Artificial intelligence for harvesting the untapped cosmological information in the Lyman-α forest

Parth Nayak

München 2025

Artificial intelligence for harvesting the untapped cosmological information in the Lyman-α forest

Parth Nayak

Dissertation
an der Fakultät für Physik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Parth Nayak
aus Surat, Indien

München, den 06.11.2025

Erstgutachter: Prof. Dr. Daniel Grün
Zweitgutachter: Prof. Dr. Jochen Weller
Tag der mündlichen Prüfung: 19.12.2025

The work was made possible with the support of a scholarship from the German Academic Exchange Service (DAAD).

Contents

Zusammenfassung

Spektroskopische Himmelsdurchmusterungen wie SDSS/eBOSS und DESI haben die Astronomie durch die Bereitstellung einer gigantischen Datenmenge revolutioniert. Es besteht ein dringender Bedarf an effizienten Analysealgorithmen, um dieser Datenflut gerecht zu werden. Als Antwort darauf werden Techniken der künstlichen Intelligenz (KI) immer beliebter. Die Aufgabe der Parameterinferenz in der Kosmologie ist ein Anwendungsbereich, in dem KI zunehmend verbreitet ist. Dies liegt daran, dass die traditionellen Methoden auf analytischen zusammenfassenden Statistiken beruhen, die zwar leicht zu interpretieren sind, aber nicht effizient genug sind, um die Informationen in den Daten zu kodieren. Prominente Beispiele hierfür sind Zwei-Punkt-Korrelationen oder Leistungsspektren, die alle Gaußschen Informationen in den Daten enthalten, aber deren nicht-Gaußsche Aspekte nicht erfassen können. KI, insbesondere Deep Learning (DL), bietet leistungsstarke Lösungen für dieses wachsende Problem.

In dieser Arbeit habe ich das erste solcher DL-Frameworks für den speziellen Fall des Lyα-Waldes entwickelt, welcher die Absorption der Emissionskontinua entfernter Quasare durch das intergalaktische Medium (IGM) im expandierenden Universum darstellt. Als solches ist der Lyα-Wald ein einzigartiger und empfindlicher Indikator für einen weiten Bereich der kosmischen Physik, einschließlich des thermischen Zustands des IGM, der hier im Mittelpunkt steht. Das Lyα-Transmissionsfeld hat von Natur aus eine nicht-gaußsche Verteilung, was es zu einem hervorragenden Kandidaten für die DL-Parameterinferenz macht.

Die Aussagekraft von DL hängt stark von der Verfügbarkeit großer Trainingsdatensätze ab. Ich habe riesige kosmologische Hydrodynamik-Simulationen verwendet, um gelabelte Mock-Datensätze für das Training neuronaler Netze (NN) mittels Supervised Learning zu generieren. Daher habe ich zunächst die Erzeugungsverfahren dieser Mock-Datensätze auf ihre Genauigkeit hin überprüft und dabei verschiedene Näherungen, die üblicherweise zur Reduzierung des Rechenaufwands verwendet werden, genau untersucht. Anschließend habe ich eine ResNet-CNN-Architektur unter Verwendung dieser unkontaminierten Mock-Datensätze trainiert, um eine optimale Extraktion von Merkmalen aus dem Transmissionsfeld zu erreichen. In diesem theoretischen Fall habe ich eine signifikante Verbesserung der Präzision im Posterior gegenüber herkömmlichen Zusammenfassungen um einen Faktor von einigen wenigen beobachtet. Nach diesem Erfolg habe ich Anpassungen des Inferenzframeworks zur Nutzung realistischerer Daten vorgenommen. Konkret habe ich eine aktualisierte Architektur mit verrauschten Spektren niedrigerer Auflösung trainiert

und den Einfluss des Rauschpegels auf die Parameterpräzision charakterisiert. Obwohl die Verbesserung der Genauigkeit nicht ganz so stark ist wie im theoretischen Szenario, zeigt sie dennoch das Potenzial von DL, mit realistischen Datensätzen über die herkömmliche Zweipunktstatistik hinaus zu arbeiten.

Diese Dissertation demonstriert somit, dass DL ein sehr effizientes und hochrelevantes Werkzeug für die Informationsextraktion und Parameterinferenz mit dem Lyα-Wald unter Verwendung großer Datensätze sein kann. Gleichzeitig liefert sie wertvolle Erkenntnisse über die Herausforderungen bei der Entwicklung einer voll funktionsfähigen und vertrauenswürdigen DL-Inferenzpipeline für spektroskopische Untersuchungen der nächsten Generation und mögliche Wege zu deren Bewältigung.

Abstract

Spectroscopic sky surveys such as SDSS/eBOSS and DESI have revolutionized astronomy by delivering a gargantuan amount of data. An urgent need for finding efficient analysis algorithms to complement this data inflation has arisen. Artificial intelligence (AI) techniques are becoming increasingly popular in response. The task of parameter inference in cosmology is an avenue with increasingly widespread applications of AI. This is because the traditional methods rely on analytically well-defined summary statistics that are easily interpreted but are not efficient enough to encode the information in the data. Prominent examples of those are two-point correlations or power spectra that include all Gaussian information in the data, but fail to capture their non-Gaussian aspects. AI, specifically deep learning (DL), offers powerful remedies to this growing problem.

In this thesis, I have developed the first of such DL frameworks for the specific case of the Lyα forest, which is a consequence of the interaction of distant quasars' emission continua with the intergalactic medium (IGM) in the expanding Universe. As such the Lyα forest is a unique and sensitive probe of a wide range of cosmic physics, including the IGM thermal state, which is the focus here. The Lyα transmission field has an inherently non-Gaussian nature, making it a great candidate for DL parameter inference.

The power of DL heavily relies on the availability of large training datasets. I employed giant cosmological hydrodynamic simulations to source labeled mock datasets for training neural networks (NN) via supervised learning. Therefore, I first scrutinized the generation of those mocks for their fidelity, closely examining different approximations commonly made to reduce the computational cost. I then trained a ResNet convolutional architecture using those pure mocks for an optimal extraction of features from the transmission field. In this theoretical, uncontaminated case, I observed a significant improvement of posterior precision over traditional summaries by factors of a few. Following this success, I made adaptations to the framework for inference with more realistic data. Namely, I trained an updated architecture with noisy, lower resolution spectra and characterized the impact of the noise level on the posterior precision. Although the tightening of the constraints is not quite as strong as for the theoretical scenario, it is still indicative of the potential of DL to work beyond conventional two-point statistics with realistic datasets.

This thesis hence motivates DL as a very efficient and highly relevant tool for information extraction and parameter inference with the Lyα forest using large datasets. Simultaneously, it also provides valuable insights into the challenges for the development of a fully functional and trustworthy DL inference pipeline for the next generation of spectroscopic

surveys and potential ways to tackle them.

Author's note

This dissertation contains original research led by me and in collaboration with my coauthors as and when appropriately listed. Any typographical, grammatical or material errors are my own.

Use of large language models (LLM)

While the scientific content of this thesis is entirely human generated, I have on several occasions relied on OpenAI's public LLM ChatGPT (in particular GPT-4o and 5) as well as Google's Gemini (2.5 Flash) for assistance with the following tasks: for making pretty, publication-ready plots in Matplotlib; while fixing minor code bugs for breaking down complicated snippets into simpler pieces; for finding the correct syntax of Python and LaTeX tools; for drawing vocabularic inspiration; for finding more formal alternative phrasings. I have also used GitHub Copilot integrated with VS Code for suggesting and fixing code snippets. I have written the abstract originally in English and translated into German using DeepL Translator.

In all these cases, the outputs of the LLMs have been thoroughly vetted by me before use in this thesis.

Mathematical notations

In the mathematical text of this thesis, the following global notations have been adopted, unless otherwise specified. Boldface letters such as $\boldsymbol{\pi}$ denote vector (or matrix) quantities. Einstein summation convention is implied in the case of Greek μ, ν indices. I follow the "East Coast" convention for the Minkowski metric, namely, $\eta_{\mu\nu} = \mathrm{diag}(-1, 1, 1, 1)$. The natural logarithm is denoted by simple log, whereas the logarithm with any other basis is denoted specifically, such as $\log_{10}$. The determinant of a matrix is denoted by $\det[\cdot]$ as well as $|\ldots|$, depending on the chapters. $\langle \ldots \rangle$ denote ensemble averages. Derivatives *w.r.t.* x ($\mathbf{x}$) have been explicitly denoted by $d[\ldots]/dx$ in case of ordinary derivatives and $\partial[\ldots]/\partial x$ in case of partial derivatives ($\boldsymbol{\nabla}_{\mathrm{x}}[\ldots]$ in case of gradients). Derivatives *w.r.t.* t (as in time) are denoted by $\dot{[...]}$ (first derivative) and $\ddot{[\ldots]}$ (second derivative). Standard spectroscopists' notations are adopted for ionic species of an element, e.g., HeI denotes neutral Helium, HeII singly ionized Helium, and so on.

Outline of this thesis

This document is organized as follows. I provide the necessary background in cosmology, statistical inference, and deep learning in Chapters 1 and 2. Not only do these chapters serve as primers on the topics discussed, they also lay a more gradual path leading to the core agenda of this dissertation (Chapters 3, 4, 5).

In Chapter 3 I discuss the process of mock Lyα forest data creation from hydrodynamic simulations for training our deep learning inference framework. I then describe this framework by reproducing two papers I led (Nayak et al. 2024, 2025) in Chapters 4 and 5 as a proof of concept and for more realistic, noisy data, respectively. Only minor stylistic changes have been applied to the original manuscripts while preserving the main subject matter. For this reason, some redundancies in the content may appear. The appendices of both the original papers have been moved to the end of the thesis for a logical flow of the contents.

For the reader in a hurry, I provide a concise summary of this text in Chapter 6. The Outlook at the end of that chapter outlines prospective avenues that may yet attract further exploration.

Chapter 1

Cosmology

"Our whole universe was in a hot, dense state
Then nearly 14 billion years ago, expansion started [...]"
— Ed Robertson & Steven Page, "The History of Everything" (2007)

This line from the Canadian musicians has widely popularized the Big Bang paradigm of modern cosmology. It aptly conveys our current understanding of the origins of our Universe that has been cultivated over decades of careful scientific investigation. Since the development of Einstein's general theory of relativity (Einstein 1915, 1917), the discipline of physical cosmology saw a renewed interest both from a theoretical and an observational perspective. The concept of an expanding universe (e.g., Lemaître 1931)—and therefore a smaller, denser universe in the past—started to be accepted by the scientific community as evidence emerged that distant galaxies were receding away from us (Hubble 1929). With the milestone $\alpha\beta\gamma$ paper (Alpher et al. 1948), the theory of Big Bang nucleosynthesis was born, positing that light chemical elements were formed in the hot, primordial Universe in the aftermath of the Big Bang. The discovery of the cosmic microwave background (CMB) by Penzias and Wilson (1965) drove the final nail in the coffin of previously proposed alternative cosmological theories (e.g., Bondi and Gold 1948).

The inflationary paradigm (e.g., Guth 1981) proposed that in the first 10^{-32} s of the Big Bang, the Universe underwent an infinitesimally brief period of exponential expansion by at least 60 e-folds. This came as a potential solution to two outstanding problems with the Big Bang picture: (i) the horizon problem, the observation that regions of the CMB at large angular separations have nearly identical properties despite their never having been in causal contact, and (ii) the flatness problem, the apparent serendipity of the density of the Universe being extremely fine-tuned to produce a spatially flat Universe today, despite any minute early deviations being expected to grow with time. Inflation also provided a means for the primordial quantum fluctuations in the density distribution to grow rapidly and seed the large scale structure we see today (Mukhanov and Chibisov 1982).

Pioneering measurements of the rotation curves of spiral galaxies led by Vera Rubin posed an unavoidable need for dark matter, a component not showing any electromagnetic footprint yet comprising the majority of the Universe's mass (Rubin 1983). The High-

Z Supernova Search Team and the Supernova Cosmology Project found groundbreaking evidence that the expansion of the Universe is accelerating (e.g., Riess et al. 1998; Goobar et al. 2000), reinvoking Einstein's cosmological constant Λ as a key ingredient of the Universe. Thus began the modern era of the standard ΛCDM cosmological model.

Meanwhile, observations of intergalactic absorption by Hydrogen and Helium in quasar spectra (e.g., Gunn and Peterson 1965; Bahcall and Salpeter 1965) had provided evidence that the intergalactic medium (IGM), the biggest reservoir of baryonic matter in the Universe, underwent a major phase transition from completely neutral to completely ionized—a period called the epoch of reionization that occurred in the wake of the formation of first stars and galaxies. The ionized gas also attenuated the CMB via Thomson scattering, leaving a measurable imprint on the polarization angular power spectrum (see, e.g., Planck Collaboration et al. 2020). This phase transition is thought to have deposited large amounts of heat into the IGM, governing the thermal evolution of the gas in its aftermath. The Lyα forest, a unique imprint of the IGM on quasar spectra, has emerged as a powerful probe of the thermal state of the gas leading up to the end of reionization.

In the following, I briefly discuss the fundamentals of the modern cosmological theory, including the ΛCDM model. I also delve a little into the epoch of reionization and the physics of the Lyα forest. The material presented here is of a typical graduate level cosmology textbook, and I have particularly relied on Weinberg (2008) and Draine (2011) for reference.

1.1 Geometry of the universe

In Einstein's general relativity (GR), gravity is manifest as the geometry of the spacetime. In the 3+1 dimensional Minkowski framework, an infinitesimal squared distance element (henceforth "line element") is

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu, \tag{1.1}$$

where $x^0 := ct$ and $x^i := \mathbf{x} \cdot \hat{\boldsymbol{\eta}}_i$ for $i \in \{1, 2, 3\}$ where $\hat{\boldsymbol{\eta}}_i$ are the unit spatial basis vectors. The tensor $g_{\mu\nu} = g_{\mu\nu}(t, \mathbf{x})$ is called the metric of the spacetime.

Cosmological Principle: Astronomical observations suggest that on very large scales ($d \gtrsim 300$ Mpc) the Universe is homogeneous and isotropic. In cosmology, therefore, the metric of the background spacetime (on which all the contents of the Universe sit) is only a function of the cosmic time t and not spatial coordinates $\mathbf{x}$. Hence, $g_{\mu\nu} = g_{\mu\nu}(t)$. This is known as the Cosmological Principle.

The only degree of freedom for such geometries is the time-dependent scaling of the physical coordinates, i.e. $\mathbf{x} \to a(t)\mathbf{x}$. A generalized line element of our spacetime can thus be expressed, in spherical polar coordinates, as

$$ds^2 = -c^2 dt^2 + a^2(t) \left[\frac{dr^2}{1 - Kr^2} + r^2 (d\theta^2 + \sin^2 \theta d\phi^2) \right], \tag{1.2}$$

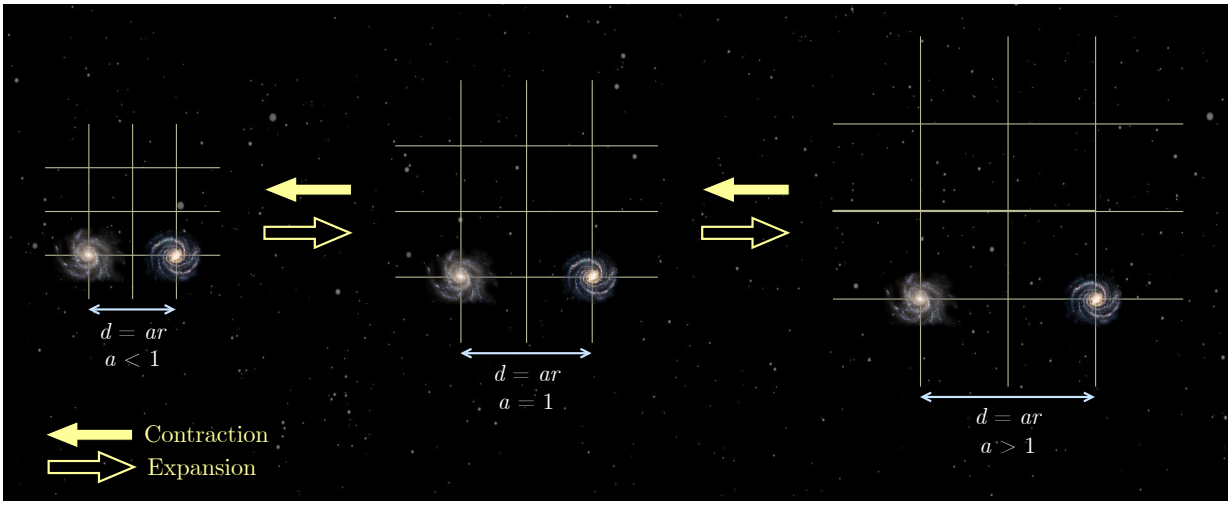

Figure 1.1: A visualization of the Robertson-Walker metric for a simplified case of flat 2D space. Notice here that the galaxies stay at the fixed points on the grid while the grid itself expands or contracts such that the physical distance between those galaxies changes. The comoving distance between those galaxies is $r \sim 2$ grid points.

where

$$K = \begin{cases} 0 & \text{flat space} \\ +1 & \text{spherical space} \\ -1 & \text{hyperspherical space.} \end{cases} \tag{1.3}$$

The flat space is the usual Euclidean 3D space, the spherical space is a spherical 3-surface in the 4D Euclidean space, and the hyperspherical space is a hyperbolic 3-surface in the 4D Euclidean space (Weinberg 2008). The metric pertaining to the above geometry is commonly known as the Robertson-Walker (RW) metric. In this formalism, r is the "comoving" coordinate which is constant at all cosmic times and $a(t)r$ the physical coordinate at time t (see Figure 1.1). $a(t)$ is called the scale factor and conveniently takes the value $a(t = t_0) = 1$ at the present cosmic time t_0. In the rest of this work, we shall assume a flat RW spacetime.

1.2 Cosmological redshift and Hubble's law

In 1929 Edwin Hubble (Hubble 1929) observed that the spectra of distant galaxies are redshifted and that a recessional velocity $v(\ll c)$ associated with a galaxy's redshift is linearly proportional to its distance from us. A redshift parameter can be defined as

$$z := \frac{\lambda_{\text{obs}} - \lambda_{\text{em}}}{\lambda_{\text{em}}} \approx \frac{v}{c}, \tag{1.4}$$

where the latter equality is valid only for small v/c.

For a light wave traveling in a flat universe, $ds^2 = 0$ and thus $cdt = \pm a(t)dr$. For the scenario of dt representing the frequency of the said light wave $\nu \equiv 1/dt$, its wavelength $\lambda = a(t)dr$ (adopting the positive sign because of obvious physical reasons). If, further, this light ray is emitted by a distant source at $t = t_{\rm em}$ and observed at $t = t_{\rm obs} = t_0$ (present),

$$1 + z = \lambda_{\rm obs}/\lambda_{\rm em} = a(t_0)/a(t) = 1/a(t). \tag{1.5}$$

This is known as the cosmological redshift, and it is physically different from the Doppler shift, which concerns the relative motion of objects *through* space. In accordance with the one-to-one mapping in Eq. (1.5), the cosmological redshift z is often used in lieu of t as a lookback parameter. It explains the observation by Hubble to an arbitrarily large z as a consequence of the expansion of the Universe. In this picture, the recessional velocity, called the "Hubble flow" is merely a proxy for the rate of expansion; it carries no other physical meaning. In the RW formalism, the proper distance to a galaxy is $d(t) = a(t)r$ and the Hubble flow velocity is $v(t) = \dot{d}(t) = \dot{a}(t)r$. Thus, we have

$$v = \left(\frac{\dot{a}}{a}\right) d =: H(t)d, \tag{1.6}$$

where $H(t)$ is called the Hubble parameter and determines the rate of expansion. The present value of the Hubble parameter is conveniently expressed in the literature as $H_0 = 100h$ km s^{-1}Mpc^{-1}. The determination of the value of h through observations is still a topic of active research, especially since its current best estimate from late time probes (such as Supernova Ia) and early time probes (such as CMB) seem to be in tension with each other (see Schöneberg et al. 2022, for an interesting review).

1.3 Friedmann equations

Einstein's field equation of GR can be written as

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = \frac{8\pi G}{c^4} T_{\mu\nu}, \tag{1.7}$$

where $R_{\mu\nu}$ is the Ricci tensor, R is the Ricci scalar, $T_{\mu\nu}$ is the energy-momentum tensor and G is the gravitational constant. The Robertson-Walker metric in Eq. (1.2) can be shown to be a solution to this equation. Plugging the Robertson-Walker $g_{\mu\nu}$ and $T_{\mu\nu} = -\mathrm{diag}(-\rho c^2, P, P, P)$ of an ideal fluid into Eq. (1.7) yields the two Friedmann equations

$$\left(\frac{\dot{a}}{a}\right)^2 + \frac{Kc^2}{a^2} = \frac{8\pi G}{3}\rho, \tag{1.8}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3P}{c^2}\right), \tag{1.9}$$

where ρ is the energy density of the ideal fluid and P is its relativistic pressure. The equation of state for such a fluid is $P = w\rho c^2$. From these, a further continuity equation follows:

$$\dot{\rho} + 3H(t)(1+w)\rho = 0, \tag{1.10}$$

which can be solved to obtain a general expression of density evolution, $\rho \propto a^{-3(1+w)}$.

1.4 The ΛCDM cosmology

In the standard ΛCDM model of cosmology, our Universe has a flat (Einstein-de Sitter) geometry ($K = 0$) and is made up of the following three main components:

(i) Cold (non-relativistic) matter with $w = 0$, hence $\rho_{\rm m} \propto a^{-3} = (1+z)^3$. This component is further classified into ordinary (baryonic) matter ($\rho_{\rm b}$) and dark matter[1] ($\rho_{\rm CDM}$).

(ii) Relativistic matter (radiation) with $w = 1/3$, hence $\rho_\gamma \propto a^{-4} = (1+z)^4$.

(iii) Vacuum energy or dark energy with $w = -1$, hence a time-independent energy density $\rho_\Lambda \equiv \frac{\Lambda}{8\pi G}$. The notation Λ is attributed to the equivalence of this component with Einstein's cosmological constant (Einstein 1917).

Together, the total density of the Universe is $\rho = \rho_{\rm m} + \rho_\gamma + \rho_\Lambda$. We can now rearrange the first Friedmann equation Eq. (1.8) in terms of a critical density $\rho_{\rm c} = \frac{3H_0^2}{8\pi G}$ to read

$$\frac{H^2(t)}{H_0^2} = \frac{\rho(t)}{\rho_{\rm c}}. \tag{1.11}$$

Further, defining the density parameters as $\Omega_i := \rho_i/\rho_{\rm c}$ for the component $i \in \{{\rm m}, \gamma, \Lambda\}$, we can write a parametric Friedmann equation as

$$H^2 = H_0^2(\Omega_{\rm m} + \Omega_\gamma + \Omega_\Lambda), \tag{1.12}$$

with a special case of the present time, $\sum_i \Omega_{i,0} = 1$. This is the central equation of the standard ΛCDM model. The current concordance model has the following values of the density parameters: $\Omega_{\rm b,0} = 0.04$, $\Omega_{\rm CDM,0} = 0.23$, $\Omega_{\Lambda,0} = 0.73$ and $\Omega_{\gamma,0} \sim \mathcal{O}(10^{-5})$, with $\Omega_{\rm m} = \Omega_{\rm b} + \Omega_{\rm CDM}$.

1.5 Epoch of reionization

Approximately 380,000 years after the Big Bang ($z \sim 1100$), the Universe had expanded and cooled enough so that protons and electrons in the primordial plasma could combine to form neutral atoms, and the hitherto trapped radiation (due to Thomson scattering off

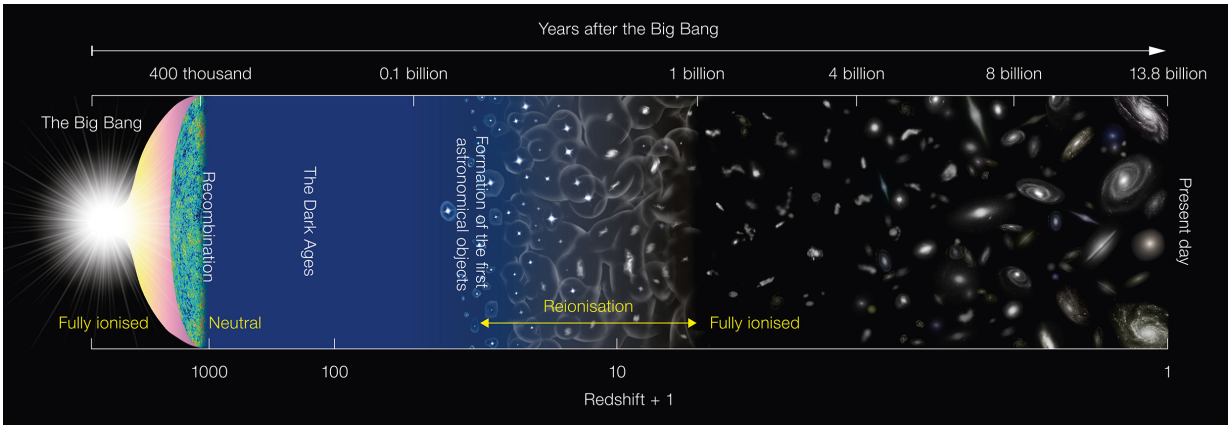

Figure 1.2: The history of the Universe in the Big Bang model, depicting major epochs such as recombination, dark ages, and reionization. The epoch of reionization progresses in an inside-out fashion with individual galaxies forming ionized bubbles that grow over cosmic time and merge with each other to eventually result in the entirety of the Universe being ionized. Credits: NAOJ.

charged particles) could escape to travel freely into space. That radiation is still seen from Earth as the 2.73 K blackbody CMB radiation.

After this event known as the epoch of recombination, the Universe underwent a prolonged phase of expansion and structure formation. During this period, there were no new sources of radiation other than the CMB—the Universe was essentially dark. During these Dark Ages, the Universe was matter-dominated and gravity drove collapse of dark matter into halos. The baryonic matter (gas) then fell onto the potential wells created by dark matter overdensities and formed the first stars and galaxies, by $z \sim 15$, launching the "cosmic dawn." As the number density of those galaxies grew, their ultraviolet (UV) radiation started ionizing their surrounding intergalactic gas that was predominantly made of neutral Hydrogen and Helium. The ionization potential of Hydrogen is $E_{\rm H}^{\rm ion} = 13.6$ eV corresponding to a transition of electron from the ground state to an unbound state. For Helium, the ionization potentials are $E_{\rm He,1}^{\rm ion} = 24.6$ eV and $E_{\rm He,2}^{\rm ion} = 54.4$ eV corresponding to the first and second electron, respectively. The ionizing UV background radiation spectrum of the early galaxies allowed the ionization of H (HI→ HII) and the first ionization of He (HeI→ HeII). Individual stellar and galactic clusters sent ionization fronts and formed bubbles of ionized gas surrounding themselves, which grew and started merging with each other. Eventually, by $z \sim 6$ almost all of the intergalactic medium (IGM) was ionized with the mass fraction of neutral H as small as $x_{\rm HI} \sim 10^{-4}$. Since the previously recombined cosmic gas was now ionized again, this period in the history of the Universe is called the "epoch of reionization" and marks a very important phase transition. During reionization, the background UV radiation also photo-heated the IGM up to $T \sim 10^4$ K. Figure 1.2 shows an illustration of the history of the Universe, marking the above phase transitions and epochs.

[1]This distinction is based on the interaction of the material with photons: baryons interact with electromagnetic radiation such that we can "see" them whereas dark matter does not.

The stellar and galactic UV background was unfortunately not hard enough to doubly ionize Helium (HeII→ HeIII). This is thought to have happened much later, between $z \sim 4$ and $z \sim 3$ thanks to the hard UV radiation of quasars. By $z \sim 3$, almost all of the baryonic intergalactic gas was ionized with HII and HeIII. This second Helium reionization event further dumped heat into the IGM.

Following the end of reionization, the IGM started cooling off predominantly adiabatically as the Universe expanded. This gave rise to a power law relationship between the gas density and temperature for the diffuse gas. This temperature-density relation (TDR) can be written as

$$T = T_0 \left(\frac{\rho_\mathrm{b}}{\bar{\rho}_\mathrm{b}} \right)^{\gamma - 1}, \tag{1.13}$$

where T_0 characterizes a temperature of the IGM with the mean density $\bar{\rho}_\mathrm{b}$ and γ is the power law index. Constraining these two parameters from $z \sim 2$ up to $z \sim 5$ provides crucial evidence for the reionization timescale and the heat deposited into the IGM. In practice, however, the temperature and density of the ionized IGM are not measurable directly. Fortunately, the expanding Universe provides us with a unique probe of this gas called the Lyα forest.

1.6 Lyα forest

The electronic transition in neutral Hydrogen (HI) between the ground state ($n = 1$) and the first excited state ($n = 2$) is called a Lyman-α transition and leads to absorption or emission of a photon of wavelength $\lambda_{\mathrm{Ly}\alpha} = 1215.67$ Å in the rest frame. Hydrogen is the most abundant element in the universe and that is certainly true for massive accretion disks around quasars as well as the intergalactic medium (IGM). Indeed, the broad Lyα emission line is one of the most iconic features of a quasar's spectral energy distribution (SED). The Lyα is an ultraviolet radiation, however, and is inaccessible to ground-based observatories due to Earth's atmosphere. These Lyα photons, while traveling through the expanding Universe before being collected by a telescope on Earth, experience cosmological redshift proportional to their distance of travel. Hence, the observed emission peak is seen at $\lambda_{\mathrm{Ly}\alpha}^{\mathrm{obs}} = 1215.67$ Å $(1 + z_{\mathrm{QSO}})$ for a quasar of redshift z_{QSO}. At $z_{\mathrm{QSO}} \gtrsim 2$, the Ly$\alpha$ peak enters the optical spectrum and we can detect it using ground-based telescopes.

Electronic transitions between the ground state and higher excited states form the Lyman series (Lyα $(1 \to 2)$, Lyβ $(1 \to 3)$, Lyγ $(1 \to 4)$, …) and a bound-free transition from the ground state leads to the Lyman limit $(1 \to \infty)$. Photons with higher energy than the Ly∞ make up the Lyman continuum (LyC). From the Rydberg formula

$$\lambda_n = 911.75 \text{ Å} \left(1 - \frac{1}{n^2} \right)^{-1}, \quad n = 2, 3, 4, \ldots,$$

one can see that a higher Lyman series line has a shorter wavelength.

The neutral Hydrogen in the IGM attenuates the Lyα radiation of the quasars via $I_{\mathrm{obs}} = I_{\mathrm{em}} e^{-\tau}$, where τ is called the optical depth of the medium. However, due to the

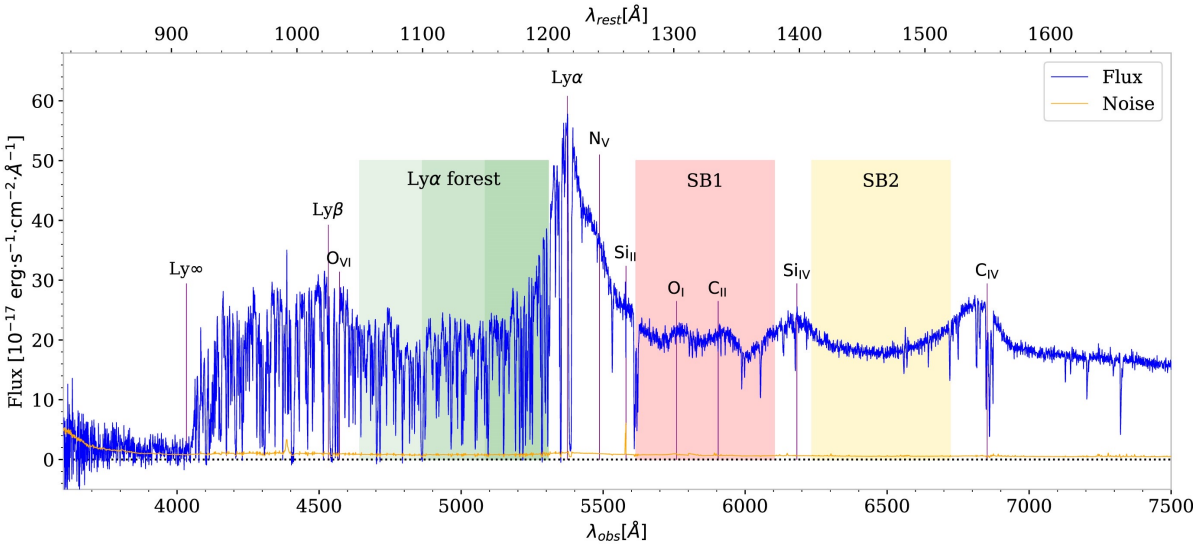

Figure 1.3: A quasar spectrum observed by the Dark Energy Spectroscopic Instrument (DESI) at $z \sim 3.4$. The Lyα forest can be seen on the blue side of the quasar's Lyα emission peak at 1216Å. Adopted from Ravoux et al. (2023), DESI Collaboration.

constantly redshifting quasar spectrum, this attenuation is spread over a wide range of observed wavelengths ($\tau \to \tau(z)$) blueward ($z < z_{\rm QSO}$) of the intrinsic Lyα emission of the quasar (Gunn and Peterson 1965). The local Lyα optical depth at redshift z scales linearly with the number density of neutral H $n_{\rm H\,I}$ as

$$\tau(z) = \frac{\pi q_{\rm e}^2 f_{\ell u} \lambda_0}{m_{\rm e} c H(z)} n_{\rm H\,I}(z) \sim 1.856 \left(\frac{n_{\rm H\,I}}{10^{-10}{\rm cm}^{-3}} \right) \tag{1.14}$$

around $z \sim 2$, where $q_{\rm e}$ is the electronic charge, $f_{\ell u} = 0.416$ is the Lyα oscillator strength, and $\lambda_0 = \lambda_{\rm Ly\alpha}$ is the restframe Lyα wavelength. Neutral H densities of $n_{\rm H\,I} \gtrsim 10^{-10}{\rm cm}^{-3}$ lead to a complete absorption of the incident radiation and an absorption trough can be seen in the spectra of some high-z quasars blueward of the intrinsic Lyα. It is called a Gunn-Peterson trough.

The gas distribution in the IGM is not devoid of any structure, however. Density spikes along the line of sight due to, e.g., cosmic sheets, filaments, and halos lead to enhanced attenuation by absorption lines scattered in the spectrum (Bahcall and Salpeter 1965). Thus, the Lyα forest owes its name to the appearance of such closely packed Lyα absorption lines in high-resolution quasar spectra. Similarly, higher order Lyman series forests also exist blueward of their respective emission wavelengths for a given quasar. For instance, in the observed wavelengths $\lambda < \lambda_{\rm Ly\beta}$ a superposition of Lyα and -β forests is featured, for $\lambda < \lambda_{\rm Ly\gamma}$ a superposition of Lyα, -β and -γ forests, and so on. Physically, a Lyα forest in a $z \sim 3$ quasar extends over hundreds of Mpc in front of the quasar and provides a continuous trace of the IGM along the sightline. Figure 1.3 shows a quasar spectrum observed by the Dark Energy Spectroscopic Instrument (DESI).

