mathrm{d}\eta$ to read

$$\begin{aligned} u'' + \left(\frac{H'}{H} - 1\right) u' + \left[\frac{k^2}{a^2H^2} - \left(2 - 2\left(\frac{H'}{H}\right)^2 - 4\frac{H'}{H}\frac{\varphi''}{\varphi'} + 5\frac{H'}{H} - \frac{\partial_\varphi^2 V}{H^2}\right)\right] u = 0 \\ v'' + \left(\frac{H'}{H} - 1\right) v' + \left[\frac{k^2}{a^2H^2} - \left(2 - \frac{H'}{H}\right)\right] v = 0. \end{aligned} \tag{6.1}$$

Similar to the Bunch-Davies initial conditions in conformal time, see eqn. (4.26), we can find an expression for the initial conditions of the modes in $\ln a$ as

$$\begin{aligned} \mathrm{Re}(u_k) = \mathrm{Re}(v_k) &= \frac{1}{\sqrt{2k}} \\ \mathrm{Im}(u_k) = \mathrm{Im}(v_k) &= 0 \\ \mathrm{Re}(\partial_{\ln a} u_k) = \mathrm{Re}(\partial_{\ln a} v_k) &= 0 \\ \mathrm{Im}(\partial_{\ln a} u_k) = \mathrm{Im}(\partial_{\ln a} v_k) &= \frac{-kaH}{\sqrt{2k}}. \end{aligned} \tag{6.2}$$

The evolution of the mode in terms of the logarithmic scale factor is shown in Figure 6.2. Using this time variable the oscillations do not appear to have a fixed period, and the Mukhanov-Sasaki variable in the left panel grows more gradually. The right panel validates the approach to stop the evolution soon after horizon crossing, as the mode stops evolving. In all numerical calculations we stop computing the evolution of the mode when $aH = 50k$ similar to the approach in CLASS. When this condition is fulfilled the curvature mode is far outside the horizon and its evolution is frozen.

This behavior can also be inferred directly from the differential equation for $\tilde{u}_k = \frac{u}{z}$ and

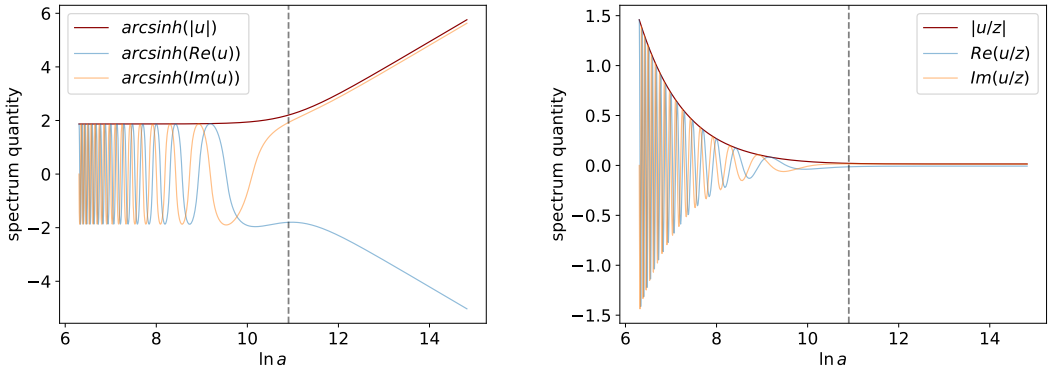

Figure 6.2: Evolution of the curvature perturbation mode u_{k_*} for $k_* = 0.05\ \mathrm{Mpc}^{-1}$ in $\ln a$. Blue lines depict the real part of the mode equations, while orange is the imaginary part. Red lines show their absolute value, while the gray line marks where the mode crosses the horizon. The left panel shows the evolution of the Mukhanov-Sasaki modes, while the right panel shows the evolution of the curvature modes directly.

$\tilde{v}_k = \frac{v}{a}$, which can be expressed as

$$\begin{aligned} \tilde{u}'' + \left(3 + \frac{H'}{H} + 2\frac{\varphi''}{\varphi'}\right)\tilde{u}' + \frac{k^2}{a^2H^2}\tilde{u} = 0 \\ \tilde{v}'' + \left(3 + \frac{H'}{H}\right)\tilde{v}' + \frac{k^2}{a^2H^2}\tilde{v} = 0. \end{aligned} \tag{6.3}$$

The changing period in oscillations is an effect of the comoving horizon appearing in the third term. Additionally, for large aH, corresponding to late times the last term in the equations is approximately zero and a constant is a solution to the differential equation. The initial conditions for this set of differential equations can be derived from the initial conditions for u and v in eqn. (6.2).

6.1.2 Madelung transformation

The previous section is concerned with the evolution of the real and imaginary parts of the curvature modes during inflation. However, after the evolution stops only the radial part of the complex-valued mode is used in the computation of the power spectrum, see eqn. (4.29). This section derives a differential equation for the radial part and the phase velocity of the modes similar to the derivation of the Madelung equation [Madelung, 1926, 1927] in quantum mechanics.

The differential equations describing the curvature evolution share the common shape

$$y''(x) + f(x)y'(x) + g(x)y(x) = 0. \tag{6.4}$$

Here x can be any time variable and primes denote derivatives with respect to it. Additionally, $f(x)$ and $g(x)$ are real-valued such that the differential equation for the complex conjugate is the same. The Mukhanov-Sasaki potentials, and their derivatives,

are expressed as

$$\begin{aligned} y &= \rho e^{i\alpha} \\ y' &= \left(\rho' + i\rho\alpha'\right) e^{i\alpha} \\ y'' &= \left(\rho'' + 2i\rho'\alpha' + i\rho\alpha'' - \rho\alpha'^2\right) e^{i\alpha}. \end{aligned} \tag{6.5}$$

This description is used to construct two independent differential equations from the symmetric and antisymmetric combinations of the modes y and their derivatives y'. In terms of the radial part ρ and the phase α the antisymmetric combination reads

$$y^* y' - y(y^*)' = 2i\rho^2\alpha'. \tag{6.6}$$

In the following derivation y^* denotes complex conjugate of y. Taking the derivative and inserting the information from 6.4 yields

$$\begin{aligned} \left(2i\rho^2\alpha'\right)' &= y^* y'' - y(y^*)'' \\ &= y^*(-fy' - gy) - y(-f(y^*)' - gy^*) \\ &= -f(y^* y' - y(y^*)') = -2if\rho^2\alpha'. \end{aligned} \tag{6.7}$$

The symmetrical analogue reads

$$y^* y' + y(y^*)' = 2\rho\rho' = (\rho^2)', \tag{6.8}$$

resulting in the differential equation

$$\begin{aligned} (\rho^2)'' &= y^* y'' + y(y^*)'' + 2y'(y^*)' \\ &= 2y'(y^*)' - f(y^* y' + y(y^*)') - 2gy^* y \\ &= 2\left(\rho' + i\rho\alpha'\right)\left(\rho' - i\rho\alpha'\right) - f(\rho^2)' - 2g\rho^2 \\ &= 2(\rho')^2 - 2\rho^2(\alpha')^2 - f(\rho^2)' - 2g\rho^2. \end{aligned} \tag{6.9}$$

Evaluating the derivatives yields the second differential equation

$$\begin{aligned} 2\rho''\rho + 2(\rho')^2 &= 2(\rho')^2 - 2\rho^2(\alpha')^2 - 2f\rho'\rho - 2g\rho^2 \\ \implies \rho'' &= -\rho(\alpha')^2 - f\rho' - g\rho. \end{aligned} \tag{6.10}$$

The full mode equations then read

$$\begin{aligned} (\rho^2\alpha')' + f\rho^2\alpha' &= 0 \\ \rho'' + f\rho' + \left((\alpha')^2 + g\right)\rho &= 0. \end{aligned} \tag{6.11}$$

For any integrable function $f(x)$ with primitive $F(x)$ the first differential equation can be used to identify a constant $\exp(F)\rho^2\alpha'$. Substituting in the Bunch-Davies initial conditions yields

$$-\exp(F(x))\rho^2(x)\alpha'(x) = \frac{1}{2}. \tag{6.12}$$

While the initial conditions fix $\alpha'(x_0) < 0$, the conservation equation ensures that the sign of α' never changes. For $y = u$ and $x = \eta$ the constant can be identified as $\rho^2(\eta)\partial_\eta\alpha(\eta) = -\frac{1}{2}$. This is similar to angular momentum conservation in classical mechanics. The solution to the mode equation rotates around $\rho = 0$ in the complex plane. The derivative of the complex phase plays the role of angular velocity. The modulus of

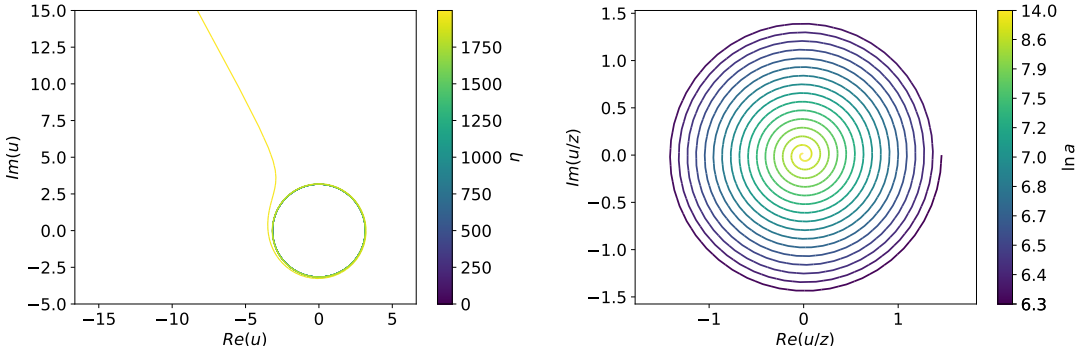

Figure 6.3: Evolution of the curvature perturbation mode u_{k_*} for $k_* = 0.05\ \mathrm{Mpc}^{-1}$ in the complex plane. The left panel depicts the evolution of the Mukhanov-Sasaki potential u_{k_*} in conformal time directly. The right panel is the evolution of the curvature mode u_{k_*}/z.

the complex number is analogous to the radius in angular movement.

Figure 6.3 gives a visual representation of the conserved quantities. The left panel depicts the case of $y = u$ and $x = \eta$. As the angular velocity approaches zero the radial part of the mode diverges. Similarly, the right panel shows the case of $y = \frac{u}{z}$ and $x = \ln a$. In this case eqn. (6.3) suggests $f(\ln a) = \left(3 + \frac{H'}{H} + 2\frac{\varphi''}{\varphi'}\right)$ and consequently $F(\ln a) = 3\ln a + \ln H + 2\ln\varphi'$. The conservation equation then reads

$$a^3 H(\varphi')^2 \rho^2 \alpha' = -\frac{1}{2}. \tag{6.13}$$

As the scale factor increases during inflation both the radial part and the angular velocity decrease.