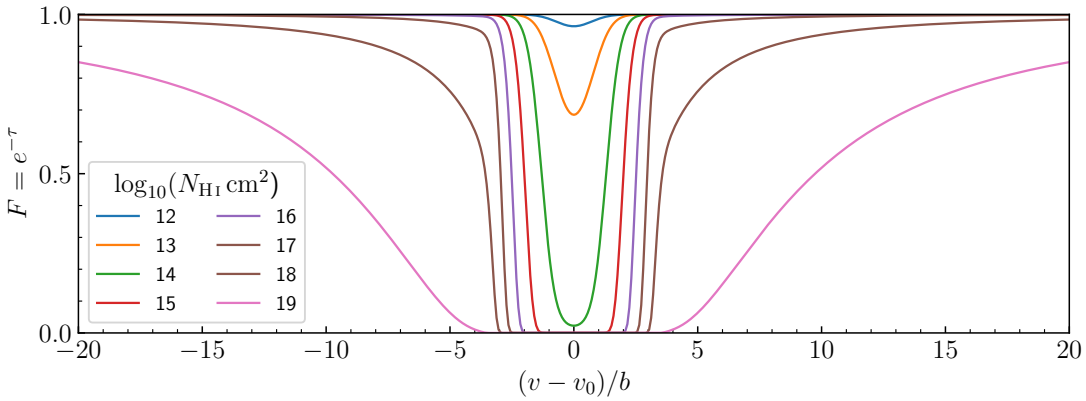

Figure 1.4: Voigt profile for a Lyα absorption line with $b = 20$ km/s for a range of absorber column densities.

For studying the intervening Universe with the Lyα forest, the intrinsic shape of the quasar's spectrum is of little importance and as such we can factor it out by defining a dimensionless Lyα transmission field as $F(\lambda_{\rm obs}) := I_{\rm obs}(\lambda_{\rm obs})/I_{\rm em}(\lambda_{\rm obs}) = e^{-\tau(\lambda_{\rm obs})}$, where $I_{\rm em}(\lambda_{\rm obs})$ is the intrinsic emission spectrum of the quasar, redshifted to $z_{\rm QSO}$.

Even though the transition wavelength is well-defined, the individual absorption lines in the Lyα forest are not sharp δ-function dips in the spectra. Instead they have a specific line shape expressed by a normalized profile function ϕ such that $\int \phi(\nu)d\nu = \int \phi(\lambda)d\lambda = 1$. Two principal physical effects govern the line profile:

(i) Stochastic thermal motion of atoms in an absorber cloud smooths out the absorption line via Doppler broadening. This effect has an intrinsic Gaussian line profile of the form

$$\phi_{\rm D}(\nu) = \frac{1}{\Delta\nu_{\rm D}\sqrt{\pi}} e^{-(\nu-\nu_0)^2/\Delta\nu_{\rm D}^2}, \tag{1.15}$$

with the Doppler width $\Delta\nu_{\rm D} = b\nu_0/c$ that controls the width of the Gaussian, where $b = \sqrt{2k_{\rm B}T/m_{\rm H}}$. This is a manifestation of the Maxwellian velocity distribution of the gas particles.

(ii) Quantum mechanical uncertainty in the transition timescale gives rise to a Lorentzian profile

$$\phi_{\rm L} = \frac{\gamma_{\rm L}}{\pi(\Delta\nu^2 + \gamma_{\rm L}^2)}, \quad \gamma_{\rm L} = \frac{2\pi q_{\rm e}^2 f_{\ell u}}{3m_{\rm e}c\lambda_0^2}, \tag{1.16}$$

where $\gamma_{\rm L}$ is known as the Lorentz damping width.

Thus, the Lyα optical depth at Hubble velocity v along the line of sight can be expressed as $\tau(v) \sim n_{\rm H\,I}(v)\phi_{\rm V}(v)$, where $n_{\rm HI}$ is the number density of neutral Hydrogen and $\phi_{\rm V}$ is called the Voigt profile, which is a convolution of the Gaussian and Lorentzian profiles.

Curve of growth: The broadening of absorption lines can be quantified by their equivalent width (see, e.g., Draine 2011). For any given absorption line, the equivalent

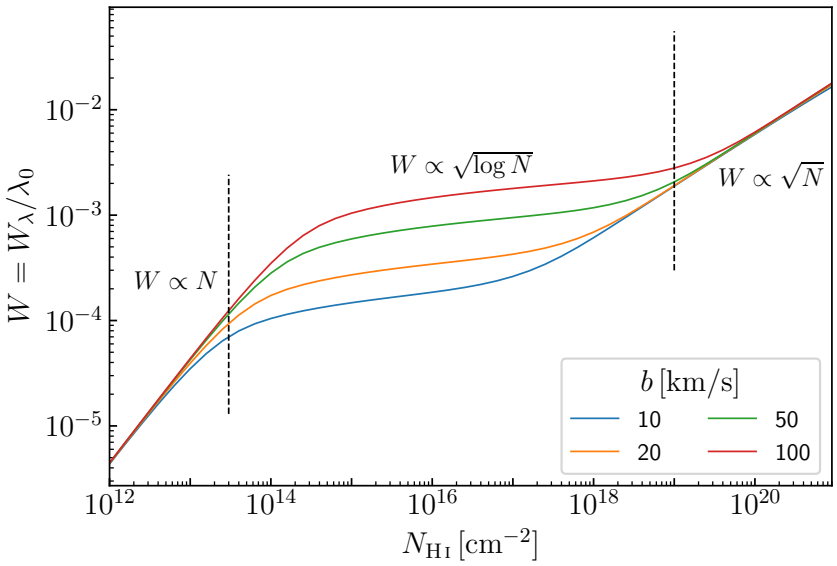

Figure 1.5: The curve of growth for a Lyα absorption line (λ_0 =1216 Å) and a variety of Doppler parameters b typical of the diffuse IGM that gives rise the the Lyα forest feature.

width is defined as the width of a rectangle of height $F = 1$ (i.e., $I = I_{\rm em}$) whose area equals that under the given absorption line. Mathematically, this is equivalent to

$$W_\lambda = \int d\lambda\,[1 - F(\lambda)]. \tag{1.17}$$

In the wings of an absorption line, the Doppler profile attenuates fast but the Lorentzian is a slowly decaying function. For very low column densities $N \sim \int n \cdot {\rm d}v$, i.e., optically thin absorbers, the effect of the wing absorption is negligible. Therefore the equivalent width increases rapidly with increasing density (linearly, $W \propto N$) and quickly achieves saturation. After that, the equivalent width varies slowly with density as $W \propto \sqrt{\log N}$. At very high column densities, the absorption in the wings starts to matter and the equivalent width starts increasing again with density as $W \propto \sqrt{N}$. An example of the Voigt line profile for varying column densities is shown in Figure 1.4. This behavior of equivalent width of an absorption line *w.r.t.* the absorber column density is known as the "curve of growth" and is shown in Figure 1.5 for the resonant Lyα line.

Eqs. (1.14) and (1.15) establish that the absorption lines in the Lyα forest are extremely sensitive to the temperature and density of the diffuse IGM, which can be exploited for constraining the TDR (T_0, γ) parameters. The concepts developed in this subsection shall be applied to the generation of mock Lyα forest data in Chapter 3.

Chapter 2

Statistical methods

One of the core themes of this work is the use of statistics to make sense of complex, high-dimensional data. In this chapter I discuss the fundamental statistical techniques that are predominantly employed in this thesis. For the discussion of deep learning in Section 2.3 I have drawn inspiration from a classic text by Goodfellow et al. (2016).

2.1 Bayesian inference

A scientific observable from an experiment is often fit with a theoretical model with some free parameters. For instance, the angular power spectrum of the CMB can be fit with the standard ΛCDM model and in the process the values of cosmological parameters can be inferred.

A primary goal of any scientific experiment is to measure or infer the value of a set of parameters $\boldsymbol{\pi}$ of interest. In most cases in astrophysics and cosmology, it is not possible to make a direct measurement and we must rely on observable proxies with a known physical relationship to the parameter(s) of interest. For instance, to infer the total mass of a galaxy, its rotation curve (rotational velocity as a function of radius) can be used since it is sensitive to the galactic mass[1]. In an astrophysical observation, Starting from our prior belief of the likely values of those parameters and using data of an observable $\mathbf{d}$, we can apply Bayes' theorem of conditional probabilities to infer the probability distribution of $\boldsymbol{\pi}$ *a posteriori*. The prior can be a wide flat (uniform) uninformative distribution or it may be influenced by any underlying physical constraints, or we may choose to trust a previous experiment and use the posterior distribution as our prior. Bayesian probability assigns a degree of belief or confidence in a measurement of a random variable ($\boldsymbol{\pi}$ in our case), as opposed to a frequentist probability that relies on repeated experiments to chart a representative distribution of the random variable. Having a prejudice over the parameters *a priori* implies a Bayesian standpoint and henceforth all the probabilities should be interpreted as such.

[1]The existence of dark matter was hypothesized to explain "missing" mass in spiral galaxies (Rubin 1983).

With Bayes' theorem, we can express the joint distribution of the observed data and the parameters in terms of their conditional distributions as

$$\mathcal{P}(\mathbf{d}, \boldsymbol{\pi}) = \mathcal{P}(\boldsymbol{\pi}|\mathbf{d})\mathcal{P}(\mathbf{d}) = \mathcal{P}(\mathbf{d}|\boldsymbol{\pi})\mathcal{P}(\boldsymbol{\pi}), \tag{2.1}$$

where $\mathbf{d}$ is an observable data vector of size N_d. Intuitively, this means the act of observing given values of $\mathbf{d}$ and $\boldsymbol{\pi}$ together can be decomposed into observing one quantity independently and then the other, conditional on the first having taken a specific value. We can express the posterior probability $\mathcal{P}(\boldsymbol{\pi}|\mathbf{d})$ as

$$\mathcal{P}(\boldsymbol{\pi}|\mathbf{d}) = \frac{\mathcal{P}(\mathbf{d}, \boldsymbol{\pi})}{\mathcal{P}(\mathbf{d})} = \frac{\mathcal{P}(\mathbf{d}|\boldsymbol{\pi})\mathcal{P}(\boldsymbol{\pi})}{\mathcal{P}(\mathbf{d})}. \tag{2.2}$$

In the terminology of Bayesian inference, $\mathcal{P}(\mathbf{d}|\boldsymbol{\pi})$ is called the "likelihood" and is a function of $\boldsymbol{\pi}$, $\mathcal{P}(\mathbf{d}|\boldsymbol{\pi}) =: \mathcal{L}(\boldsymbol{\pi})$, for the given value $\mathbf{d}$ of the observable. It should be noted that $\mathcal{L}$ is not a probability density function (PDF) of $\boldsymbol{\pi}$. $\mathcal{P}(\mathbf{d})$ is called the Bayesian evidence and is useful for comparing the adequacy of competing models. Since in this work we focus exclusively on one model, however, the evidence term merely plays the role of a normalization constant.

In cosmological literature, the likelihood function is often assumed to be a (multivariate) Gaussian function of the model prediction for the observable $\mathbf{d}_\mathrm{m}(\boldsymbol{\pi})$ centered around the data $\mathbf{d}$, such as

$$\log \mathcal{L}(\boldsymbol{\pi}) \propto -\frac{1}{2} \log\det \mathbf{C} - \frac{\chi^2}{2}, \tag{2.3}$$

with

$$\chi^2 = \boldsymbol{\Delta}\mathbf{C}^{-1}\boldsymbol{\Delta}^T, \tag{2.4}$$

where $\boldsymbol{\Delta} = \mathbf{d}_\mathrm{m}(\boldsymbol{\pi}) - \mathbf{d}$ and $\mathbf{C}$ is the covariance matrix of the observable. In experiments, the covariance matrix may be estimated from repeated observations of $\mathbf{d}$ (called the data covariance) or, if available, a theoretical model may be used to compute it. In the latter, $\mathbf{C}$ may be dependent on $\boldsymbol{\pi}$, $\mathbf{C} \equiv \mathbf{C}(\boldsymbol{\pi})$. Then, samples can be drawn from the posterior PDF using Markov chain Monte Carlo (MCMC) methods. Each MCMC step requires evaluation of the likelihood function and the prior PDF, which in turn requires an evaluation of $\mathbf{d}_\mathrm{m}$. Within ΛCDM, a closed-form expression of the CMB angular power spectrum model can be derived using Boltzmann equations. At late cosmological epochs, however, nonlinear gravitational structure formation renders analytical modeling infeasible. To model the Universe at late times, expensive N-body or hydrodynamical simulations need to be run. A simulator may be defined as a machine or process that generates mock data vectors for any given input parameter values, $\mathbf{d}_\mathrm{sim}(\boldsymbol{\pi})$. Conventional N-body or hydrodynamical simulations are computationally expensive and challenging to mass-produce. Usually, a small number of simulation boxes are generated with a variety of underlying parameters and mock $\mathbf{d}_\mathrm{sim}(\boldsymbol{\pi})$ vectors are extracted from them. A fast and computationally inexpensive surrogate simulator can then be fit to the known set $\{\mathbf{d}^i_\mathrm{sim}, \boldsymbol{\pi}^j\}$ that emulates the original simulation process, for instance, by using Gaussian processes or machine learning. These "emulators" are then used directly in the likelihood evaluations during MCMC.

2.1.1 Likelihood-free inference (LFI)

In many practical applications, the exact form of the likelihood function $\mathcal{P}(\mathbf{d}|\boldsymbol{\pi})$ is analytically intractable, especially when the data-generating process involves complex simulations and/or nonlinear systematics. In such cases, common approximations such as Gaussian likelihoods may fail to capture the true distribution of residuals. Likelihood-free inference (LFI) encompasses a class of methods that bypass the need for an explicit form of the likelihood function by instead relying on our ability to forward simulate data from the physical model. Prominent examples include approximate Bayesian computation (ABC), density-estimation likelihood-free inference (DELFI), and more recently simulation-based inference (SBI) techniques employing neural networks. In this work, I experimented with DELFI (see, e.g., Alsing et al. 2018), wherein a surrogate model of the joint distribution $\mathcal{P}(\mathbf{d}, \boldsymbol{\pi})$ is trained using samples from the simulator. Then, the posterior PDF is expressed as a slice through the joint PDF at the observed data $\mathbf{d}_*$. Using the first equality in Eq. (2.2),

$$\mathcal{P}(\boldsymbol{\pi}|\mathbf{d} = \mathbf{d}_*) \propto \mathcal{P}(\mathbf{d}, \boldsymbol{\pi})\big|_{\mathbf{d}=\mathbf{d}_*}. \tag{2.5}$$

A surrogate model for the joint PDF can be fit using Gaussian mixture models or more sophisticated density estimation methods. MCMC is then used again to draw samples from the posterior PDF.

2.2 Summary statistics

Astrophysical observables $\mathbf{d}$ are often extremely high-dimensional, containing a vast number of degrees of freedom—for example, Lyα forest transmission profiles from spectroscopic observations of quasars. These are commonly termed field-level observables. For stochastic, high-dimensional fields like these, evaluating a likelihood function becomes infeasible, since the complexity of the data precludes meaningful comparison to model predictions. This high dimensionality also poses a significant obstacle for fitting a surrogate model to the joint density $\mathcal{P}(\mathbf{d}, \boldsymbol{\pi})$ in DELFI.

However, the subspace of features in the field that are sensitive to the parameters of interest is often much lower-dimensional. It is therefore reasonable to compress the full field into lower-dimensional vectors that summarize the relevant information in various statistical ways. These summary statistics allow us to extract essential, informative features of the data while marginalizing over other, potentially irrelevant or noisy degrees of freedom. We can then use summary statistics in lieu of the field as our observables for performing Bayesian inference.

Different summary statistics of the same field can exhibit different degrees of sensitivity to the parameters of interest. This sensitivity directly affects the tightness of our posterior constraints and degeneracies, if any, in the parameter space. In general, higher sensitivity manifests itself as tighter posterior constraints and vice versa. Formally, this notion is

captured by the Fisher information matrix, defined for a likelihood function $\mathcal{L}(\boldsymbol{\pi})$ as

$$\mathcal{I}_{ij}(\boldsymbol{\pi}) = -\left\langle \frac{\partial^2 \log \mathcal{L}(\boldsymbol{\pi})}{\partial \pi_i \partial \pi_j} \right\rangle, \tag{2.6}$$

which quantifies the amount of information a statistic carries about the given parameters. Selecting summary statistics that preserve Fisher information is therefore a crucial step in simulation-based inference.

With the advent of deep learning based data compression, a new classification of summary statistics has emerged: (i) conventional, intuitive and analytical summaries, and (ii) machine-learned, weakly interpretable (black-box) summaries. In this work, I developed a framework of the latter kind for the Lyα forest. By training a deep neural network to compress the field and make point predictions of the parameters of interest, we can construct a custom summary statistic optimized to preserve the field-level information pertinent to those parameters. Because of this, such methods are sometimes referred to as field-level approaches. An important caveat, however, is that compressions of this sort tend to be largely insensitive to parameters not included in training. Expanding the parameter set $\boldsymbol{\pi}$ typically requires retraining a more complex model, although techniques such as pre-training and transfer learning may help alleviate this problem. In contrast, conventional summary statistics are often sensitive to a broader range of cosmological and astrophysical parameters, albeit typically with reduced per-parameter sensitivity. In this work, we trade off broader sensitivity for higher precision of a targeted subset of parameters.

In the following, I introduce two of the commonly used conventional summary statistics of the Lyα forest transmission field. In later chapters, I compare the posterior constraints obtained from my deep learning framework with those derived from these conventional statistics and their combination.

2.2.1 1D power spectrum

The line-of-sight power spectrum, more commonly referred to as the 1D power spectrum of the Lyα forest is one of the most well-studied summary statistics of the transmission field. It is a measure of the variance of the Fourier modes of the transmission contrast $\delta_F(v) = F(v)/\langle F \rangle_v - 1$,

$$\Delta_F^2(k) = k P_F(k)/\pi, \tag{2.7}$$

with $P_F(k) \sim \langle \tilde{\delta}_F(k)^* \cdot \tilde{\delta}_F(k) \rangle$, where v is the Hubble flow velocity along the line of sight and $k = 2\pi/v$. At large scales, i.e., small k, P_F exhibits a power-law increase. At larger k, the thermal and pressure smoothing of the gas leads to suppression of power. At these scales, P_F is highly sensitive to the thermal state of the IGM as illustrated in Figure 2.1 (left). An effective increase in IGM temperature amounts to an enhanced damping of power at large k and further a shift of the P_F turnover scale toward smaller k. The 1D power spectrum of the Lyα forest has been measured by large cosmological surveys with very high precision. Because of its sensitivity to a wide range of physics, it has been used to constrain cosmological parameters such as σ_8, Ω_m, and n_s as well.

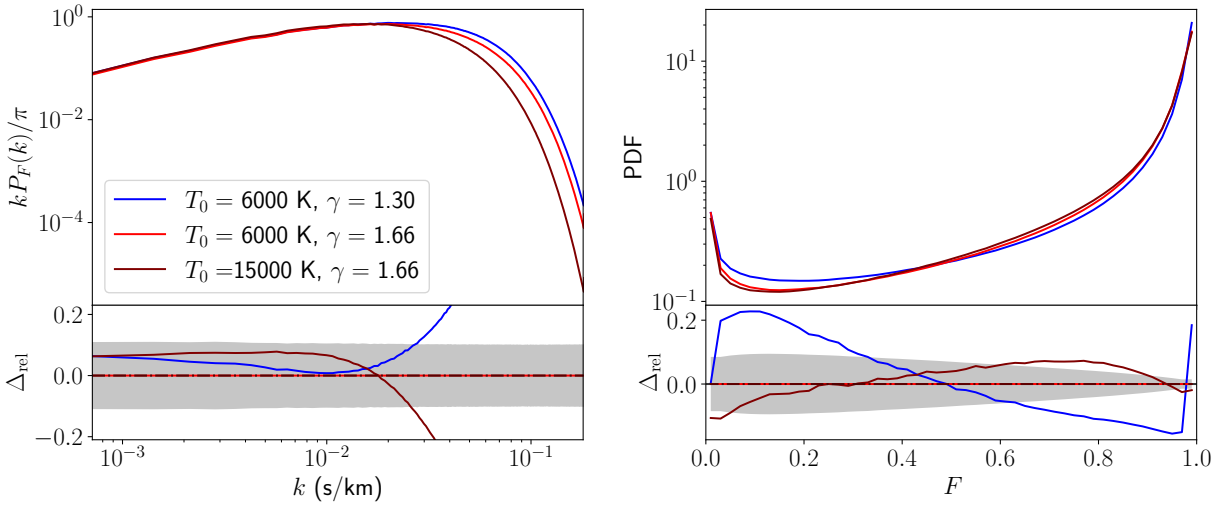

Figure 2.1: 1D power spectrum (left) and PDF (right) of the Lyα forest transmission from hydrodynamical simulations with varying TDR parameters T_0 and γ.

2.2.2 Transmission PDF

The PDF of the Lyα transmission, $p(F)$, is another common statistic of the Lyα forest. At low redshifts ($z \sim 2$) the absorption is small and the forest is sparse, yielding the maximum transmission probability at $F \sim 1$. Near $F \sim 0$ saturated absorption lines lead to another peak in $p(F)$. As the right panel of Figure 2.1 shows, the transmission PDF (TPDF) is also sensitive to the thermal state of the IGM. An increase in temperature – implying a broadening of lines – amounts to a transfer of probablity from $F \sim 0.2$ to $F \sim 0.8$. The TPDF is shown to be more sensitive to variations in γ than T_0. In practice, the TPDF is measured in $\sim \mathcal{O}(10)$ discrete transmission bins in $0 \leq F \leq 1$ and normalized such that $\int_{-\infty}^{\infty} dF\, p(F) = 1$. Because of the normalization, this statistic is highly degenerate, and while estimating the covariance matrix, one of the bins should be left out in order to avoid singularities.

2.3 Deep learning

Deep learning refers to a specific class of artificial intelligence (AI) algorithms that is rooted in learning representations in a hierarchical fashion from data. Training proceeds in a layered manner, from some input through a set of tiers stacked onto each other to the output, each intermediate tier a (usually smaller dimensional) representation of the previous one. As such, these models are deep, the depth controlled by the number of layers.

Feedforward models: These are deep models in which the information flows strictly forward from the input to the output without any intermediate feedback circuits.

Usually, the goal of such models is to learn an unknown complex, high-dimensional

relationship or a function from one vector-valued variable to another using an available representative set of examples called the training data. Say we wish to learn an unknown target function $\mathbf{y} = \hat{\mathbf{f}}(\mathbf{x})$, where the sizes of $\mathbf{x}$ and $\mathbf{y}$ may be different. Deep learning aims to find a parametric representation of this function,

$$\mathbf{y}' = \mathbf{f}(\mathbf{x}; \boldsymbol{\phi}), \tag{2.8}$$

with a vector of "trainable" parameters $\boldsymbol{\phi}$ and a measure of distance or similarity between $\mathbf{f}$ and $\hat{\mathbf{f}}$ in terms of a cost function. Provided is a representative set of example data $\{\mathbf{x}^i, \mathbf{y}^i\}$ drawn from the underlying data generating distribution $p(\mathbf{x}, \mathbf{y})$. Training entails optimizing the cost function *w.r.t.* $\boldsymbol{\phi}$; in other words, fitting $\mathbf{f}$ with the training data. Since we have known function evaluations at specific data points, this constitutes supervised learning wherein the network is supervised by the example labels $\{\mathbf{y}^i\}$ of the underlying unknown function to be learned. The term label is borrowed from a classification problem where the different target categories are labeled and the task of the machine is to classify the input data vector into one of them. In a regression task such as the one described here, the known values of the function serve a similar purpose.

The choice of a cost function, or more commonly a loss function in deep learning terminology, is somewhat dependent on the task at hand. Binary cross-entropy is a common type of loss function used in classification tasks. For regression, the mean squared error (MSE) is one of the most typical loss functions. Other examples are mean absolute error, χ^2, negative log likelihood, etc. The programmer may choose a specific loss function that must then be used throughout the learning process. Training constitutes starting with random values of $\boldsymbol{\phi}$, evaluating the function $\mathbf{y}' = \mathbf{f}(\mathbf{x}; \boldsymbol{\phi})$ for the training examples $\{\mathbf{x}^i\}$ (called model prediction), calculating the distance between the model prediction $\{\mathbf{y}'^i\}$ and the true labels $\{\mathbf{y}^i\}$ in terms of the loss $L(\boldsymbol{\phi})$, and then finding optimal values of the parameters $\boldsymbol{\phi}^*$ that minimize the loss function such that $\boldsymbol{\phi}^* = \arg\min_{\boldsymbol{\phi}} L(\boldsymbol{\phi})$. The algorithm for optimization is also the choice of the programmer and typical optimizers include stochastic gradient descent (SGD) and Adam.

2.3.1 Neural networks

We start with the simplest form of adaptable representation *viz.* a linear mapping,

$$\begin{aligned} \mathbf{y}' &= \mathbf{f}_{\text{lin}}(\mathbf{x}; \boldsymbol{\phi}) \\ &= \mathbf{W}\mathbf{x} + \mathbf{b}, \end{aligned} \tag{2.9}$$

where $\mathbf{W}$ is a matrix of weight coefficients and $\mathbf{b}$ is an offset or bias vector, and together they form the set of trainable parameters $\boldsymbol{\phi} \equiv \{\mathbf{W}, \mathbf{b}\}$. Linear fitting may be a good representation in many simpler, lower dimensional cases following Occam's razor, however, it does not always serve when the target function is more complex and high dimensional. Introduction of nonlinearity may help ameliorate this problem,

$$\mathbf{y}' = \mathbf{a}(\mathbf{W}\mathbf{x} + \mathbf{b}),$$

where **a** is a nonlinear function of the programmer's choosing. However, in this case, the form of the fit is entirely governed by **a**. To allow for more fitting freedom a further linear mapping layer may be introduced:

$$\begin{aligned} \mathbf{y}' &= \mathbf{W}_2\mathbf{h} + \mathbf{b}_2, \\ \mathbf{h} &= \mathbf{a}(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1), \end{aligned} \tag{2.10}$$

where **h** is an intermediate nonlinear mapping. The model in Eq. (2.10) is called a "perceptron" and is the simplest nonlinear learning algorithm. One may interpret it as implicitly learning a nonlinear **h** and then finding a linear combination of its units to form the final representation. In this spirit, **h** is a hidden layer of learning. The parameters of the hidden layer are not directly influenced by the training data and the model must find the optimal values of them that best represent the data. The fitting degrees of freedom can be greatly augmented by chaining multiple such mappings

$$\mathbf{y}' = \mathbf{f}^M(\mathbf{x}; \{\mathbf{W}_m, \mathbf{b}_m\}), \quad m = 1, \ldots, M+1$$

such as

$$\begin{aligned} \mathbf{y}' &= \mathbf{W}_{M+1}\mathbf{h}_M + \mathbf{b}_{M+1}, \\ \mathbf{h}_M &= \mathbf{a}(\mathbf{W}_M\mathbf{h}_{M-1} + \mathbf{b}_M), \\ &\vdots \\ \mathbf{h}_2 &= \mathbf{a}(\mathbf{W}_2\mathbf{h}_1 + \mathbf{b}_2), \\ \mathbf{h}_1 &= \mathbf{a}(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1). \end{aligned} \tag{2.11}$$

In this model, each hidden layer learns an intermediate nonlinear representation and together they help better adapt the full representation to the actual target function. Eq. (2.11) corresponds to a multilayer perceptron (MLP) model and is the quintessential deep learning algorithm. It is more commonly known as a (artificial) **neural network** (NN) due to its structural and functional similarity to biological neural networks from where it draws the inspiration. When biological neurons receive an input signal, they either stay passive or they fire and activate the passage of that signal to other connected neurons depending on the strength of the signal. Each unit of a hidden layer $\mathbf{h}_m$ acts as a neuron: it receives a signal from the previous layer $\mathbf{h}_{m-1}$, the nonlinear function **a** acts as the activation and this "activated" signal is then passed on to all the units of the next layer $\mathbf{h}_{m+1}$ connected with it. The number of neurons in a layer (input, hidden or output) determines the width of the network and the number of hidden layers determines the depth. Figure 2.2 shows an illustration of a fully-connected NN with a depth of 2 hidden layers. The specific structure of a NN, e.g., depth, width of individual hidden layers, etc, is called the network's "architecture."

Gradient based optimizers require that derivatives of the vector valued functions be computed *w.r.t.* the input variable(s). In the case of chained representations such as neural networks, this implies that the chain rule of derivatives be employed for computing

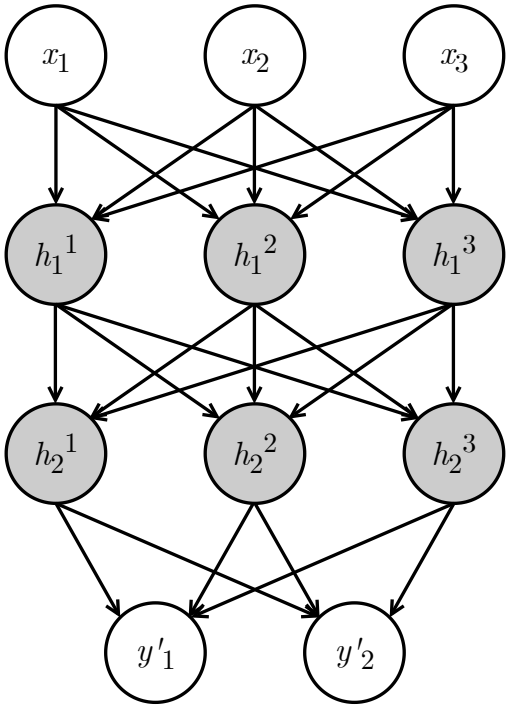

Figure 2.2: An illustration of a multilayer perceptron model with two hidden layers $\mathbf{h}_1$ and $\mathbf{h}_2$. The input dimension is 3 and the output dimension is 2. All the neurons in each hidden layer are fully connected with the previous and the following layers. The information is fed from $\mathbf{x}$ to $\mathbf{y}$ in one direction without any intermediate feedback loops. For this reason, it can also be called a fully-connected feedforward NN.

all the intermediate differentiations leading back to the input. For a simpler case of a perceptron in Eq. (2.10) with the set of parameters $\boldsymbol{\phi} = \{\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1, \mathbf{b}_2\}$ we may compute the gradient of a component y'_i *w.r.t.* $\boldsymbol{\phi}$ using the chain rule as

$$\boldsymbol{\nabla}_{\boldsymbol{\phi}} y'_i = \left(\frac{\partial \mathbf{h}}{\partial \boldsymbol{\phi}}\right)^T \boldsymbol{\nabla}_{\mathbf{h}} y'_i,$$

where $\partial \mathbf{h}/\partial \boldsymbol{\phi}$ is the Jacobian matrix of partial derivatives of $\mathbf{h}$ *w.r.t.* $\boldsymbol{\phi}$. For deeper networks, it requires the computation of the Jacobian of each hidden layer *w.r.t.* the previous layer, or a propagation of derivatives from the output through the intermediate mappings all the way back to the input. For this reason, the computation of gradients in deep models and a step of the optimizer is called "back-propagation."

Deep learning models usually come with a set of background "hyperparameters," separate from $\boldsymbol{\phi}$, that do not change over the course of model training, but rather they influence the model architecture, loss function, optimizers, etc. Tuning the values of these hyperparameters is crucial for obtaining a model with optimal capacity.

Capacity: The capacity is a somewhat abstract concept that refers to the ability of any given deep learning model to fit a large variety of target functions. Alternatively, it may be interpreted as the adaptability of the model to fit the given training data. Models with too low capacity struggle to find a best fit solution, or in other words their optimal training loss is too high. This indicates underfitting. For instance, a linear model in Eq. (2.9) would underfit a polynomial function with a degree $n > 1$. On the other hand, models with too high capacity fit the training set too well so that the optimal training loss is very small, however, they perform poorly on new, previously unseen data. This failure in generalization of high capacity models indicates overfitting. Models with optimal capacity achieve small error values on both the training data and previously unseen test data.

We would like our model to make good predictions for new data that the model hasn't seen during the learning process. Learning here comprises training with back-propagation *and* hyperparameter-tuning. For this purpose, a small part of the training data is usually set aside before starting the learning process to monitor the generalization of the model during hyperparameter-tuning. This set is called the "validation" dataset that guides the choice of optimal hyperparameter values. In practice, the two steps of the learning process can be completed within a single nested model training program.

2.3.2 Convolutional neural networks

A convolutional neural network (CNN) is a special type of NN that employs the convolution operation to detect useful features of data with a specific topology like 1D time series or 2D images. CNNs have revolutionized the field of Computer Vision. They have been widely used in image processing applications such as detecting cats and dogs, or recognizing handwritten digits. CNNs are also deep networks that learn important features of data in a hierarchical fashion, starting from low level features (sometimes called atomic elements) in the initial layers, through repeated submotifs and high-level motifs in the later layers.

In 1D, the discrete convolution operation between a function $\mathbf{f}$ and a kernel $\mathbf{k}$ can be expressed as

$$(\mathbf{f} \otimes \mathbf{k})_i := \sum_j f_j \cdot k_{i-j}. \tag{2.12}$$

It can be seen intuitively as sliding the kernel over the given function (signal) to estimate a local correlation of the signal with the kernel. CNNs incorporate this operation in place of the matrix multiplications in fully-connected NNs (Eq. (2.11)). In this analogy, the weights $\mathbf{W}$ can be replaced with the convolutional kernels or filters $\mathbf{k}$ and the hidden layers $\mathbf{h}$ are then called feature maps since they represent correlation maps of the input with the given filters. A key difference to note here is that while in fully-connected NNs the hidden layers are vectors, in CNNs a hidden layer constitutes multiple filters of the same size leading to multiple feature maps. The aim of training such a network is to learn distinct features of the data through trainable filters at different scales and eventually use the presence and strength of those features to make a final decision (regression, classification, etc). Heuristically, hierarchical data such as images or spectra consist of a small number of

atomic elements and the higher level features are different combinations of the lower level features. Thus, in a typical CNN architecture the number of trainable filters increases with the depth of the convolutional layer. Convolution may not necessarily result in a compression of data into smaller sizes. For deep CNNs with a large number of filters in deeper layers, this may practically result in a computational overload. To tackle such numerical challenges, **pooling layers** are usually added to help keep the data dimension in check. Most common examples of pooling layers are max-pooling and average pooling layers.

Residual learning

The depth of a NN architecture is an important hyperparameter. Deeper models have larger degrees of freedom, allowing us to fit more complex representations. However, is increasing the depth, i.e., stacking more layers always beneficial while the model is still below optimal capacity? Curiously, the answer turns out to be no. In this low capacity regime, not all of the kernels in a newly added layer may need to learn new features and they only need to relay the features of the previous layer. When nonlinear activations are involved, training individual kernels to emulate identity mappings becomes extremely difficult, usually leading to a rise in the training loss upon increasing the depth. He et al. (2015a) came up with a clever solution to this problem by introducing linear skip connections into a sequential (stacked) architecture. The layers learn residual mappings relative to the identity that are easier to push to null with nonlinear layers. Further discussion on residual learning can be found in Section 4.3.2. This revolutionary invention facilitated training deeper CNNs and led to major advances in Computer Vision. These residual networks, or more commonly "**ResNets**" have since become the canonical CNN architectures. I thus employed a 1D ResNet architecture for the field-level inference framework that is at the heart of this thesis.

Having established the necessary theoretical background, I now turn to the central investigations of this thesis in the following chapters.

Chapter 3

Synthesis of mock Lyα forest spectra

Deep learning models are only as good as their training data. The success of a deep learning model for the inference of the IGM TDR parameters (T_0, γ) from the Lyα forest lies within a large labeled training dataset of mock spectra. For the production of those mocks, I used a suite of hydrodynamic simulations by Walther et al. (2025) run with the Nyx code (Almgren et al. 2013) that is suitable for Lyα forest analyses. This is motivated by some convergence studies such as Walther et al. (2019) and Chabanier et al. (2023) at the level of 1D power spectrum. Although they lend a sufficient degree of reliability to the mocks for inference with traditional summary statistics, they make no similar assertions regarding the trustworthiness of the mocks at the full transmission field level. Such an extensive investigation into the physical fidelity of the full hydrodynamic simulations is beyond the scope of this thesis, however, an equivalent excursion into the post-processing mock generation is not only realizable but also one of necessity. Since a field-level comparison of mocks with real observations is highly impractical, an internal consistency check is the next most crucial step for asserting that the mocks carry the same degree of fidelity as the underlying hydrodynamic simulations if they are to be utilized for training NN models. Existing tools for the generation of mocks, used e.g., by Walther et al. (2025) rely on physically motivated assumptions at various stages of this process, however, often with little evidentiary justification. In this chapter, I examine those approximations critically and compare them with their counterparts involving exact calculations. I also evaluate the computational cost for the approximate as well as the full methods and find the best tradeoffs in different cosmological regimes.

3.1 Introduction

A cosmological hydrodynamic simulation models the evolution of a Universe with certain cosmological and astrophysical parameters and with arbitrary initial conditions under gravity and hydrodynamics. The Nyx hydrodynamics code—used for running the simulations used in this thesis—models dark matter as gravitating N-body Lagrangian particles coupled with baryons simulated as an ideal fluid on an Eulerian grid. It offers adaptive

mesh refinement (AMR) capability for a better treatment of overdense regions, however, for analyses involving the Lyα forest that primarily traces only mild overdensities AMR is inefficient and as such not adopted while running this simulation. For a more detailed description of the physical processes accounted for by Nyx I refer the reader to Almgren et al. (2013) and Lukić et al. (2015). The simulation used for the analyses discussed in this chapter has a box side length of 80 Mpc/h (comoving) and contains 4096^3 volumetric cells (hereafter voxels) and dark matter particles. It is the fiducial model of Walther et al. (2025) with the cosmological parameters from Planck Collaboration et al. (2020), namely, $h = 0.6732$, $\Omega_\mathrm{m} = 0.3144$, $\Omega_\mathrm{b} = 0.0494$, $A_\mathrm{s} = 2.101 \times 10^{-9}$, $n_\mathrm{s} = 0.9660$.

A snapshot of a simulation refers to the state of the entire simulation box at a fixed redshift z (or cosmic time). It contains the information of dark matter density ρ_m, density of the baryonic gas ρ_b, peculiar velocity of the gas $\mathbf{v}_\mathrm{pec}$, and temperature of the gas T (among others) for every voxel of the box. The box follows periodic boundary conditions on all surfaces. A line of sight can be realized as a pencil beam penetrating all consecutive cells along one of the three axes of the box while the other two are kept fixed. Such a mock line of sight is often called a "skewer" as a culinary analogy. If we imagine that a skewer points to a bright source of Lyα photons such as a quasar at a high enough redshift ($z_\mathrm{QSO} > z_\mathrm{box}$), we can use the hydrodynamic fields provided by the simulation to mock up the attenuation of that radiation by this part of the universe. Figure 3.1 shows an illustration of this process.

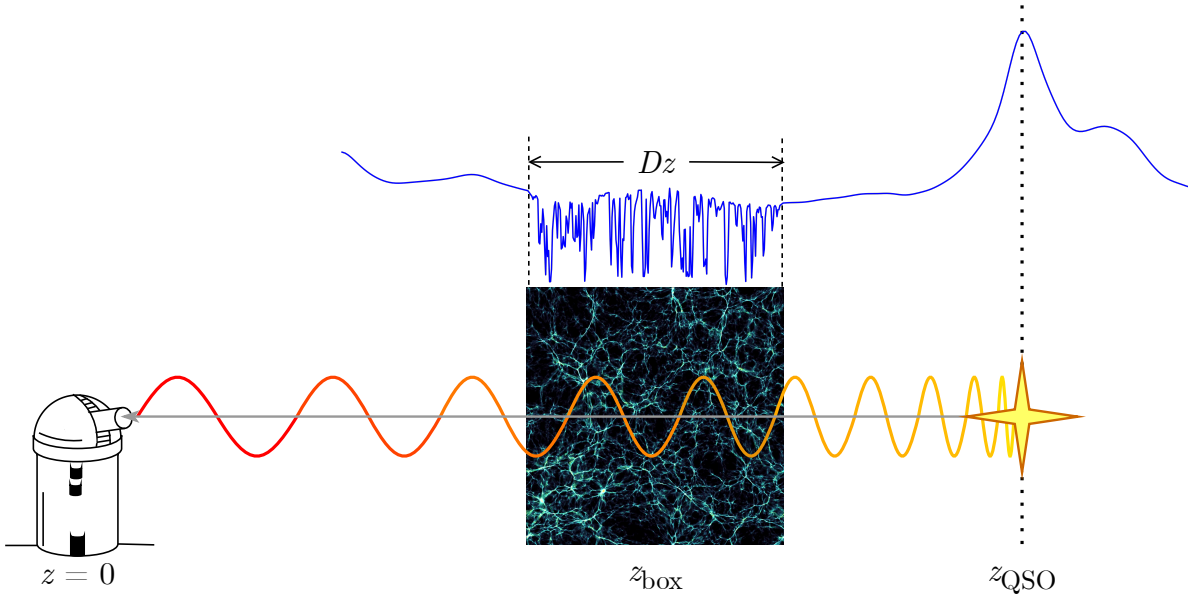

Figure 3.1: An illustration (not to scale) of the creation of a mock Lyα forest spectrum. The part of the universe simulated by a cosmological simulation centered at z_box shown here imprints a Lyα forest absorption feature on the corresponding part of the quasar spectrum. Quasar continuum credits: LyCAN / Wynne Turner.

As a result of the interaction of the Lyα radiation of a bright background source with

the gas along a skewer through the box, a unique Lyα forest transmission spectrum is generated, represented by $F(v) = e^{-\tau(v)}$, where v is equivalent to a redshifted Lyα wavelength. The optical depth is given by

$$\tau(v) = \frac{\pi q_{\rm e}^2 f_{\ell u} \lambda_0}{m_{\rm e} c H(z)} \int_{\rm LoS} n_{\rm HI}(v') \phi'(v) dv', \tag{3.1}$$

where $\phi'(v)$ is the line profile centered at v' and evaluated at v. Here, $H(z)$ is assumed to be fixed along the skewer to $H(z_{\rm box})$. Such a mock spectrum has an infinite spectral resolution and may be estimated at an arbitrarily high pixel resolution ($\Delta v_{\rm p} \to 0$). However, since the underlying skewer has a limited spatial resolution, I chose the pixel size corresponding to the size of the voxel, i.e., $\Delta v_{\rm p} \equiv H(z) a(z) \Delta l_{\rm com} = \frac{H(z)}{1+z} \Delta l_{\rm com}$, where $\Delta l_{\rm com}$ is the lateral comoving extent of a voxel. The discretized version of the expression in Eq. (3.1) is

$$\tau_i = C_0 \sum_j n_{{\rm HI},j} \phi_j(v_i) \Delta v', \tag{3.2}$$

where $C_0 = \frac{\pi q_{\rm e}^2 f_{\ell u} \lambda_0}{m_{\rm e} c H(z)}$, $\Delta v'$ is the size of a voxel in velocity units, the index i runs over the output pixels of a spectrum, and index j runs over the voxels in the input skewer. The profile ϕ_j is centered at the net velocity (Hubble flow + peculiar) at the voxel j, e.g., the Doppler profile $\phi_j^{\rm D}(v_i) \sim \exp\left[-(v_i - v_j' - v_{\parallel,j})^2 / b_j^2\right]$ where v_j' is the Hubble velocity at voxel j and $v_{\parallel,j}$ is the line-of-sight component of the peculiar velocity at the same voxel. I shall occasionally refer to the voxels as belonging to the "real space" and the pixels to the "redshift space."

As indicated by Eq. (3.2), the determination of τ for a mock spectrum requires two key ingredients at every voxel j along the underlying skewer: (i) the local $n_{{\rm HI},j}$, and (ii) the correct line profile $\phi_j(v_i)$. The local $n_{\rm HI}$ depends on the baryon density and temperature as well as the underlying UV background. The line profile is a function of the local temperature (through Doppler broadening) and peculiar velocity, the latter of which shifts the absorption lines in the redshift space. I describe this two-step methodology in the following two sections. I implemented this complete procedure in an open-source Python pipeline SynTH, standing for Synthesis of Transmission from Hydrodynamical simulations.

3.2 Determination of $n_{\rm HI}$

The IGM can be modeled as primarily comprising two components: atomic Hydrogen and Helium (neutral as well as ionized) with mass abundances $X \approx 0.75$ and $Y \approx 0.25$, respectively. An atomic species can be ionized collisionally as well as radiatively and the same holds for the inverse (recombination) process. The IGM can be reasonably assumed to be in ionization equilibrium among the different neutral and ionized species (HI, HII, HeI, HeII, HeIII) and free electrons (e^-). This leads to the following three equilibrium equations:

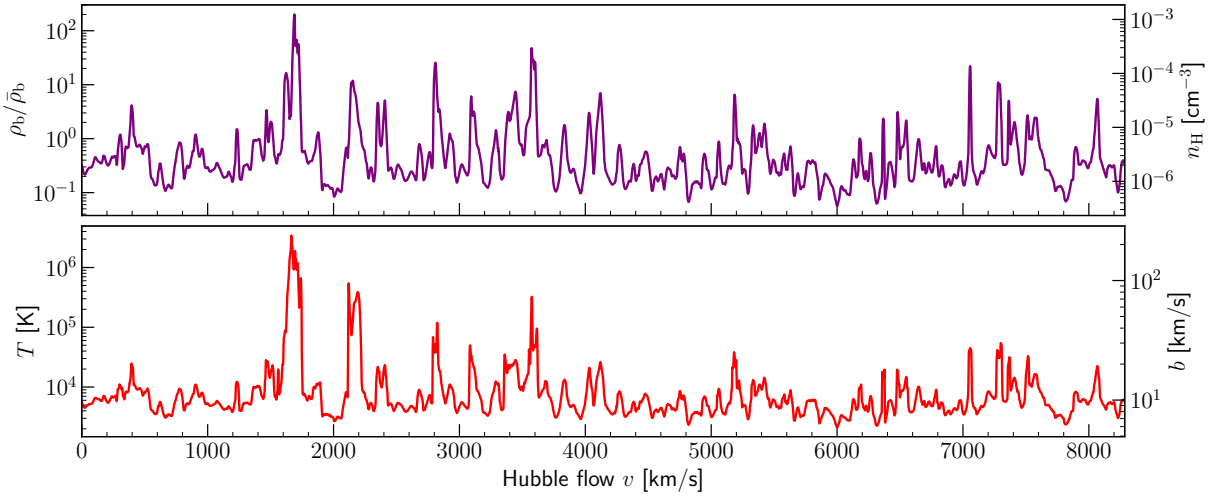

Figure 3.2: A randomly chosen skewer through the simulation at $z \sim 2.2$. **Top**: baryon density in units of cosmic mean baryon density and the number density of H. **Bottom**: temperature of the gas and the Doppler broadening parameter b.

$$\Gamma_{e,HI} n_e n_{HI} + \Gamma_{\gamma,HI} n_{HI} = \alpha_{r,HII} n_e n_{HII}, \tag{3.3}$$

$$\Gamma_{e,HeI} n_e n_{HeI} + \Gamma_{\gamma,HeI} n_{HeI} = (\alpha_{r,HeII} + \alpha_{d,HeII}) n_e n_{HeII}, \tag{3.4}$$

$$\Gamma_{e,HeII} n_e n_{HeII} + \Gamma_{\gamma,HeII} n_{HeII} = \alpha_{r,HeIII} n_e n_{HeIII}, \tag{3.5}$$

where, for a given species S, $\Gamma_{e,S}$ is the collisional ionization coefficient, $\Gamma_{\gamma,S}$ is the photoionization coefficient (sometimes called the UV background), and $\alpha_{r,S}$ and $\alpha_{d,S}$ are the radiative and dielectronic recombination coefficients. The left-hand side of each of the above equations constitutes the ionization processes and the right-hand side the recombination processes.

We also have the closure and charge conservation equations:

$$n_{HII} = A - n_{HI}, \tag{3.6}$$

$$n_{HeIII} = \frac{B}{2} - n_{HeI} - n_{HeII}, \tag{3.7}$$

$$n_e = n_{HII} + n_{HeII} + 2n_{HeIII}, \tag{3.8}$$

where $A := \frac{\rho_b X}{m_H}$ $(= n_H)$ and $B := \frac{\rho_b Y}{2m_H}$ $(= 2n_{He})$.

An example skewer through the simulation is shown in Figure 3.2.

In practice, solving the full set of coupled ionization equations is computationally expensive, especially if we need to synthesize mock spectra in bulk, for instance, for estimating the model power spectrum or for training a deep learning machinery. However, in different

redshift regimes, we may be able to rely on some fast approximate methods for the estimation of $n_{\rm H\,I}$. I describe two such methods below and compare their results against a full solution of the coupled equations in Section 3.4.

3.2.1 A low redshift approximation

At low cosmological redshifts ($z < 2.5$), it is safe to assume that the second Helium reionization episode is over and hence Helium is almost entirely doubly-ionized, i.e., $x_{\rm He\,III} \approx 1$ (e.g., Gaikwad et al. 2025). This greatly simplifies our analytical calculations and leaves us with a single quadratic equation in $n_{\rm H\,I}$ that is much faster to solve. If we call $y \equiv n_{\rm H\,I}$, then the simplified quadratic equation is

$$ay^2 + by + c = 0, \tag{3.9}$$

where

$$a = \Gamma_{\rm e,H\,I} + \alpha_{\rm r,H\,II}, \tag{3.10}$$
$$b = -\big(\Gamma_{\rm e,H\,I}(A+B) + \alpha_{\rm r,H\,II}(2A+B) + \Gamma_{\gamma,\rm H\,I}\big), \tag{3.11}$$
$$c = \alpha_{\rm r,H\,II}(A^2 + AB). \tag{3.12}$$

The two analytical solutions of this equations are then $y^{\pm} = \frac{-b \pm \sqrt{b^2-4ac}}{2a}$, however, due to certain physical constraints on $n_{\rm H\,I}$ (e.g., $n_{\rm H\,I} > 0$ and $n_{\rm H\,I}/n_{\rm H} < 1$), only one of them ($y^-$) is of interest to us. Thus, we can uniquely estimate the neutral H density by simply solving a quadratic equation in a single variable. Henceforth I will refer to this method as the "approximate" method. We would expect this solution to become increasingly inaccurate as we move to higher redshifts, however, since the underlying assumption ($x_{\rm He\,III} \approx 1$) may no longer be valid.

3.2.2 Intermediate redshifts: an iterative approach

At redshifts $z \gtrsim 3$ it becomes more important to consider all three ionization equilibrium equations since the second Helium reionization is still underway. Some assumptions may still hold, such as $x_{\rm H\,I,He\,I} \approx 0$, that results in simplification of the coupled equations and an approximate estimation of $n_{\rm e}$ and $x_{\rm He\,II,He\,III}$ that can then inform the estimation of $x_{\rm H\,I}$. We may iteratively correct this initial estimate of the neutral fraction, dropping the initial assumptions. I describe this approach in detail below.

For the sake of brevity, I first relabel quantities. The different species are assigned numbers respectively as $\{{\rm e^-}, {\rm H\,I}, {\rm H\,II}, {\rm He\,I}, {\rm He\,II}, {\rm He\,III}\} \equiv \{0,1,...,5\}$. Then, the ionization and recombination rates for species S: $\Gamma_{\rm e,S} \to \Gamma_{0,s}$, $\Gamma_{\gamma,\rm S} \to \Gamma_{1,s}$, $\alpha_{\rm r,S} \to \alpha_{0,s}$, and $\alpha_{\rm d,S} \to \alpha_{1,s}$ $\forall s \in \{1,2,...,5\}$. In this subsection, I work with mass fractions instead of number densities, $x_{1,2} = n_{1,2}/n_{\rm H} = n_{1,2}/A$ and $x_{3,4,5} = n_{3,4,5}/n_{\rm He} = 2n_{3,4,5}/B$. Now we can rewrite the

ionization equilibrium equations as

$$(\Gamma_{0,1}n_0 + \Gamma_{1,1})x_1 - \alpha_{0,2}n_0x_2 = 0, \tag{3.13}$$
$$(\Gamma_{0,3}n_0 + \Gamma_{1,3})x_3 - (\alpha_{0,4} + \alpha_{1,4})n_0x_4 = 0, \tag{3.14}$$
$$\big[\Gamma_{1,4} + n_0(\Gamma_{0,4} + \alpha_{0,4} + \alpha_{1,4})\big]x_4 - (\Gamma_{0,3}n_0 + \Gamma_{1,3})x_3 - \alpha_{0,5}n_0x_5 = 0, \tag{3.15}$$

where the last equation combines the HeI and HeII ionization processes. The closure and charge conservation equations are

$$x_1 + x_2 = 1, \tag{3.16}$$
$$x_3 + x_4 + x_5 = 1, \tag{3.17}$$
$$n_0 = A(1 - x_1) + B(1 - x_3 - x_4/2), \tag{3.18}$$

where in Eq. (3.18), we have already used Eqs. (3.16) and (3.17).

As the first approximation, we assume that the gas is almost completely ionized, i.e., the neutral fractions are negligible,

$$x_1, x_3 \approx 0. \tag{3.19}$$

Under this assumption, Eq. (3.15) leads, in combination with Eqs. (3.16)-(3.18), to a quadratic equation in x_4 of the form

$$ax_4^2 + bx_4 + c = 0, \tag{3.20}$$

where,

$$a = B(\Gamma_{0,4} + \alpha_{0,4} + \alpha_{1,4} + \alpha_{0,5})/2, \tag{3.21}$$
$$b = -\big[\Gamma_{1,4} + (A + 3B/2)\alpha_{0,5} + (A + B)(\Gamma_{0,4} + \alpha_{0,4} + \alpha_{1,4})\big], \tag{3.22}$$
$$c = (A + B)\alpha_{0,5}. \tag{3.23}$$

The next steps are:

- Solve Eq. (3.20) and adopt the more physically sensible of the two roots as x_4 (e.g., a constraint such as $0 \leq x_4 \leq 1$).
- Using that, solve Eq. (3.18) for n_0 under the assumption in Eq. (3.19).
- Using that, solve Eq. (3.13) in combination with Eq. (3.16) for x_1,

$$x_1 = \frac{\alpha_{0,2}n_0}{\Gamma_{1,1} + (\alpha_{0,2} + \Gamma_{0,1})n_0}. \tag{3.24}$$

The steps until this point can be called the naive first iteration. This first iteration can, however, be refined by the next 3 steps:

- Solve Eq. (3.14) for x_3 using n_0 computed in the naive first iteration.

- Recalculate n_0 from Eq. (3.18) without any assumptions.
- Thence recalculate x_1 from Eq. (3.13) in combination with Eq. (3.16).

This concludes the full first iteration of this method ($N_{\mathrm{iter}} = 1$). Further iterations with the following steps can be made as necessary:

- Solve Eq. (3.15) in combination with Eqs. (3.14) and (3.17) for x_4 as

$$x_4 = \frac{\alpha_{0,5} n_0 x_5}{(\Gamma_{0,4} n_0 + \Gamma_{1,4})}. \tag{3.25}$$

- Find x_3 using Eq. (3.14) from all the rest of the existing values of the quantities.
- Recalculate n_0 from Eq. (3.18) without any assumptions.
- Recalculate x_1 from Eq. (3.13) in combination with Eq. (3.16).

This iterative method is still computationally efficient since we are at most only solving a quadratic equation in a single variable at a time. A full solution, on the contrary, constitutes three coupled quadratic equations in three independent variables. For the sake of completeness, I show those full equations below.

3.2.3 Full set of equations

We may use the closure and charge conservation equations to reduce the full set of ionization equations to three polynomials in as many independent variables. I shall reuse the shorthand notations from Section 3.2.2, however, in number densities n_s instead of fractions x_s.

First ionization equilibrium equation

$$\begin{aligned}
&\Gamma_{0,1}(A + B - n_1 - 2n_3 - n_4)n_1 + \Gamma_{1,1} n_1 \\
&\quad - \alpha_{0,2}(A + B - n_1 - 2n_3 - n_4)(A - n_1) = 0. \\
\Rightarrow\quad &n_1^2(\Gamma_{0,1} + \alpha_{0,2}) + 2n_1 n_3(\Gamma_{0,1} + \alpha_{0,2}) + n_1 n_4(\Gamma_{0,1} + \alpha_{0,2}) \\
&\quad - n_1\{(\Gamma_{0,1} + \alpha_{0,2})(A + B) + \alpha_{0,2} A + \Gamma_{1,1}\} \\
&\qquad - 2n_3\alpha_{0,2}A - n_4\alpha_{0,2}A + \alpha_{0,2}A(A + B) = 0.
\end{aligned}$$

This is a second-order polynomial in $\mathbf{n} = (n_1, n_3, n_4)$ that can alternatively be expressed as

$$\mathcal{P}(\mathbf{n}) := \mathbf{n}\mathbf{P}\mathbf{n}^T + \mathbf{p}\mathbf{n}^T + p = 0, \tag{3.26}$$

where

$$\mathbf{P}_{3\times3} = \begin{bmatrix} \Gamma_{0,1}+\alpha_{0,2} & \Gamma_{0,1}+\alpha_{0,2} & (\Gamma_{0,1}+\alpha_{0,2})/2 \\ & 0 & 0 \\ \text{sym.} & & 0 \end{bmatrix},$$

$$-\mathbf{p}^T_{1\times3} = \begin{bmatrix} (\Gamma_{0,1}+\alpha_{0,2})(A+B)+\alpha_{0,2}A+\Gamma_{1,1} \\ 2\alpha_{0,2}A \\ \alpha_{0,2}A \end{bmatrix},$$

and

$$p = \alpha_{0,2}A(A+B).$$

Second ionization equilibrium equation

$$\begin{aligned} &\Gamma_{0,3}(A+B-n_1-2n_3-n_4)n_3+\Gamma_{1,3}n_3 \\ &\quad -(\alpha_{0,4}+\alpha_{1,4})(A+B-n_1-2n_3-n_4)n_4 = 0. \\ \Rightarrow\quad &-n_1n_3\Gamma_{0,3}+n_1n_4(\alpha_{0,4}+\alpha_{1,4})-2n_3^2\Gamma_{0,3} \\ &\quad +n_3n_4\{2(\alpha_{0,4}+\alpha_{1,4})-\Gamma_{0,3}\}+n_4^2(\alpha_{0,4}+\alpha_{1,4}) \\ &\qquad +n_3\{\Gamma_{0,3}(A+B)+\Gamma_{1,3}\}-n_4(\alpha_{0,4}+\alpha_{1,4})(A+B)=0. \end{aligned}$$

This is another second-order polynomial in $\mathbf{n}$ that can alternatively be expressed as

$$\mathcal{Q}(\mathbf{n}) := \mathbf{n}\mathbf{Q}\mathbf{n}^T+\mathbf{q}\mathbf{n}^T+q=0, \tag{3.27}$$

where

$$\mathbf{Q}_{3\times3} = \begin{bmatrix} 0 & -\Gamma_{0,3} & \alpha_{0,4}+\alpha_{1,4} \\ & -2\Gamma_{0,3} & 2(\alpha_{0,4}+\alpha_{1,4})-\Gamma_{0,3} \\ \text{sym.} & & \alpha_{0,4}+\alpha_{1,4} \end{bmatrix},$$

$$\mathbf{q}^T_{1\times3} = \begin{bmatrix} 0 \\ \Gamma_{0,3}(A+B)+\Gamma_{1,3} \\ -(\alpha_{0,4}+\alpha_{1,4})(A+B) \end{bmatrix}, \quad \text{and} \quad q=0.$$

Third ionization equilibrium equation

$$\begin{aligned} &\Gamma_{1,4}n_4+\Delta_4n_4(A+B-n_1-2n_3-n_4) \\ &\quad -\alpha_{0,5}(A+B-n_1-2n_3-n_4)\left(\frac{B}{2}-n_3-n_4\right) \\ &\qquad -\Gamma_{0,3}n_3(A+B-n_1-2n_3-n_4)-\Gamma_{1,3}n_3=0. \\ \Rightarrow\quad &2n_3^2(\Gamma_{0,3}-\alpha_{0,5})-n_4^2(\Delta_4+\alpha_{0,5})+n_1n_3(\Gamma_{0,3}-\alpha_{0,5})-n_1n_4(\Delta_4+\alpha_{0,5}) \\ &\quad +n_3n_4(\Gamma_{0,3}-2\Delta_4-3\alpha_{0,5})+n_3\{\alpha_{0,5}(A+2B)-\Gamma_{0,3}(A+B)-\Gamma_{1,3}\} \\ &\qquad +n_1\alpha_{0,5}B/2+n_4\{\Gamma_{1,4}+\Delta_4(A+B)+\alpha_{0,5}(A+3B/2)\}-\alpha_{0,5}B(A+B)/2=0. \end{aligned}$$

This is the final second-order polynomial in $\mathbf{n}$ that can alternatively be expressed as

$$\mathcal{R}(\mathbf{n}) := \mathbf{n}\mathbf{R}\mathbf{n}^T + \mathbf{r}\mathbf{n}^T + r = 0, \tag{3.28}$$

where

$$\mathbf{R}_{3\times3} = \begin{bmatrix} 0 & \Gamma_{0,3} - \alpha_{0,5} & -(\Delta_4 + \alpha_{0,5}) \\ & 2(\Gamma_{0,3} - \alpha_{0,5}) & \Gamma_{0,3} - 2\Delta_4 - 3\alpha_{0,5} \\ \text{sym.} & & -(\Delta_4 + \alpha_{0,5}) \end{bmatrix},$$

$$\mathbf{r}^T_{1\times3} = \begin{bmatrix} \alpha_{0,5}B/2 \\ \alpha_{0,5}(A + 2B) - \Gamma_{0,3}(A + B) - \Gamma_{1,3} \\ \alpha_{0,5}(A + 3B/2) + \Delta_4(A + B) + \Gamma_{1,4} \end{bmatrix},$$

and

$$r = -\alpha_{0,5}B(A + B)/2.$$

In SynTH, I solve these coupled equations numerically with `scipy.optimize.root` routine of Python. This additionally requires the Jacobian of the set of functions involving partial derivatives $\partial_i[\cdot] := \partial[\cdot]/\partial n_i$,

$$\mathbb{J} = \begin{bmatrix} \partial_1\mathcal{P} & \partial_3\mathcal{P} & \partial_4\mathcal{P} \\ \partial_1\mathcal{Q} & \partial_3\mathcal{Q} & \partial_4\mathcal{Q} \\ \partial_1\mathcal{R} & \partial_3\mathcal{R} & \partial_4\mathcal{R} \end{bmatrix},$$

for which we may find an analytical closed-form expression using Eqs. (3.26)-(3.28).

3.3 Determination of $\tau_{\mathrm{Ly}\alpha}$

The second major step in the computation of τ for a mock spectrum is to use local temperatures and peculiar velocities along the corresponding skewer to properly inform the absorption line profiles. As described in Section 1.6, absorption lines are broadened by two primary effects—intrinsic quantum mechanical (pressure) broadening and thermal (Doppler) broadening—that result in a Voigt profile, which is the convolution of the Doppler and Lorentz profiles. The intrinsic Lorentz broadening width (half width at half maximum or HWHM) in units of velocity is given by

$$\gamma_{\mathrm{L},v} = \frac{2\pi q_\mathrm{e}^2 f_{\ell u}}{3m_\mathrm{e}c\lambda_0} = 6.055 \times 10^{-3}\ \mathrm{km/s}, \tag{3.29}$$

whereas the HWHM of the Doppler profile is

$$b\sqrt{\log 2} = \sqrt{\frac{2k_\mathrm{B}T\log 2}{m_\mathrm{H}}} = 12.853\,\sqrt{T_4}\ \mathrm{km/s}, \tag{3.30}$$

where $T_4 = T/(10^4\mathrm{K})$ is typically of order unity in the diffuse IGM that gives rise to the Lyα forest. Clearly, Doppler broadening is the dominant broadening effect of the

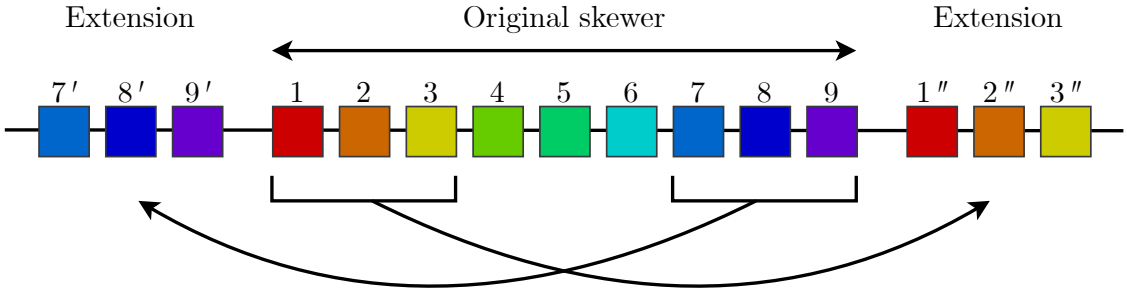

Figure 3.3: An illustration of the partial extension of a skewer necessary for an adequate estimation of the line profile, especially at the edge pixels. This is possible due to the skewers' periodic boundary conditions. This process may be likened to notes on a piano. In this toy model, the original skewer is 9 voxels long (equivalent to an octave on the keyboard) and is extended on each end by $N_{\rm ext} = 3$ voxels from the opposite end (keys from the following and the preceding octaves in the musical analogy). The estimation of the line profile (and thus τ) would benefit the most at pixels corresponding to voxels 1 and 9, and it would benefit the least at the pixel corresponding to voxel 5 (in a hypothetical scenario with no peculiar motion of the gas).

two. Only when the column density is quite large, $N_{\rm H\,I} \sim \mathcal{O}(10^{19}{\rm cm}^{-2})$, do the wings of the Lorentz profile start showing up in the absorption lines. Examples of such high column density systems (HCDs) are galaxies or proto-galaxies along the line of sight. The absorption in the damping wings extends out to much larger Δv from the central Hubble velocity of the absorber and overlaps heavily with the Lyα forest signal over the surrounding wavelengths. For studies involving the forest, such damped Lyα systems (DLAs) are undesirable contaminants in our analyses, and hence we would like to omit any spectra that contain them. Additionally, our simulations do not possess high enough fidelity to model galaxy formation, and therefore the baryon overdensities $\rho_{\rm b}/\bar{\rho}_{\rm b}$ are artificially clipped to 10^3 before estimating local $n_{\rm H\,I}$ to avoid inaccurately simulating HCDs. This does not introduce any errors since the Lyα transmission $F \sim 0$ in the regime of $\rho_{\rm b}/\bar{\rho}_{\rm b} \sim 10^3$ and is thus insensitive to any density fluctuations.

For every pixel in the output spectrum, the integral in Eq. (3.1)—and consequently the sum in Eq. (3.2)—runs in principle along the entire line of sight, i.e., from the observer to the source. However, depending on the local broadening widths and line center shifts due to peculiar velocities, the integrand is expected to be negligible far outside the line core and the integral may be truncated at sufficiently large Δv from the pixel(s). An issue arises at the pixels corresponding to the skewer boundaries due to a lack of voxels on one side of the line center. Since the skewers follow periodic boundary conditions, however, at one end of the skewer we may append an arbitrary number of voxels from the opposite end, effectively extending the skewer to be able to properly estimate the line profile at the edge pixels. This process is illustrated in Figure 3.3. The number of voxel in the extension, $N_{\rm ext}$, depends on the peculiar velocities and temperatures toward the edges of the skewers. I examined the effect of this hyperparameter on the line profile accuracy; the results are shown in Section 3.4.

The bulk of the IGM at low redshifts $z \sim 2$ that are involved in our studies of deep learning inference (Chapters 4 and 5) is optically thin to Lyα radiation and the absorption line profiles can be reasonably well approximated by the Doppler profile. This is especially useful since the evaluation of the full Voigt profile is computationally expensive. In this work, I investigated the impact of such approximations while synthesizing mock Lyα forest spectra. In particular, I implemented the full Doppler profile as well as a version of it involving the error function as used by Lukić et al. (2015), and made comparisons with a full Voigt profile.

Error function version of the Doppler profile

Rather than approximating the profile as $\phi_j(v_i)$ for voxel j in the discretized version, we may break the integral in Eq. (3.1) down into a piecewise integral over the voxels, i.e.,

$$\int_{\mathrm{LoS}} n_{\mathrm{H\,I}}(v')\phi'(v)\,dv' \equiv \sum_j n_{\mathrm{H\,I},j} \int_{j-1/2}^{j+1/2} \phi'(v)\,dv', \tag{3.31}$$

where $j \pm 1/2$ denote the two edges of voxel j along the line of sight, and the local $n_{\mathrm{H\,I}}$ is fixed over the extent of the voxel (so are all other hydrodynamic fields in the simulation, such as temperature and peculiar velocity). If we further assume a Doppler profile, this expression reduces to error functions. Applying a change of variables $y' = (v - v')/b$,

$$\begin{aligned}
\sum_j \int_{j-1/2}^{j+1/2} \phi'_{\mathrm{D}}(v)\,dv' &= \frac{1}{b\sqrt{\pi}} \sum_j \int_{j-1/2}^{j+1/2} e^{-(v-v')^2/b^2}\,dv' \\
&= \frac{1}{\sqrt{\pi}} \sum_j \int_{j+1/2}^{j-1/2} e^{-y'^2}\,dy' \\
&= \frac{1}{\sqrt{\pi}} \sum_j \left[\int_0^{j-1/2} e^{-y'^2}\,dy' - \int_0^{j+1/2} e^{-y'^2}\,dy'\right] \\
&= \frac{1}{2} \sum_j \left[\mathrm{erf}(y_{j-1/2}) - \mathrm{erf}(y_{j+1/2})\right],
\end{aligned}$$

where the error function is

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\,dt,$$

and $y_{j\pm1/2} = (v_i - v'_{j\pm1/2} - v_{||,j})/b_j$. Since we estimated the full integral (in a piecewise manner) here, this is a slightly better approximation of the Doppler profile for the computation of τ than Eq. (3.2). However, it involves two evaluations of the error function as opposed to one evaluation of the Gaussian function, and may add additional computational load.

3.3.1 Rescaling τ to adapt the mean transmission

The mean Lyα transmission $\bar{F}$ in a set of spectra is one of the most important statistical measures of the Lyα forest. An effective optical depth corresponding to this, $\tau_{\rm eff} := -\log \bar{F}$, is a quantification of the overall optical depth of the IGM. Its value is set by observations of a large number of quasars (e.g., Becker et al. 2013) and it is imperative for our mocks to reproduce this value as a zeroth order check. For optically thin gas in photoionization equilibrium (for which the contribution of collisional ionization can be ignored), the fluctuating Gunn-Peterson approximation (FGPA; see, e.g., Croft et al. 1998) relates the optical depth to the intergalactic HI UV background Γ (shorthand for $\Gamma_{\gamma,\rm HI}$) as

$$\tau \propto n_{\rm HI} \propto \frac{\rho_{\rm b}^2 T^{-0.7}}{\Gamma}. \tag{3.32}$$

The value of Γ is very poorly constrained, however, and the underlying model for the UV background may lead to a slightly different optical depth $\tau_{\rm eff}$ than its observed value. A common practice is to apply a constant factor to scale all the τ values in the mocks to correct for any mismatch *post facto*. For a set of spectra, we would like to estimate a correction factor A such that

$$\langle F \rangle = \langle e^{-A\tau} \rangle = \bar{F}_{\rm obs},$$

where the average is taken over the entire given set of spectra. The UV background may then be adjusted to $\sim \Gamma/A$. We can estimate the factor A by solving for $f(A^*) = 0$, where $f(A) := \langle e^{-A\tau} \rangle - \bar{F}_{\rm obs}$. In my implementation in SynTH, I achieve this using Newton's root finding method (`scipy.optimize.newton`).