In PARALLIZIS the evolution of the perturbation modes is computed using the Madelung transformed mode equations (6.11) with $y = \frac{u}{z}$ and $x = \ln a$. The differential equation is further rewritten as an ODE in terms of $\ln\rho$ since the radial part of $\frac{u}{z}$ is always positive and crosses orders of magnitude in its evolution. Using a similar argument the angular velocity enters the differential equation as $\ln(-\alpha')$. The equations used in the numerical computations then read

$$\begin{aligned} 2(\ln\rho)' + (\ln(-\alpha'))' + f &= 0 \\ (\ln\rho)'' + (\ln\rho)'\left[(\ln\rho)' + f\right] + g - \exp(2\ln(-\alpha')) &= 0. \end{aligned} \tag{6.14}$$

Here $f = \left(3 + \frac{H'}{H} + 2\frac{\varphi''}{\varphi'}\right)$ and $g = \frac{k^2}{a^2H^2}$. The first equation in combination with the constant in eqn. (6.13) is used to express the angular velocity in terms of the radial part. With this choice of variables the initial conditions read

$$\begin{aligned} \rho &= \frac{1}{\sqrt{2k}}\frac{1}{(a\varphi')} \\ (\ln\rho)' &= -\left(1 + \frac{\varphi''}{\varphi'}\right) \\ \alpha' &= -\frac{k}{aH}. \end{aligned} \tag{6.15}$$

Remaining in conformal time and using a naive SCIPY implementation Madelung transformed mode equations lead to a speed improvement of roughly a factor two, averaged over modes spaced linearly in $\ln k \in [-4, 0]$. While the speed improves for all wavenumbers smaller wavenumbers gain a larger improvement. This can be attributed to the longer integration times needed for smaller wavenumbers, see Figure 6.6. Using $\ln a$ as a time variable for the Madelung transformed equations does not have a large impact on the computation time.

6.2 PARALLelIZed Inflation Solver

This section is designed as a tour through the parallelized inflation solver PARALLIZIS. It starts with the definition of the inflationary potential and follows the steps performed in the solver to arrive at a numerical solution for the primordial power spectrum.

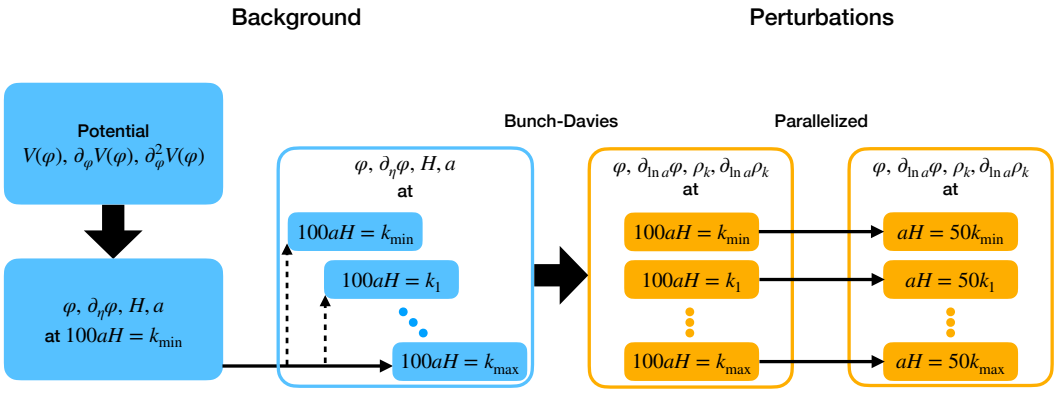

Figure 6.4: PARALLIZIS graphical overview. The inflaton potential and background evolution are described in section 6.2.1. The perturbations and the parallelization are described in section 6.2.2.

6.2.1 Potential to initial conditions

The inflaton potential

The inflationary paradigm implies a shrinking comoving horizon. This translates to a period in the cosmic expansion history where the universe is dominated by a fluid with an equation of state $w_\varphi < -\frac{1}{3}$. While there are further constraints from observational data, such as the CMB anisotropies and Large Scale Structures (LSS), the shape of the inflaton potential is not known. Descriptions of current models and observational constraints on them can be found in [Akrami et al., 2020].

For all numerical experiments in this chapter the inflaton potential is characterized as

$$V(\varphi) = v_0 + v_1(\varphi - \varphi_*) + \frac{v_2}{2}(\varphi - \varphi_*)^2 + \frac{v_3}{6}(\varphi - \varphi_*)^3 + \frac{v_4}{24}(\varphi - \varphi_*)^4, \tag{6.16}$$

the parameters are described using the potential slow-roll parameters [Leach et al., 2002]

$$\begin{aligned}
\epsilon_V &= \frac{1}{2}\frac{m_{pl}^2}{8\pi}\left(\frac{\partial_\varphi V}{V}\right)^2\Bigg|_{\varphi=\varphi_*} \\
\eta_V &= \frac{m_{pl}^2}{8\pi}\frac{\partial_\varphi^2 V}{V}\Bigg|_{\varphi=\varphi_*} \\
\xi_V^2 &= \frac{m_{pl}^4}{(8\pi)^2}\frac{\partial_\varphi V \partial_\varphi^3 V}{V^2}\Bigg|_{\varphi=\varphi_*} \\
\omega_V^3 &= \frac{m_{pl}^6}{(8\pi)^3}\frac{(\partial_\varphi V)^2 \partial_\varphi^4 V}{V^3}\Bigg|_{\varphi=\varphi_*}.
\end{aligned} \tag{6.17}$$

Since the inflaton field has units of Planck mass $m_{pl} = (G)^{-1/2}$ it is explicitly written out to emphasize that these parameters are dimensionless. They are evaluated at the inflaton field value at horizon crossing of the pivot scale. The amplitude of both the potential and the primordial power spectrum is determined by the additional parameter [Lesgourgues and Valkenburg, 2007]

$$\frac{128\pi}{3m_{pl}^6}\frac{V^3}{(\partial\varphi V)^2}\Bigg|_{\varphi=\varphi_*}. \tag{6.18}$$

This parametrization of the inflaton potential already avoids specifying a model. However, all calculations performed on this section can be performed for any inflaton potential as long as its first and second derivatives are implemented as a python function and the potential results in an inflating universe until all length scales of interest leave the horizon.

Background evolution

The background evolution of the inflaton is performed in conformal time η. Accordingly, the differential equation governing the background evolution, eqn. (4.16), is transformed using the prescription $dt = ad\eta$, to read

$$\begin{aligned}
\frac{\partial^2\varphi}{\partial\eta^2} + 2aH\frac{\partial\varphi}{\partial\eta} &= -a^2\frac{\partial V}{\partial\varphi} \\
\frac{\partial H}{\partial\eta} &= -\frac{8\pi}{2m_{pl}^2 a}\left(\frac{\partial\varphi}{\partial\eta}\right)^2 \\
\frac{\partial a}{\partial\eta} &= a^2 H.
\end{aligned} \tag{6.19}$$

Here, the evolution of the Hubble function is encoded in the second line through the derivative of the Friedmann equation. To ensure that the solution fulfills the Friedmann equations the initial conditions must fulfill

$$H^2 = \frac{8\pi}{3m_{pl}^2}\left(\frac{1}{2a^2}\left(\frac{\partial\varphi}{\partial\eta}\right)^2 + V(\varphi)\right). \tag{6.20}$$

With this time variable both the Friedmann equation and the evolution equation of the inflaton contain the scale factor explicitly. Its evolution in conformal time, the third line of eqn. (6.19), is determined from the definition of the Hubble function $H = \frac{\dot{a}}{a}$.

Since this differential equation does not explicitly depend on conformal time, the equations are invariant under conformal time translation. While solving the equation requires a choice of initial condition the conformal time at which they describe the physical system is always set to $\eta = 0$. The equations are also invariant under rescaling of the scale factor when taking into account that this also rescales conformal time. Additionally, the initial value of the inflaton field is degenerate with the normalization of the potential at $V(\varphi(0))$.

The initial conditions of the background equations are found using the same approach as in CLASS [Blas et al., 2011]. The field value of the inflaton at horizon crossing of the pivot scale is chosen as $\varphi_* = 0$. Starting from this the other function values are chosen such that $(\varphi_*, \partial_\eta\varphi_*, H_*, a_0 = 1)$ constitute an attractor solution. Numerically this is done by finding the derivative of the inflaton field and the Hubble function from the slow-roll approximation to the background equation (4.20). These values are then used to evolve backward in time, still using the slow-roll approximation. Next, the values at this earlier time are evolved forward using eqn. (6.19) until $\varphi(\eta) = \varphi_*$. This yields updated values of $\partial\varphi$ and the Hubble function when $\varphi(\eta) = \varphi_*$. This process is repeated until the value of $\partial\varphi$ at the end of the forward evolution does not vary by more than a tunable precision parameter. To ensure that the pivot scale k_* crosses the horizon at these field values the scale symmetry of the scale factor is used to set $a_* = \frac{k_*}{H_*}$.

This set of conditions anchors the background equation in time. To capture the evolution of the smallest comoving wavenumbers we evolve backward in time until $100aH = k_{\mathrm{min}}$. At this time we find the initial conditions of the background differential equation $(\varphi_0, \partial_\eta\varphi_0, H_0, a_0)$. It is worth noting that the translation invariance with conformal time is used in this derivation to start any evolution of the background at $\eta = 0$. Once the field value at horizon crossing of the pivot scale is fixed, the scale factor, or the comoving horizon, indicates the state of the system and the passage time between different solutions of the differential equation.

Initial conditions of the perturbation equations

The evolution of the perturbation equations is determined by eqn. (4.22) and eqn. (4.24), while their initial conditions are set using the Bunch-Davies vacuum in eqn. (4.26). Since the background fields enter the description of the perturbation equations both sets of differential equations are solved at the same time. To capture the evolution of a mode with comoving wavenumber k it is evolved from the time when $100aH = k$ where the Bunch-Davies initial conditions can be used. Consequently, the full set of initial conditions in conformal time is obtained by evolving eqn. (6.19) forward starting from the background initial conditions and appending the Bunch-Davies vacuum at $100aH = k$. In a parallelized setting the initial conditions are found for each comoving wavenumber k in the observable range.

From a numerical point of view ending the integration of the differential equation when a specific set of conditions is fulfilled requires keeping track of these conditions in an event function. The ODE solver TORCHDIFFEQ [Chen et al., 2018] offers this capability [Chen et al., 2021]. Additionally, since the solvers are PyTorch based it is possible to differentiate through them using automatic differentiation.

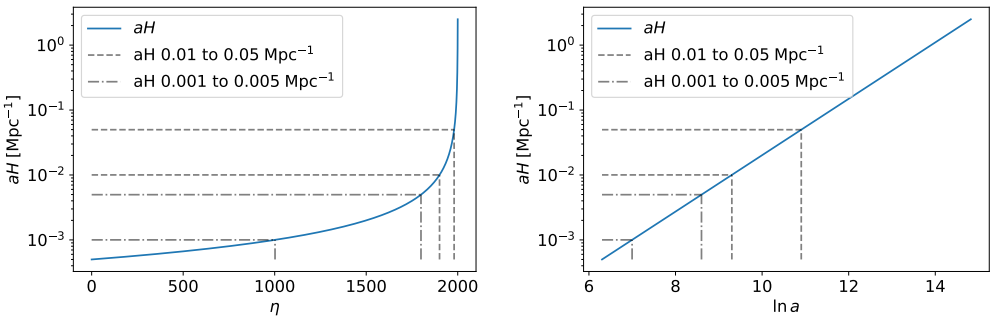

Figure 6.5: Inverse comoving horizon evolution. The left panel depicts the evolution in terms of conformal time. The right panel shows the evolution in $\ln a$. The dashed lines show a comoving horizon interval such that $\frac{k}{aH} \in [1, 5]$ for $k = 5 \cdot 10^{-2}$ Mpc^{-1}. The dot-dashed lines show the same interval for $k = 5 \cdot 10^{-4}$ Mpc^{-1}.