3.4 Benchmark tests

Table 3.1: Computation time (in ms) for estimation of $n_{\rm HI}$ on a single skewer using three different methods: (i) approximate, (ii) iterative, and (iii) full. The numbers in this table correspond to estimates of a single evaluation, $t_{\rm l}/N_{\rm l}$, with $N_{\rm l} = 100$ and $N_{\rm r} = 100$ (see main text for an explanation). This was performed on a single CPU with 32 GB of random-access memory.

Method	Best	Average	St.dev.
Approximate	0.272	0.274	0.002
Iterative			
$N_{\rm iter} = 1$	0.903	0.910	0.005
$N_{\rm iter} = 2$	0.975	0.982	0.005
$N_{\rm iter} = 3$	1.045	1.052	0.005
$N_{\rm iter} = 4$	1.115	1.123	0.007
Full	470.867	473.483	1.832

Using SynTH, I performed a set of extensive tests of the efficiency and robustness of the different (approximate) methods for the estimation of local $n_{\rm HI}$ and spectral τ. In

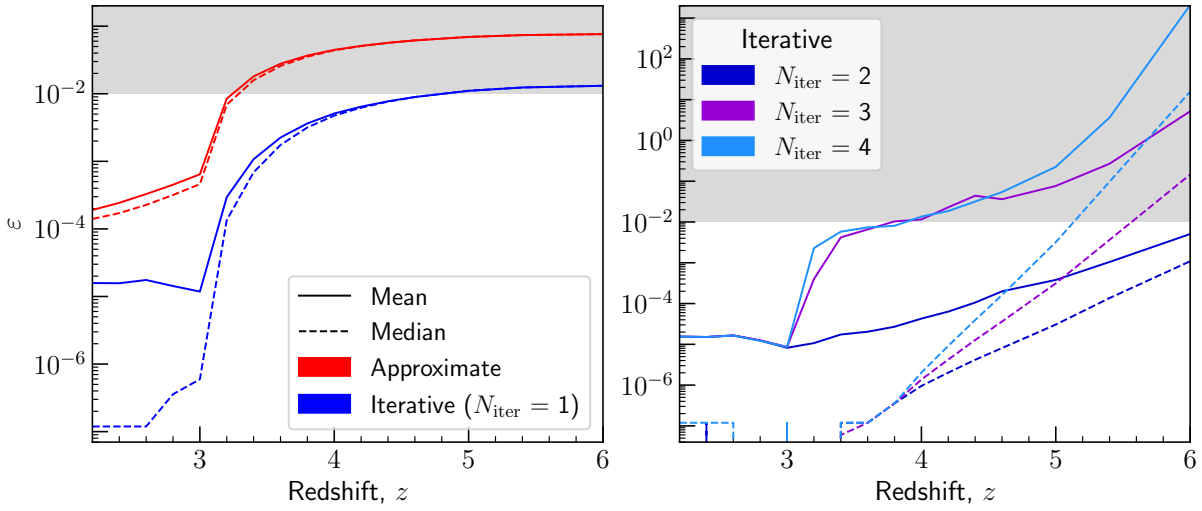

Figure 3.4: Relative error ε in $n_{\rm H\,I}$ *w.r.t.* the full solution across snapshot redshifts for the test set of skewers. The gray-shaded region indicates $\varepsilon > 1\%$. In all cases shown here, the relative errors increase with redshift. For the iterative method with $N_{\rm iter} = 2$, the mean ε stays under 1% at all redshifts.

Table 3.2: Computation time (in ms) for estimation of τ for a single spectrum of length 4096 pixels (with precomputed $n_{\rm H\,I}$ and fixed $f_{\rm ext} = 0.1$) using three different profile implementations: (i) discrete Doppler, (ii) error function, and (iii) discrete Voigt. The numbers in this table correspond to estimates of a single evaluation, $t_{\rm l}/N_{\rm l}$, with $N_{\rm l} = 100$ and $N_{\rm r} = 100$ (see main text for an explanation). This was performed on a single CPU with 32 GB of random-access memory, the same as for Table 3.1.

Profile	Best	Average	St.dev.
Discrete Doppler	260.017	264.530	11.431
Error function	539.054	561.970	86.469
Discrete Voigt	682.593	696.048	57.925

particular, I made comparisons of the approximate and iterative methods (with a range of $N_{\rm iter}$) with the full solution for the computation of $n_{\rm H\,I}$, and of the three line profile cases with each other for the computation of τ, namely the discrete Doppler profile, the error function version of the Doppler profile, and the discrete Voigt profile. Additionally, I analyzed the importance of the skewer extension by varying the hyperparameter $N_{\rm ext}$ and quantifying its impact on the accuracy of τ estimation. For each of the above analyses, I ran benchmarks of the computation time of the different methods and their respective error(s) in the fields $n_{\rm H\,I}$ and τ as well as in the 1D power spectrum of the Lyα forest, $P(k)$. I used a set of 1000 randomly chosen skewers (along a fixed Cartesian axis of the box) through the simulations here and I henceforth collectively call them the "test set." (The physical location of the skewers was kept fixed across snapshot redshifts.) The results of this comprehensive study provide the key basis for selecting the optimal method(s) to produce mock spectra in bulk, directly impacting the reliability of neural network inference. I present those results in the following.

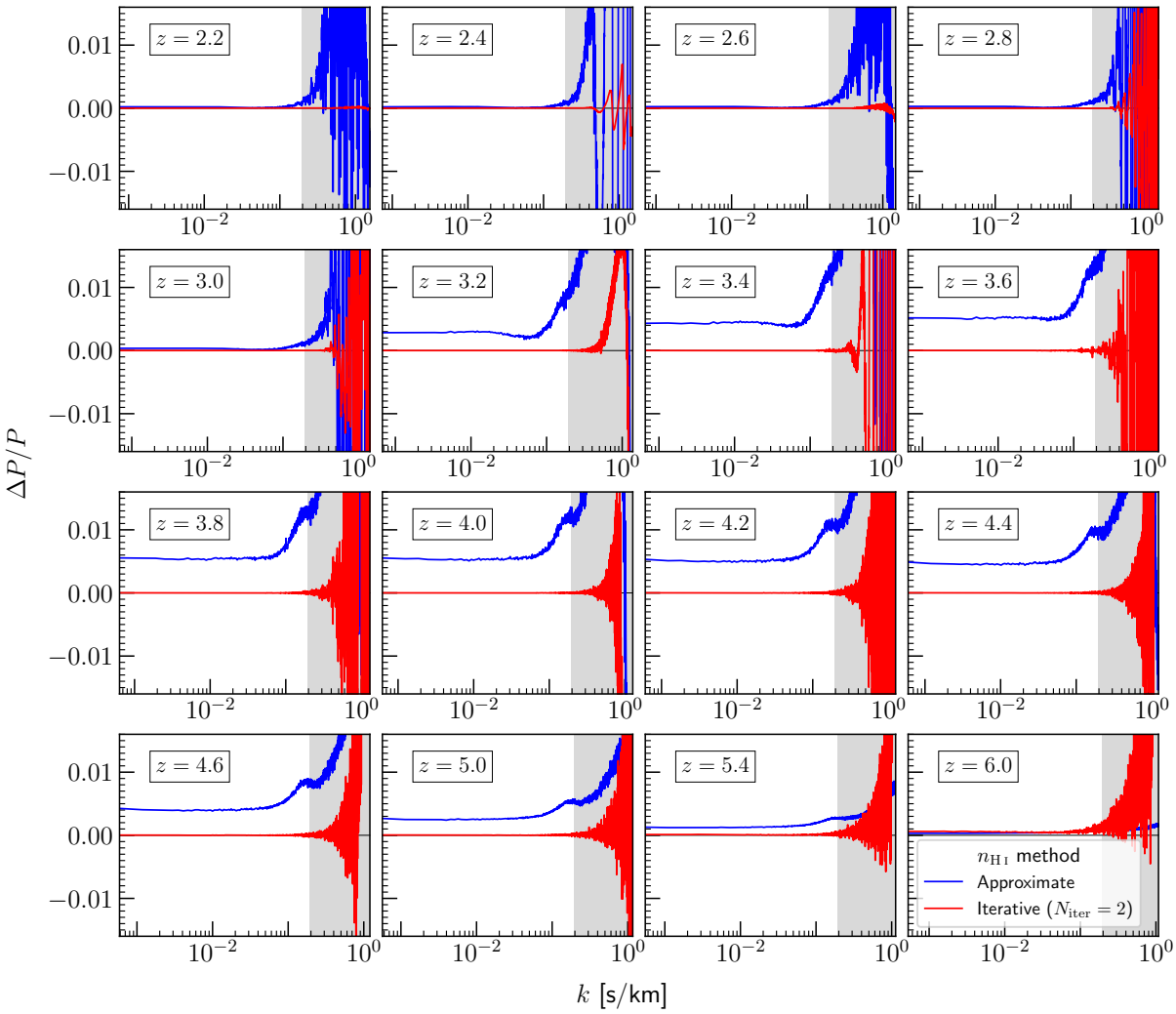

Figure 3.5: A power spectrum comparison at all available snapshot redshifts of the approximate and the iterative methods of $n_{\mathrm{H\,I}}$ estimation *w.r.t.* the full solution for the test set of skewers. For the iterative method, only the best case of $N_{\mathrm{iter}} = 2$ is shown. For all of them, the error function profile was used to compute $F = e^{-\tau}$ and the τ values were rescaled to match the mean transmission to the observed value. The gray-shaded area indicates $k > 0.195$ s/km, a realistic cutoff on the Fourier modes.

$n_{\mathrm{H\,I}}$ estimation

One of the main bottlenecks to the estimation of $n_{\mathrm{H\,I}}$ is the computation time. Especially, the full solution involving three coupled matrix polynomials is anticipated to be quite slow. I used the PYTHON routine `timeit` to benchmark the run-time of the three $n_{\mathrm{H\,I}}$ estimation methods (approximate, iterative, and full) by repeatedly running evaluation on a *single* skewer. `timeit` provides two controllable hyperparameters, namely, N_{l} (loops) and N_{r} (repetitions). The total time t_{l} of running N_{l} iterations of the desired evaluation

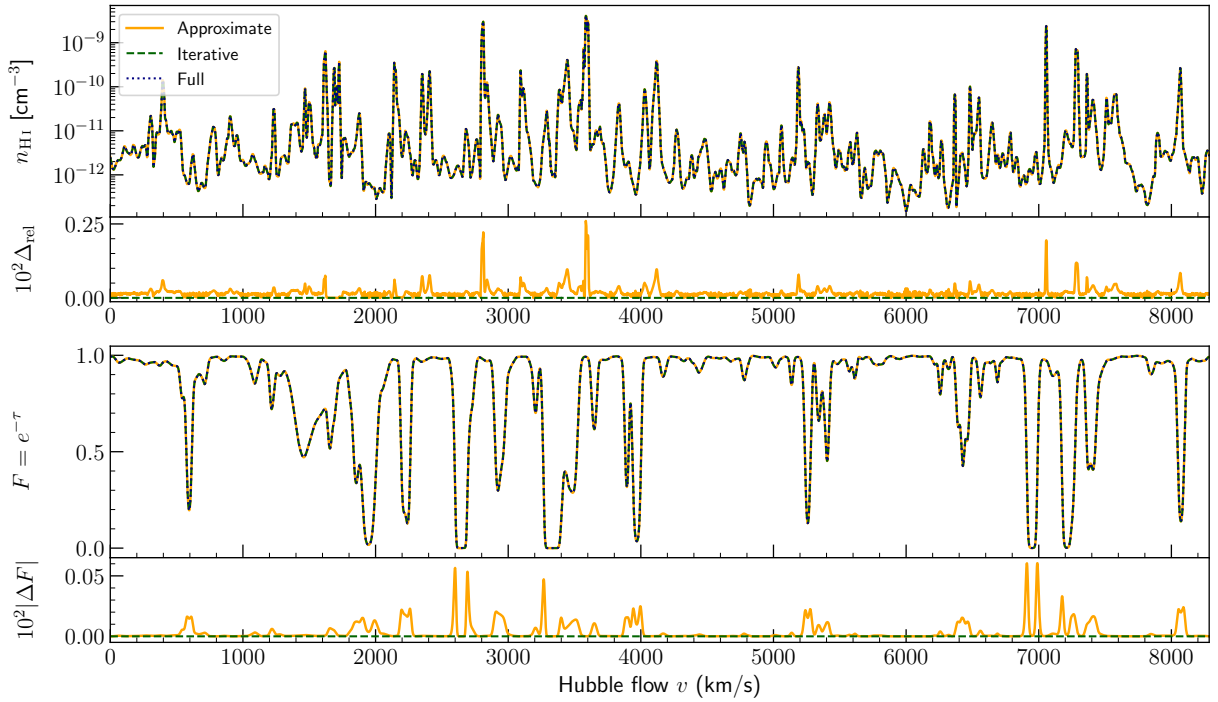

Figure 3.6: A comparison of the three different $n_{\rm HI}$ estimation methods at the field level on a randomly chosen skewer from the test set (the same as in Figure 3.2) at $z \sim 2.2$. The iterative case is with $N_{\rm iter} = 2$. The top panel(s) show $n_{\rm HI}$ along the skewer estimated with the three methods and their relative differences $\Delta_{\rm rel}$. The bottom panels show the Lyα transmission F for a spectrum corresponding to this skewer and the absolute differences $|\Delta F|$. For this skewer, $\Delta_{\rm rel} \leq 0.25\%$ and the largest $\Delta_{\rm rel}$ occur at voxels with $n_{\rm HI} > 10^{-9}{\rm cm}^{-3}$; the transmission F and the absolute differences are all close to zero for the corresponding pixels in the redshift space.

is measured and this process is repeated $N_{\rm r}$ times[1]. In Table 3.1 I show the statistics of those repetitions (it is noteworthy that no other processes were running parallelly on the same CPU used for these measurements). The approximate method is by far the fastest of them all, as expected. The iterative method with only one iteration is slower than the approximate method by a factor ~ 3.3, and gets slower with a rate ~ 0.07 ms per iteration per skewer. The full solution is significantly slower than the approximate and the iterative methods, by a factor ranging from 400 to 1700, underscoring the need for approximate methods in the first place.

I then ran $n_{\rm HI}$ evaluation on the test set. Figure 3.4 shows the relative error $\varepsilon := |n_{\rm HI} - n_{\rm HI,full}|/n_{\rm HI,full}$ across snapshot redshifts $z_{\rm box} \in [2.2, 6.0]$. For low redshifts $z \leq 3$, the approximate and the iterative methods agree with the full solution to within 0.1%. At higher redshifts ε becomes larger than 1% for the approximate solution. For the iterative method, we may make the following observations: (i) For one iteration ($N_{\rm iter} = 1$), the error remains close to or under 1% at all redshifts. (ii) One more iteration ($N_{\rm iter} = 2$)

[1]The best (smallest) value found by this repeated analysis is an estimate of the upper limit of the device's speed for the given computation and is the most reliable measure of the computation time.

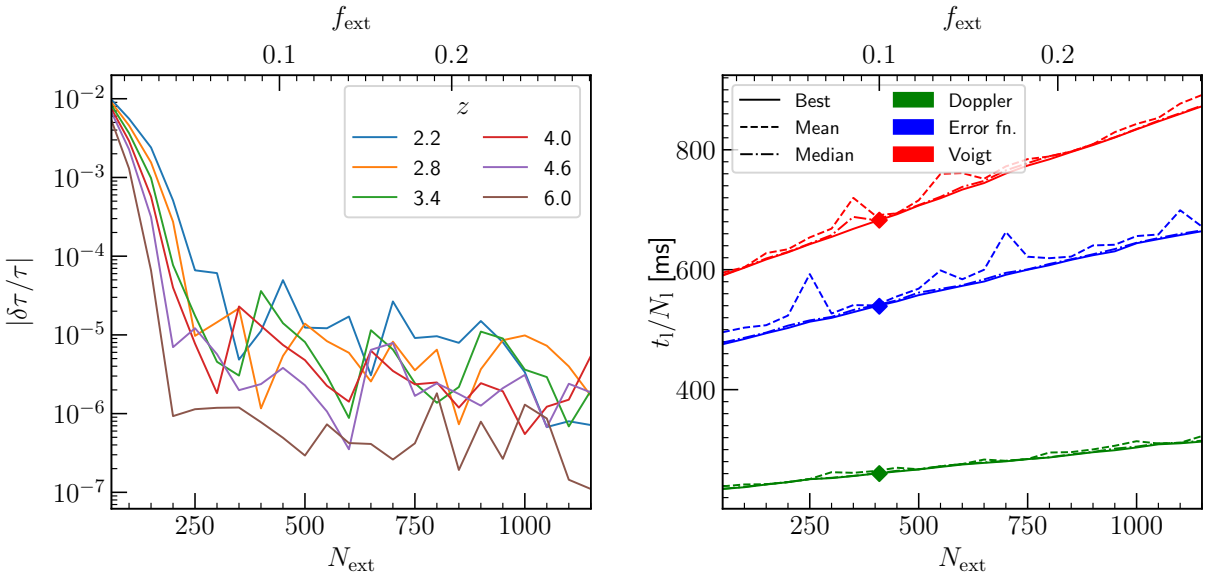

Figure 3.7: The impact of the hyperparameter $N_{\rm ext}$ (the number of voxels in the extension on each side) on the accuracy of τ estimation and the computation time. **Left**: the mean fractional change in τ *w.r.t.* $N_{\rm ext}$ for a handful of different snapshot redshifts. **Right**: the computation time per spectrum (with precomputed $n_{\rm H\,I}$) as a function of $N_{\rm ext}$ for all three line profile cases at $z \sim 2.2$.

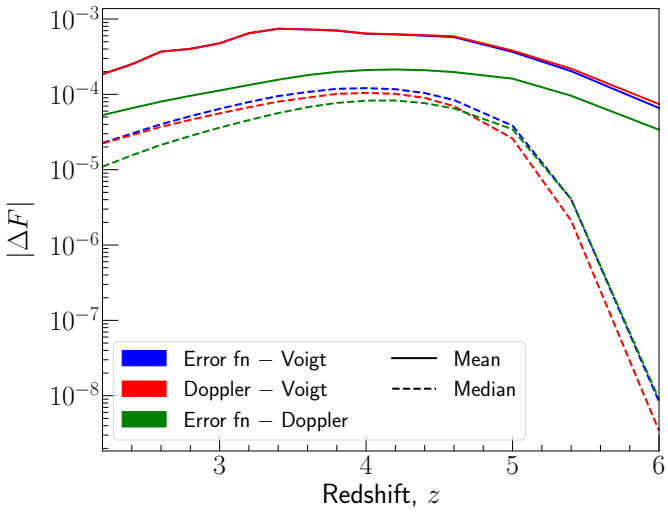

Figure 3.8: Absolute differences in transmission, $|\Delta F|$, in pairs of the profile cases described in the main text across snapshot redshifts in the test set. A gradual increase in $|\Delta F|$ from $z \sim 2$ through 4-5 and a drop thereafter can be observed for all three pairs.

helps reduce the error at high redshifts, bringing ε at all redshifts to within 1%. (iii) Further iterations ($N_{\rm iter} \geq 3$) increase the error at redshifts $z > 3$ dramatically, surpassing 100% around $z \sim 5$. A potential explanation to these observations is that the iterative method starts with an arguably more accurate assumption than the approximate method

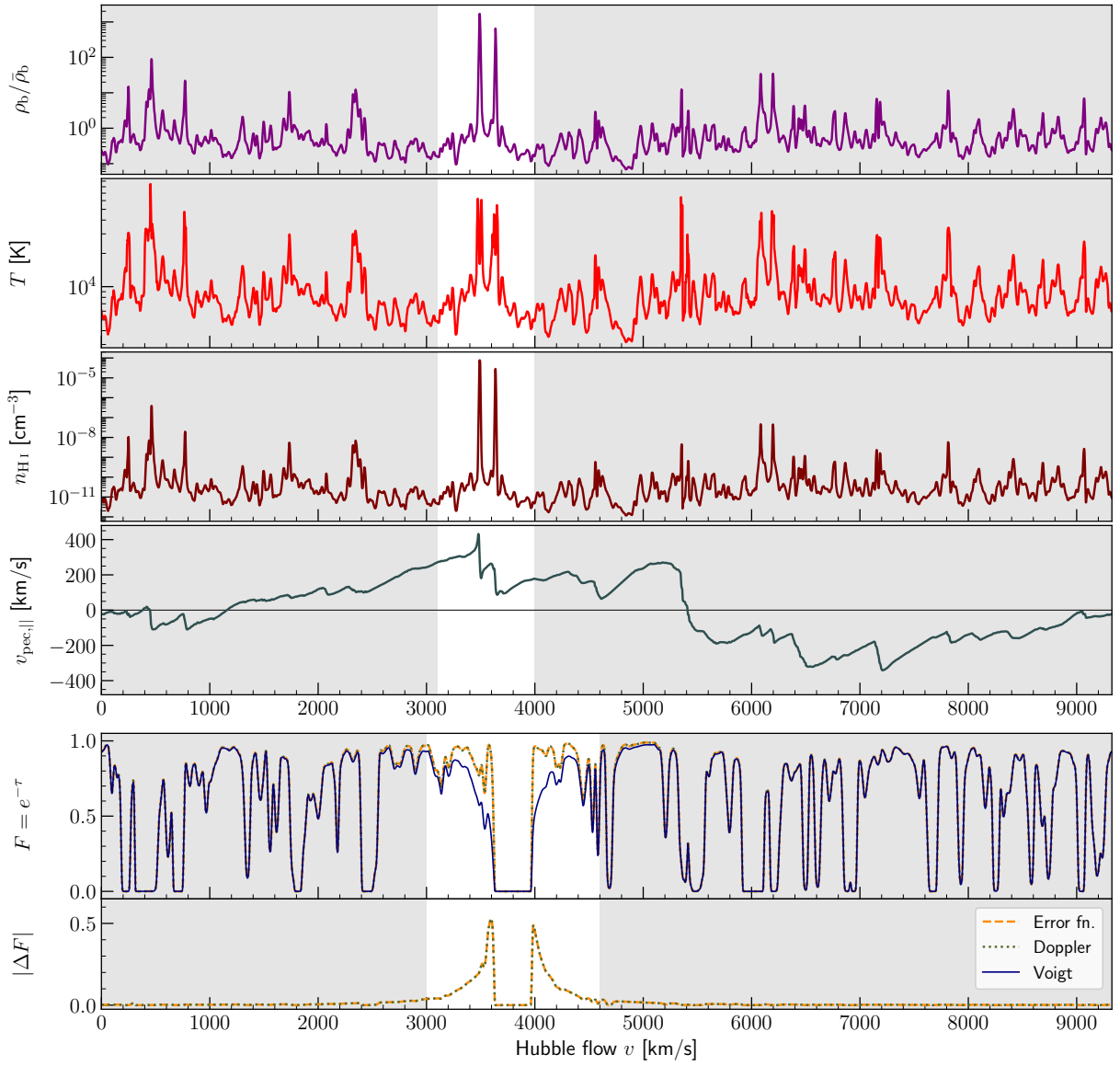

Figure 3.9: A skewer (from the test set) containing an HCD absorber system at $z \sim 3.2$, highlighted by the non-gray-shaded area (the width(s) of the highlighted regions are arbitrary). The top four panels show quantities in the real space along the skewer and the bottom two panels show transmission and differences thereof in the redshift space in the corresponding spectrum. A shift in the line center for this HCD system due to a high peculiar velocity may be noticed. The differences in F between the Voigt profile and the Doppler approximation(s) rise sharply from 0 as we transition from the saturated line core to the wings and then drops back to 0 in the wings over $\Delta v \sim 1500$ km/s on each side.

and hence results in an increasingly less erroneous estimation of $n_{\mathrm{H\,I}}$ over the first two iterations, especially at high redshifts. However, the error made in each iteration by making inaccurate approximations starts to accumulate, potentially causing the overall error for $N_{\mathrm{iter}} \geq 3$ to rise.

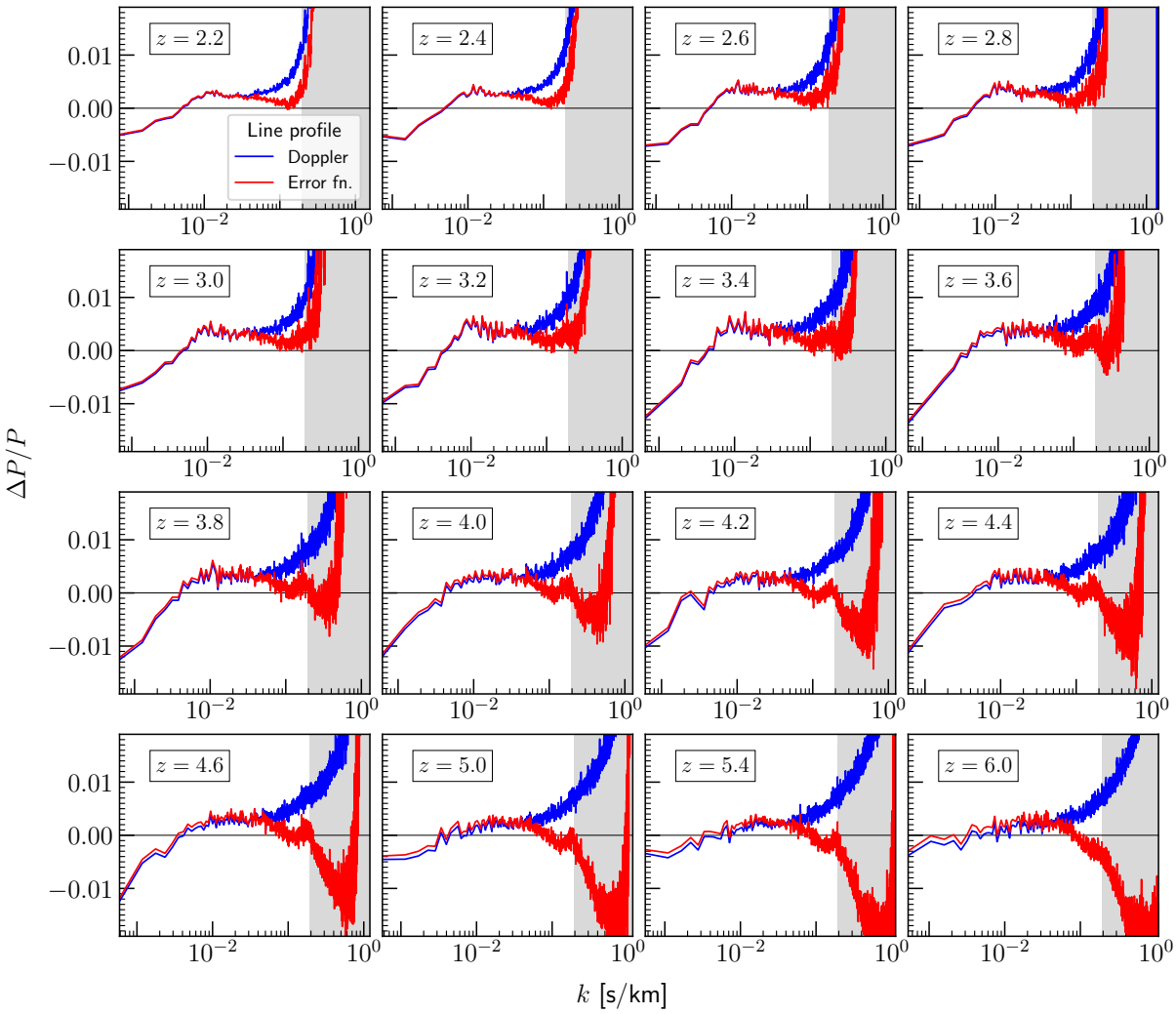

Figure 3.10: A power spectrum comparison at all available snapshot redshifts of the discrete Doppler profile and the error function version thereof *w.r.t.* the discrete Voigt profile for the test set of skewers. For all of them, the full solution was used for $n_{\mathrm{H\,I}}$ estimation and the τ values were rescaled to match the mean transmission to the observed value. The gray-shaded areas indicate $k > 0.195$ s/km, a realistic limit on the power spectrum accuracy in our simulations.

Another indicative test of the accuracy of our mocks is a comparison of different methods of $n_{\mathrm{H\,I}}$ estimation at the level of 1D power spectrum, $P(k)$. I first used the error function profile to produce mock spectra corresponding to the test set of skewers for all available snapshot redshifts and then computed $P(k)$ (having rescaled all τ values to match the corresponding observed $\bar{F}$). Figure 3.5 shows a comparison of $P(k)$ among the three $n_{\mathrm{H\,I}}$ estimation methods at all redshifts. In the range $k \leq 0.195$ s/km, the relative differences in $P(k)$, $\Delta P/P$, for the approximate method *w.r.t.* the full solution are under 0.15% for $z \leq 3$. At higher redshifts those differences grow larger, still staying under 1.5% in the same k range. For $z \geq 5$, the differences in $P(k)$ get smaller again. The iterative

method with $N_{\rm iter} = 2$ agrees with the full solution at the $P(k)$ level to within 0.25% at all redshifts. At higher k values than 0.195 s/km (smaller scales) the power spectrum estimate becomes noisier due to numerical inaccuracies in the simulations and the three cases start to diverge. Furthermore, spectral resolution effects in real-world observations put a realistic constraint on the scales that can be accurately probed by the Lyα forest. Hence, in this work we shall disregard the scales $k > 0.19$ s/km.

In Figure 3.6 (top panel), I show $n_{\rm H\,I}$ along a randomly chosen skewer at $z \sim 2.2$ (the same skewer as in Figure 3.2) estimated with the three methods described above. Not only is the relative error ε under 1% at every voxel, but the largest errors along the skewer are made at voxels with $n_{\rm H\,I} > 10^{-9}{\rm cm}^{-3}$ (recall Eq. (1.14)) leading to a complete absorption in any case. The absolute error in transmission F (computed using the error function profile) is under 0.0006 for this particular skewer (bottom panel), indicating an excellent agreement with the full solution, and the largest errors do not correspond to the saturated lines ($F \sim 0$).

Based on these results, I make the following recommendations for the choice of method of $n_{\rm H\,I}$ estimation when producing mocks in bulk:

(i) The approximate method is the fastest and accurate to within 1% of the full solution up to $z \sim 3$, and hence, is an excellent choice in that redshift regime.

(ii) At higher redshifts $z > 3$, the iterative method with one or two iterations works quite well, and the difference in the computation time between the two is rather small. Hence, in this redshift range, the iterative method with $N_{\rm iter} = 2$ is preferable.

Skewer extension

As discussed in Section 3.3, it is necessary to extend the skewers exploiting their periodic boundary conditions in order to accurately estimate line profiles. The number of voxels needed to extend the skewer on each end, $N_{\rm ext}$, depends on the local temperatures, baryon densities and peculiar velocities. I investigated the impact of this parameter on the accuracy of τ estimation and the computation time. (By definition, $N_{\rm ext}$ is a comoving length measure.) I computed the optical depth τ spectra for a varying $N_{\rm ext}$ (alternatively $f_{\rm ext} := N_{\rm ext}/4096$, the fraction of the skewer length) corresponding to the test set of skewers, and measured the fractional change in τ *w.r.t.* $N_{\rm ext}$, $|\delta\tau/\tau|$ over the sampled values of $N_{\rm ext}$. This is expected to converge to an extremely small value after a sufficiently large extension of the skewers. For this investigation I used the full Voigt profile as the line profile because it exhibits an extended wing feature (generally absent in the Doppler profile) that bears the most importance for $N_{\rm ext}$. I performed a benchmark analysis for this hyperparameter $N_{\rm ext}$ with all three profile cases using the `timeit` module with $N_{\rm l} = 100$ and $N_{\rm r} = 10$. In Figure 3.7 I present the results of this investigation. As seen on the left panel, for all redshifts shown, the mean $|\delta\tau/\tau|$ drops steeply from $\sim 10^{-2}$ per voxel for no extension through $< 10^{-4}$ per voxel and converges after $N_{\rm ext} \approx 400$ voxels, which is less than 10% of the skewer length. On the right panel, we may see a linear increase in the

computation time for all three profile cases, which is the expected behavior. The best values of $t_\mathrm{l}/N_\mathrm{l}$ at $f_\mathrm{ext} \sim 0.1$ are $\sim 15\%$ larger than those at $f_\mathrm{ext} \sim 0$ (no extension). Based on these results, $f_\mathrm{ext} = 0.1$ appears to be sufficient for an accurate computation of τ without adding a large computational overhead, and therefore I fix that value for the remainder of this work. I list the computation time values for the three profile cases at fixed $f_\mathrm{ext} = 0.1$ in Table 3.2. Clearly, the discrete Doppler profile case is the fastest of all since it only requires evaluation of a single Gaussian function per pixel per term in the summation in Eq. (3.2). The error function version requires two evaluations of the error function on an equivalent level and hence is approximately a factor of 2 slower. The discrete Voigt profile case is the slowest of all.

Line profile

For the fixed $f_\mathrm{ext} = 0.1$, I analyzed the differences in τ and F among the three line profile cases for the test set. In Figure 3.8 I show the absolute deviations in F for the three distinct pairs across snapshot redshifts. I make the following observations: (i) The mean differences between the error function and the discrete Doppler profile cases are extremely small at all redshifts, suggesting that the error function version does not particularly offer any more accurate estimation of τ than the discrete Doppler profile. (ii) The mean and median deviations increase from low redshifts $z \sim 2$ through intermediate $z \sim 4$ and then drop through high $z \sim 6$. (iii) The median absolute deviations in all three pairs are smaller than the mean deviations at all redshifts. This hints at a high $|\Delta F|$ tail. Indeed, the differences between the Doppler and the Voigt profile cases reach as high as 0.7 for isolated skewers, in short pixel ranges, due to the presence of HCDs. An example of such a skewer from the test set at $z \sim 3.2$ is shown in Figure 3.9. Since this skewer contains an HCD absorber system, the Doppler profile (or a variation thereof, such as the error function case) is no longer a good approximation at the wings of the absorption line, leading to large differences in F. In the studies of neural network inference, I shall omit skewers that contain one or more such HCDs.

In Figure 3.10, I show a comparison of the different profile cases at the $P(k)$ level for the test set at all redshifts. I use the full solution for the n_HI here and rescale all τ values to match the corresponding observed $\bar{F}(z)$. In the range $k \leq 0.195$ s/km, the relative differences in $P(k)$ *w.r.t.* the Voigt profile case stay close to or under 1%. It is important to note that, since the skewers with potential HCDs are included here, large errors $\Delta P/P$ *w.r.t.* the Voigt profile case at large scales (small k) are expected. Under these conditions, a percent level agreement is very promising. The differences between the discrete Doppler and the Error function cases appear to be negligible at large scales, but they start to diverge for $k > 0.03$ s/km at all redshifts.

τ rescaling

In my implementation of the synthesis of mock spectra, I use the photoionization rates ($\Gamma_{\gamma,\mathrm{S}}, \forall\, \mathrm{S} \in \{\mathrm{HI}, \mathrm{HeI}, \mathrm{HeII}\}$) from the late reionization model of Oñorbe et al. (2017),

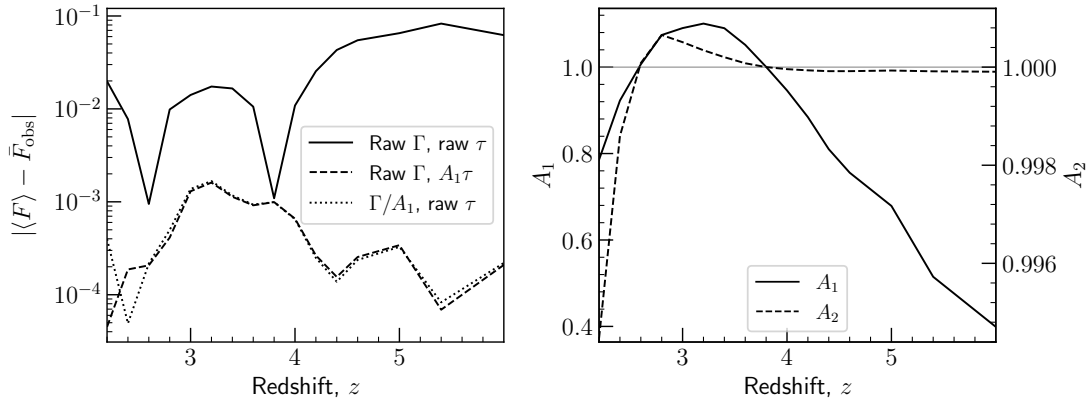

Figure 3.11: The effect of rescaling τ values (and the UV background Γ exploiting the FGPA) to match the observed mean transmission $\bar{F}_{\rm obs}$ of Becker et al. (2013) at all snapshot redshifts. **Left**: absolute deviation from $\bar{F}_{\rm obs}$ using the raw Γ values from Oñorbe et al. (2017) and the subsequent raw τ values (solid line); raw Γ but the rescaled τ values (dashed); and the rescaled Γ/A_1 with the subsequent raw τ values (dotted). **Right**: the two rescaling factors $A_{1,2}$.

assuming a spatially uniform UV background. As discussed in Section 3.3.1, however, those values are not very well-constrained and lead to a different mean Lyα transmission from the observed value, $\bar{F}_{\rm obs}$. Therefore I rescale all the newly computed τ values by a factor $A_1(z)$ to correct for this mismatch in the mocks. In order to test the validity of this operation, I applied the resulting correction factor to the HI UV background as $\Gamma \to \Gamma/A_1$ and ran the spectral extraction anew for the test set of skewers. The resultant mean transmission is expected to be less divergent from $\bar{F}_{\rm obs}$ than that of the raw first iteration if the rescaling is valid. As a convergence check, I computed a further factor $A_2(z)$ for scaling the τ values derived from the rescaled UV background Γ/A_1. If the rescaled UV background is a better estimate of the true value than originally, the new τ values should more closely reproduce $\bar{F}_{\rm obs}$ and A_2 should be closer to 1. I show the deviation from $\bar{F}_{\rm obs}$ against redshift for the raw τ values in the original (raw) and rescaled Γ cases, as well as for the rescaled, $A_1\tau$ values in the raw Γ case in the left panel of Figure 3.11. In the right panel therein are the two rescaling factors $A_{1,2}(z)$. In Table 3.3, I list those factors at all snapshot redshifts along with the target $\bar{F}_{\rm obs}(z)$ derived using the fitting formula of Becker et al. (2013). Indeed, our expectations hold to an excellent degree, in that the second rescaling factor A_2 is very close to 1 and the deviations from $\bar{F}_{\rm obs}$ for the rescaled UV background are close to or under 10^{-3} at all redshifts.

For the purpose of neural network training on spectra, the rescaled optical depth values $A_1\tau$ of the raw Γ and the raw τ of the rescaled Γ/A_1 (dashed and dotted lines in the left panel of Figure 3.11, respectively) are equivalent and hence I utilize the former in order to minimize the computational load. Any discrepancies between the two are the result of the small but non-zero contribution of collisional ionization that is ignored in the FGPA. This contribution is a function of density. The overdensities get increasingly pronounced at smaller redshifts as the structures grow, enhancing the impact of collisional ionization

Table 3.3: The redshift-dependent scaling factors $A_{1,2}$ for the first and second τ rescalings, respectively, as described in the main text. The corresponding target $\bar{F}_{\rm obs}$ values from Becker et al. (2013) are listed for reference. A_1 ranges from 0.4 to 1.1 whereas A_2 remains consistently close to 1, eliminating the need for further rescaling.

z	$\bar{F}_{\rm obs}$	A_1	A_2
2.2	0.8629	0.7857	0.9945
2.4	0.8178	0.9226	0.9986
2.6	0.7701	1.0064	1.0001
2.8	0.7204	1.0743	1.0007
3.0	0.6692	1.0903	1.0005
3.2	0.6170	1.1007	1.0003
3.4	0.5646	1.0900	1.0002
3.6	0.5125	1.0515	1.0001
3.8	0.4614	1.0005	1.0000
4.0	0.4117	0.9460	1.0000
4.2	0.3642	0.8843	0.9999
4.4	0.3191	0.8101	0.9999
4.6	0.2769	0.7558	0.9999
5.0	0.2024	0.6788	0.9999
5.4	0.1418	0.5148	0.9999
6.0	0.0763	0.3990	0.9999

at the level of τ. However, in those optically thick regimes, sub-percent changes in τ are negligible at the level of F.

3.5 Summary

The synthesis of mock Lyα forest spectra is a crucial task for providing the most important ingredient of a NN inference machinery: the training data. As discussed in this chapter, it involves multiple computational steps corresponding to a set of physical effects that play a role in giving rise to the Lyα forest feature in QSO spectra. Since neutral Hydrogen is the progenitor atomic species of this spectral phenomenon, we first and foremost require the information of its number density, $n_{\rm HI}$, from the contents of a hydrodynamical simulation. At the cosmological redshifts of interest in this work, the epoch of reionization is over and the gas is almost entirely ionized. The remaining neutral species of H and He are in ionization equilibrium with the ionized species and free electrons, and this results in a set of equilibrium equations that need to be solved in order to extract the necessary information. In Section 3.2 I describe the physical and computational details of this procedure.

Peculiar motions (along the lines of sight) in the intergalactic gas due to gravity as well as turbulence move the resonant absorption lines in a spectrum via Doppler shift and the

density and temperature of the absorber clouds lead to broadening of the lines via Doppler and quantum mechanical pressure broadening profiles, culminating in a Voigt profile. The shapes and locations of the Lyα absorption lines in a forest are governed by these two effects combined. Furthermore, a poorly constrained value of intergalactic ionizing background(s) could lead to a mismatch in the effective Lyα optical depth, $\tau_{\rm eff}$, of the simulated IGM from its observed value, however, an overall scalar factor may be applied *a posteriori* to correct for this, as is common practice in the literature. In Section 3.3 I discuss all of these effects that are necessary for an accurate computation of Lyα optical depth per pixel in the synthetic spectra from the existing information of $n_{\rm H\,I}$.

Due to the heavy computational resource demands of the various steps in the synthesis of mocks, cheaper alternatives are often found based on certain assumptions that may be applicable at some but not all cosmological epochs. I describe some of the popular approximations for computing $n_{\rm H\,I}$ and the line profiles in the aforementioned sections. It becomes crucial to analyze the discrepancies of those approximate methods $w.r.t$ their exact counterparts and weigh them in light of their gain in computational time. In Section 3.4 I present a thorough investigation of this nature involving the popular approximations described earlier and make recommendations as to the suitable techniques in various cosmological regimes based on the findings. Those are also the techniques I have employed for producing vast amounts of mock Lyα forest datasets for training our field-level inference machinery that is the focus of the following two chapters.

Chapter 4

Deep learning inference: proof of concept

Bibliographic and copyright information

P. Nayak, M. Walther, D. Gruen, and S. Adiraju (Sept. 2024); "LyαNNA: A deep learning field-level inference machine for the Lyman-α forest." A&A 689, A153.

DOI: 10.1051/0004-6361/202348485

In this work, we demonstrate the potential of deep learning at the field-level for the inference of the power law parameters (T_0, γ) of the IGM temperature density relation. From pure (uncontaminated) mock spectra, our machinery puts more stringent constraints on the said parameters than the traditional summary statistics such as the 1D power spectrum and PDF of the Lyα transmission. As the lead author, I completed the following key tasks: creating all the mock training/validation/test datasets, developing the core inference framework, writing the vast majority of code in Python, training neural networks and optimizing hyperparameters, running tests with the trained networks and the traditional summaries that I computed from the mocks. I also wrote the majority of the manuscript and produced the various illustrations/plots. My coauthors M. Walther and D. Gruen contributed significantly to the development of the project idea and also provided valuable scientific discussions. M. Walther ran the suite of simulations used in this work. S. Adiraju contributed to the development of the basic initial codebase of the neural machinery with Keras upon which I built this inference framework. All coauthors reviewed the draft of the manuscript and suggested improvements.

4.1 Introduction

The characteristic arrangement of Lyα absorption lines in the spectra of distant quasars, commonly known as the "Lyα forest" (Lynds 1971), has been shown to be a unique probe of the physics of the Universe at play over a wide window of cosmic history ($z \lesssim 6$). As the continua of emission by the quasars traverse the diffuse intergalactic gas, resonant scattering by the neutral Hydrogen leads to a net absorption of the radiation at the wavelength of the Lyα transition (Gunn and Peterson 1965). In an expanding Universe where spectral redshift z is a proxy of distance, a congregation of absorber clouds in the intergalactic medium (IGM) along a quasar sight line imprints a dense forest of Lyα absorption lines on their spectra. Due to cosmic reionization of Hydrogen being largely complete by $z \sim 6$ (e.g., McGreer et al. 2015), its neutral fraction $x_{\rm HI}$ within the IGM is extremely minute, yet sufficient to produce this unique feature that enables a continuous trace of the cosmic gas.

The observations of the Lyα forest, through the advent of high-resolution instruments such as Keck/HIRES and VLT/UVES as well as large-scale structure surveys such as the extended Baryon Oscillation Spectroscopic Survey (eBOSS; Dawson et al. 2013) of the Sloan Digital Sky Survey (SDSS; Blanton et al. 2017) and the Dark Energy Spectroscopic Instrument (DESI; DESI Collaboration et al. 2022), have delivered a wealth of information about the nonlinear matter distribution on submegaparsec scales, thermal properties of the intergalactic gas, and large-scale structure. Not only is the Lyα forest an extremely useful tool to study the thermal evolution of the IGM and reionization (as demonstrated, e.g., by Becker et al. 2011, Walther et al. 2019, Boera et al. 2019, Gaikwad et al. 2021), but it has also opened up avenues for constraining fundamental cosmic physics. Chief among those are the baryon acoustic oscillation (BAO) scale and cosmic expansion (e.g., Slosar et al. 2013, Busca et al. 2013, du Mas des Bourboux et al. 2020, Gordon et al. 2023, Cuceu et al. 2023), the nature and properties of dark matter (e.g., Viel et al. 2005, Viel et al. 2013, Iršič et al. 2017b, Armengaud et al. 2017, Rogers and Peiris 2021), and in combination with the cosmic microwave background (CMB, e.g., Planck Collaboration et al. 2020) also inflation and neutrino masses (e.g., Seljak et al. 2006, Palanque-Delabrouille et al. 2015, Yèche et al. 2017, Palanque-Delabrouille et al. 2020).

The classical way of carrying out parameter inference with the Lyα forest, as for any other cosmic tracer, relies on summary statistics of the underlying field, as they conveniently pick out a small number of relevant features from a much larger number of degrees of freedom of the full data. For the Lyα forest, a number of summary statistics exists that have been accurately measured and effectively used for cosmological and astrophysical parameter inference. These include the line-of-sight (1D) transmission power spectrum (TPS hereinafter; e.g., Croft et al. 1998, Chabanier et al. 2019, Walther et al. 2019, Boera et al. 2019, Ravoux et al. 2023, Karaçaylı et al. 2024), transmission probability density function (TPDF hereinafter; e.g., McDonald et al. 2000, Bolton et al. 2008, Viel et al. 2009, Lee et al. 2015), wavelet statistics (e.g., Meiksin 2000, Theuns and Zaroubi 2000, Zaldarriaga 2002, Lidz et al. 2010, Wolfson et al. 2021), curvature statistics (e.g., Becker et al. 2011, Boera et al. 2014), distributions of absorption line fits (e.g., Schaye et al. 2000, Bolton et al.

2014, Hiss et al. 2019, Telikova et al. 2019, Hu et al. 2022), and combinations thereof (e.g., Garzilli et al. 2012, Gaikwad et al. 2021). While these provide accurate measurements of parameter values, they fail to capture all of the information contained in the transmission field, thereby resulting in a loss of constraining power the full spectral datasets have to offer.

Recently, deep learning approaches have become popular in the context of cosmological simulations and data analysis. Complex and resource-heavy conventional problems in cosmology have started to see fast, efficient, and demonstrably robust solutions in neural network (NN) based algorithms (see, e.g., Moriwaki et al. 2023 for a recent review). Artificial intelligence has opened up a broad avenue for studies of the Lyα forest as well. Cosmological analyses with the Lyα forest generally demand expensive hydrodynamic simulations for an accurate modeling of the small-scale physics of the IGM. Deep learning offers alternative, light-weight solutions to such problems. For instance, Harrington et al. 2022 and Boonkongkird et al. 2023 recently built U-Net based frameworks for directly predicting hydrodynamic quantities of the gas from computationally much less demanding, dark-matter-only simulations. A super-resolution generative model of Lyα-relevant hydrodynamic quantities is presented in Jacobus et al. (2023), based on conditional generative adversarial networks (cGANs). These works greatly accelerate the generation of mock data for Lyα forest analyses. Deep learning is also demonstrated to be a very effective methodology for a variety of tasks involving spectral, 1D datasets. Ďurovčíková et al. (2020) introduced a deep NN to reconstruct high-z quasar spectra containing Lyα damping wings. Melchior et al. (2023) and Liang et al. (2023) describe a framework for generating, analyzing, reconstructing, and detecting outliers from SDSS galaxy spectra consisting of an autoencoder and a normalizing flow architecture. Recent works have shown immense potential of various deep NN methods for the analysis of the Lyα forest. For example, a convolutional neural network (CNN) model to detect and characterize damped Lyα systems (DLAs) in quasar spectra was introduced by Parks et al. (2018). Similarly, Busca and Balland (2018) applied a deep CNN called "QuasarNET" for the identification (classification) and redshift estimation of quasar spectra. Huang et al. (2021) constructed a deep learning framework to recover the Lyα optical depth from noisy and saturated Lyα forest transmission. Later, Wang et al. (2022b) applied the same idea to the reconstruction of the line-of-sight temperature of the IGM and detection of temperature discontinuities (e.g., hot bubbles). In neighboring disciplines, deep learning is already identified as a reliable tool for field-level inference. For instance, a set of recent works (Gupta et al. 2018, Fluri et al. 2018, Ribli et al. 2019, Kacprzak and Fluri 2022 among others) has established the superiority of deep learning techniques for cosmological inference directly from weak gravitational lensing maps over the classical two-point statistics of the cosmic density-field proxies.

In this work we present LyαNNA – short for "Lyα Neural Network Analysis" – a deep learning framework for the analysis of the Lyα forest. Here, we have implemented a 1D ResNet (a special type of CNN with skip-connections between different convolutional layers to learn the residual maps; He et al. 2015a) called "Sansa" for inference of model parameters with Lyα forest spectral datasets, harvesting the full information carried by

the transmission field. We perform nonlinear regression on the thermal parameters of the IGM directly from the spectra containing the Lyα forest absorption features that are extracted efficiently by our deep model. This architecture was trained in a supervised fashion using a large set of mock spectra from cosmological hydrodynamic simulations with known parameter labels to not only distinguish between two distinct parameter combinations but also pinpoint the exact location of a given spectral set in the parameter space. For better statistical reliability of our results, we employed a committee of 20 neural networks for the inference, combining the outputs via bootstrap aggregation (Breiman 2004). Finally, we built a likelihood model to perform inference on mock datasets via Markov chain Monte Carlo (MCMC) and compared with classical summary statistics, namely a combination of TPS and TPDF, showcasing the improvement we gain by working at the field level.

This paper is structured as follows. Section 4.2 describes the simulations, the mock Lyα forest spectra we use for training and testing our methodology, and the summary statistics we compare to. In Section 4.3 we introduce the inference framework of Sansa with details of the architecture and its training. Our results of doing inference with Sansa and a comparison with the traditional summary statistics are presented and discussed in Section 4.4. We conclude in Section 4.5 with a précis of our findings and an outlook.

4.2 Simulations

In this section we introduce the hydrodynamic simulation used throughout this work as well as the post-processing approach we adopt to generate mock Lyα forest spectra.

4.2.1 Hydrodynamic simulations

We used a Nyx cosmological hydrodynamic simulation snapshot generated for Lyα forest analyses (see Walther et al. 2021) to create the mock data used for various purposes in this work. Nyx is a relatively novel hydrodynamics code based on the AMReX framework and simulates an ideal gas on an Eulerian mesh interacting with dark matter modeled as Lagrangian particles. While adaptive mesh refinement (AMR) is possible and would allow better treatment of overdense regions, we used a uniform grid here as the Lyα forest only traces mildly overdense gas, rendering AMR techniques inefficient. Gas evolution was followed using a second-order accurate scheme (see Almgren et al. 2013 and Lukić et al. 2015 for more details). In addition to solving the Euler equations and gravity, Nyx also models the main physical processes required for an accurate model of the Lyα forest. The chemistry of the gas was modeled following a primordial composition of H and He. Inverse Compton cooling of the CMB was taken into account as well as the updated recombination, collisional ionization, dielectric recombination and cooling rates from Lukić et al. 2015. All cells were assumed to be optically thin to ionizing radiation and a spatially uniform ultraviolet background (UVB) was applied according to the late reionization model of Oñorbe et al. 2017, where heating rates were modified by a fixed factor $A_{\rm UVB}$ affecting the thermal history and thus pressure smoothing of the gas. Here, we used a simulation box

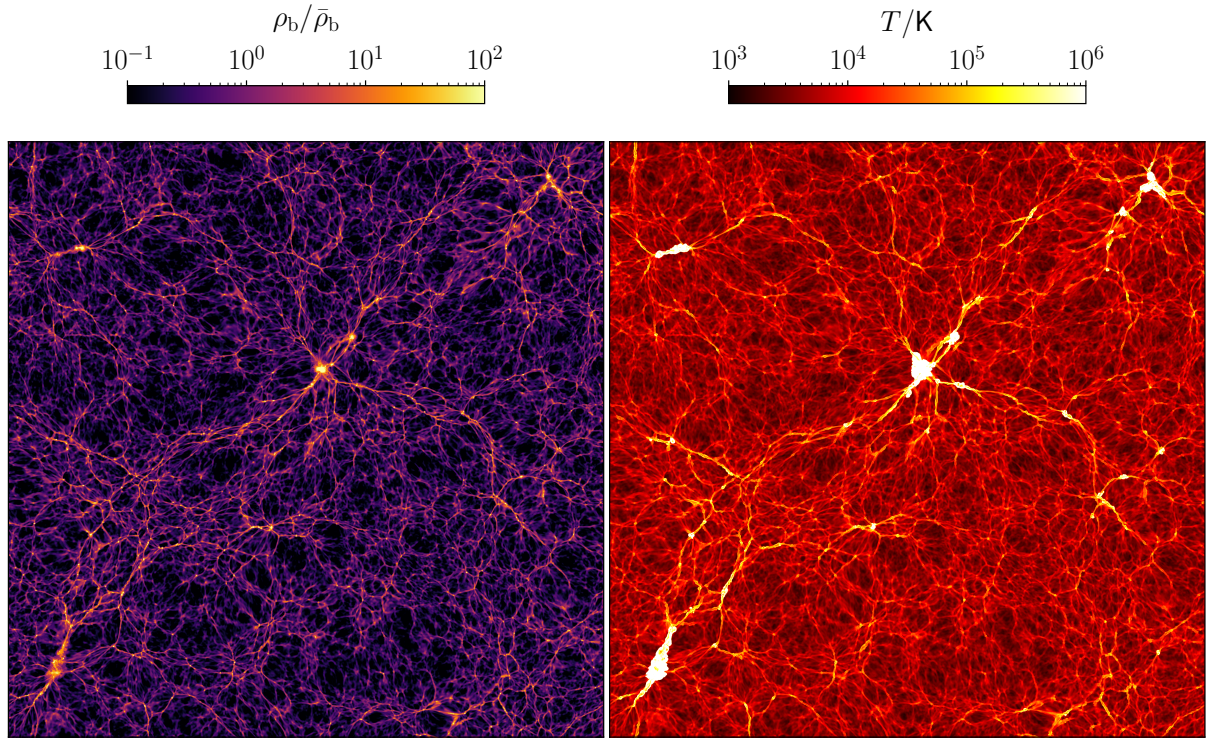

Figure 4.1: Slice (29 kpc thick) through our Nyx simulation box (120 Mpc side length, with 4096 cells along a side) at $z = 2.2$ shown here for the gas overdensity (left) and temperature (right). A systematic relationship between both the hydrodynamic fields can be seen.

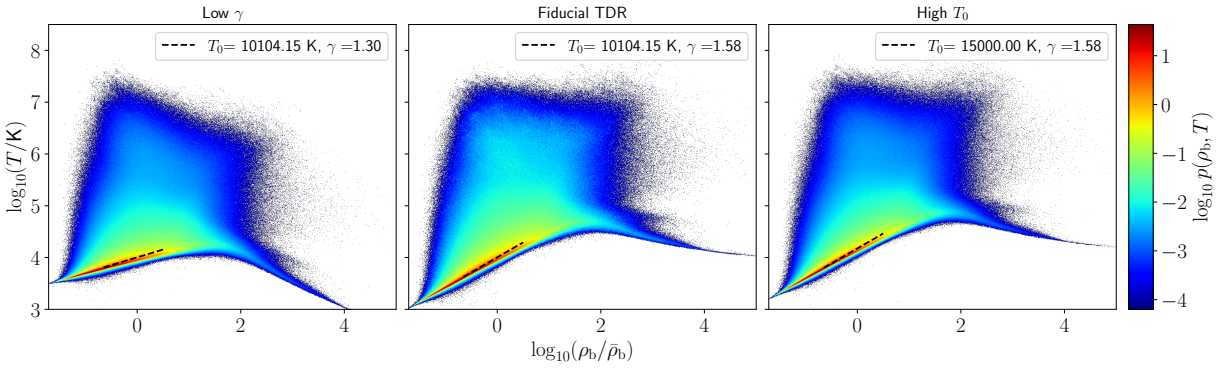

Figure 4.2: Joint, volume-weighted distribution of the temperature and density of baryons in our simulation at $z = 2.2$. **Center**: the fiducial (T_0, γ) parameter case (values obtained by fitting a power-law through gas temperatures and densities). **Left**: the case with rescaled temperatures for a lower γ than the fiducial and the same T_0. This can be seen to affect the slope of the TDR by a twisting of the 2D distribution. **Right**: the case with rescaled temperatures for a higher T_0 than the fiducial and the same γ. The entire distribution is shifted along the $\log_{10}(T/\mathrm{K})$ axis, keeping the shape the same. (Note that the color-bar label is using shorthand for $p(\log_{10}(\rho_b/\bar{\rho}_b), \log_{10}(T/\mathrm{K}))$.)

at $z = 2.2$ with 120 Mpc side length and 4096^3 volumetric cells ("voxels") and dark matter particles, motivated by recent convergence analyses (Walther et al. 2021 and Chabanier et al. 2023). The cosmological parameters of the box are $h = 0.7035$, $\omega_\mathrm{m} = \Omega_\mathrm{m}h^2 = 0.1589$, $\omega_\mathrm{b} = \Omega_\mathrm{b}h^2 = 0.0223$, $A_\mathrm{s} = 1.4258 \times 10^{-9}$, $n_\mathrm{s} = 1.0327$, $A_\mathrm{UVB} = 0.9036$.

During the epoch of reionization the ionizing UV radiation from star-forming galaxies heats up the intergalactic gas as well. Afterward, as the universe expands, the IGM cools down mostly adiabatically with subdominant nonadiabatic contribution, for instance, from inverse Compton scattering off of the CMB and recombination of the ionized medium, as well as heating due to photoionization and gravitational collapse. The bulk of this gas is diffuse (relatively cool with $T < 10^5$ K and mildly overdense with $\log_{10}(\rho_\mathrm{b}/\bar{\rho}_\mathrm{b}) < 2$, contained mostly in cosmic voids, sheets, and filaments; see, e.g., Martizzi et al. 2019) and imprints the Lyα forest absorption features on to quasar spectra. The IGM at $z \sim 2.2$ can be further classified into subdominant phases such as warm-hot intergalactic medium (WHIM; $T > 10^5$ K and $\log_{10}(\rho_\mathrm{b}/\bar{\rho}_\mathrm{b}) < 2$), condensed halo ($T < 10^5$ K and $\log_{10}(\rho_\mathrm{b}/\bar{\rho}_\mathrm{b}) > 2$) and warm halo or circumgalactic medium (WCGM; $T > 10^5$ K and $\log_{10}(\rho_\mathrm{b}/\bar{\rho}_\mathrm{b}) > 2$). At this redshift, the effects due to the inhomogeneous UVB of the reionization of H and He are considered to be small (e.g., Oñorbe et al. 2019; Upton Sanderbeck and Bird 2020) and we ignore them for this work. The diffuse IGM component in our cosmological simulation exhibits a tight power-law relation in temperature and density (Hui and Gnedin 1997; McQuinn and Upton Sanderbeck 2016) that is classically characterized by

$$T = T_0 \left(\frac{\rho_\mathrm{b}}{\bar{\rho}_\mathrm{b}} \right)^{\gamma - 1}, \tag{4.1}$$

where $\bar{\rho}_\mathrm{b}$ is the mean density of the gas, and T_0 (a temperature at the mean gas density) and γ (adiabatic power-law index) are the two free parameters of the model. Indeed, a strong systematic ρ_b-T relationship is visually apparent in a slice through our simulation box (Figure 4.1). We performed a linear least-squares fit of the above relation through our simulation in the range $-0.5 < \log_{10}(\rho_\mathrm{b}/\bar{\rho}_\mathrm{b}) < 0.5$ and $\log_{10}(T/\mathrm{K}) < 4$. The best-fit (fiducial) values are $T_0 = 10104.15$ K and $\gamma = 1.58$. While a range of works have demonstrated the potential of using different summary statistics of the Lyα forest as probes to measure T_0 and γ (see, e.g., Gaikwad et al. 2020), in this work we highlight a first field-level framework for inference of these two thermal parameters of the IGM.

The following strategy was adopted for sampling the parameter space of (T_0, γ) to produce labeled data for the supervised training of the inference machine. Both the parameters were varied by a small amount at a time, $\log T_0 \to \log T_0 + \log x$ and $\gamma \to \gamma + y$ to obtain a new temperature-density relation (TDR). We then rescaled the simulated temperatures at every cell of the simulation by $T \to x \cdot (\rho_\mathrm{b}/\bar{\rho}_\mathrm{b})^y \cdot T$ at fixed densities ρ_b to appropriately incorporate the scatter off the TDR into our mock data, effectively conserving the underlying T-ρ_b distribution rather than drawing from a pure power-law. This procedure is illustrated in Figure 4.2 with the help of the full 2D histograms of temperature and density for two individual parameter rescalings as well as the fiducial case.

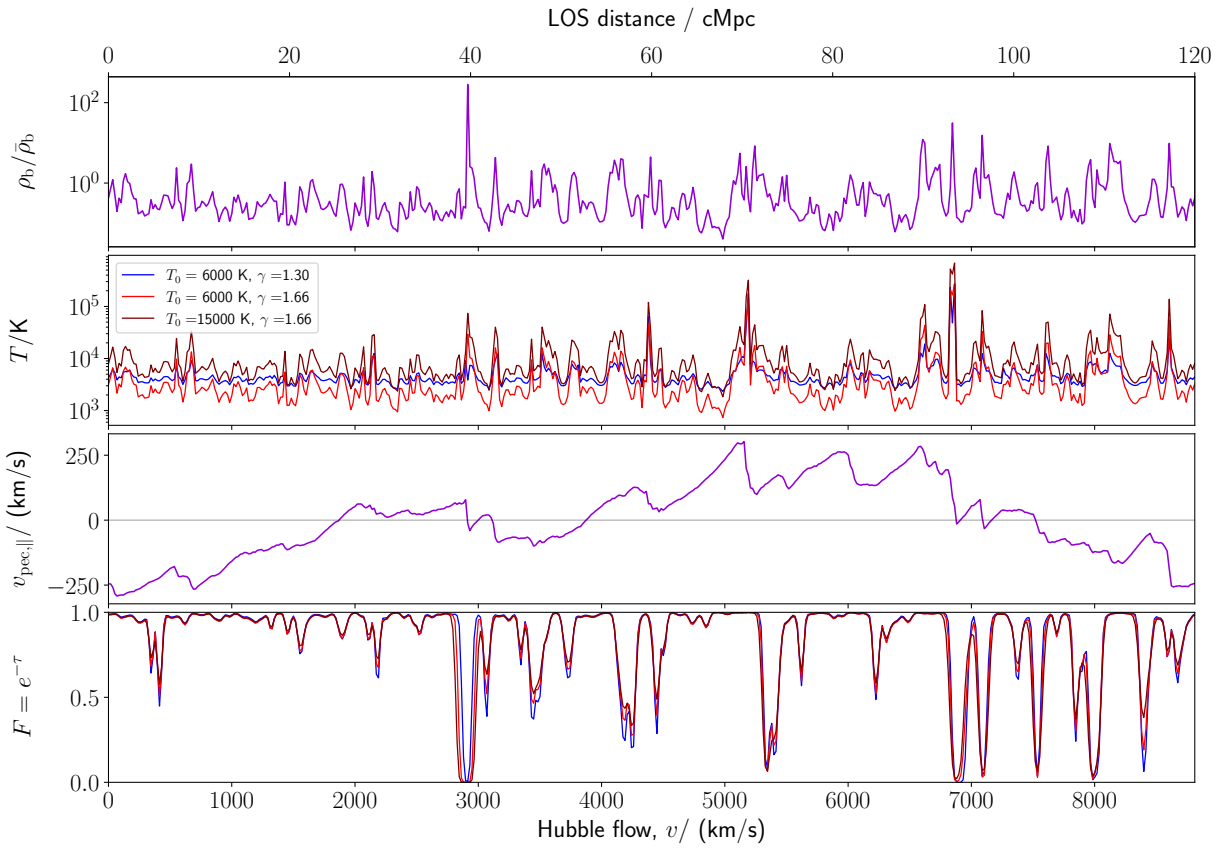

Figure 4.3: Simulated baryon overdensities, temperatures, line-of-sight peculiar velocity, and transmission along an example skewer through our box for three different temperature-density relations – one pair each for fixed T_0 and fixed γ – to indicate characteristic variations with respect to the two TDR parameters. The line shifts due to the peculiar velocity component ($v_{\rm pec,\parallel}$) can be easily noticed. Since the rescaling of the temperatures depends on the densities, inhomogeneous differences are seen in the skewer temperatures in the cases with varying γ. The absorption lines are broader where the temperatures are higher and the amount of broadening is also a function of densities being probed through γ.

4.2.2 Mock Lyman-α forest

In order to simulate the Lyα forest transmission $F = e^{-\tau}$ (τ being the optical depth), we first chose 10^5 random lines of sight (LOS, a.k.a. skewers) parallel to one of the Cartesian sides (e.g., Z-axis) of the box by picking all consecutive 4096 voxels along that axis while keeping the other two coordinates (X and Y) fixed at a time. The Lyα optical depth at an output pixel in a spectrum was calculated from the information of the density, temperature, and the LOS component of gas peculiar velocity ($v_{\rm pec,\parallel}$) at each corresponding voxel along the given skewer. Here, the gas was reasonably assumed to be in ionization equilibrium among the different species of H and He and further that He is almost completely (doubly) ionized at $z \sim 2.2$ (i.e., $x_{\rm HeIII} \approx 1$; Miralda-Escudé et al. 2000, Becker et al. 2015) in order to estimate the neutral H density, $n_{\rm HI}$ for each of those voxels. The Lyα optical depth at

a pixel with Hubble velocity v and gas peculiar velocity $v_{\mathrm{pec},\parallel}$ was estimated as

$$\tau(v) = \frac{\pi e^2 \lambda_{\mathrm{Ly}\alpha} f_{lu}}{m_{\mathrm{e}} c H(z)} \int n_{\mathrm{H\,I}}(v') \phi_{\mathrm{D}}\big(v'; v + v_{\mathrm{pec},\parallel}, b\big) dv', \tag{4.2}$$

where the rest-frame $\lambda_{\mathrm{Ly}\alpha} = 1215.67$ Å, the Lyα oscillator strength $f_{lu} = 0.416$, and

$$\phi_{\mathrm{D}}(v; v_0, b) \equiv \frac{1}{b\sqrt{\pi}} \exp\left[- \left(\frac{v - v_0}{b} \right)^2 \right] \tag{4.3}$$

is the Doppler line profile with the temperature-dependent broadening parameter $b = \sqrt{2 k_{\mathrm{B}} T / m_{\mathrm{H}}}$. These τ values were additionally rescaled by a constant factor such that the mean Lyα transmission in our full set of 10^5 skewers matched its observed value of $\bar{F}_{\mathrm{obs}} = 0.86$, compatible with Becker et al. (2013). For our simulation, we performed tests with a simplified, approximate model of lightcone evolution along the LOS (Appendix A) and found negligible impact on the performance of our inference framework. Therefore, we ignored any small lightcone evolution along our skewers and used the snapshot of the simulation for creating mock Lyα forest spectra, assuming a constant TDR for simplicity.

To mimic observational limitations and minimize the impact of numerical noise on small scales in the simulations, we restricted the Fourier modes within the spectra to $k < k^*$, $k^* = 0.18$ s km^{-1}. This was effectively achieved by smoothing them with a spectral resolution kernel of $R_{\mathrm{FWHM}} \approx 11,000$ and additionally rebinning them by 8-pixel averages, matching the Nyquist sampling limit. The final size of a spectrum in our analysis is thus 512 pixels.

When sampling the (T_0, γ) space, new mock spectra were produced for each parameter combination with the new (rescaled) temperatures and the original densities and line-of-sight peculiar velocities along the same set of skewers. An example skewer is shown in Figure 4.3 for three different TDRs. Since the temperature rescaling is a function of $\rho_{\mathrm{b}}/\bar{\rho}_{\mathrm{b}}$ through γ, characteristic differences in skewer temperatures and Lyα transmission between cases with varying γ are visible. Changes in T_0 result in a homogeneous broadening of the absorption lines, whereas changes in γ (for $\gamma > 1$)[1], depending on the underlying overdensity being probed, result in larger or smaller broadening (i.e., generally, the shallower lines are broadened less than the deeper ones). We expect our convolutional architecture to be able to pick up such features in order to discriminate between thermal models. In this work, we sampled a grid of 11×11 (T_0, γ) combinations as shown in Figure 4.4 – for each of which we have the same set of 10^5 physical skewers – for training and testing our deep learning machinery. This grid is oriented in a coordinate system that captures the well-known degeneracy direction in the (T_0, γ) space as identified in many TPS analyses (e.g., Walther et al. 2019) and is motivated by the heuristic argument that it is easiest to train a neural network for inference with an underlying parametrization that captures

[1] $\gamma = 1$ would mean the diffuse IGM is roughly isothermal and $\gamma < 1$ would lead to an "inverted TDR," where underdensities are hotter than overdensities as proposed by Bolton et al. 2008, for instance.