6.2.2 Parallel perturbations

The time evolution of the Mukhanov-Sasaki potentials is sensitive to the size of the comoving horizon. This can be seen in the simplifications of the differential equations at $aH \gg k$ where the Bunch-Davies initial conditions are implemented. Additionally, at $aH \ll k$ the corresponding mode freezes. While TORCHDIFFEQ requires only some modification to allow for parallelized event functions it is designed to work with only one time variable.

The left plot in Figure 6.5 demonstrates the evolution of the comoving horizon in conformal time. For each mode the interval between the starting condition and the stopping condition is of similar size in $\ln aH$. However, the size of the intervals in conformal time can be vastly different. A parallel integration of all modes would require integrating over the whole time interval needed for the smallest mode. The left-hand side of Figure 6.6 depicts the time evolution of three different modes. They show an oscillating behavior at different time scales. A solver addressing all modes in parallel would require time steps resolving these oscillations for all modes at once. This makes the parallel numerical computation in conformal time using one time axis computationally challenging.

This problem is addressed by choosing a time variable more closely aligned to the comoving horizon. Since the evolution of the curvature modes happens during inflation, the Hubble function varies slowly, following equation (4.18). The evolution of the comoving horizon is driven by the growth of the scale factor. Following this intuition we choose $\ln a$ as a time variable similar to [Mortonson et al., 2011]. The right-hand side of Figure 6.5 demonstrates that this leads to $\ln a$ intervals of approximately equal size for different modes.

The background differential equations can be transformed using the prescription $\mathrm{d} \ln a = aH \mathrm{d}\eta$ to read

$$\begin{aligned} \varphi'' + \left(\frac{H'}{H} - 3\right)\varphi' &= -\frac{1}{H^2}\partial_\varphi V \\ H^2 &= \frac{2V(\varphi)}{\frac{6m_{pl}^2}{8\pi} - (\varphi')^2}. \end{aligned} \tag{6.21}$$

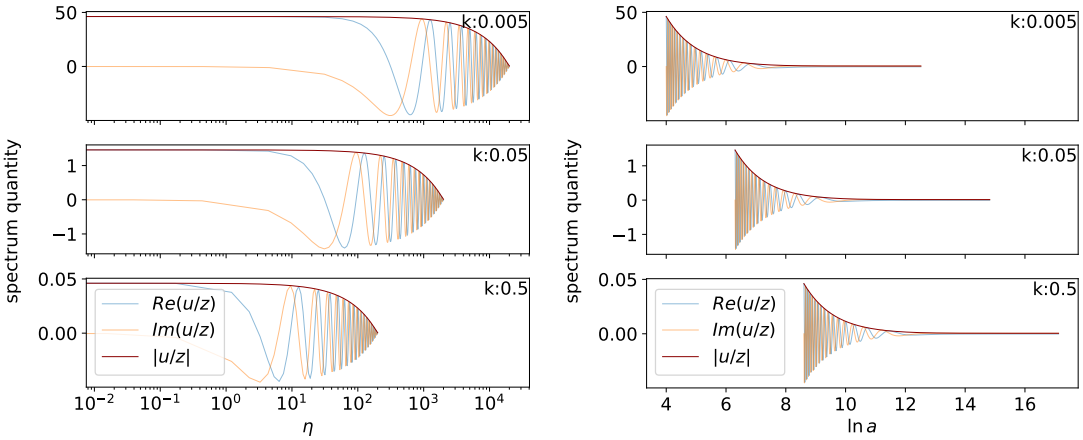

Figure 6.6: Conformal time and $\ln a$ evolution of three different modes. In conformal time, left panel, the numerical computation always starts $\eta = 0$ and different modes need to be evolved for different amounts of conformal time, until u/z freezes. In $\ln a$, right panel, the starting scale factor is shifted, however, the time intervals are of equal size.

Where the primes denote derivatives with respect to $\ln a$. Additionally, the derivative of the Hubble function is expressed as

$$H' = \frac{8\pi}{2m_{pl}^2}(\varphi')^2 H. \tag{6.22}$$

Similarly, the mode equations are transformed into the shape in eqn. (6.1). Note that the time variable explicitly appears in these equations. They are not invariant under translation in $\ln a$.

This behavior is depicted on the right side of Figure 6.6. While the intervals of interest are of similar size they have different starting points. The intervals are brought to the same starting point by shifting in $\ln a$ while explicitly shifting the scale factor a by different amounts for each mode $a \to a - a_{\text{ini}}(k)$. Here $a_{\text{ini}}(k) = \frac{k}{100H}$ is determined by the evolution of the comoving horizon and the comoving wavenumber of the mode.

Computing the modes in parallel on the GPU leads to a significant speedup compared to solving the same equation using TORCHDIFFEQ on the CPU. However, the parallelized computation is roughly a factor two slower than the primordial module of CLASS which uses CPU parallelization and a C based differential equation solver. Solving the perturbation equations without parallelizing with the python based TORCHDIFFEQ is an order of magnitude slower than SCIPYs `solve_ivp`, depending on the tolerances required of the solvers.

6.2.3 Primordial power spectrum

The parallel calculation in the previous section yields all the Mukhanov-Sasaki modes at the end of inflation. From these the primordial power spectrum is computed as in eqn. (4.29). This numerical calculation of the primordial power spectrum can then be tested against analytical approximations of the power spectrum. To additionally allow comparison to CLASS the potential of the inflaton is chosen as in eqn. (6.16), parametrized

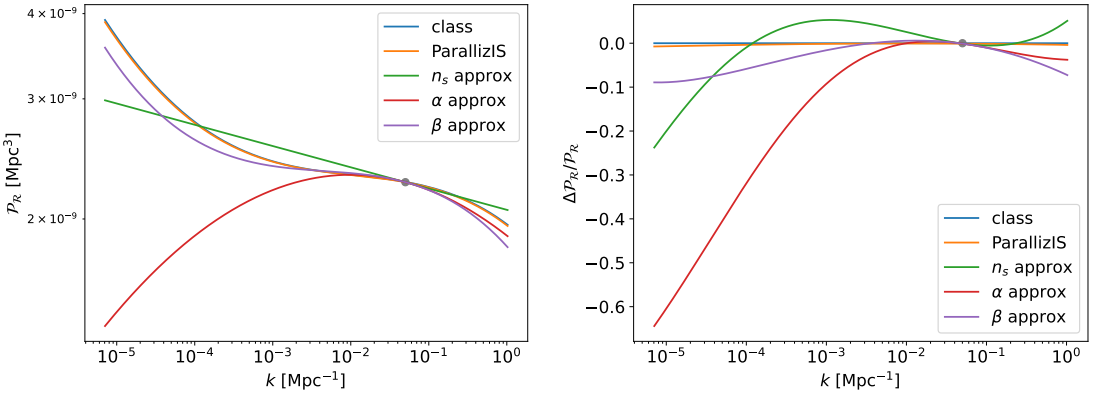

Figure 6.7: The left plot shows a comparison of the primordial power spectrum obtained from a simulation with PARALLIZIS, CLASS and approximations using eqn. (6.23). The right plot depicts the relative difference between them.

by the potential slow-roll parameters. Based on them the scale dependence of the primordial power spectrum is computed as [Kohri et al., 2013]

$$\begin{aligned} n_s - 1 =& -6\epsilon_V + 2\eta_V + \left(-\frac{10}{3} + 24C\right)\epsilon_V^2 + \frac{2}{3}\eta_V^2 \\ & - (2 + 16C)\epsilon_V\eta_V + \left(\frac{2}{3} + 2C\right)\xi_V^{(2)} \\ \alpha =& -24\epsilon_V^2 + 16\epsilon_V\eta_V - 2\xi_V^{(2)} \\ \beta =& -192\epsilon_V^3 + 192\epsilon_V^2\eta_V - 32\epsilon_V\eta_V^2 + (-24\epsilon_V + 2\eta_V)\xi_V^{(2)} + 2\omega_V^{(3)}, \end{aligned} \tag{6.23}$$

where $C = 4(\ln(2) + \gamma_e) - 5$ and $\gamma_e \approx 0.577$.

Figure 6.7 depicts the primordial power spectra predicted by PARALLIZIS, CLASS and the analytical approximations. While the predictions of CLASS and PARALLIZIS agree at the percent level the analytical approximations struggle to capture the shape of the spectrum away from the pivot scale. Note that while the deviations of the analytical approximations are smaller close to the pivot scale they can reach a level of 10% in the regions Figure 6.11 identifies as relevant to the Planck 2018 data.

6.3 Evolution after horizon reentry

While the primordial power spectrum is not directly observable it can be constrained based on observations of the angular power spectra of the CMB and the matter power spectrum. They are related through transfer functions as specified in section 4.3. Traditionally, this evolution is performed numerically using Boltzmann codes such as CAMB [Lewis et al., 2000] and CLASS [Blas et al., 2011]. More recently emulators for these codes that replace either part of or the full calculation with a neural net have been published in [Spurio Mancini et al., 2022, Nygaard et al., 2023, Günther et al., 2022]. In the most direct approach a fully connected network maps cosmological parameters to observable power spectra. The drawback of this approach is that each emulator is restricted to a specific model of inflation. This section describes an emulation approach directly based on the primordial power spectrum instead of a set of parameters describing a model. In

this work the cosmological parameters after the end of inflation are fixed to the default values of CLASS. We train a separate emulator each for the matter power spectrum, and the TT, TE, EE spectra of the CMB. While it is possible to emulate the BB spectrum, based on the primordial tensor power spectrum we leave this to future work.

For all of these emulation tasks the scalar power spectrum is determined by the parameters $[A_s, n_s, \alpha_s, \beta_s]$ through eqn. (4.32), using a pivot scale of $k_* = 0.005$ Mpc^{-1}. The parameters are sampled from one-dimensional Gaussian distributions with the means given in [Akrami et al., 2020]. This spectrum is discretized on the range $[10^{-6}, 1]$ Mpc^{-1} using 600 equally spaced logarithmic wavenumbers. These primordial power spectra are used as the input for CLASS to generate both the matter power spectrum at $z = 0$ and the angular power spectra of the CMB. These constitute the labels in the network training. For all emulation tasks we use a heteroscedastic loss (4.56), the trained networks are more precise than their counterparts trained with an MSE loss, similar to the observation in 5.2.

6.3.1 Matter power spectrum

The map between the primordial power spectrum and the matter power spectrum is described schematically in eqn. (4.43). This links the primordial power spectrum to the matter power spectrum at a scale k in a scale-dependent way. Consequently, the emulator is designed to map from the tuple $(k, P_{\mathcal{R}}(k))$ to $P_m(k)$, learning the transfer function directly. The matter power spectrum emulator uses a fully connected neural network with four hidden layers of width 256. The network is trained for 200 epochs in a batch learning setup with batches of size 256.

In practice both the network inputs and outputs are preprocessed such that they are distributed around a mean of zero with a width of one at each wavenumber k. We generate 900 spectra and split them into a training, validation and test set containing 300 spectra each. The training set is used to find the mean and standard deviation of the logarithm in both the training data and the labels. These are then used to define a preprocessing and deprocessing operation for both the training data and the labels.

The left panel of Figure 6.8 depicts the distribution of the matter power spectra from the test set in the left panel in gray. The dark gray band depicts the standard deviation while the light gray is twice the standard deviation. This matches almost perfectly with the orange error band depicting a deprocessed standard normal. Since the training data is generated using Planck constraints this gives an intuition in what wavenumber ranges Planck is the most constraining. Note that we have disregarded correlations between the spectral parameters and restricted ourselves to a specific representation of the power spectrum.

Figure 6.9 shows the precision of the matter power spectrum emulator. The left panel compares the preprocessed results of the emulator to the labels of the test data set. The right panel depicts the relative error of ten deprocessed matter power spectra, demonstrating a precision in the permille range. Note that the matter spectrum emulator covers comoving wavenumbers far below the pivot scale k_*, which are not observable with current surveys.

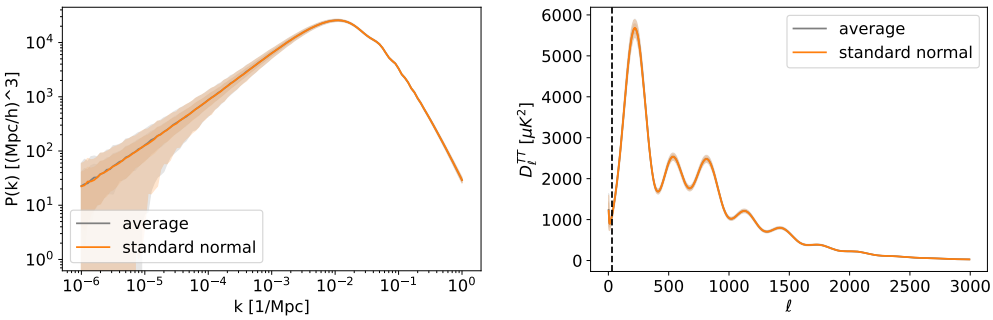

Figure 6.8: The left plot shows the average matter power spectrum in gray as well as the result of deprocessing a standard normal in orange. The right plot shows the average temperature anisotropy spectrum in gray as well as the result of deprocessing a standard normal in orange.

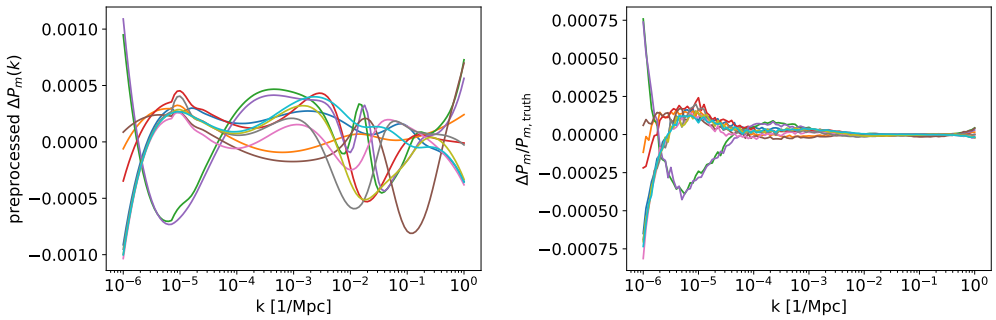

Figure 6.9: Precision of the matter power spectrum emulator. Each color represents a different sample from the test set. The left plot is the difference between the preprocessed test labels and the network output, while the right plot is the relative error on the matter power spectrum at $z = 0$.

6.3.2 Angular power spectra

In contrast to the construction of the matter power spectra the theoretical computation of the angular power spectra in eqn. (4.42) contains an integration over all scales. Each mode in the angular power spectra is influenced by the primordial power spectrum at all comoving wavenumbers. The angular spectrum emulators map from the whole primordial power spectrum $P(k)$, evaluated at 600 wavenumbers, to the whole angular power spectra. While the structure of the theoretical computation suggests using a convolutional architecture [Fukushima, 1980] it is outperformed by a sufficiently big fully connected network capturing all the interactions between different inputs. For all further numerical experiments we used an fully connected network with four hidden layers of width 256 each. The networks are trained on 3000 spectra.

Similar to the matter spectrum emulator the training data and labels are preprocessed to resemble a standard normal before training the network. The right panel in Figure 6.8 depicts the average angular temperature anisotropy spectrum, based on 200 test samples, in gray. The orange error band corresponds to a deprocessed standard normal. Varying the primordial power spectrum within the Planck 2018 limits of the spectral parameters allows for some freedom in the value of the peaks of angular power spectra but not their position. This is in line with the intuition that the position of the peaks is determined

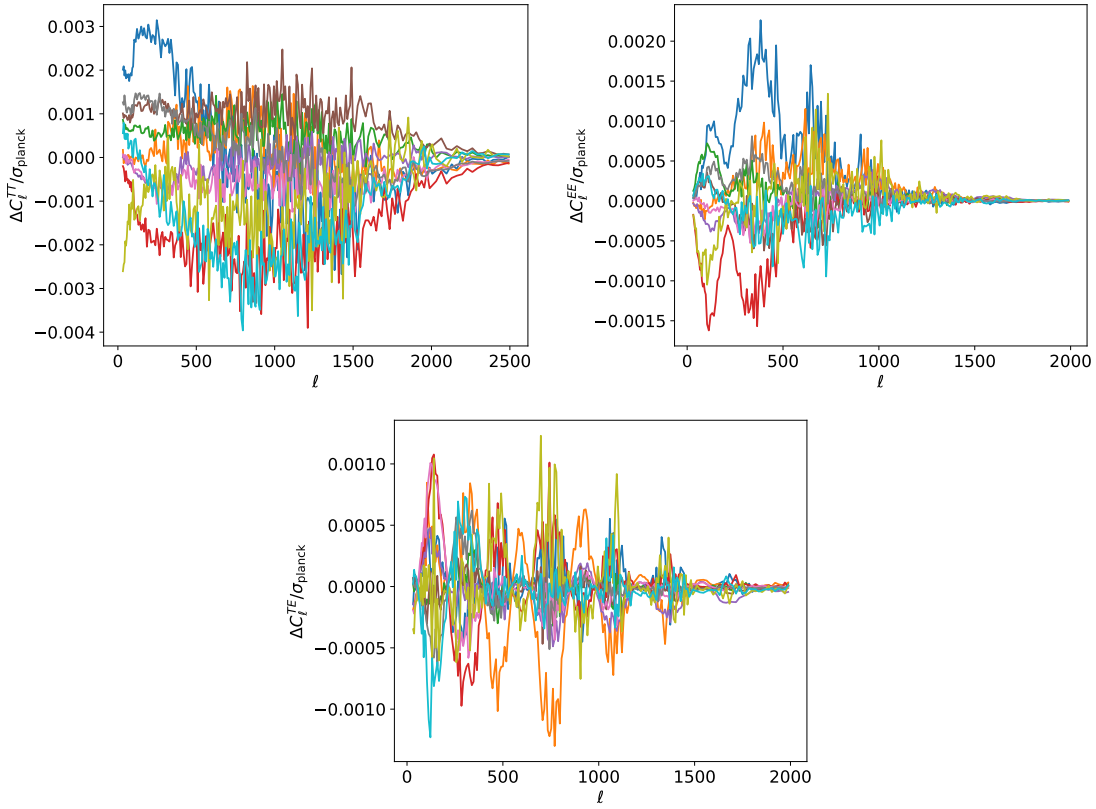

Figure 6.10: Precision of the angular power spectrum emulators. The plots show the ratio of the difference between deprocessed test labels and emulator results and the binned Planck 2018 error bars. The left plot is the TT emulator, the right for the EE emulator and the middle plot is the comparison for the TE emulator.

by baryonic acoustic oscillations, which take place after inflation ends and the observable modes have reentered the horizon.

Figure 6.10 compares the difference between the emulator solution and the labels in the test dataset to the binned Planck 2018 error bars. The angular power spectra are binned using the code released together with [Prince and Dunkley, 2019] and compared to the diagonal of the binned covariance matrix used in the Planck lite likelihoods of [Aghanim et al., 2020c]. The accuracy of the emulators is smaller than the experimental uncertainty in Planck 2018 by at least a factor of 100. While the accuracy of the emulator at each ℓ is approximately proportional to the value of the spectrum, the experimental error bars in Figure 4.1 are not. This discrepancy leads to the visible imprints of the CMB peaks in the difference plots. The behavior is particularly pronounced in the TE spectrum.

The emulator can also be used to gain an intuition at what wavenumbers of the Primordial power spectrum are best constrained by Planck data. The left plot in Figure 6.11 depicts the average primordial power spectrum generated from the Planck constraints. They constrain the primordial power spectrum well in the wavenumber range $[10^{-3}, 1]$ Mpc^{-1}. The right plot shows the derivative of the binned Planck likelihood with respect to the primordial power spectra. The primordial spectrum is passed through the angular power spectrum emulators, the resulting spectra are then used to compute the binned likelihood. Derivatives through this forward evolution are obtained using automatic differentiation.

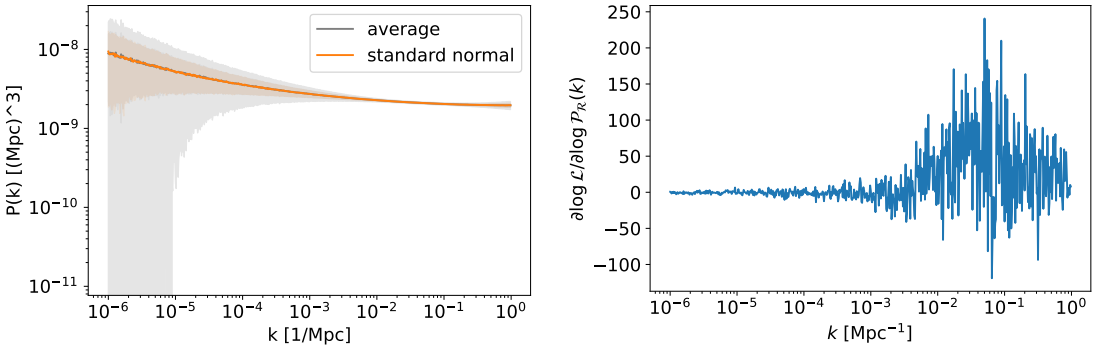

Figure 6.11: Connection between the primordial power spectrum and the Planck likelihood. The left panel shows the training data of the emulators obtained from the Planck 2018 constraints. The right panel shows influence primordial power spectrum on the Planck likelihood as a function of comoving wavenumber.