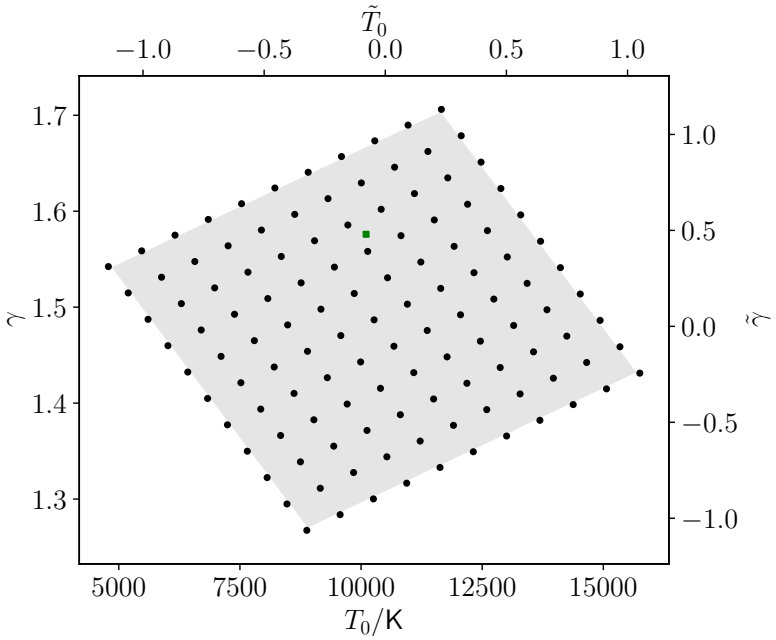

Figure 4.4: Our sample of thermal models for the various training and test purposes that contains 11×11 (121) distinct (T_0, γ) combinations. The fiducial TDR parameter combination is depicted as the green square. The rescaled $(\tilde{T}_0, \tilde{\gamma})$ axes are also shown for context (see Section 4.3.1). This is a uniform grid in a parametrization that captures the well known degeneracy direction in the (T_0, γ) space. The gray-shaded area shows our prior range in this parameter space. Please refer to Appendix B for more details regarding the strategy used for this sampling.

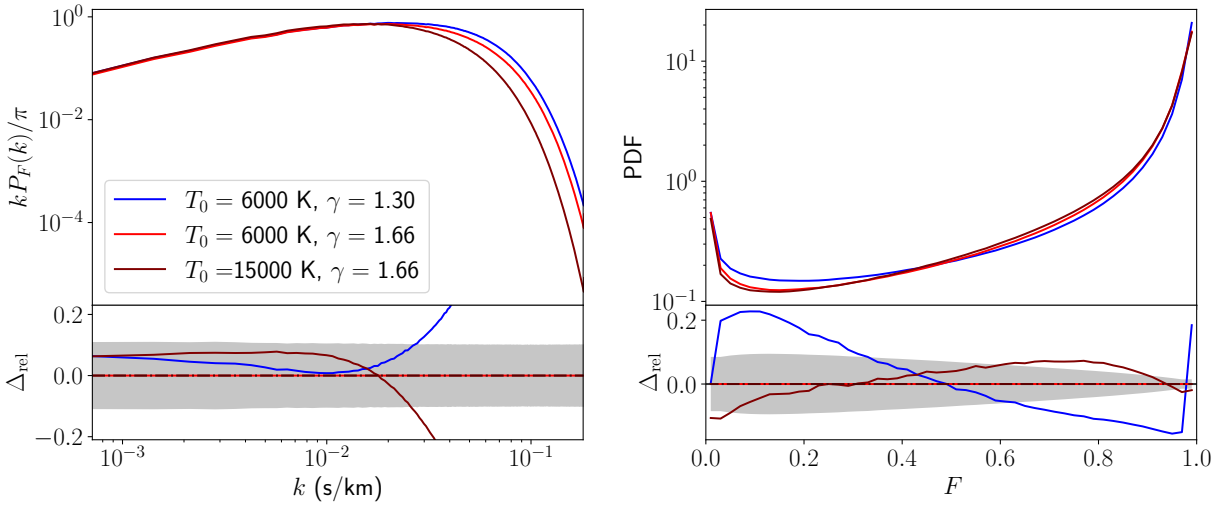

Figure 4.5: Transmission power spectrum (left) and transmitted PDF (right) computed from our set of 10^5 skewers for three TDR parameter combinations; the same as in Figure 4.3. The fractional differences between different TDR cases are shown in the corresponding bottom panels with the gray-shaded areas as 1σ uncertainty ranges (bands drawn from discrete k-modes in TPS and discrete histogram bins in TPDF), equivalent of 100 spectra.

the most characteristic variations in the data. The exact sampling strategy is further described in Appendix B. We used the gray-shaded region in Figure 4.4 as our prior range of parameters having a uniform prior distribution in all our further analyses.

4.2.3 Summary statistics

We considered two summary statistics – TPS and TPDF – of the Lyα forest in this work for demonstrating the benefit of field-level inference. The TPS is defined here as the variance of the transmitted "flux contrast" $\delta_F \equiv (F - \bar{F})/\bar{F}$ in Fourier space, that is, $P_F(k) \sim \langle \tilde{\delta}(k)^* \cdot \tilde{\delta}(k) \rangle$. For a consistent comparison of inference outcomes, we applied the same restriction $k < 0.18$ s km^{-1} as in the input to our deep learning machinery (see Section 4.2.2). To obtain the TPDF, we considered the histogram of the transmitted flux F in the full set of skewers over 50 bins of equal width between 0 and 1. For the likelihood analysis with the TPDF we left the last bin out as it is fully degenerate with the rest due to the normalization of the PDF. The mean TPS and TPDF computed from the 10^5 skewers for three different TDR parameter combinations are shown in Figure 4.5 along with the relative differences in both the statistics between pairs of TDR models. The uncertainty range shown as a gray band therein corresponds to a 1σ equivalent of 100 spectra. The TPS follows a power-law increase for small k and exhibits a suppression of power at larger k (smaller physical scales) due to deficiency of structures as well as the thermal broadening of lines. The variations in the thermal parameters that effectively result in the broadening of absorption lines amount to a shift in this turnover scale toward

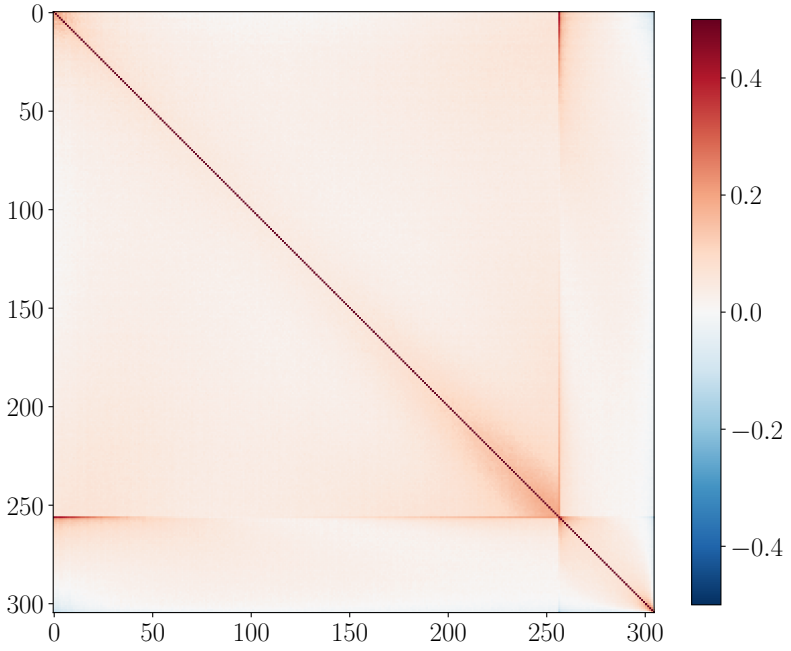

Figure 4.6: Correlation matrix of the joint summary vector (the first 256 entries being TPS and the later 49 being TPDF) estimated from our set of 10^5 mock spectra for the fiducial thermal model. Notice the mild correlations within the individual summaries and the cross-correlations between the two summaries.

smaller k-modes[2]. Equivalently, the broader lines have shallower depths (in this low density regime of the curve of growth), which in turn results in a transfer of probability from smaller ($F \lesssim 0.4$) to larger ($F \gtrsim 0.6$) transmission. We also computed the joint covariance matrix of the concatenated summary vector of the TPS and TPDF from our full set of mock spectra by the estimator

$$\mathbf{C}_\mathbf{s} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{s}_i - \mathbf{s})^T(\mathbf{s}_i - \mathbf{s}), \tag{4.4}$$

where $N = 10^5$ and $\mathbf{s}$ is a vector of which the first 256 entries are the Fourier modes in the TPS ($k < k^*$) and the later 49 entries are the bins in the TPDF, $F \in [0, 1)$. The joint correlation matrix for a thermal model from our sample is shown in Figure 4.6. Mild correlations among relatively close entries within the TPS or the TPDF can be observed as well as moderate cross-correlations between the two summary statistics.

For inference with these summary statistics, we cubic-spline interpolated both (the TPDF per histogram bin and the TPS per discrete k-mode) as a function of the parameters (T_0, γ) to obtain an emulator over our prior range as depicted in Figure 4.4 where we assumed a flat (uniform) prior in both the parameters. We verified that choosing a different

[2]We recall that the mean transmission $\bar{F}$ is kept fixed for our spectra; for fixed UVB – hence varying $\bar{F}$ – the quantitative variations in the TPS with respect to the thermal parameters would be different, especially for small k.

interpolation scheme, such as linear, does not strongly affect the results of the inference.

4.3 Field-level inference machinery

As described in Section 4.2, we have simulated Lyα forest absorption profiles (spectra) from hydrodynamic simulations having known thermal (T_0, γ) parameter values. The aim of our machinery is to learn the characteristic variations in the spectra (i.e., at the field level) with respect to those parameters in order first to distinguish between two adjacent thermal models and ultimately also to provide an uncertainty estimate as well as a point estimate of the parameter values whereby Bayesian inference can be performed. Thus framed, this is a very well-suited problem for application of supervised machine learning. The output of a fully trained deep neural network can be used as a model (emulator) for a newly learned, optimal "summary statistic" of the Lyα transmission field that is fully degenerate with the thermal parameters[3], hence carrying most of the relevant information about them that the full field offers. In the following we describe our framework in detail, with a special focus on the neural network architecture and training.

4.3.1 Overview

The general structure of inference with LyαNNA entails a feed-forward 1D ResNet neural network called "Sansa" that connects an underlying input information vector (transmission field) to an output "summary vector" that can be conveniently mapped to the thermal parameters (T_0, γ). Ideally, we expect this summary vector to be a direct actual estimate of the parameters itself, however, due to a limited prior range of thermal models available for training, a systematic (quantifiable) bias was observed in the pure network estimates (see Appendix C). Nonetheless, these estimates can be mapped to the parameters via a tractable linear transformation. For brevity our network encompasses this mapping[4] such that its output is a direct estimate of the parameters $\tilde{\boldsymbol{\pi}} = (\tilde{T}_0,\ \tilde{\gamma})$. As an estimate of its own uncertainty of a given prediction, the network also returns a parameter covariance matrix $\tilde{\mathbf{C}}$. Since our two parameters have dynamic ranges different by orders of magnitude, we linearly rescaled them as $T_0 \to \tilde{T}_0$ and $\gamma \to \tilde{\gamma}$, to fall in the same range $\sim (-1, 1)$. This bijective mapping ensures numerical stability of the point estimates by the network and is a common practice for deep learning regression schemes. The output covariance matrix (and its inverse) must be positive-definite as a mathematical requirement. This was ensured in our framework in a way similar to Fluri et al. 2019 by a Cholesky decomposition of the form

$$\tilde{\mathbf{C}}^{-1} = \mathbf{L}\mathbf{L}^T, \tag{4.5}$$

[3]Meaning that the machine can summarize the field most optimally (informed by the full data) into $N_{\rm par}$ values that can be directly mapped to the actual parameters of interest.

[4]Although this linear map is part of our network, it is not fitted during back-propagation in order to avoid the bias due to the prior limits.

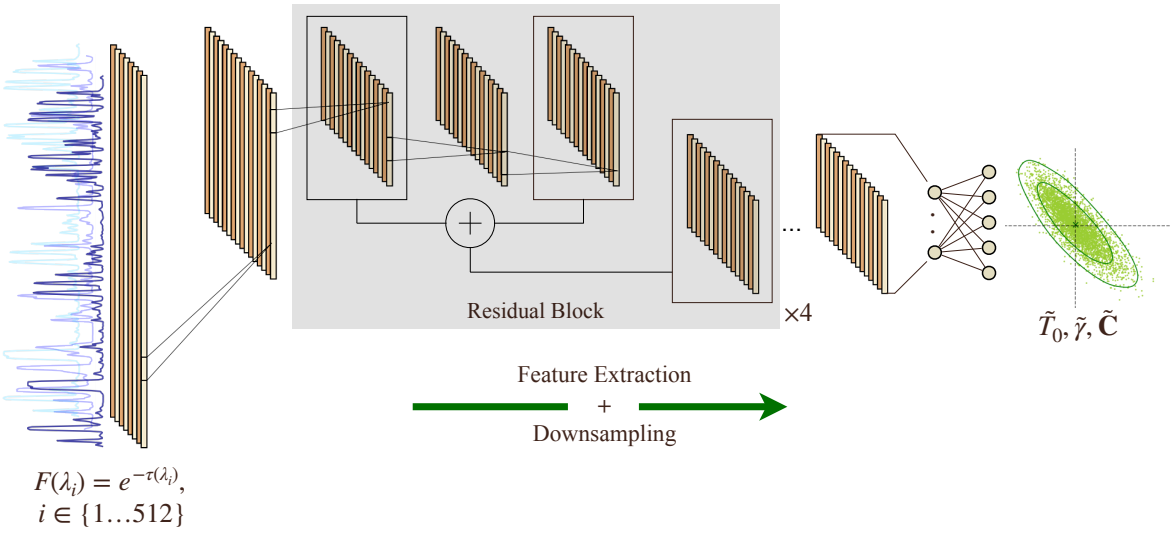

Figure 4.7: Schematic representation of the architecture of Sansa. It comprises a 1D residual convolutional neural network with four residual blocks in series for extracting crucial features from the spectra of size 512 pixels, followed by a fully connected layer to map the outcome to the parameter point predictions $\tilde{\boldsymbol{\pi}}$ and a covariance estimate $\tilde{\mathbf{C}}$ over 5 output neurons.

where $\mathbf{L}$ is a lower triangular matrix. Our network predicts the three independent components of that matrix ($\log L_{11}$, $\log L_{22}$, L_{12} to further ensure uniqueness[5]) rather than the covariance matrix directly. The network was optimized following a Gaussian negative log-likelihood loss (hereinafter NL3),

$$\mathcal{L}(\tilde{\boldsymbol{\pi}}) = \log|\tilde{\mathbf{C}}| + (\tilde{\boldsymbol{\pi}} - \hat{\boldsymbol{\pi}})\tilde{\mathbf{C}}^{-1}(\tilde{\boldsymbol{\pi}} - \hat{\boldsymbol{\pi}})^T, \tag{4.6}$$

where $\hat{\boldsymbol{\pi}} = (\hat{T}_0, \hat{\gamma})$ are the true parameter labels. This can be seen as an extension of the conventional mean squared error, MSE $= (\tilde{\boldsymbol{\pi}} - \hat{\boldsymbol{\pi}})(\tilde{\boldsymbol{\pi}} - \hat{\boldsymbol{\pi}})^T$, in the presence of a network estimated covariance. It can be noted that the covariance matrix does not have any labels and is primarily a way to regularize network predictions under a Gaussian likelihood assumption.

4.3.2 Architecture

We built our architecture using the open-source Python package TensorFlow/Keras (Chollet et al. 2015). The neural network for field-level inference with LyαNNA is called Sansa and consists of a 1D ResNet (He et al. 2015a), a residual convolutional neural network that extracts useful features from spectra and turns them into a "summary" vector which can then be used for inference of model parameters. Figure 4.7 shows a schematic representation of the architecture of Sansa. The input layer consists of 512 spectral units.

[5]Since $\tilde{C}_{ii}^{-1} = (\pm L_{ii})^2$, we explicitly chose the positive branch of the Cholesky coefficients (a unique, one-to-one mapping) via the lognormal transformation, for numerical stability reasons.

This input is passed through four residual blocks in series with varying numbers of input and output units, each block having the same computational structure as illustrated in Figure 4.8. Each residual block is followed by a batch normalization and an average-pooling layer to downsample the data vectors consecutively. A ResNet architecture is particularly attractive because of its ease of convergence during training owing to predominantly learning the residual mappings that could conveniently be driven to zero if identity maps are the most optimal in intermediate layers. This is achieved by introducing "skip-connections" in a sequential convolutional architecture that fulfill the role of identity (linear) functions. The residual blocks can more easily adapt to those linear mappings than having to train nonlinear layers to mimic them. A special advantage of the skip-connections is that they do not introduce more parameters than a sequential counterpart. Our neural network has a total of 136,784 trainable parameters that were tuned via back-propagation. We used the TensorFlow in-built leaky ReLU (rectified linear unit) function for all the nonlinear activations in the residual blocks with the negative-slope of 0.3. The resultant set of feature tensors is flattened into a single vector of size 128 and mapped to the output vector $(\tilde{T}_0, \tilde{\gamma}, \log L_{11}, \log L_{22}, L_{12})$ with a fully connected, unbiased linear layer. We regularized the network kernels with a very small L2 weight decay ($\mathcal{O}(10^{-8})$ in the convolutional layers, $\mathcal{O}(10^{-7})$ in the fully connected layer). We also used a dropout (Srivastava et al. 2014) of 0.09 after each residual block during training for encouraging generalization[6].

4.3.3 Training

All convolutional kernels in Sansa were initialized following the approach of Glorot and Bengio (2010) and the weights in the final linear layer were initialized similarly to He et al. (2015b). After a preliminary convergence test with respect to training dataset size, we chose a training set consisting of 10,000 distinct spectra from each thermal model in our sample. We also had a separate validation set for monitoring overfitting *viz.* 2/5 the size of the training set with an equivalent distribution of spectra among thermal models. The network was trained by minimizing the NL3 loss function in Eq. (4.6). Additionally, three other metrics were monitored during the training: $\log|\tilde{\mathbf{C}}|$, $\chi^2 = (\tilde{\boldsymbol{\pi}} - \hat{\boldsymbol{\pi}})\tilde{\mathbf{C}}^{-1}(\tilde{\boldsymbol{\pi}} - \hat{\boldsymbol{\pi}})^T$, and the MSE. We notice that the loss function is simply the sum of the first two metrics.

The training was performed by repeatedly cycling through the designated training dataset in randomly chosen batches of a fixed size. Each cycle through the data was deemed an "epoch", and each back-propagation action on a batch was termed a "step of training". Since the spectra follow periodic boundary conditions, a cyclic permutation of pixels ("rolling") is mathematically allowed and leads to no alteration of underlying physical characteristics (e.g., thermal parameters T_0, γ). This is also true for reverting the order of the pixels ("flipping"). These are some of the modifications that augment the existing training set and we expect our network to be robust against. Therefore, at every epoch we applied a uniformly randomly sampled amount (in number of pixels) of rolling

[6]We note that this value of the dropout rate p is consistent with the Keras convention, that is, the fraction of input layer units to drop, unlike the original definition by Srivastava et al. 2014 where p is the probability of the output of a given layer unit being propagated further in the network.

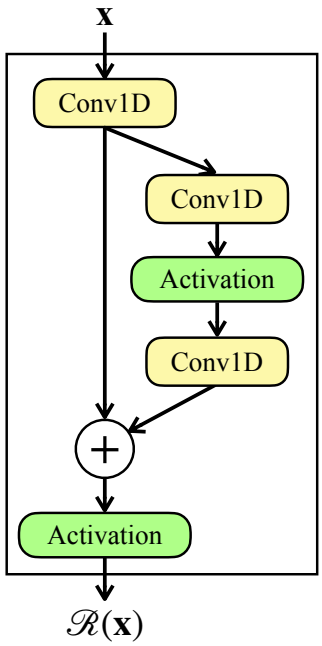

Figure 4.8: Residual block in SANSA. An input vector $\mathbf{x}$ is first passed through a convolutional layer and a copy of the output tensor is made which consecutively goes through a pair of convolutional layers introducing nonlinearity, all the while preserving the shape of the output tensor. The outcome is then algebraically added to the earlier copy (i.e., a parallel, identity function) and the sum is passed through a nonlinear activation to obtain the final outcome of the block. The latter two convolutional layers thus learn a residual nonlinear mapping. (Note that a zero-padding is applied during all convolutions in order to preserve the feature shape in the subsequent layers through the network.)

and a flipping with 50% probability to each of the training spectra, on the fly. We note that the validation set was not augmented on the fly because we would like to compare the generalization of the network predictions at different epochs for the same set of input spectra.

We expect the χ^2 metric to optimally take the value $\sim N_{\rm par}$ because of the underlying Gaussian assumption (in our case $N_{\rm par} = 2$). The improvement of the network during training is then in large parts due to a decrement in $\log|\tilde{\mathbf{C}}|$ which indicates that the network becomes less uncertain of its estimates as the training progresses. The state of a network is said to be improving if the value of NL3 decreases and the network χ^2 remains close to 2 for data unseen during back-propagation, the validation set. Therefore, we deemed the best state of the network to be occurring at the epoch during training at which the validation NL3 was minimal while the validation $|\chi^2 - 2| < \epsilon$, for a small $\epsilon = 0.05$[7]. We used the Adam optimizer (Kingma and Ba 2014) with a learning-rate of 5.8×10^{-4}. The Adam moment parameters had the values $\beta_1 = 0.97$ and $\beta_2 = 0.999$. We performed a Bayesian hyperparameter tuning for fixing the values of kernel weight decays, the dropout rate, the learning rate, and the optimizer moment β_1 parameter. We refer the reader to Appendix D for a further description of our strategy for choosing optimal hyperparameters for our network architecture and training. We present the progress of the network's training quantified by the four metrics mentioned above in Appendix E.

4.3.4 Ensemble learning

The initialization of our network weights (kernels) as well as the training over batches and epochs is a stochastic process. This introduces a bias in the network predictions that can be traded for variance in a set of randomly initialized and trained networks. Essentially, if the errors in different networks' predictions are uncorrelated, then combining the predictions of multiple such networks helps in improving the accuracy of the predictions. It has been shown that a "committee" of neural networks could outperform even the best-performing member network (Dietterich 2000). This falls under the umbrella of "ensemble learning".

Once we found an optimal set of hyperparameter values for SANSA, we trained 20 neural networks with the exact same architecture and the learning hyperparameters but initializing the network weights with different pseudo-random seeds and training with differently shuffled and augmented batches of the dataset. The output predictions by all the member networks of this committee of $N_{\rm SANSA} = 20$ neural networks were then combined in the form of a weighted averaging of the individual predictions to obtain the final outcome (this is commonly known as bootstrap aggregating or "bagging"; see, e.g., Breiman 2004). For a given input spectrum x, let $\mathcal{S}_i(x)$ denote the output point predictions by the ith network in our committee. Then the combined prediction of the committee is

$$\mathcal{S}(x) \propto \frac{1}{N_{\rm SANSA}} \sum_i \frac{1}{|\tilde{\mathbf{C}}_i(x)|} \mathcal{S}_i(x), \tag{4.7}$$

[7]The sample variance on χ^2 (at each epoch, assuming a χ^2-distribution) for the validation set can be expected to be $\sigma^2 = 2N_{\rm par}/N_{\rm validation} \sim (0.003)^2$.

where $\tilde{\mathbf{C}}_i(x)$ is the output estimate of the covariance matrix by the ith network for the input spectrum x. This combination puts more weight on less uncertain network predictions and thus is optimally informed by the individual network uncertainties. Even with such a small number of cognate members, we observed slight improvements with respect to the best-performing member as discussed in Appendix F. All the output point predictions by SANSA considered in the following part of the text are implicitly assumed to be that of the committee and not of an individual network unless specified otherwise.

4.3.5 Inference

We performed Bayesian inference of the model parameters with SANSA as well as the traditional summary statistics introduced in Section 4.2.3. In all the cases, we assumed a Gaussian likelihood and a uniform prior over the range shown in Figure 4.4. For inference with SANSA we created an emulator for a likelihood analysis in the following way. A test set of spectra[8] for a given truth $(\hat{T}_0, \hat{\gamma})$ were fed into SANSA and a corresponding set of parameter point estimates $(\tilde{T}_0, \tilde{\gamma})$ were obtained. Owing to our optimization strategy (described in Section 4.3.3) these network predictions have an inherent scatter that is consistent with a network covariance estimate $\tilde{\mathbf{C}}$. A mean point prediction $\bar{\boldsymbol{\pi}} = (\bar{T}_0, \bar{\gamma})$ and a covariance matrix were estimated from the scatter of the point estimates. This was performed for each of our 121 thermal models in the test sample. We then cubic-spline interpolated the mean network point prediction and the scatter covariance[9] over our prior range of thermal parameters $\boldsymbol{\pi}$ to obtain a model (emulator) $\big[\boldsymbol{\mu}(\boldsymbol{\pi}), \boldsymbol{\Sigma}(\boldsymbol{\pi})\big]$. The advantage of creating such a model for the likelihood is two-fold: (i) we can perform Bayesian inference with a different choice of prior (e.g., Gaussian) within the gray-shaded area of Figure 4.4, and (ii) the inference results of our machinery could, in principle, be combined with other probes of interest to further constrain our knowledge of the thermal state of the IGM. This emulator was then used to perform an MCMC analysis for getting posterior constraints with a likelihood function,

$$\log L_N(\boldsymbol{\pi}) \sim -\frac{1}{2}\Big[\log|\boldsymbol{\Sigma}_N(\boldsymbol{\pi})| - (\bar{\boldsymbol{\pi}}_N - \boldsymbol{\mu}(\boldsymbol{\pi}))\boldsymbol{\Sigma}_N{}^{-\mathbf{1}}(\boldsymbol{\pi})(\bar{\boldsymbol{\pi}}_N - \boldsymbol{\mu}(\boldsymbol{\pi}))^T\Big], \tag{4.8}$$

where $\bar{\boldsymbol{\pi}}_N$ is the mean network point prediction for a given set of N test spectra and $\boldsymbol{\Sigma}_N(\boldsymbol{\pi}) = \boldsymbol{\Sigma}(\boldsymbol{\pi})/N$ quantifies the uncertainty in the mean point estimate for the given dataset size[10]. We show this model in Figure 4.9. The model for the mean parameter values, $\boldsymbol{\mu}(\boldsymbol{\pi})$, consists of rather smoothly varying functions approximating $\mu_{1,2}(\pi_1, \pi_2) = \pi_{1,2}$, conforming to our expectation.

[8] We note that this was the same set of spectra as that used for the validation of generalization during network training (see Section 4.3.3).

[9] For computational simplicity, we actually interpolated the inverse of the scatter covariance matrix.

[10] Formulated thus (and due to the Gaussian likelihood assumption), N could be deliberately varied to mimic the inference outcome of a given size of the dataset.

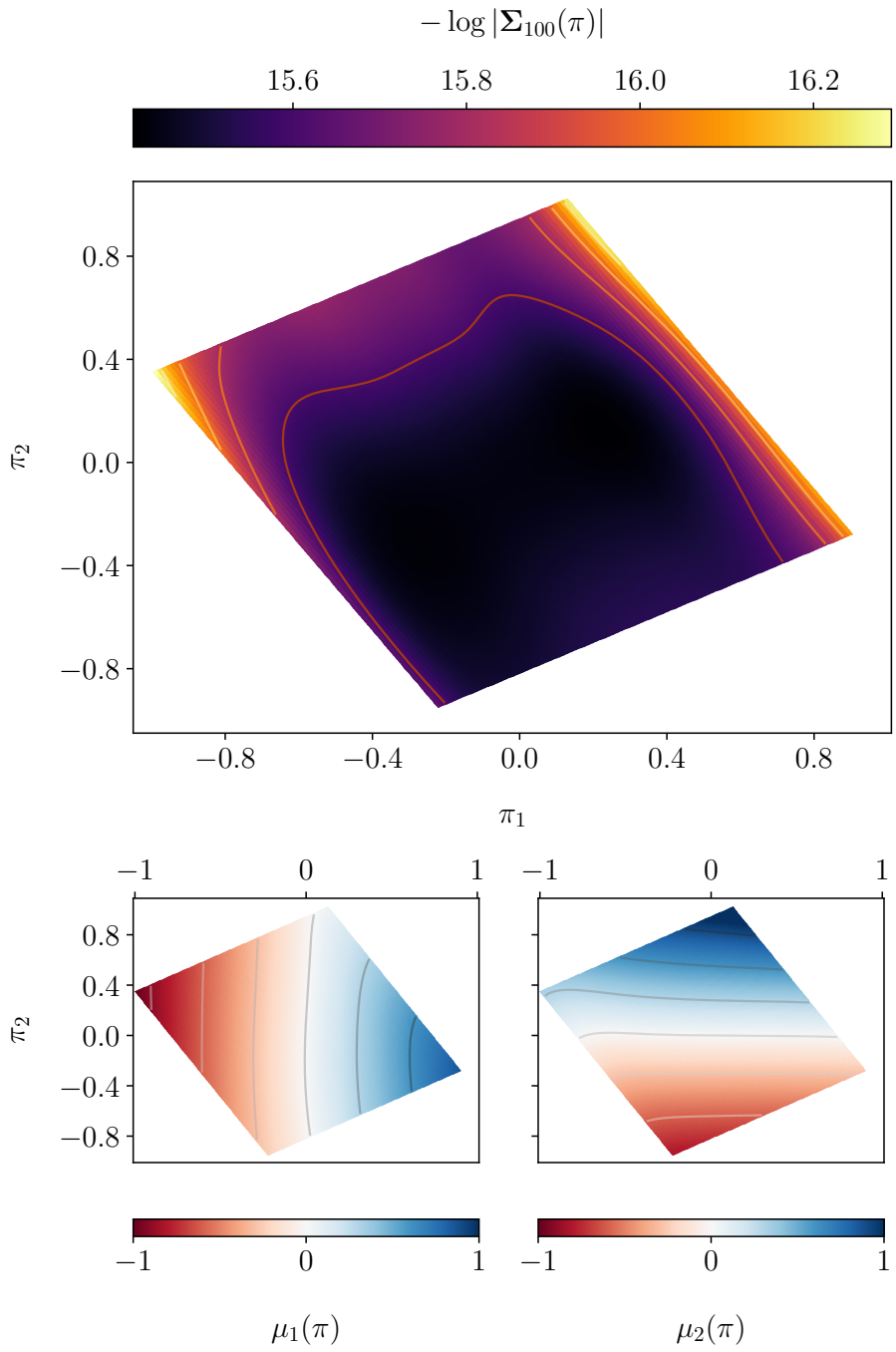

Figure 4.9: Likelihood model of inference with SANSA over our prior range for the rescaled $\boldsymbol{\pi}$ parameters. The top panel shows a measure of the covariance model $\boldsymbol{\Sigma}(\boldsymbol{\pi})$ and the bottom panels the model for the two rescaled parameters, $\mu_1 \equiv \tilde{T}_0$ and $\mu_2 \equiv \tilde{\gamma}$. The model for the mean summary vector fairly approximates $\mu_i(\boldsymbol{\pi}) = \pi_i$.

4.4 Results and discussion

In this section, we show the results of doing mock inference with our machinery as described in Section 4.3 and compare them with a summary-based approach (see Section 4.2.3 for more details on the summaries used). We investigated a few different test scenarios for establishing robustness of our inference pipeline. For each test case, we quantify our results in two chief metrics:
(i) precision, in terms of the area of posterior contours as a figure of merit (FoM),

$$\text{FoM} \sim 1/\sqrt{|\mathbf{C}|}; \tag{4.9}$$

and (ii) accuracy, in terms of a reduced χ^2,

$$\delta\chi^2_{\rm r} \equiv \langle \mathbf{\Delta C^{-1} \Delta}^T \rangle / N_{\rm par} - 1; \tag{4.10}$$

where $\mathbf{\Delta}^i = (\boldsymbol{\pi}^i - \hat{\boldsymbol{\pi}})$ is a point in the posterior MCMC sample, $\mathbf{C}$ is a covariance matrix of $\boldsymbol{\pi}$ estimated from the posterior sample and the average $\langle\rangle$ is taken over the entire sample. We note that in the two parameter case, the area of the posterior contours is proportional to $\sqrt{|\mathbf{C}|}$. We expect that the FoM improves when including more information (about the parameters of interest) in the underlying summary statistic from the transmission field, since the constraints get tighter (contours smaller) as a consequence. A smaller value of $\delta\chi^2_{\rm r}$ implies a more accurate recovery of the true parameters.

First, we considered a test set of spectra that are distinct from those used in training and validation to evaluate the performance of our inference pipeline for previously unseen data (we recall that the likelihood model for Sansa was built using the validation set). This set consisted of 4,000 spectra for the underlying true (fiducial) thermal model, $\hat{T}_0 = 10104.15$ K and $\hat{\gamma} = 1.58$ (we note that this model is off the training grid, as shown in Figure 4.4, affording us a test of the machinery's performance off-grid). To distinguish this set with the other test sets in the following, we call it the "original" set hereinafter. In Figure 4.10, we show the output scatter of point estimates by Sansa for the original test set, with contours of 68% and 95% probability. For comparison, we also plot the posterior contours (obtained by Sansa following the strategy outlined in Section 4.3.1), inflated to emulate the uncertainty equivalent of one input spectrum. A very good agreement is observed between both the cases, suggesting that a cubic-spline interpolation of the scatter covariance is a sufficiently good emulator for a likelihood analysis as discussed in Section 4.3.5. We performed inference with three further previously unseen sets of 4,000 random spectra with the fiducial TDR to establish the statistical sanity of our pipeline. We show the posterior contours obtained for those in Figure 4.11 and the metric values in Table 4.1.

Skewers can be picked along any of the three axes of the simulation box (see Section 4.2.2), each leading to a different realization of the Lyα transmission, mimicking cosmic variance. We expect our pipeline to be robust to the choice of axis along which the input skewers are extracted. We chose three different test sets of skewers along another

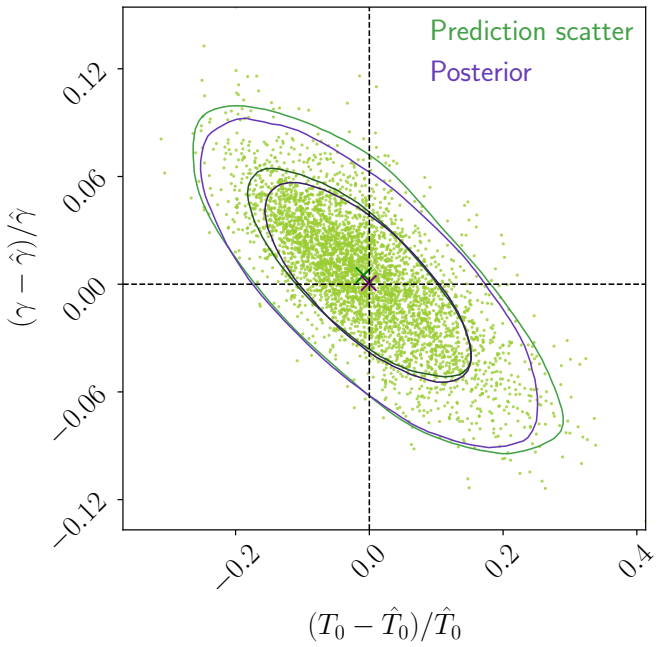

Figure 4.10: Scatter of point predictions for the original test set of spectra from our fiducial TDR model shown here with the 68% and 95% contours (green). The contours of the posterior distribution (purple) obtained by SANSA with the procedure outlined in Section 4.3.1, inflated to match the information equivalent to one spectrum, follow the scatter contours very closely. SANSA also recovers the true parameters (dashed) with a very good accuracy, as indicated by the mean of the point prediction scatter (green cross) as well as that of the posterior (purple cross).