This plot demonstrates, that the likelihood is most sensitive to changes in the primordial power spectrum in a similar wavenumber range.

It is worth noting that both of these plots are biased by the choice of primordial power spectrum parametrization. In the left plot choosing a less constraining parametrization would allow increased error bars, especially away from the well-constrained regions. The right plot is obtained through differentiating through emulators trained on the same data set, making it similarly susceptible to changes in the parametrization.

6.4 Summary and discusison

The theory part of this chapter introduces the Madelung transformation for mode equations during inflation, in section 6.1. They allow the formulation of a conservation equation for single-field inflation which is similar to angular momentum conservation in the complex plane. Together with the Bunch-Davies initial conditions we find that the radial part and the phase velocity of the modes in conformal time combine to the conserved quantity

$$\rho^2 \partial_\eta \alpha = -\frac{1}{2}. \tag{6.24}$$

This equation can be used to reduce the size of the ODE system by one per mode equation considered. Implementing the transformed mode equations in SCIPYS `solve_ivp` reduces the computation time by roughly 50%.

Section 6.2 introduces a Parallelized inflation solver using this theoretical insight in its numerical computation of the mode equations. The derivation of the primordial power spectrum starts at the definition of the inflaton potential and its first two derivatives as a python function. From this general form the solver finds the initial conditions for each mode and performs the forward evolution in a parallelized way using the ODE solver TORCHDIFFEQ to find a numerical representation of the potential. This allows to automatically differentiate through the forward simulation starting from the initial conditions of the parallel calculation. Taking derivatives through the determination of the initial conditions yields results contrasting with finite difference estimates due to the loop structure in their computation.

On a power series potential, which is implemented in CLASS, PARALLIZIS finds similar power spectra and is significantly more accurate than approximating it with the scale parameters A_s, n_s, α_s and β_s. This allows to probe inflation scenarios away from ΛCDM leading to primordial power spectra with non-zero running of the scale factor.

After the density fluctuations, sourced during inflation reenter the horizon, section 6.3 models them using MLPs instead of a numerical Boltzmann code. In contrast to existing methods the emulators in this work map from spectrum to spectrum instead of cosmological parameters to spectrum. As a trade-off they are fixed to one choice of cosmological parameters after the end of inflation. This section demonstrates the feasibility of mapping between spectra. Deploying the emulators for a range of different potential parametrizations requires training with a wide range of different primordial power spectra.

While PARALLIZIS has these new features it is slower than CLASS by about a factor of two. This result is somewhat unexpected since there should be a significant speedup upon switching to GPU parallelization. The slow overall computation appears to be driven by solving the differential equation numerically. The PYTORCH based differential equation solver is significantly slower than the solver implemented in SCIPY. This might be addressed by using a differential inflation solver based on JAX, such as DIFFRAX [Kidger, 2021]. However, that would require implementing a parallelized stopping criterion in this framework.

7 Summary and Outlook

The first part of this work applies a partition function approach to inference problems to both toy models and cosmological problems. The partition function Z is used as a cumulant generating function of the posterior. While the approach is successfully used to recover the cumulants for a posterior based on type Ia supernova data, the partition function itself is computed from a Markov chain probing the posterior. An alternative approach is to reformulate the source term with an imaginary unit and make use of existing fast Fourier transformation algorithms to reconstruct the partition function.

Markov chain Monte Carlo can be understood as a thermal random walk in an energy landscape defined by the partition function. Based on this intuition, section 3.3 defines a virialization condition $\langle \theta^\mu \partial\Phi/\partial\theta^\mu \rangle = \langle p_\mu \partial T/\partial p_\mu \rangle = \frac{n}{\beta}$ and equipartition conditions $\langle \theta^\mu \partial\Phi/\partial\theta^\mu \rangle = \delta^\mu_\nu \beta$, $\langle p^\mu \partial T/\partial p^\mu \rangle = \delta^\mu_\nu \beta$ for Hamilton Monte Carlo. Additionally, thermal equilibrium in a Markov chain can also be characterized through no net energy exchange with the heat bath. An equilibrated Markov chain has a constant average energy $\langle \mathcal{H} \rangle$. These convergence criteria are tested both on a toy example and supernova type Ia data.

The second part of the thesis moves away from the partition function approach and explores machine learning methods to reconstruct functions in cosmology. Section 5.2 focuses on PINNs to construct an emulator, speeding up the forward simulation in MCMC approaches to SN Ia data. While this section demonstrates the viability of this approach, section 3.3.5 uses the emulator to reduce computation time.

In section 5.3, the redshift-dependent Hubble function is reconstructed using a PINN approach to inverse problems. Introducing heteroscedastic error bars allows to also construct an uncertainty estimate on the Hubble function based on the data uncertainty. While this reconstruction is largely model agnostic it is possible to reintroduce some model assumptions on the energy composition of the universe to recover an uncertainty-aware network reconstruction of the dark energy equation of state.

Chapter 6 focuses on the inflationary paradigm and the numerical computation of the primordial power spectrum. Reexpressing the perturbation equations during inflation in terms of their phase velocity and radial part allows to recover a conserved quantity analogous to an angular momentum in the complex plane, $\rho^2 \partial_\eta \alpha = -\frac{1}{2}$. In addition, this reformulation allows to reduce the computation time in a straightforward implementation of the mode equation by about 50% compared to a formulation in terms of the real and imaginary parts.

Section 6.2 describes a parallelized computation of the mode equation using TORCHDIFFEQ. Their PYTORCH based differential equation solvers allow to automatically differentiate through the differential equation. However, for this system of equations, they come with a reduced performance compared to the differential equation solvers in SCIPY. The performance of the inflation solver might be improved by switching to a different parallelized differential equation solver. Finally, section 6.3 describes the forward evolution of the primordial power spectrum to observable quantities with fully connected networks.

Acknowledgments

I would like to thank the Institute for Theoretical Physics, STRUCTURES and the Astronomisches Rechen-Institut for funding my doctoral studies, under Germany's Excellence Strategy EXC 2181/1 - 390900948 (the Heidelberg STRUCTURES Excellence Cluster) and providing the infrastructure and computing power necessary for this thesis, especially through Tom and Jerry. Furthermore, I am grateful to the HGSFP for the positive environment during my studies and organizing the always interesting Graduate Days.

Moreover, I am grateful to Björn Malte Schäfer for supervising my PhD and teaching me statistics, cosmology and so many interesting aspects of physics. Your attempts to further my humanist education were greatly appreciated. I would like to thank Tilman Plehn for teaching me about machine learning and pushing me to present our work at conferences. Additionally, I am grateful to Arthur Hebecker for being my third supervisor, Tristan Bereau for refereeing my thesis and Belina von Krosigk for acting as an examiner on my committee.

I want to thank my proofreaders Benedikt Schosser, Rebecca Kuntz, Heinrich von Campe, Nikita Schmal, Jonas Spinner and Marion Eidinger for improving the quality of this work. Additionally, I want to thank Lorenz Vogel for the template that I modified for the worse.

I am especially grateful to all the current and former members of the cosmostatistics and pheno research groups for making my PhD an utterly positive experience: Tanmoy for introducing me into the world of science, Barry for showing me the best vegan eating spots, Claudius with his curiosity about every talk he attends, Maeve for bringing mugs full of motivation into the SFitter office, Victor with his relentless drive to self improve, Henning for life advice, Ayodele for sharing his curiosity about new methods, Eileen for victory hot chocolates, Michel for his kindness and humble brilliance, Emma for her attempts to make us all happier and nicer people, Pedro for helping me delve deeper into statistics, Theo for his near bottomless knowledge on all ML topics, Luigi a master of mystery, Benedikt for being my closest companion and confidant, Nina for conversations on life outside of physics, Nikita for letting me sample so many baked goods, Nathan for discussions on social issues, Sofia for her dedication to furthering equality, Jonas for his insights into physics problems, Maxi the never ending font of ideas, Heinrich for his careful considerations of past equations, Rebecca for deep conversations when we really should have been working, Pablo with his impressive management of the multitude of aspects in his life, Javier for morning chats, Giovanni for company at ML4Jets, Lorenz the design genius, Tobias able to retain everything anyone ever tells him, Adrian for helping me scout out the possibilities outside of academia and Jhananii for helping me find the best drink from the vending machine.

Finally, I would like to thank my family friends. Neither my PhD nor my physics studies would have been possible without you. And most of all I want to thank Linda for always lending an ear to all my woes, supporting me in all my endeavors and driving me to make smart decisions.

Bibliography

K. Abazajian et al. Snowmass 2021 cmb-s4 white paper. *arXiv preprint arXiv:2203.08024*, 2022. URL `https://arxiv.org/abs/2203.08024`.

T. M. C. Abbott et al. Dark Energy Survey Year 3 results: Cosmological constraints from galaxy clustering and weak lensing. *Phys. Rev. D*, 105(2):023520, 2022. doi:10.1103/PhysRevD.105.023520.

A. G. Adame et al. The Early Data Release of the Dark Energy Spectroscopic Instrument. *Astron. J.*, 168(2):58, 2024. doi:10.3847/1538-3881/ad3217.

N. Aghanim et al. Planck 2018 results. I. Overview and the cosmological legacy of Planck. *Astron. Astrophys.*, 641:A1, 2020a. doi:10.1051/0004-6361/201833880.

N. Aghanim et al. Planck 2018 results. VI. Cosmological parameters. *Astron. Astrophys.*, 641:A6, 2020b. doi:10.1051/0004-6361/201833910. [Erratum: Astron.Astrophys. 652, C4 (2021)].

N. Aghanim et al. Planck 2018 results. V. CMB power spectra and likelihoods. *Astron. Astrophys.*, 641:A5, 2020c. doi:10.1051/0004-6361/201936386.

Y. Akrami et al. Planck 2018 results. X. Constraints on inflation. *Astron. Astrophys.*, 641:A10, 2020. doi:10.1051/0004-6361/201833887.

E. Allys et al. Probing Cosmic Inflation with the LiteBIRD Cosmic Microwave Background Polarization Survey. *PTEP*, 2023(4):042F01, 2023. doi:10.1093/ptep/ptac150.

A. Almeida et al. The Eighteenth Data Release of the Sloan Digital Sky Surveys: Targeting and First Spectra from SDSS-V. *Astrophys. J. Suppl.*, 267(2):44, 2023. doi:10.3847/1538-4365/acda98.

R. Amanullah et al. Spectra and light curves of six type ia supernovae at $0.511 < z < 1.12$ and the union2 compilation. *ApJ*, 716(1):712–738, 2010. ISSN 0004-637X, 1538-4357. doi:10.1088/0004-637X/716/1/712. URL `http://arxiv.org/abs/1004.1711`.

A. Amara and T. D. Kitching. Figures of merit for testing standard models: application to dark energy experiments in cosmology. *MNRAS*, 413:1505–1514, 2011. doi:10.1111/j.1365-2966.2010.17947.x.