Table 4.1: Comparison metric values for SANSA with four distinct sets of test spectra (for information equivalent to 100 spectra).

Test set	$\delta\chi^2_{\rm r}$	FoM / FoM(orig.)
#1 (orig.)	0.002	- -
#2	0.053	0.994
#3	0.059	1.004
#4	0.004	1.003

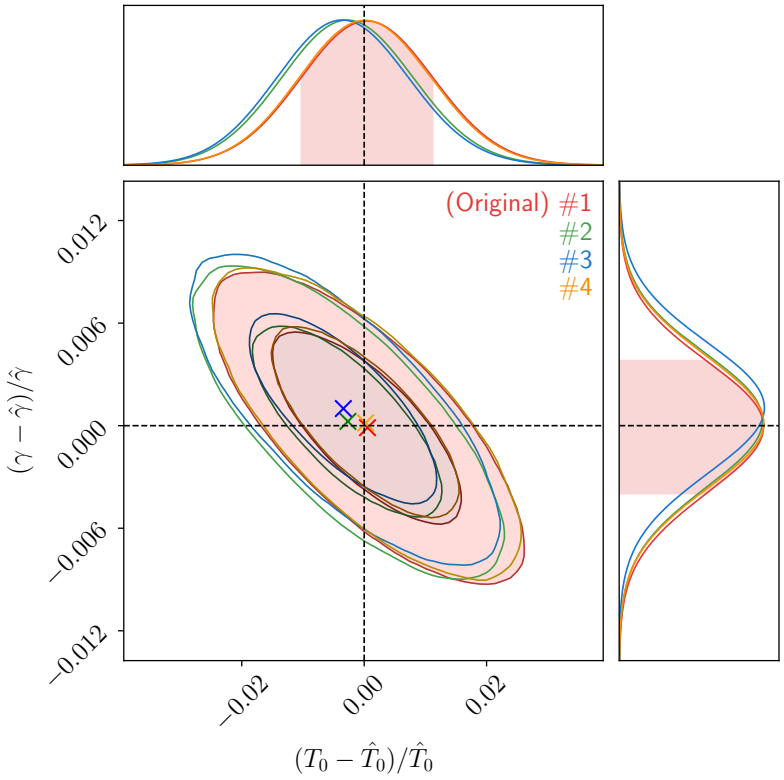

Figure 4.11: Comparison of posterior contours obtained for four different sets of 4,000 spectra, where #1 is the "original" (for information equivalent to 100 spectra).

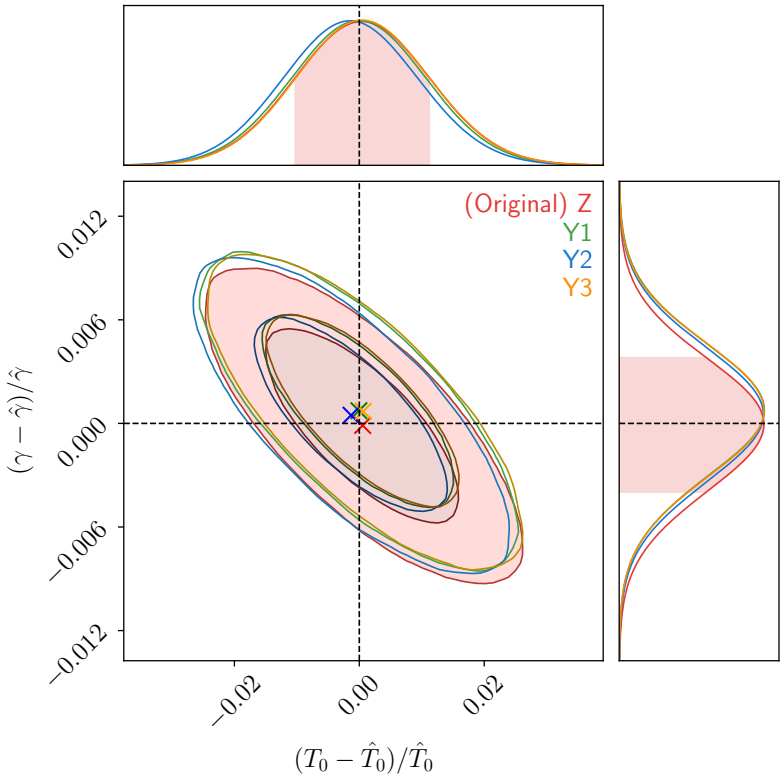

Figure 4.12: Comparison of posterior contours obtained for three different sets of 4,000 spectra computed along another axis ("Y") of the simulation box (for information equivalent to 100 spectra). "Z" corresponds to the "original."

axis ("Y") of our box that have the same underlying thermal parameters (fiducial). We estimated the posterior constraints for all three (Y1,2,3) datasets with Sansa and we show them in Figure 4.12, along with the "original" (skewers extracted along the Z-axis of the simulation box). The corresponding metric values are listed in Table 4.2. We observe a statistically consistent posterior distribution in each of the three Y-extracted test cases with the original case, indicating that Sansa is agnostic to the choice of LOS direction, even though it was trained only with one of the three possibilities.

Moreover, we tested our inference machinery with numerically modified (augmented) spectra, as discussed in Section 4.3.3. In one case, we applied a cyclic permutation of the pixels (rolling) by a random amount (between 1 and 512 pixels) to each spectrum of our original test set. We denote this by "rolled". In a second, we flipped an arbitrary 50% of the original test set of spectra, denoted by "flipped". We generated another set of spectra with a random mix of both of the above operations applied to the original set; this is labeled "mixed". We present the posterior constraints for all of these cases in Figure 4.13 and list the metric values in Table 4.3. The posterior constraints in all the augmented test scenarios agree very well with each other and with the original test case, establishing robustness of the inference against such degeneracies.

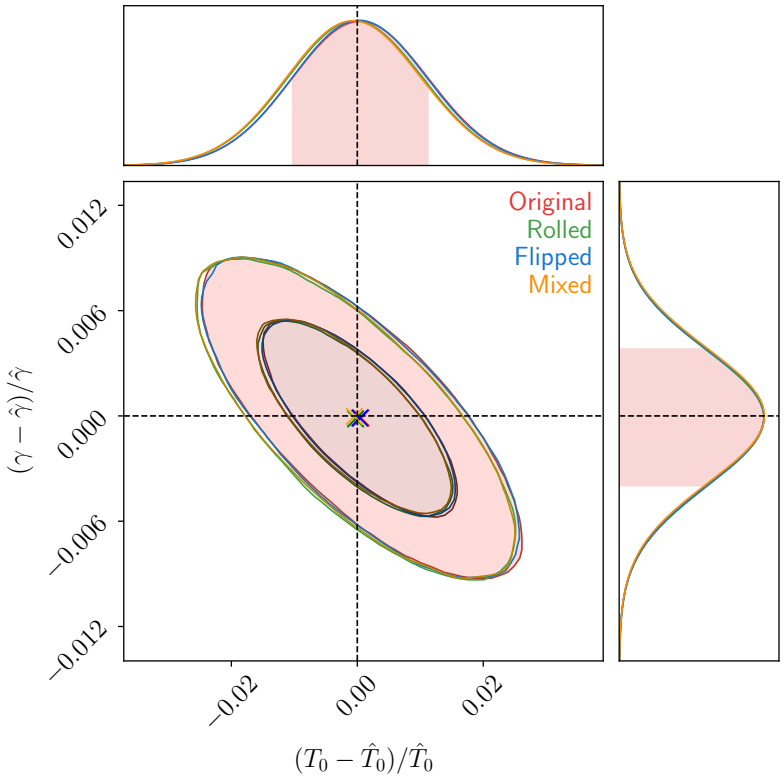

Figure 4.13: Comparison of posterior contours for differently augmented test spectra. In the "rolled" case, a uniform random amount (between 1 and 512 pixels) of cyclic permutation of the pixels is applied to each spectrum in the original test set. An arbitrary 50% of spectra from the original test set are flipped (mirrored) in the "flipped" case. A random mix of both is applied in the "mixed" case. All of the contours carry information equivalent to 100 skewers. A mean (expectation) value of all the posterior distributions is also shown with a cross of the corresponding color. The posterior contours for all the cases agree extremely well with the original test case.

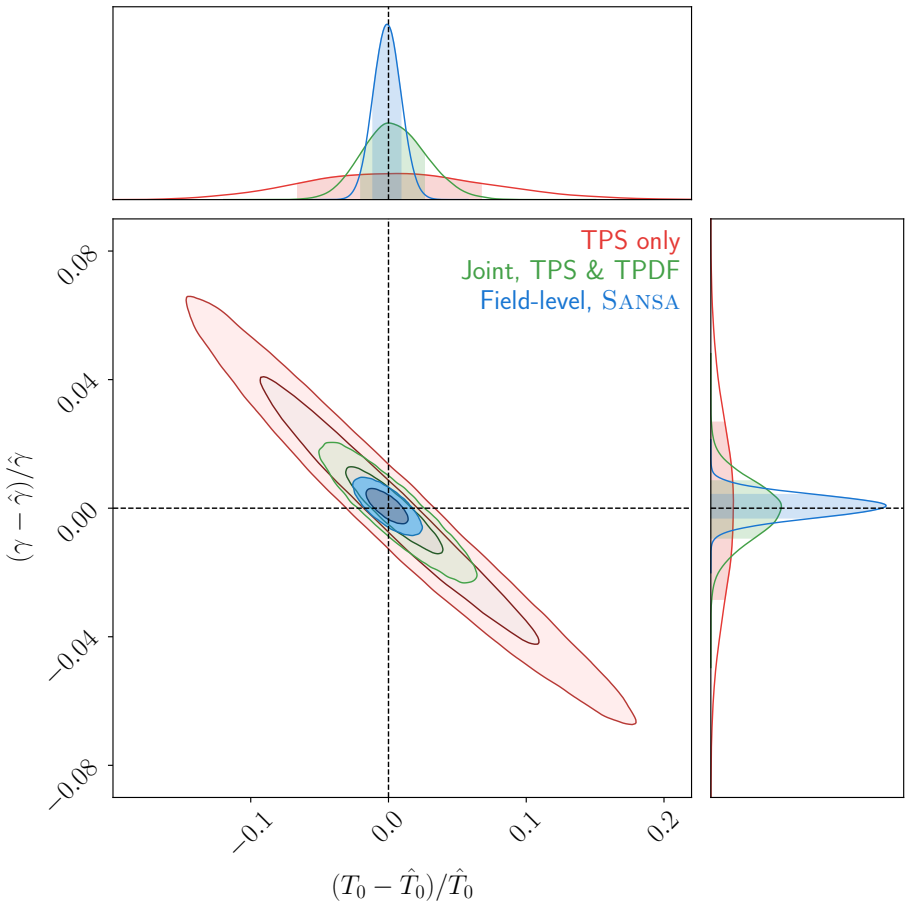

Figure 4.14: Posterior contours obtained by SANSA for the underlying fiducial thermal model. The posterior contours from the two traditional summary statistics are shown for comparison: (i) TPS only and (ii) joint constraints of TPS and TPDF where the cross-correlations of the summaries are accounted for by a joint covariance matrix. In terms of the size of the contours, the SANSA constraints are tighter than the TPS-only ones by a factor 10.92 and the joint constraints by 3.30, corroborating the claim that SANSA recovers relevant information for inference that is not carried by the TPS and/or the TPDF. All the cases carry information equivalent to 100 spectra.

Table 4.2: Comparison metric values for Sansa with spectra along different axes of the simulation (for information equivalent to 100 spectra).

Test set	$\delta\chi^2_{\rm r}$	FoM / FoM(orig.)
Original	0.002	- -
Y1	0.043	0.998
Y2	0.010	0.997
Y3	0.064	1.003

Table 4.3: Comparison metric values in data-augmentation scenarios for Sansa (for information equivalent to 100 spectra).

Test set	$\delta\chi^2_{\rm r}$	FoM / FoM(orig.)
Original	0.002	- -
Rolled	0.005	0.997
Flipped	0.001	1.001
Mixed	0.003	1.003

Finally, we compared the inference outcome of Sansa with the traditional summary statistics (TPS and TPDF) based procedure. We present the posterior constraints on $(\boldsymbol{\pi}-\hat{\boldsymbol{\pi}})/\hat{\boldsymbol{\pi}}$ obtained by a MCMC analysis of TPS only, TPS and TPDF jointly, and Sansa for the fiducial thermal model in Figure 4.14. Evidently, the joint constraints of the two summary statistics are tighter than the TPS-only case as there is more information of the thermal parameter in the former. However, by far the field-level constraints by Sansa are tighter than both the traditional summary statistics cases, namely, a factor of 10.92 compared to the TPS-only case and a factor of 3.30 compared to the joint constraints in our FoM. Indeed, the TPS is only a two-point statistic of the transmission field that has a highly non-Gaussian one-point PDF itself. Combining TPS and TPDF provides some more leverage, however, it still fails to account for some relevant parts of the information for inference. As illustrated by Figure 4.14, Sansa provides a remedy to the lost information by trying to optimally extract all the features of relevance at the field level.

4.5 Conclusion

We built a convolutional neural network called Sansa for inference of the thermal parameters (T_0, γ) of the IGM with the Lyα forest at the field level. We trained this using a large set of mock spectra extracted from the Nyx hydrodynamic simulations. For estimating posterior constraints, we created a reasonably robust pipeline that relies on the point predictions of the parameters and the uncertainty estimates by our neural network and that can in principle be easily combined with multiple other probes of the thermal state of the IGM. A comparison of our results with those of traditional summary statistics (TPS and TPDF in particular) revealed an improvement of posterior constraints in area

of the credible regions by a factor 10.92 with respect to TPS-only and 3.30 with respect to a joint analysis of TPS and TPDF. We established statistical robustness of our pipeline by performing tests with a few different sets of input spectra.

However, our neural network that is trained with noiseless mock spectra for inference fails for spectra with even very small noise (as small as having a continuum-to-noise ratio of 500 per 6 km/s). Indeed, our framework must be adapted for use with noisy spectra by retraining Sansa with datasets containing artificially added noise that varies on the fly during training (to prevent learning from the noise).

Furthermore, in this work we have assumed a fixed underlying cosmology for generating the various datasets used for training and inference. However, the Lyα forest carries information about the cosmological parameters that may be correlated with the thermal properties, and as a consequence, our machinery would exhibit a bias if the cosmology of the training data were not equal to that of the test case. A generalized pipeline would thus require marginalization over the cosmological parameters. Our proof-of-concept analysis, however, opens up an avenue for constraining cosmological parameters at the field level as well.

Baryonic feedback from AGN and supernovae (and similarly, inhomogeneous reionization) affects the phase-space distribution of the IGM and is thus expected to influence the performance of our neural machinery. Nonetheless, we anticipate this impact to be marginal since the network is more sensitive to the power-law regime of the diffuse gas which still holds to a large degree.

As described in Section 4.2.2, we used snapshots of the skewers to train our pipeline instead of accounting for a lightcone evolution of the IGM properties along the LOS pixel-by-pixel. We performed a test of this framework with an approximate model of such evolution; see Appendix A. For the length of our skewers, we expect the actual lightcone evolution of the gas properties to be marginal and as such unproblematic for network inference.

Nevertheless, in the spirit of creating a robust pipeline for highly realistic spectral datasets, a plethora of physical and observational systematic effects (such as limited spectral resolution, sky lines, metal absorption lines, continuum fitting uncertainty, damped Lyα systems on the observational side; and lightcone effects, baryon feedback, cosmological correlations on the modeling side) must also be incorporated in the training data. This warrants a further careful investigation into training supervised deep learning inference algorithms with a variety of accurately modeled systematics added to our mock Lyα forest datasets and we plan to carry it out in future works.

Chapter 5

Deep learning inference: increased data realism

Bibliographic and copyright information

Appendices to the original paper have been moved to the end of the thesis.

P. Nayak, M. Walther, and D. Gruen (Oct. 2025); "LyαNNA II: Field-level inference with noisy Lyα forest spectra". arXiv:2510.19899.

DOI: 10.48550/arXiv.2510.19899

In this work, we establish deep learning as a competitive tool for field-level inference with noisy real-world Lyα forest datasets. We characterize the impact of noise level on the posterior precision of the neural network with an adapted neural architecture and re-optimized hyperparameters. While the gain in posterior precision over traditional summaries is somewhat reduced in the realistic case, the trend of improvement remains consistent with that in the theoretical scenario of Chapter 4. We also perform density estimation likelihood-free inference (DELFI) and make comparisons with the Gaussian likelihood-based inference. As the lead author, I completed the following key tasks: creating all the mock training/validation/test datasets, developing the core inference framework, writing the vast majority of code in Python, optimizing hyperparameters and training large committees, estimating joint data-parameter density for DELFI, formulating and computing noise-corrective estimators of the traditional summaries, running tests with those estimators as well as with the NN framework. I also wrote the majority of the manuscript and produced the various illustrations/plots. My coauthors provided valuable scientific discussions to help devise the framework. They reviewed the draft of the manuscript and suggested improvements. M. Walther ran the suite of simulations used in this work.

5.1 Introduction

The interplay of the continua of quasars' emission and the intergalactic medium (IGM) imprints a unique and powerful absorption feature on their observed spectra called the Lyα forest (Lynds 1971). The term refers to a dense cluster of resonant Lyα absorption lines of Hydrogen blueward of a quasar's observed Lyα emission peak, spread over a wide range of wavelengths as a result of the expansion of the universe which causes the cosmological redshift. Since Hydrogen is the most abundant element in the intergalactic gas, the Lyα forest provides a continuous one dimensional probe of the IGM along the quasar line of sight.

Individual high resolution spectroscopic observations of quasars using state-of-the-art instruments such as VLT/UVES (e.g., Murphy et al. 2019), VLT/XSHOOTER (e.g. López et al. 2016) and Keck/HIRES (e.g., O'Meara et al. 2017), as well as large cosmological spectroscopic surveys such as the extended Baryon Oscillation Spectroscopic Survey (eBOSS, Dawson et al. 2013) and most recently the Dark Energy Spectroscopic Instrument (DESI, DESI Collaboration et al. 2022) have opened a spectroscopic treasure trove of cosmological information such as submegaparsec scale nonlinearities in the cosmic matter distribution, and the thermal and ionization state of the IGM. These Lyα forest observations have been widely used to constrain the thermal history of the IGM and the epoch of reionization (e.g., Becker et al. 2011; Walther et al. 2019; Boera et al. 2019; Gaikwad et al. 2021; Bosman et al. 2022), the baryon acoustic oscillation (BAO) scale (e.g., Slosar et al. 2013; Busca et al. 2013; du Mas des Bourboux et al. 2020; Abdul Karim et al. 2025), and the properties of dark matter (e.g., Viel et al. 2005; Iršič et al. 2017b; Armengaud et al. 2017; Kobayashi et al. 2017; Rogers and Peiris 2021). When combined with external probes such as the cosmic microwave background (CMB, e.g., Planck Collaboration et al. 2020), they provide valuable insights into cosmic inflation and the mass of neutrinos (e.g., Seljak et al. 2006; Palanque-Delabrouille et al. 2015; Yèche et al. 2017; Palanque-Delabrouille et al. 2020; Sarkar and Sethi 2024; Ivanov et al. 2025).

The classical way of gaining insights from the Lyα forest is through human defined summary statistics. For instance, the 1D power spectrum, a measure of the absorption structure (and hence the underlying IGM) across length scales, is one of the most widely used summary statistics in literature for parameter inference (e.g., Croft et al. 1998; Theuns et al. 2000; McDonald et al. 2000; Chabanier et al. 2019; Walther et al. 2019; Boera et al. 2019; Ravoux et al. 2025; Karaçaylı et al. 2025b). Other examples of summary statistics include the probability density function (PDF) of the Lyα transmission (e.g., Theuns et al. 2000; McDonald et al. 2000; Bolton et al. 2008; Viel et al. 2009; Lee et al. 2015), curvature statistics (e.g., Becker et al. 2011; Boera et al. 2014), and wavelet statistics (e.g., Meiksin 2000; Theuns and Zaroubi 2000; Zaldarriaga 2002; Lidz et al. 2010; Wolfson et al. 2021; Tohfa et al. 2024). However, those summary statistics do not efficiently compress the information of the non-Gaussian Lyα forest transmission field, restricting the constraining power of the full data.

In recent years, machine learning methods are gaining popularity for the task of parameter inference in cosmology due to their ability to optimally extract and compress relevant

non-Gaussian features from the underlying field. Presently, weak gravitational lensing is one of the most common cosmological probes with machine learning inference applications in the literature (e.g., Gupta et al. 2018; Fluri et al. 2019; Ribli et al. 2019; Lu et al. 2022; Fluri et al. 2022; Lucas Makinen et al. 2025; Jeffrey et al. 2025). For the Lyα forest, the avenue of field-level neural network inference remains relatively unexplored with only a handful of works (e.g., Nayak et al. 2024; Maitra et al. 2024; Nasir et al. 2024). In Nayak et al. (2024, hereafter N24), we demonstrated the potential of deep learning for parameter inference from the Lyα forest spectra for the first time. The field-level neural network (NN) framework we presented therein (called LYαNNA) surpasses the traditional, human defined summaries in the constraining power significantly, as substantiated by Chang et al. (2025, hereafter C25).

The factors-of-a-few improvement by using LYαNNA as presented in N24 and C25, however, is a proof-of-concept, in the regime with no realistic nuisances modeled on top of the pure simulations of the Lyα forest. In real observations, a plethora of astrophysical and instrumental systematics go hand in hand with the cosmological signal of interest. These include, but are not limited to, instrumental noise and spectral resolution, sky lines, uncertainties in the estimation of the intrinsic quasar continua, high column density systems (HCDs), and intergalactic metal absorption (see, e.g., Karaçaylı et al. 2025a, for a discussion of systematics in the context of DESI). Each of these effects acts as a contaminant to the Lyα forest, potentially compromising the full constraining power of the signal. They may add uncertainty to the data (thereby reducing inference precision) and they may also bias the summary vectors (and thus the posterior distributions). With a theoretical understanding of the impact of some of these effects (e.g. noise, resolution) on some of the traditional summaries, their interference may be corrected, allowing us to make the most of our contaminated datasets. For instance, noise in the Lyα forest spectra adds a measurable bias to the power spectrum, and spectral resolution dampens the power on small scales with a known kernel (see, e.g., Walther et al. 2018). However, for inference with machine-learned summaries such as LYαNNA, we lack such understanding.

In this work, we explore field-level inference with deep learning using contaminated Lyα forest datasets. In particular, we analyze the LYαNNA framework in the presence of noise and limited spectral resolution—two of the strongest contaminants in spectroscopic datasets —and investigate whether the same amount of improvement may be gained from such observations as for pure simulations while inferring the amplitude and slope of the power-law temperature density relation of the IGM. We retrained the pipeline described in N24 with a range of signal-to-noise ratio values in the training data and an adapted architecture to better process them. After an extensive hyperparameter search rooted in Bayesian optimization, we trained a committee of 2200 NNs with the best performing architecture for ensemble learning (Dietterich 2000). We built an emulator of our NN compressed summary vector for likelihood-based inference as well as a Gaussian mixture model (GMM) surrogate of the joint density of the NN summary and the true parameter values for likelihood-free inference (LFI, see e.g. Grazian and Fan 2019 for a review) and made comparisons between them. For a Gaussian likelihood case, we compared the inference outcome of our machinery to that of the 1D power spectrum and the transmission

PDF to assess the improvement of NN for noisy spectra.

We organize this paper as follows. In Section 5.2 we describe the hydrodynamical simulations and the mock datasets we used for training our machinery, including the realistic effects modeled. In Section 5.3 we present a detailed overview of the NN machinery used in this work. In Section 5.4 we describe our approach to Bayesian inference with our machinery including LFI. We show and discuss our results of using this framework in Section 5.5 with a comparison to some of the traditional statistics. We conclude with a summary of this work and an outlook in Section 5.6.

5.2 Simulations

5.2.1 Hydrodynamic simulations and mock spectra

For creating the mock Lyα forest spectra used in this work, we used a Nyx cosmological hydrodynamic simulation snapshot from the Lyssa suite described in Walther et al. (2025) generated primarily for eBOSS and DESI Lyα forest analyses. Nyx is a hydrodynamics code that simulates baryons as an ideal fluid on an Eulerian grid, gravitationally coupled with N-body dark matter particles. In our simulation, initial conditions are generated at $z = 99$ using second order Lagrangian perturbation theory. The gas is evolved by solving the Euler equations using a second-order accurate scheme. Nyx models all the key physical processes affecting the IGM and hence the Lyα forest such as collisional and photoionization, recombination, photoheating and cooling of the gas that has a primordial composition of H and He (for more details see Almgren et al. 2013 and Lukić et al. 2015). In our simulation, a spatially flat, time-varying ionizing ultraviolet background (UVB) was applied following the late reionization model of Oñorbe et al. (2017), and the radiative and dielectric recombination, collisional ionization and cooling rates from Lukić et al. (2015) were used. Our snapshot has a comoving side length of 80 Mpc/h and contains 4096^3 volumetric cells (aka voxels) and dark matter particles. This snapshot is the "fiducial" run from Walther et al. (2025) at $z = 2.2$ with the cosmological parameters based on Planck Collaboration et al. (2020), namely, $h = 0.6732$, $\Omega_\mathrm{m} = 0.3144$, $\Omega_\mathrm{b} = 0.0494$, $A_\mathrm{s} = 2.101 \times 10^{-9}$, $n_\mathrm{s} = 0.9660$.

The diffuse IGM that constitutes the bulk of baryonic gas in the universe follows a tight power law temperature-density relation (TDR) characterized by

$$T = T_0(\rho_\mathrm{b}/\bar{\rho}_\mathrm{b})^{\gamma-1}, \tag{5.1}$$

where T_0 is a temperature at the cosmic mean gas density $\bar{\rho}_\mathrm{b}$ and γ is the adiabatic power-law index (Hui and Gnedin 1997). The fiducial values of these two parameters in our snapshot are $T_0 = 10765$ K and $\gamma = 1.57$ (estimated by a linear least-squares fitting in the range $-0.5 < \log_{10}(\rho_\mathrm{b}/\bar{\rho}_\mathrm{b}) < 0.5$ and $\log_{10}(T/\mathrm{K}) < 4$).

We used an original Python package called SynTH[1] for the generation of mock spectra from the simulation snapshot that is different from the pipelines used by Lukić et al.

[1]Synthesis of Transmission from Hydrodynamical simulations; https://github.com/par-nay/synth.

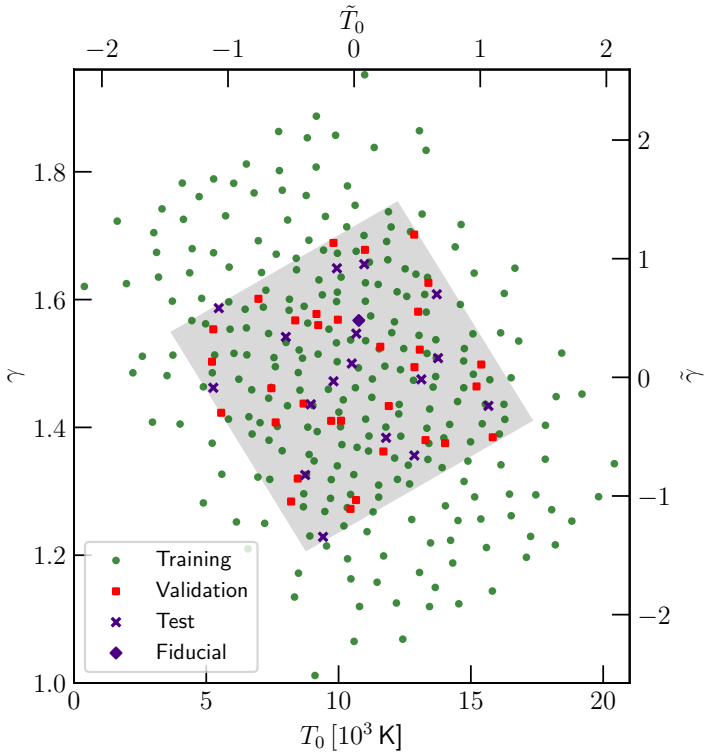

Figure 5.1: The sample of training, validation, and test labels in our mock dataset along with the fiducial TDR model in the (T_0, γ) as well as the rescaled $(\tilde{T}_0, \tilde{\gamma})$ space. The gray shaded region indicates our prior for the likelihood analysis as well as for the density estimation likelihood free inference. The exact sampling strategy is described in Appendix G.

(2015) and Walther et al. (2025). We rescaled the gas temperatures at fixed densities as a post-processing step to populate the (T_0, γ) parameter space—and account for the scatter off the power law relation—for creating labeled datasets for our supervised learning framework, same as N24. We also applied a rescaling to our parameter labels $T_0 \to \tilde{T}_0$ and $\gamma \to \tilde{\gamma}$, similar to N24—such that $\tilde{T}_0$ and $\tilde{\gamma}$ have the same dynamic range—for ensuring numerical stability during NN training. We sampled the training set, a validation set for hyperparameter tuning, and a smaller test set in the joint $\tilde{\boldsymbol{\pi}} = (\tilde{T}_0, \tilde{\gamma})$ space, all using Sobol' sequences (Sobol' 1967). The training set contains 256 distinct parameter labels, the validation set 32, and the test set 16. For each of these labels, we produced 100,000 spectra (from the same 100,000 physical skewer positions across labels) for various training and testing purposes using the exact same procedure as in N24. Figure 5.1 shows a scatter of our training, validation, and test labels. We describe the sampling strategy in more detail in Appendix G. We rescaled all the optical depth τ with a constant factor (per distinct thermal model) determined by a linear least-squares fitting in order to match the resulting mean transmission with its observed value of $\bar{F}_{\rm obs}(z = 2.2) = 0.86$ as measured by Becker et al. (2013).

5.2.2 Spectral resolution and noise

We applied a Gaussian filter in Fourier space to our spectra corresponding to a spectroscopic resolving power of $R_{\rm FWHM} = \lambda/\Delta\lambda_{\rm FWHM} = c/(2\sigma_v\sqrt{2\log 2}) = 6000$ with the kernel $\mathcal{R}_k = \exp(-k^2\sigma_v^2/2)$. We additionally applied a Heaviside step function kernel to null the Fourier modes $k > k^*$, $k^* \approx 0.094$ s/km corresponding to spectra observable with, e.g., VLT/XSHOOTER. We then rebinned the spectra (in real space) with a 16-pixel average to have the final spectrum size of $N_{\rm pixels} = 256$ and each pixel of size $\Delta v_{\rm p} = 32.37$ km/s.

In this work we consider a simple model of homoscedastic[2] Gaussian noise in our spectra, $\tilde{F} = F + \epsilon$, $\epsilon \sim \mathbb{N}(0; \sigma_{\rm p})$. (We note that $\sigma_{\rm p}$ may vary across spectra in our datasets.) We consider a range of continuum-to-noise ratio (CNR) values between 10 and 500 per 6 km/s, with a uniform prior distribution. This range encompasses the noise levels typical of TPS measurements from targeted quasar samples (e.g., Iršič et al. 2017a; Walther et al. 2019). In the Poisson photon counting limit, the noise level $\sigma_{\rm p} = \sigma_6\sqrt{6/\Delta v_{\rm p}}$ and $\sigma_6 = 1/{\rm CNR}_6$, where σ_6 and ${\rm CNR}_6$ correspond to the values for a pixel of size 6 km/s. We note that $\log\sigma_{\rm p} \in [-7.06, -3.14]$.

5.2.3 Traditional summary statistics

We consider two widely used traditional summary statistics of the Lyα forest, the transmission power spectrum (TPS) and the transmission PDF (TPDF). We perform individual and joint analyses of these statistics with noise and spectral resolution effects taken into account and obtain posterior constraints on the parameters (T_0, γ) for comparison with our NN machinery.

[2]The noise level on our pixels $\sigma_{\rm p}$ is fixed within each spectrum.

The TPS is defined as the variance of the Fourier modes of the transmission contrast field, $\tilde{\delta}_F(k)$, properly normalized. We first measured the TPS from low resolution, noisy spectra (called $\hat{\mathbb{P}}_{\rm raw}$) and then applied the following corrections to obtain an estimate $\hat{\mathbb{P}}_{\rm pure}$ of the pure TPS $\mathbb{P}_{\rm pure}$,

$$\hat{\mathbb{P}}_{\rm pure}(k) = \frac{\hat{\mathbb{P}}_{\rm raw} - \hat{\mathbb{P}}_{\rm noise}}{\mathcal{W}_k^2 \mathcal{R}_k^2}, \tag{5.2}$$

where the noise power spectrum $\hat{\mathbb{P}}_{\rm noise}$ is measured from 10^6 random realizations of the noise vector with a known noise level and $\mathcal{W}_k = \sin(k\Delta v_{\rm p}/2)/(k\Delta v_{\rm p}/2)$ is the pixel window function. Since we cut off the modes $k > k^*$, the summary vector $\hat{\mathbb{P}}_{\rm pure}$ is of size 122.

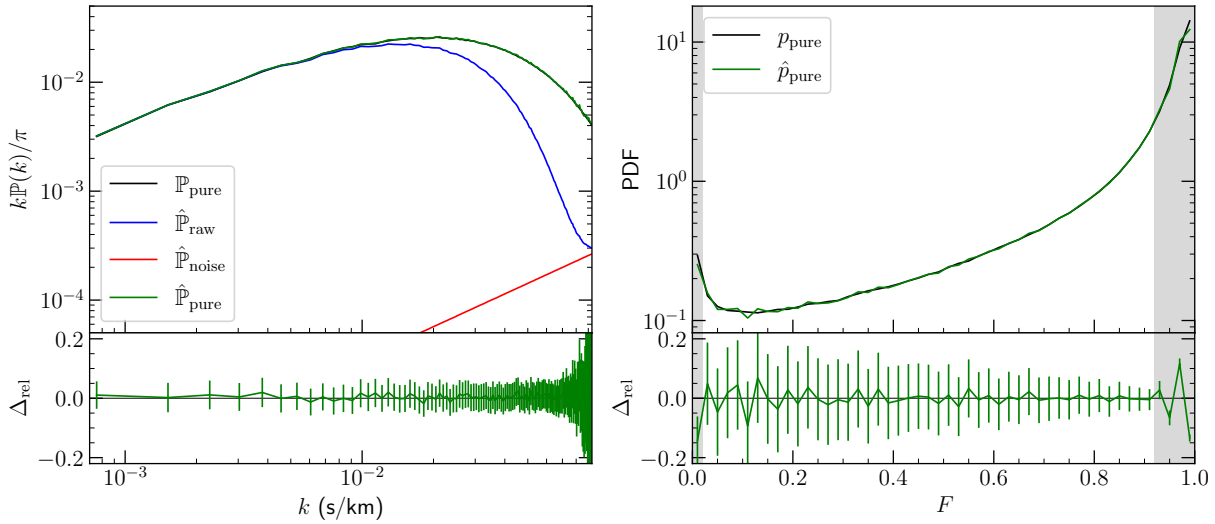

Figure 5.2: Traditional summary statistics (TPS on the left, TPDF on the right) estimated for the fiducial thermal state of our simulation box and with the mean transmission fixed to its observed value. The raw estimators correspond to a noise level of $\mathrm{CNR}_6 = 30$ ($\sigma_{\rm p} = 0.014$). The corresponding pure statistics are computed from 100,000 noiseless spectra (with infinite spectral resolution in TPS, with a resolution $R_{\rm FWHM} = 6000$ for TPDF). The errors correspond to $N_{\rm s} = 100$ spectra. The gray regions in TPDF correspond to our cuts due to edge effects in the deconvolved estimator.