S.-i. Amari. *Information Geometry and Its Applications*, volume 194 of *Applied Mathematical Sciences*. Springer Japan, 2016. ISBN 978-4-431-55977-1 978-4-431-55978-8. doi:10.1007/978-4-431-55978-8. URL `http://link.springer.com/10.1007/978-4-431-55978-8`.

J. Y. Araz, J. C. Criado, and M. Spannowsky. Elvet–a neural network-based differential equation and variational problem solver. *arXiv preprint arXiv:2103.14575*, 2021.

R. Arutjunjan, B. M. Schäfer, and C. Kreutz. Constructing exact confidence regions on parameter manifolds of non-linear models. *to be submitted to JRSSB*, Oct. 2022.

D. J. Bacon et al. Cosmology with Phase 1 of the Square Kilometre Array: Red Book 2018: Technical specifications and performance forecasts. *Publ. Astron. Soc. Austral.*, 37:e007, 2020. doi:10.1017/pasa.2019.51.

J. C. Baez and T. Fritz. A bayesian characterization of relative entropy. *arXiv preprint arXiv:1402.3067*, 2014.

B. A. Bassett, Y. Fantaye, R. Hlozek, and J. Kotze. Fisher4cast users' manual. *arXiv preprint arXiv:0906.0974*, 2009.

B. A. Bassett, Y. Fantaye, R. Hlozek, and J. Kotze. Fisher Matrix Preloaded – Fisher4Cast. *Int. J. Mod. Phys. D*, 20:2559–2598, 2011. doi:10.1142/S0218271811020548.

D. Baumann. Tasi lectures on inflation, 2012.

D. Baumann et al. CMBPol Mission Concept Study: Probing Inflation with CMB Polarization. *AIP Conf. Proc.*, 1141(1):10–120, 2009. doi:10.1063/1.3160885.

M. Bellagente, M. Haussmann, M. Luchmann, and T. Plehn. Understanding Event-Generation Networks via Uncertainties. *SciPost Phys.*, 13(1):003, 2022. doi:10.21468/SciPostPhys.13.1.003.

C. Bennett et al. Preliminary separation of galactic and cosmic microwave emission for the cobe differential microwave radiometer. *The Astrophysical Journal*, 396:L7–L12, 08 1992. doi:10.1086/186505.

C. L. Bennett et al. Four year COBE DMR cosmic microwave background observations: Maps and basic results. *Astrophys. J. Lett.*, 464:L1–L4, 1996. doi:10.1086/310075.

S. Berkowitz and F. J. Garner. The calculation of multidimensional hermite polynomials and gram-charlier coefficients. *Mathematics of Computation*, 24(111):537–545, 1970.

M. Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017.

D. Blas, J. Lesgourgues, and T. Tram. The cosmic linear anisotropy solving system (class). part ii: Approximation schemes. *Journal of Cosmology and Astroparticle Physics*, 2011 (07):034–034, Jul 2011. ISSN 1475-7516. doi:10.1088/1475-7516/2011/07/034. URL `http://dx.doi.org/10.1088/1475-7516/2011/07/034`.

S. Bollweg, M. Haußmann, G. Kasieczka, M. Luchmann, T. Plehn, and J. Thompson. Deep-Learning Jets with Uncertainties and More. *SciPost Phys.*, 8(1):006, 2020. doi:10.21468/SciPostPhys.8.1.006.

S. Brook and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, 1997. URL `https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=602344`.

S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, editors. *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, may 2011. doi:10.1201/b10905. URL `https://doi.org/10.1201%2Fb10905`.

D. Brout et al. The pantheon+ analysis: Cosmological constraints. *The Astrophysical Journal*, 938(2):110, oct 2022. doi:10.3847/1538-4357/ac8e04. URL `https://dx.doi.org/10.3847/1538-4357/ac8e04`.

T. S. Bunch and P. C. Davies. Quantum field theory in de sitter space: renormalization by point-splitting. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 360(1700):117–134, 1978.

A. Butter, T. Heimel, S. Hummerich, T. Krebs, T. Plehn, A. Rousselot, and S. Vent. Generative networks for precision enthusiasts. *SciPost Phys.*, 14(4):078, 2023. doi:10.21468/SciPostPhys.14.4.078.

E. Cameron and A. N. Pettitt. Approximate Bayesian Computation for astronomical model analysis: a case study in galaxy demographics and morphological transformation at high redshift. *Monthly Notices of the Royal Astronomical Society*, 425(1):44–65, 09 2012. ISSN 0035-8711. doi:10.1111/j.1365-2966.2012.21371.x. URL `https://doi.org/10.1111/j.1365-2966.2012.21371.x`.

J. Carron, A. Amara, and S. Lilly. Probe combination in large galaxy surveys: application of fisher information and shannon entropy to weak lensing. *Monthly Notices of the Royal Astronomical Society*, 417(3):1938–1951, 2011.

A. T. Chantada, S. J. Landau, P. Protopapas, C. G. Scóccola, and C. Garraffo. Cosmology-informed neural networks to solve the background dynamics of the Universe. *Phys. Rev. D*, 107(6):063523, 2023. doi:10.1103/PhysRevD.107.063523.

R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 2018.

R. T. Q. Chen, B. Amos, and M. Nickel. Learning neural event functions for ordinary differential equations. *International Conference on Learning Representations*, 2021.

N. Chernikov and E. Tagirov. Quantum theory of scalar field in de sitter space-time. In *Annales de l'institut Henri Poincaré. Section A, Physique Théorique*, volume 9, pages 109–141, 1968.

M. Chevallier and D. Polarski. Accelerating universes with scaling dark matter. *International Journal of Modern Physics D*, 10:213–223, 2001. doi:10.1142/S0218271801000822.

D. Coe. Fisher matrices and confidence ellipses: a quick-start guide and software. *arXiv preprint arXiv:0906.4123*, 2009.

H. Cramér. *Mathematical methods of statistics*. Princeton university press, 1999.

S. Cuomo, V. S. Di Cola, F. Giampaolo, G. Rozza, M. Raissi, and F. Piccialli. Scientific machine learning through physics–informed neural networks: Where we are and what's next. *Journal of Scientific Computing*, 92(3):88, 2022.

F. D'Angelo and V. Fortuin. Repulsive deep ensembles are bayesian. In *Neural Information Processing Systems*, 2021. URL `https://api.semanticscholar.org/CorpusID:235593134`.

A. K. David M. Blei and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi:10.1080/01621459.2017.1285773. URL `https://doi.org/10.1080/01621459.2017.1285773`.

S. Duane, A. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid monte carlo. *Physics Letters B*, 195(2):216–222, 1987. ISSN 0370-2693. doi:https://doi.org/10.1016/0370-2693(87)91197-X. URL `https://www.sciencedirect.com/science/article/pii/037026938791197X`.

R. Durrer. *The Cosmic Microwave Background*. Cambridge University Press, 2 edition, 2020.

D. J. Eisenstein and W. Hu. Baryonic features in the matter transfer function. *The Astrophysical Journal*, 496(2):605–614, Apr 1998. ISSN 1538-4357. doi:10.1086/305424. URL http://dx.doi.org/10.1086/305424.

F. Elsner and B. D. Wandelt. Fast calculation of the fisher matrix for cosmic microwave background experiments. *Astronomy & Astrophysicas*, 540:L6, 2012. ISSN 0004-6361, 1432-0746. doi:10.1051/0004-6361/201218985. URL http://arxiv.org/abs/1202.4898.

D. J. Fixsen. The temperature of the cosmic microwave background. *The Astrophysical Journal*, 707(2):916, nov 2009. doi:10.1088/0004-637X/707/2/916. URL https://dx.doi.org/10.1088/0004-637X/707/2/916.

D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman. emcee: The MCMC hammer. *Publications of the Astronomical Society of the Pacific*, 125(925):306–312, 2013. ISSN 00046280, 15383873. doi:10.1086/670067. URL http://arxiv.org/abs/1202.3665.

K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202, 1980. URL https://api.semanticscholar.org/CorpusID:206775608.

Y. Gal. Uncertainty in deep learning. 2016. URL https://api.semanticscholar.org/CorpusID:86522127.

A. Gelman and D. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992. URL https://www.jstor.org/stable/2246093.

E. Giesel, R. Reischke, B. M. Schäfer, and D. Chia. Information geometry in cosmological inference problems. *JCAP*, 2021(1):005–005, 2021. ISSN 1475-7516. doi:10.1088/1475-7516/2021/01/005. URL http://arxiv.org/abs/2005.01057.

L. Gong and J. M. Flegal. A practical sequential stopping rule for high-dimensional mcmc and its application to spatial-temporal bayesian models. *arXiv preprint arXiv:1403.5536*, 2014.

A. Goobar and B. Leibundgut. Supernova cosmology: legacy and future. *Annual Review of Nuclear and Particle Science*, 61(1):251–279, 2011. ISSN 0163-8998, 1545-4134. doi:10.1146/annurev-nucl-102010-130434. URL http://arxiv.org/abs/1102.1431.

S. Grandis, S. Seehars, A. Refregier, A. Amara, and A. Nicola. Information gains from cosmological probes. *JCAP*, 2016(5):034–034, 2016. ISSN 1475-7516. doi:10.1088/1475-7516/2016/05/034. URL http://arxiv.org/abs/1510.06422.

A. Graves. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011.

A. Griewank and A. Walther. *Evaluating Derivatives*. Society for Industrial and Applied Mathematics, second edition, 2008. doi:10.1137/1.9780898717761. URL https://epubs.siam.org/doi/abs/10.1137/1.9780898717761.

S. Günther, J. Lesgourgues, G. Samaras, N. Schöneberg, F. Stadtmann, C. Fidler, and J. Torrado. Cosmicnet ii: emulating extended cosmologies with efficient and accurate neural networks. *Journal of Cosmology and Astroparticle Physics*, 2022, 2022. URL https://api.semanticscholar.org/CorpusID:250451184.

A. H. Guth. Inflationary universe: A possible solution to the horizon and flatness problems. *Phys. Rev. D*, 23:347–356, Jan 1981. doi:10.1103/PhysRevD.23.347. URL https://link.aps.org/doi/10.1103/PhysRevD.23.347.

Z. Hao, S. Liu, Y. Zhang, C. Ying, Y. Feng, H. Su, and J. Zhu. Physics-informed machine learning: A survey on problems, methods and applications. *arXiv preprint arXiv:2211.08064*, 2022.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2334940.

T. Heimel, R. Winterhalder, A. Butter, J. Isaacson, C. Krause, F. Maltoni, O. Mattelaer, and T. Plehn. MadNIS - Neural multi-channel importance sampling. *SciPost Phys.*, 15(4):141, 2023. doi:10.21468/SciPostPhys.15.4.141.

T. Heimel, N. Huetsch, F. Maltoni, O. Mattelaer, T. Plehn, and R. Winterhalder. The MadNIS reloaded. *SciPost Phys.*, 17(1):023, 2024. doi:10.21468/SciPostPhys.17.1.023.

M. P. Herzog, H. von Campe, R. M. Kuntz, L. Röver, and B. M. Schäfer. Partition function approach to non-Gaussian likelihoods: macrocanonical partitions and replicating Markov-chains. *The Open Journal of Astrophysics*, 7, oct 25 2024. doi:10.33232/001c.125132.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. ISSN 0893-6080. doi:https://doi.org/10.1016/0893-6080(89)90020-8. URL https://www.sciencedirect.com/science/article/pii/0893608089900208.

F. Hou, J. Goodman, D. Hogg, J. Weare, and C. Schwab. An affine-invariant sampler for exoplanet fitting and discovery in radial velocity data. *The Astrophysical Journal*, 745 (2):198, 2012. ISSN 0004-637X, 1538-4357. doi:10.1088/0004-637X/745/2/198. URL http://arxiv.org/abs/1104.2612.

W. Hu and M. White. A cmb polarization primer. *New Astronomy*, 2(4):323–344, Oct 1997. ISSN 1384-1076. doi:10.1016/s1384-1076(97)00022-5. URL http://dx.doi.org/10.1016/S1384-1076(97)00022-5.

P. Izmailov, S. Vikram, M. D. Hoffman, and A. G. G. Wilson. What are bayesian neural network posteriors really like? In *International conference on machine learning*, pages 4629–4640. PMLR, 2021.

A. Jaffe. H0 and odds on cosmology. *The Astrophysical Journal*, 471(1):24, nov 1996. doi:10.1086/177950. URL https://dx.doi.org/10.1086/177950.

J. Jasche and F. S. Kitaura. Fast hamiltonian sampling for large scale structure inference. *MNRAS*, 407(1):29–42, 2010. ISSN 00358711. doi:10.1111/j.1365-2966.2010.16897.x. URL http://arxiv.org/abs/0911.2496.

E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106 (4):620–630, 1957. ISSN 0031-899X. doi:10.1103/PhysRev.106.620. URL https://link.aps.org/doi/10.1103/PhysRev.106.620.

W. P. Johnson. The curious history of faà di bruno's formula. *The American Mathematical Monthly*, 109(3):217–234, 2002. ISSN 0002-9890. doi:10.2307/2695352. URL https://www.jstor.org/stable/2695352. Publisher: Mathematical Association of America.

G. L. Jones, M. Haran, B. S. Caffo, and R. Neath. Fixed-width output analysis for markov chain monte carlo. *Journal of the American Statistical Association*, 101(476): 1537–1547, 2006.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37:183–233, 1999.

R. Juszkiewicz, D. H. Weinberg, P. Amsterdamski, M. Chodorowski, and F. Bouchet. Weakly Nonlinear Gaussian Fluctuations and the Edgeworth Expansion. *ApJ*, 442:39, Mar. 1995. doi:10.1086/175420.

G. Kasieczka, M. Luchmann, F. Otterpohl, and T. Plehn. Per-Object Systematics using Deep-Learned Calibration. *SciPost Phys.*, 9:089, 2020. doi:10.21468/SciPostPhys.9.6.089.

P. Kidger. *On Neural Differential Equations*. PhD thesis, University of Oxford, 2021.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

F. S. Kitaura and T. A. Ensslin. Bayesian reconstruction of the cosmological large-scale structure: methodology, inverse algorithms and numerical optimization. *MNRAS*, 389: 497–544, 2008. doi:10.1111/j.1365-2966.2008.13341.x.

K. Kohri, Y. Oyama, T. Sekiguchi, and T. Takahashi. Precise Measurements of Primordial Power Spectrum with 21 cm Fluctuations. *JCAP*, 10:065, 2013. doi:10.1088/1475-7516/2013/10/065.

M. Kowalski et al. Improved cosmological constraints from new, old, and combined supernova data sets. *ApJ*, 686:749–778, 2008. ISSN 0004-637X. doi:10.1086/589937. URL `https://ui.adsabs.harvard.edu/abs/2008ApJ...686..749K`. ADS Bibcode: 2008ApJ...686..749K.

R. M. Kuntz, M. P. Herzog, H. von Campe, L. Röver, and B. M. Schäfer. Partition function approach to non-Gaussian likelihoods: partitions for the inference of functions and the Fisher-functional. *Monthly Notices of the Royal Astronomical Society*, 527 (3):8443–8458, 11 2023. ISSN 0035-8711. doi:10.1093/mnras/stad3661. URL `https://doi.org/10.1093/mnras/stad3661`.

B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Neural Information Processing Systems*, 2016. URL `https://api.semanticscholar.org/CorpusID:6294674`.

Q. V. Le, A. J. Smola, and S. Canu. Heteroscedastic gaussian process regression. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 489–496, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. doi:10.1145/1102351.1102413. URL `https://doi.org/10.1145/1102351.1102413`.

S. M. Leach, A. R. Liddle, J. Martin, and D. J. Schwarz. Cosmological parameter estimation and the inflationary cosmology. *Physical Review D*, 66(2), Jul 2002. ISSN 1089-4918. doi:10.1103/physrevd.66.023515. URL `http://dx.doi.org/10.1103/PhysRevD.66.023515`.

J. Lesgourgues and W. Valkenburg. New constraints on the observable inflaton potential from WMAP and SDSS. *Phys. Rev. D*, 75:123519, 2007. doi:10.1103/PhysRevD.75.123519.

J. Lesgourgues, A. A. Starobinsky, and W. Valkenburg. What do wmap and sdss really tell us about inflation? *Journal of Cosmology and Astroparticle Physics*, 2008 (01):010, Jan 2008. ISSN 1475-7516. doi:10.1088/1475-7516/2008/01/010. URL http://dx.doi.org/10.1088/1475-7516/2008/01/010.

A. Lewis. GetDist: a Python package for analysing Monte Carlo samples, 2019. URL https://getdist.readthedocs.io.

A. Lewis and S. Bridle. Cosmological parameters from CMB and other data: a monte-carlo approach. *PRD*, 66(10), 2002. ISSN 0556-2821, 1089-4918. doi:10.1103/PhysRevD.66.103511. URL http://arxiv.org/abs/astro-ph/0205436.

A. Lewis, A. Challinor, and A. Lasenby. Efficient computation of CMB anisotropies in closed FRW models. *Astrophys. J.*, 538:473–476, 2000. doi:10.1086/309179.

Z. Li, H. Zheng, N. B. Kovachki, D. Jin, H. Chen, B. Liu, K. Azizzadenesheli, and A. Anandkumar. Physics-informed neural operator for learning partial differential equations. *CoRR*, abs/2111.03794, 2021. URL https://arxiv.org/abs/2111.03794.

E. V. Linder. Exploring the expansion history of the universe. *Phys. Rev. Lett.*, 90(9): 091301, 2003. doi:10.1103/PhysRevLett.90.091301.

E. V. Linder. The dynamics of quintessence, the quintessence of dynamics. *General Relativity and Gravitation*, 40:329–356, 2008. doi:10.1007/s10714-007-0550-z.

J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Nature, 2004.

Q. Liu and D. Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

E. Lukacs. *Characteristic Functions*. Griffin books of cognate interest. Hafner Publishing Company, 1970. ISBN 9780852641705. URL https://books.google.de/books?id=uGEPAQAAMAAJ.

D. J. MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

E. Madelung. Eine anschauliche deutung der gleichung von schrödinger. *Naturwissenschaften*, 14:1004, 1926. URL https://api.semanticscholar.org/CorpusID:39430240.

E. Madelung. Quantentheorie in hydrodynamischer form. *Zeitschrift für Physik*, 40: 322–326, 1927. URL https://api.semanticscholar.org/CorpusID:121537534.

J. Maldacena. Non-gaussian features of primordial fluctuations in single field inflationary models. *Journal of High Energy Physics*, 2003(05):013–013, May 2003. ISSN 1029-8479. doi:10.1088/1126-6708/2003/05/013. URL http://dx.doi.org/10.1088/1126-6708/2003/05/013.

P. A. Mazzali, F. K. Ropke, S. Benetti, and W. Hillebrandt. A Common Explosion Mechanism for Type Ia Supernovae. *Science*, 315:825, 2007. doi:10.1126/SCIENCE.1136259.

Y. Mellier, J. Barroso, A. Achúcarro, J. Adamek, R. Adam, G. Addison, N. Aghanim, M. Aguena, V. Ajani, Y. Akrami, et al. Euclid. i. overview of the euclid mission. *arXiv preprint arXiv:2405.13491*, 2024.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953. ISSN 0021-9606. doi:10.1063/1.1699114. URL https://aip.scitation.org/doi/10.1063/1.1699114. Publisher: American Institute of Physics.

T. Modak, L. Röver, B. M. Schäfer, B. Schosser, and T. Plehn. Cornering extended Starobinsky inflation with CMB and SKA. *SciPost Phys.*, 15:047, 2023. doi:10.21468/SciPostPhys.15.2.047. URL https://scipost.org/10.21468/SciPostPhys.15.2.047.

M. J. Mortonson, H. V. Peiris, and R. Easther. Bayesian Analysis of Inflation: Parameter Estimation for Single Field Models. *Phys. Rev. D*, 83:043505, 2011. doi:10.1103/PhysRevD.83.043505.

M. J. Mortonson, D. H. Weinberg, and M. White. Dark energy: A short review. *arXiv preprint arXiv:1401.0046*, 2013.

V. Mukhanov, H. Feldman, and R. Brandenberger. Theory of cosmological perturbations. *Physics Reports*, 215(5):203–333, 1992. ISSN 0370-1573. doi:https://doi.org/10.1016/0370-1573(92)90044-Z. URL https://www.sciencedirect.com/science/article/pii/037015739290044Z.

R. M. Neal. Mcmc using hamiltonian dynamics. *arXiv preprint arXiv:1206.1901*, 2012.

A. Nicola, A. Amara, and A. Refregier. Consistency tests in cosmology using relative entropy. *JCAP*, 01:011, 2019. doi:10.1088/1475-7516/2019/01/011.

A. Nygaard, E. B. Holm, S. Hannestad, and T. Tram. CONNECT: a neural network based framework for emulating cosmological observables and cosmological parameter inference. *JCAP*, 05:025, 2023. doi:10.1088/1475-7516/2023/05/025.

E. Paillas et al. Optimal reconstruction of baryon acoustic oscillations for desi 2024. *arXiv preprint arXiv:2404.03005*, 2024.

A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

A. Paszke et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

R. G. Patel and O. Desjardins. Nonlinear integro-differential operator regression with neural networks. *arXiv preprint arXiv:1810.08552*, 2018.

S. Perlmutter and et al. Supernovae, dark energy, and the accelerating universe. *Physics today*, 56(4):53–62, 2003. URL http://academic.evergreen.edu/z/zita/teaching/SciSem/ss2013/articles/PT/DE_SP03.pdf.

S. Perlmutter et al. Measurements of Ω and Λ from 42 High Redshift Supernovae. *Astrophys. J.*, 517:565–586, 1999. doi:10.1086/307221.

A. M. Pinho, R. Reischke, M. Teich, and B. M. Schäfer. Information entropy in cosmological inference problems. *MNRAS*, 503(1):1187–1198, 2021. doi:10.1093/mnras/stab561.

M. L. Piscopo, M. Spannowsky, and P. Waite. Solving differential equations with neural networks: Applications to the calculation of cosmological phase transitions. *Phys. Rev. D*, 100(1):016002, 2019. doi:10.1103/PhysRevD.100.016002.

T. Plehn, A. Butter, B. Dillon, T. Heimel, C. Krause, and R. Winterhalder. Modern machine learning for lhc physicists. *arXiv preprint arXiv:2211.01421*, 2022.

B. A. Powell and W. H. Kinney. Limits on primordial power spectrum resolution: An inflationary flow analysis. *JCAP*, 0708:006, 2007. doi:10.1088/1475-7516/2007/08/006.

H. Prince and J. Dunkley. Data compression in cosmology: A compressed likelihood for Planck data. *Phys. Rev. D*, 100(8):083502, 2019. doi:10.1103/PhysRevD.100.083502.

M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics informed deep learning (part I): data-driven solutions of nonlinear partial differential equations. *CoRR*, abs/1711.10561, 2017. URL `http://arxiv.org/abs/1711.10561`.

B. Ratra and P. J. E. Peebles. Cosmological consequences of a rolling homogeneous scalar field. *Phys. Rev. D*, 37:3406–3427, Jun 1988. doi:10.1103/PhysRevD.37.3406. URL `https://link.aps.org/doi/10.1103/PhysRevD.37.3406`.

M. Raveri, M. Martinelli, G. Zhao, and Y. Wang. Cosmicfish implementation notes v1. 0. *arXiv preprint arXiv:1606.06268*, 2016.

A. Refregier, A. Amara, T. D. Kitching, and A. Rassat. iCosmo: an interactive cosmology package. *AAP*, 528:A33+, 2011. doi:10.1051/0004-6361/200811112.

R. Reischke, A. Kiessling, and B. M. Schäfer. Variations of cosmic large-scale structure covariance matrices across parameter space. *MNRAS*, 465:4016–4025, 2017. URL `http://arxiv.org/abs/1607.03136`.

A. G. Riess et al. Observational evidence from supernovae for an accelerating universe and a cosmological constant. *The Astronomical Journal*, 116(3):1009–1038, sep 1998.

A. G. Riess et al. A Comprehensive Measurement of the Local Value of the Hubble Constant with 1 km s^{-1} Mpc^{-1} Uncertainty from the Hubble Space Telescope and the SH0ES Team. *Astrophys. J. Lett.*, 934(1):L7, 2022. doi:10.3847/2041-8213/ac5c5b.

A. Riotto. Inflation and the theory of cosmological perturbations. *ICTP Lect. Notes Ser.*, 14:317–413, 2003.

G. O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351 – 367, 2001. doi:10.1214/ss/1015346320. URL `https://doi.org/10.1214/ss/1015346320`.

G. O. Roberts et al. Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120, 1997. URL `https://www.jstor.org/stable/2245134`.

L. Röver, B. M. Schäfer, and T. Plehn. Pinnferring the hubble function with uncertainties. *arXiv preprint arXiv:2403.13899*, 2024.

D. Rubin et al. Union through unity: Cosmology with 2,000 sne using a unified bayesian framework. *arXiv preprint arXiv:2311.12098*, 2023.

L. Röver, L. C. Bartels, and B. M. Schäfer. Partition function approach to non-Gaussian likelihoods: formalism and expansions for weakly non-Gaussian cosmological inference. *Monthly Notices of the Royal Astronomical Society*, 523(2):2027–2038, 05 2023a. ISSN 0035-8711. doi:10.1093/mnras/stad1471. URL `https://doi.org/10.1093/mnras/stad1471`.

L. Röver, H. von Campe, M. P. Herzog, R. M. Kuntz, and B. M. Schäfer. Partition function approach to non-Gaussian likelihoods: physically motivated convergence criteria for Markov chains. *Monthly Notices of the Royal Astronomical Society*, 526 (1):473–482, 09 2023b. ISSN 0035-8711. doi:10.1093/mnras/stad2726. URL https://doi.org/10.1093/mnras/stad2726.

L. Röver, B. M. Schäfer, and T. Plehn. Parallizis: Parallelized diffrerentiable inflation solver, 2024.

K. Sato. First-order phase transition of a vacuum and the expansion of the Universe. *Monthly Notices of the Royal Astronomical Society*, 195(3):467–479, 07 1981. ISSN 0035-8711. doi:10.1093/mnras/195.3.467. URL https://doi.org/10.1093/mnras/195.3.467.

B. M. Schäfer and R. Reischke. Describing variations of the fisher-matrix across parameter space. *MNRAS*, 460(3):3398–3406, 2016. ISSN 0035-8711, 1365-2966. doi:10.1093/mnras/stw1221. URL http://arxiv.org/abs/1603.03626.

C. M. Schafer and P. E. Freeman. Likelihood-free inference in cosmology: Potential for the estimation of luminosity functions. In *Statistical Challenges in Modern Astronomy V*, pages 3–19. Springer, 2012.

B. Schosser, T. Röspel, and B. M. Schaefer. Markov walk exploration of model spaces: Bayesian selection of dark energy models with supernovae. *arXiv preprint arXiv:2407.06259*, 2024.

D. Scolnic et al. The pantheon+ analysis: The full data set and light-curve release. *The Astrophysical Journal*, 938(2):113, Oct. 2022. ISSN 1538-4357. doi:10.3847/1538-4357/ac8b7a. URL http://dx.doi.org/10.3847/1538-4357/ac8b7a.

M. Seitzer, A. Tavakoli, D. Antic, and G. Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=aPOpXlnV1T.

E. Sellentin. A fast, always positive definite and normalizable approximation of non-Gaussian likelihoods. *Mon. Not. Roy. Astron. Soc.*, 453(1):893–898, 2015a. doi:10.1093/mnras/stv1671.

E. Sellentin. A fast, always positive definite and normalizable approximation of non-gaussian likelihoods. *MNRAS*, 453(1):893–898, 2015b. ISSN 0035-8711, 1365-2966. doi:10.1093/mnras/stv1671. URL http://arxiv.org/abs/1506.04866.

E. Sellentin, M. Quartin, and L. Amendola. Breaking the spell of gaussianity: forecasting with higher order fisher matrices. *MNRAS*, 441(2):1831–1840, 2014. ISSN 0035-8711. doi:10.1093/mnras/stu689. URL https://academic.oup.com/mnras/article/441/2/1831/1077545.

K. Shukla, P. C. di Leoni, J. L. Blackshire, D. M. Sparkman, and G. E. Karniadakis. Physics-informed neural network for ultrasound nondestructive quantification of surface breaking cracks. *Journal of Nondestructive Evaluation*, 39, 2020. URL https://api.semanticscholar.org/CorpusID:218537894.

D. Sivia and J. Skilling. *Data analysis: a Bayesian tutorial*. OUP Oxford, 2006.

J. Skilling. Nested sampling for general bayesian computation. 2006.

G. Smoot et al. Structure in the cobe differential microwave radiometer first-year maps. *The Astrophysical Journal*, 396:L1–L5, 09 1992. doi:10.1086/186504.

A. Spurio Mancini, D. Piras, J. Alsing, B. Joachimi, and M. P. Hobson. CosmoPower: emulating cosmological power spectra for accelerated Bayesian inference from next-generation surveys. *Mon. Not. Roy. Astron. Soc.*, 511(2):1771–1788, 2022. doi:10.1093/mnras/stac064.

A. Starobinsky. A new type of isotropic cosmological models without singularity. *Physics Letters B*, 91(1):99–102, 1980. ISSN 0370-2693. doi:https://doi.org/10.1016/0370-2693(80)90670-X. URL https://www.sciencedirect.com/science/article/pii/037026938090670X.

N. Suzuki et al. The hubble space telescope cluster supernova survey. v. improving the dark-energy constraints above z> 1 and building an early-type-hosted supernova sample. *The Astrophysical Journal*, 746(1):85, 2012.

M. Takada and B. Jain. Cosmological parameters from lensing power spectrum and bispectrum tomography. *Mon. Not. Roy. Astron. Soc.*, 348:897, 2004. doi:10.1111/j.1365-2966.2004.07410.x.

M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.

M. Tegmark, A. Taylor, and A. Heavens. Karhunen-loeve eigenvalue problems in cosmology: how should we tackle large data sets? *The Astrophysical Journal*, 480(1):22–35, 1997. ISSN 0004-637X, 1538-4357. doi:10.1086/303939. URL http://arxiv.org/abs/astro-ph/9603021.

L. Tierney. Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics*, 22(4):1701 – 1728, 1994. doi:10.1214/aos/1176325750. URL https://doi.org/10.1214/aos/1176325750.

R. Trotta. Applications of Bayesian model selection to cosmological parameters. *Monthly Notices of the Royal Astronomical Society*, 378(1):72–82, 05 2007. ISSN 0035-8711. doi:10.1111/j.1365-2966.2007.11738.x. URL https://doi.org/10.1111/j.1365-2966.2007.11738.x.

R. Trotta. Bayes in the sky: Bayesian inference and model selection in cosmology. *Contemporary Physics*, 49(2):71–104, 2008. ISSN 0010-7514, 1366-5812. doi:10.1080/00107510802066753. URL http://arxiv.org/abs/0803.4089.

S. Tsujikawa. Quintessence: A review. *Classical and Quantum Gravity*, 30(21):214003, 2013. ISSN 0264-9381, 1361-6382. doi:10.1088/0264-9381/30/21/214003. URL http://arxiv.org/abs/1304.1961.

T. Van Erven and P. Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

D. Vats and C. Knudson. Revisiting the gelman–rubin diagnostic. *Statistical Science*, 36 (4):518–529, 2021.

D. Vats, J. M. Flegal, and G. L. Jones. Multivariate output analysis for markov chain monte carlo. *Biometrika*, 106(2):321–337, 2019.

S. Wang, S. Sankaran, H. Wang, and P. Perdikaris. An expert's guide to training physics-informed neural networks. *arXiv preprint arXiv:2308.08468*, 2023.

C. Wetterich. Cosmology and the fate of dilatation symmetry. *Nuclear Physics B*, 302 (4):668–696, 1988. ISSN 0550-3213. doi:https://doi.org/10.1016/0550-3213(88)90193-9. URL https://www.sciencedirect.com/science/article/pii/0550321388901939.

A. Weyant, C. Schafer, and W. M. Wood-Vasey. Likelihood-free cosmological inference with type ia supernovae: Approximate bayesian computation for a complete treatment of uncertainty. *The Astrophysical Journal*, 764(2):116, jan 2013. doi:10.1088/0004-637X/764/2/116. URL https://dx.doi.org/10.1088/0004-637X/764/2/116.

L. Wolz, M. Kilbinger, J. Weller, and T. Giannantonio. On the validity of cosmological fisher matrix forecasts. *JCAP*, 9:9, 2012. doi:10.1088/1475-7516/2012/09/009.

L. Yang, X. Meng, and G. E. Karniadakis. B-pinns: Bayesian physics-informed neural networks for forward and inverse pde problems with noisy data. *Journal of Computational Physics*, 425:109913, Jan. 2021. ISSN 0021-9991. doi:10.1016/j.jcp.2020.109913. URL http://dx.doi.org/10.1016/j.jcp.2020.109913.