We estimate the TPDF from noisy spectra in the following way. We first compute the PDF $\hat{p}_{\rm raw}$ of the noisy $\tilde{F}$ as a histogram per spectrum in equal width bins ($\Delta F = 0.02$) in the range $-0.1 \leq \tilde{F} \leq 1.1$, which is a convolution of the pure PDF $p_{\rm pure}$ and $p_{\rm noise} \equiv \mathbb{N}(0; \sigma_{\rm p})$. It is normalized so that $\int_{-\infty}^{\infty} \hat{p}(F)\, dF = 1$. We then perform Wiener deconvolution (Wiener 1949) of $\hat{p}_{\rm raw}$ and $p_{\rm noise}$ to obtain an estimate of the pure PDF, $\hat{p}_{\rm pure}$. Wiener deconvolution aims to estimate a filter $\mathcal{G}$ in the Fourier space such that

$$\hat{\mathcal{P}}_{\rm pure} = \mathcal{G} \cdot \hat{\mathcal{P}}_{\rm raw} = \frac{\mathcal{N}^*}{|\mathcal{N}|^2 + \gamma} \hat{\mathcal{P}}_{\rm raw}, \tag{5.3}$$

where $\mathcal{P}$, $\mathcal{G}$, $\mathcal{N}$ are quantities in Fourier space and γ is a regularization parameter that we tune empirically to 10^{-2} to suppress numerical noise amplification. Finally, we crop the

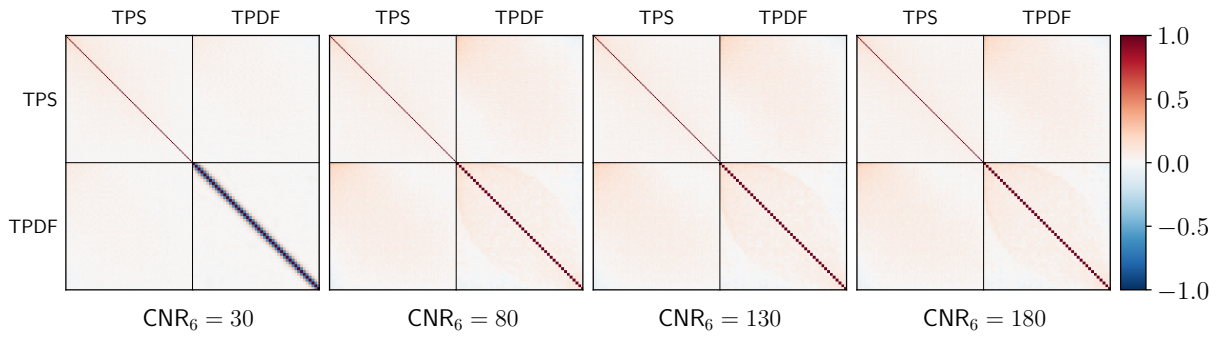

Figure 5.3: The full correlation matrices of the concatenated summary vector for four different noise levels. The TPS is the corrected vector $\hat{\mathbb{P}}_{\rm pure}$ with 122 k-modes and the TPDF is the deconvolved, cropped $\hat{p}_{\rm pure}$ with 45 bins. For each of the four matrices, the two blocks on the principal diagonal are the individual correlation matrices of TPS (top left) and TPDF (bottom right) and the other two blocks show the cross correlation of TPS and TPDF. The different blocks have been resized differently to have an equal area on the plot.

PDF to the range $F \in [0.02, 0.92]$ for removing edge effects such that the final $\hat{p}_{\rm pure}$ vector contains 45 bins.

In Figure 5.2 we show our estimates of the TPS (left) and TPDF (right) for our fiducial thermal model and for $\mathrm{CNR}_6 = 30$. For the joint analysis of TPS and TPDF we concatenate the two corrected vectors to have a full summary vector of size 167. We show the full correlation matrices of this joint summary vector (corrected) for four different noise levels in Figure 5.3. For a high noise level of $\mathrm{CNR}_6 = 30$, a significant (anti-)correlation among closeby F bins in the TPDF can be seen. This is potentially a result of numerical side effects introduced by the deconvolution with a wide noise PDF. The cross correlations between TPS and TPDF, although mild for all noise levels, are important to account for in the likelihood analysis as found by C25.

5.3 NN machinery

5.3.1 Architecture

Our machinery is based on a 1D ResNet (He et al. 2015a) convolutional architecture. Over a sequence of six residual blocks similar to the ones in N24, useful features are extracted from noisy input spectra and turned into a flattened feature vector. The (rescaled[3]) value of $\sigma_{\rm p}$ is concatenated with this vector as a query component and it is then fed into a multilayer perceptron (MLP) of one nonlinear hidden layer with $N_{\rm d} = 85$ dense nodes and the final linear output layer of 5 nodes for the summary point estimates and the independent components of their covariance matrix. We build this architecture using TensorFlow/Keras and name it nSansa. It contains 344,952 trainable parameters.

[3]For ensuring numerical stability during training, we use $\hat{\sigma}_{\rm p} := (2 \log \sigma_{\rm p} + 10)/5$ here.

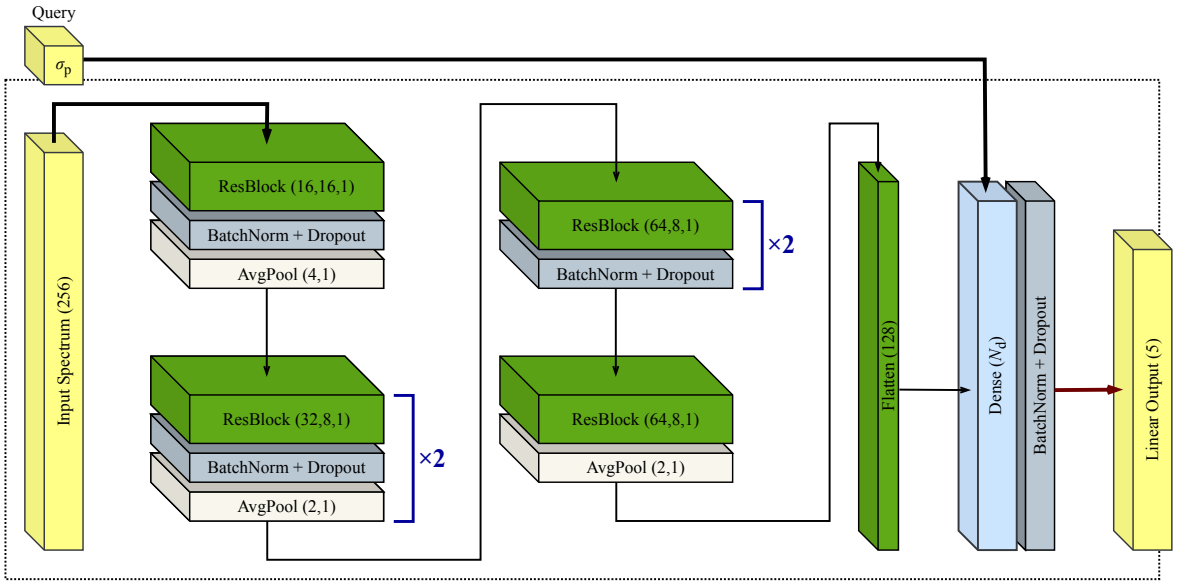

Figure 5.4: Architecture of nSansa. An input spectrum of size 256 pixels is fed into the network that contains a total of 6 residual blocks and extracts useful features from the field. Batch normalization and dropout are used for regularization and average pooling is used for downsampling. The output of the residual part is flattened, concatenated with the σ_p query, and then fed into a hidden nonlinear layer with N_d units. Finally, a linear layer with 5 nodes (2 for the summary vector, 3 for its covariance) acts as the output layer.

We also trained an architecture without the σ_p query and a different number of hidden units ($N_d = 100$) in the dense nonlinear layer but otherwise identical to nSansa. The detailed architecture of nSansa is shown in Figure 5.4.

We incorporated various regularization techniques into nSansa for promoting generalization (avoiding overfitting to the training set). These are namely, dropout, batch normalization, and L2 kernel- and weight-decay. We used a dropout of $p_1 = 8.59 \times 10^{-3}$ after each residual block (except the last) and $p_2 = 8.63 \times 10^{-3}$ after the single hidden layer in the MLP (values defined in Keras terms). All the convolutional layers in the residual part of the architecture feature a kernel decay with $l_2^{\text{conv}} = 6.61 \times 10^{-7}$ and the hidden layer in the MLP features a weight decay with $l_2^{\text{dense}} = 5.14 \times 10^{-6}$. The convolutions have a stride of 2 pixels in the first two residual blocks and a stride of 1 in the rest of them.

5.3.2 Training

We use the training and validation sets as described in Section 5.2 for training our machine in a supervised way. We use 20,000 spectra per label for training and validation (with fixed skewer positions across labels, different for the two datasets). We additionally augment the training set on the fly with random cyclic permutations and flipping of the spectra as well as randomly drawn noise realizations with properties described in Section 5.2.2. We do not augment the validation set, and we add a fixed set of noise realizations to it (that

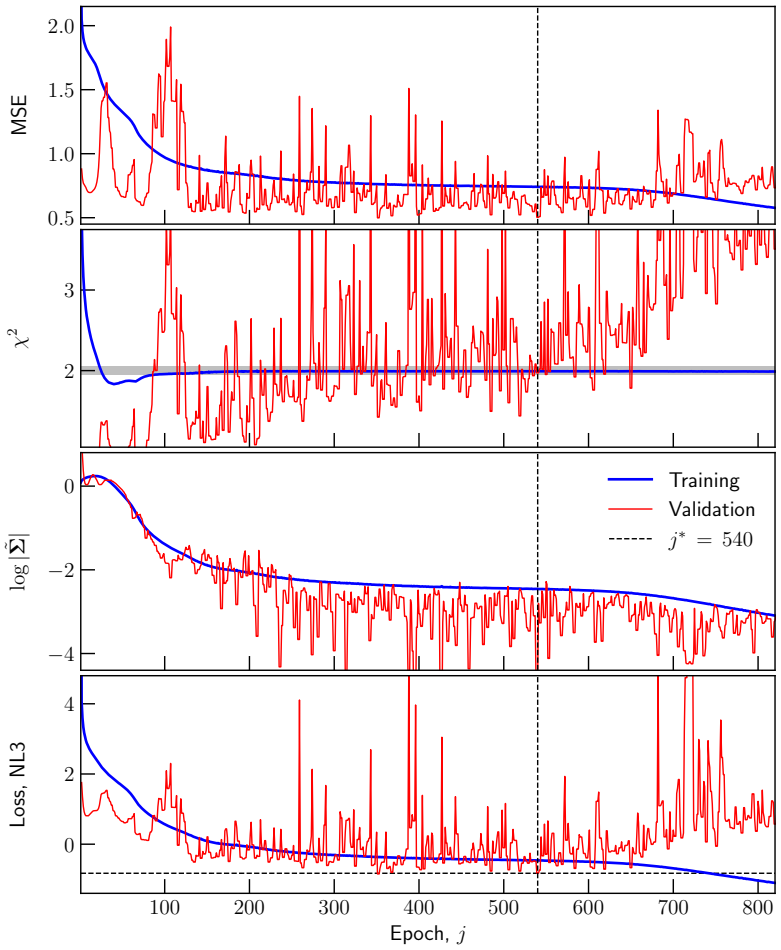

Figure 5.5: An example of typical learning curves of the nSansa architecture. The gray band in the χ^2 panel indicates our tolerance of $\epsilon = 0.05$. Here the minimum of the validation loss is reached at epoch $j^* = 540$ and the χ^2 is simultaneously within our tolerance.

is not varied on the fly), because we would like to compare the state of the network at different epochs of training with the same set of unseen spectra.

The training was performed by minimizing the negative log-likelihood loss (NL3) function of the point predictions *w.r.t.* the true labels,

$$\mathcal{L}(\tilde{\boldsymbol{\pi}}) = \log|\tilde{\boldsymbol{\Sigma}}| + \chi^2, \tag{5.4}$$

with $\chi^2 = (\tilde{\boldsymbol{\pi}} - \hat{\boldsymbol{\pi}})\tilde{\boldsymbol{\Sigma}}^{-1}(\tilde{\boldsymbol{\pi}} - \hat{\boldsymbol{\pi}})^T$, where $[\hat{\cdot}]$ are the (rescaled) parameter labels and $[\tilde{\cdot}]$ are quantities output by nSansa. We used the Adam optimizer (Kingma and Ba 2014) for this purpose with a fixed learning rate of $\alpha = 5 \times 10^{-4}$. The Adam moments parameters assumed the values $\beta_1 = 0.885$ and $\beta_2 = 0.999$.

We monitor three metrics during training besides the loss value, namely, χ^2, $\log|\tilde{\boldsymbol{\Sigma}}|$, and MSE $= (\tilde{\boldsymbol{\pi}} - \hat{\boldsymbol{\pi}})(\tilde{\boldsymbol{\pi}} - \hat{\boldsymbol{\pi}})^T$. For a meaningful estimation of $\tilde{\boldsymbol{\Sigma}}$ by nSansa we expect the average of χ^2 over the training and the validation sets to converge to $N_{\text{params}} = 2$

as the quality of the network inference improves over the course of training. Figure 5.5 shows an example of the learning curves for nSansa. Indeed, χ^2 quickly converges to 2 for the training set and for the validation set it fluctuates close to 2 with slow overall rise. The improvement in the network's state is therefore largely attributed to the minimization of $\log|\tilde{\mathbf{\Sigma}}|$ or the network's uncertainty in its predictions and of the scatter of the point predictions for any underlying true label. The validation loss $\mathcal{L}_{\rm val}$ improves over the course of the first few hundred epochs and starts getting worse after $\sim$ 600 epochs, implying overfitting to the training set. We deem the network to have the best possible state for the given hyperparameter set at the epoch j^* at which the validation loss is minimal, $\mathcal{L}_{\rm val} = \mathcal{L}^*_{\rm val}$, and simultaneously $|\chi^{*2} - 2| \leq \epsilon = 0.05$.

We performed a Bayesian optimization for tuning the values of the hyperparameters $N_{\rm d}$, $l_2^{\rm conv,\,dense}$, $p_{1,2}$, and Adam β_1. This procedure is discussed in detail in Appendix H.

5.3.3 Ensemble learning

NNs are stochastic processes with a large degree of freedom. The choice of hyperparameters, initial weights, random batching of the training data are some of the primary factors that affect the optimal weights of the fully trained machine. In such a scenario, in order to get the best outcome of the infrastructure, we use ensemble learning, which is a powerful tool in machine learning (see, e.g., Dietterich 2000). After finding a set of optimal hyperparameters, we train a committee of 2200 NNs with varying initial weights and on-the-fly stochasticity. We then employ bootstrap aggregation ("bagging," Breiman 2004) to combine the individual network predictions in form of an averaging weighted by $1/\sqrt{|\tilde{\mathbf{\Sigma}}|}$. The bagged results of the committee are expected to be statistically more stable than individual networks *w.r.t.* variations in the input vectors. In the following, the predictions of the machine are assumed to be the bagged results of the committee unless otherwise specified.

5.4 Inference

Our model nSansa acts as a compression of the data into optimal summary statistics of the same size as the number of parameters, $N_{\rm params} = 2$. The resulting summary vector $\hat{\mathcal{S}}$ can be used to infer the parameters (T_0, γ). We explore two different methods for estimating the posterior constraints with nSansa in this work: (i) Gaussian likelihood analysis and (ii) density estimation likelihood free inference (DELFI).

5.4.1 Gaussian likelihood

In the likelihood case, we chose a multivariate Gaussian likelihood function of the NN-compressed summary vector $\mathcal{S}$ (as well as the traditional summaries TPS, TPDF and their combination) of the form

$$\log L_{N_{\rm s}}(\boldsymbol{\pi}) \sim -\frac{1}{2}\log|\mathbf{C}_{N_{\rm s}}| + \frac{(\bar{\mathcal{S}}_{N_{\rm s}} - \boldsymbol{\mu})\mathbf{C}_{N_{\rm s}}{}^{-1}(\bar{\mathcal{S}}_{N_{\rm s}} - \boldsymbol{\mu})^T}{2}, \tag{5.5}$$

where N_s refers to the number of sight lines the amount of information corresponds to, $\mathbf{C}$ is the covariance matrix of $\mathcal{S}$, $\bar{\mathcal{S}}$ refers to the average summary vector for any given input data, and $\boldsymbol{\mu} \equiv \boldsymbol{\mu}(\boldsymbol{\pi}; \sigma_\mathrm{p})$ is our model (emulator) for this summary statistic.

We built an emulator of the nSansa statistic with a linear interpolation in the 3D space of $(\boldsymbol{\pi}, \sigma_\mathrm{p})$ using the mean predictions of nSansa from 20,000 previously unseen spectra per distinct $\hat{\boldsymbol{\pi}}$ of the joint training and validation sets. (These correspond to the same skewer positions as those used for validation.) For the traditional summaries we created emulators via cubic spline interpolation of k-modes (for TPS) and F bins (for TPDF) independently using the respective pure estimators over the training labels, computed using 100,000 spectra per label.

We compute the necessary covariance matrices via bootstrapping in the following way. We estimate mean summary vectors $\{\mathbf{s}_i\}$ over mutually exclusive subsets of size $N_\mathrm{s} = 100$, exhausting the underlying test set (described in more detail in Section 5.5). We then repeat this stratified subsampling 1,000 times by introducing randomness in selecting the subsets. The covariance is then estimated by

$$\mathbf{C}_\mathrm{s} = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{s}_i - \mathbf{s})^T (\mathbf{s}_i - \mathbf{s}), \tag{5.6}$$

where N is the total number of subsets and $\mathbf{s}$ is the global mean summary vector. The errors thus correspond to the information content of $N_\mathrm{s} = 100$ spectra (or equivalently a Lyα forest of length 8 cGpc/h).

5.4.2 Density estimation likelihood free inference (DELFI)

LFI, as the name indicates, gets rid of any assumptions about the form of the likelihood function for Bayesian inference. In this work we employed a popular approach to LFI, namely, density estimation likelihood free inference (DELFI; e.g., Alsing et al. 2018) with nSansa. (It is important to note that we do not perform DELFI with the traditional summaries here.) The posterior is expressed in terms of the joint distribution of summary vectors and parameters evaluated at the summary values of the underlying data,

$$p(\boldsymbol{\pi}|\mathbf{d} = \mathbf{d}_*) \propto p(\mathbf{d}_*|\boldsymbol{\pi}) p(\boldsymbol{\pi}) = p_\mathrm{joint}(\boldsymbol{\pi}, \mathbf{d})\big|_{\mathbf{d}=\mathbf{d}_*}. \tag{5.7}$$

We use a Gaussian mixture model (GMM) as a surrogate for the estimation of the joint density $p_\mathrm{joint}(\boldsymbol{\pi}, \mathbf{d})$. We treat σ_p of each spectrum as part of the data vector such that $\mathbf{d} := (\hat{\mathcal{S}}, \sigma_\mathrm{p})$. Hence, the size of the joint vector for density estimation, $(\boldsymbol{\pi}, \mathbf{d})$, is 5. We created a large dataset for training the GMM surrogate as follows. We first selected all the training and validation labels within our prior range (as indicated by the gray region in Figure 5.1) amounting to 177 and obtained nSansa predictions for 10,000 spectra per label in that joint set (a subset of those used for covariance estimation in the Gaussian likelihood case). We iterated this procedure for 19 different values of CNR_6 linearly spaced between 20 and 200. For each distinct CNR_6, we generated a large set of mean summary vectors

over subsets of $N_s = 100$ spectra with 1,000 repetitions of the stratified subsampling as described in Section 5.4.1, thus completing the GMM training set generation. The size of this dataset is 336,300,000 joint $(\boldsymbol{\pi}, \mathbf{d})$ vectors.

We used the PYTHON package SCIKIT-LEARN for fitting the GMM surrogate, with a regularization factor of 10^{-5} added to the diagonal of the component covariance matrices to ensure numerical sanity, stability, and prevent overfitting, the latter of which was asserted through a convergence analysis. We fit 30 GMMs through our collective dataset, each with a unique number of mixture components $K = 1$ through 30. Based on the results of this convergence analysis, we empirically chose the GMM surrogate with $K = 23$ for inference with DELFI.

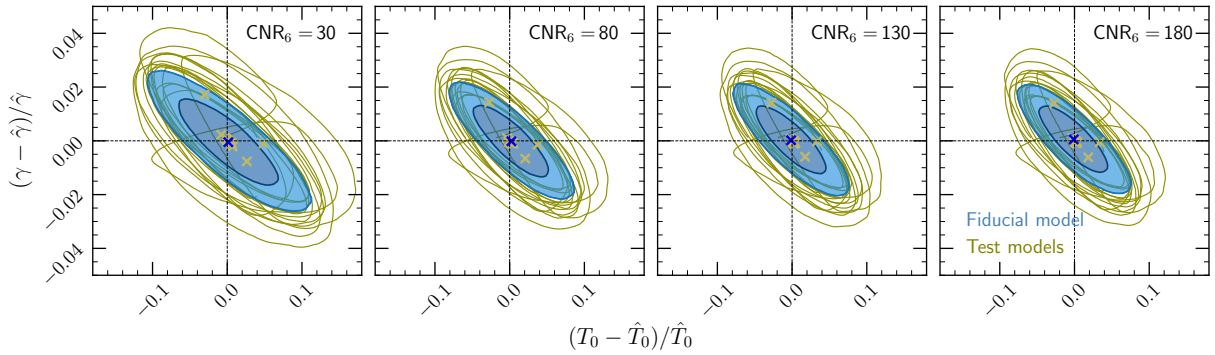

Figure 5.6: A comparison of posterior constraints on $(\boldsymbol{\pi} - \hat{\boldsymbol{\pi}})/\hat{\boldsymbol{\pi}}$ using NSANSA likelihood among 16 different test $\hat{\boldsymbol{\pi}}$ models and the fiducial model for four different noise levels. For the test models only the 95% credibility contours are shown, whereas for the fiducial model the 68% and 95% credibility contours are shown. For all of them the posterior means are also shown with crosses of the corresponding colors.

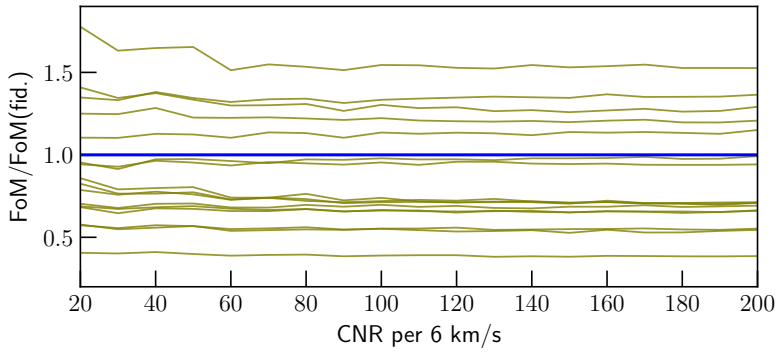

Figure 5.7: The ratio of the FoM between the test models and the fiducial model against the noise level for the NSANSA likelihood.

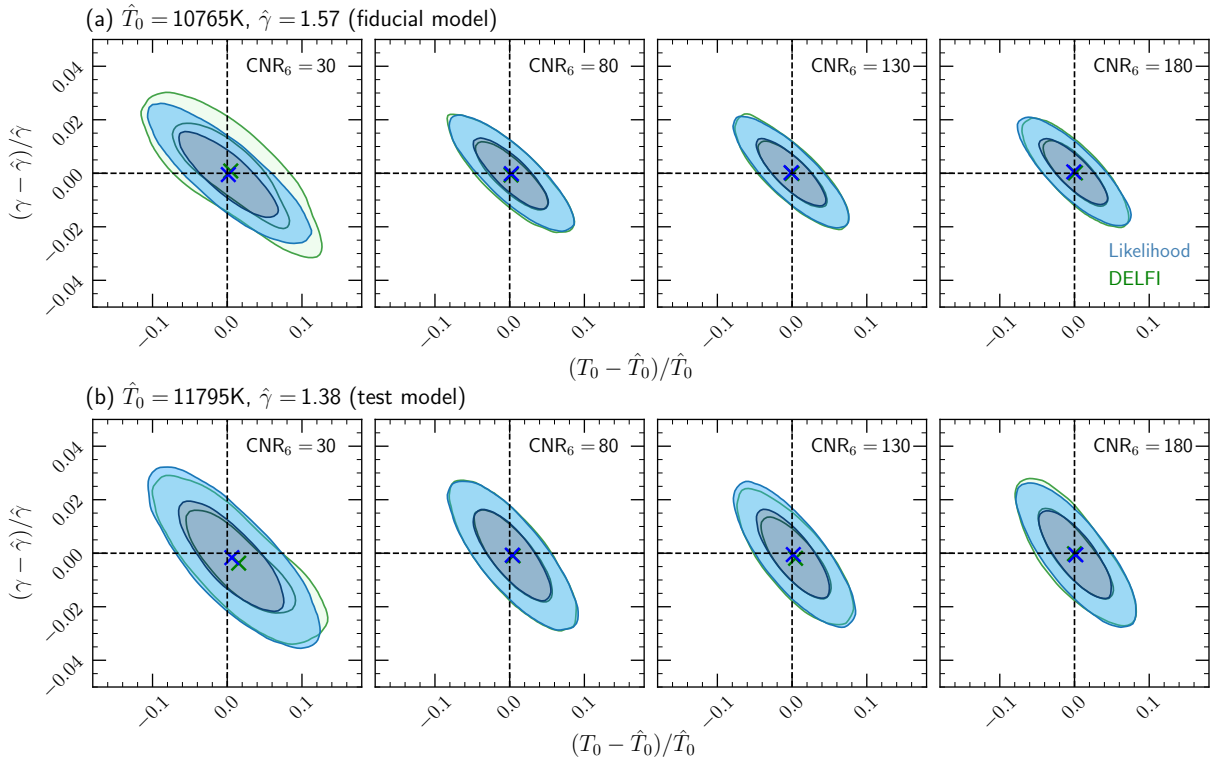

Figure 5.8: A comparison of nSansa Gaussian likelihood analysis and DELFI in terms of posterior constraints on $(\boldsymbol{\pi} - \hat{\boldsymbol{\pi}})/\hat{\boldsymbol{\pi}}$ for the fiducial TDR model (top) and a test model (bottom) for four different noise levels. In each case, a cross marker of the corresponding color indicates the posterior mean. For DELFI, a GMM of $K = 23$ components was used to estimate the joint $(\boldsymbol{\pi}, \mathbf{d})$ density.

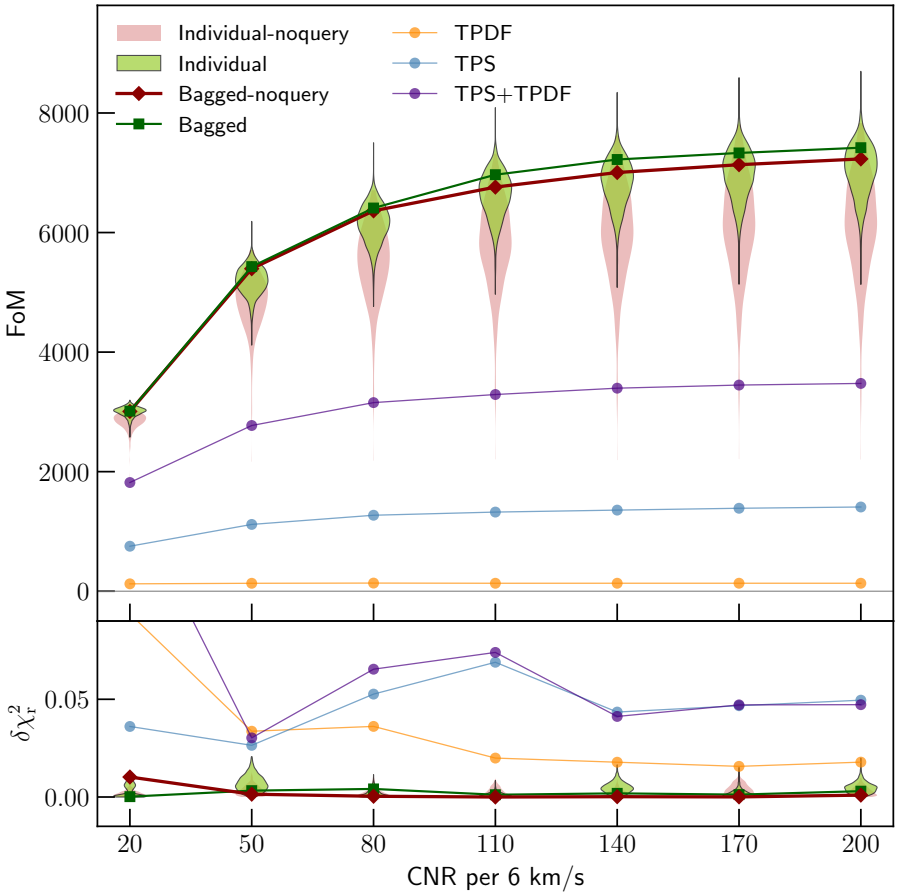

Figure 5.9: Posterior FoM (top) and $\delta\chi^2_r$ (bottom) against CNR for nSansa and an equivalent architecture without the σ_p query, compared to the TPS, TPDF and joint TPS+TPDF constraints for the fiducial thermal model. All values correspond to the Gaussian likelihood inference. The violins show the distributions of the metrics of the individual members of the ensemble(s).

5.5 Results and discussion

We performed tests of our machinery for a set of different noise levels and determined the behavior of the inference outcome *w.r.t.* CNR. Our full test dataset contains 10,000 spectra from each of the 16 test models (labels) described in Section 5.2.1 as well as the fiducial thermal model. These spectra are generated using a completely different set of skewers than those used for training, validation, and model-building. We additionally sampled 19 linearly spaced CNR_6 values between 20 and 200 in total for various test purposes[4]. For each distinct value of CNR_6 in our sample, we added homoscedastic Gaussian noise with the corresponding σ_p to the test spectra, all with random independent phases across pixels, spectra, and labels. We then fed those into nSansa and obtained a set of (bagged)

[4]This is a subset of the full range of CNR_6 used during training to ensure that tests fall well inside our training prior boundaries.

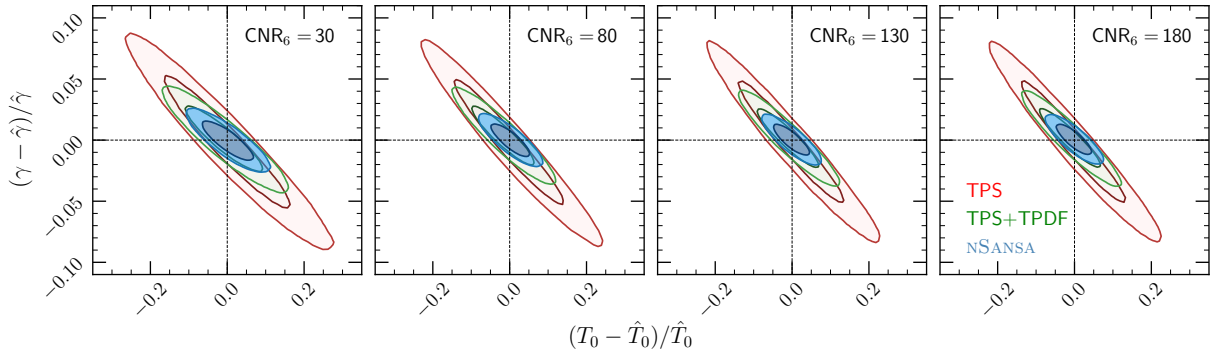

Figure 5.10: Posterior constraints on $(\boldsymbol{\pi} - \hat{\boldsymbol{\pi}})/\hat{\boldsymbol{\pi}}$ using nSansa for four different noise levels compared to the constraints by TPS and TPS+TPDF. A Gaussian likelihood was used for inference in all these cases.

predictions that we treat as summary vectors and performed Bayesian inference with them, with a Gaussian likelihood as well as DELFI. For a comparison, we also performed inference with the TPS and TPDF (combined) estimated with the same set of test spectra for all the sampled noise levels. In each case, we sampled from the posterior distribution of $(\boldsymbol{\pi} - \hat{\boldsymbol{\pi}})/\hat{\boldsymbol{\pi}}$ using MCMC with affine invariant sampling[5].

We characterize the precision and accuracy of our posterior constraints using the following two metrics:

(i) We employ the reciprocal of the contour size as our figure of merit (FoM) of the posterior precision, such that

$$\text{FoM} := 1/|\mathbf{C}_{\text{post}}|^{1/N_{\text{params}}}, \tag{5.8}$$

where the covariance $\mathbf{C}_{\text{post}}$ is estimated from the posterior MCMC sample. A larger FoM implies tighter constraints.

(ii) We quantify the accuracy of our constraints with a reduced χ^2 metric,

$$\delta\chi^2_{\text{r}} = \langle\chi^2\rangle/N_{\text{params}} - 1, \tag{5.9}$$

where $\chi^2 = (\boldsymbol{\pi} - \hat{\boldsymbol{\pi}})\mathbf{C}^{-1}_{\text{post}}(\boldsymbol{\pi} - \hat{\boldsymbol{\pi}})^T$ and the average is taken over the MCMC posterior sample. A $\delta\chi^2_{\text{r}}$ of 0 implies that the mean of the posterior exactly recovers the truth and $\delta\chi^2_{\text{r}}$ of 2 implies that the truth is 2σ away from the mean of the posterior.

We first show a comparison of the posterior constraints using the nSansa likelihood for all the 16 test labels $\hat{\boldsymbol{\pi}}$ as well as the fiducial model for four different noise levels in Figure 5.6. The contours for all the test models (except three models at the edges of our prior) show a statistically similar behavior, lending credibility to the claim that the performance of nSansa is stable across our prior in $(\boldsymbol{\pi}, \sigma_{\text{p}})$. This is further evidenced

[5]Performed using the Python package `emcee`, https://emcee.readthedocs.io.

by the ratios of the FoM between those test models and the fiducial model as shown in Figure 5.7.

We then compare the nSansa posterior constraints between the likelihood and the DELFI cases in Figure 5.8. The top panel thereof shows the posterior contours for our fiducial thermal model and the bottom panel a test model of the 16. The contours in the two cases agree with each other to an excellent degree for both the models at most noise levels shown. The DELFI precision evolves with the CNR very similarly to the likelihood case and the DELFI contours only exhibit a marginally non-Gaussian behavior at $\mathrm{CNR}_6 = 30$. This demonstrates that a Gaussian function is a reasonable approximation of the likelihood for nSansa over the entire prior range considered in this work.

In Figure 5.9 we show our posterior metrics against noise level for nSansa as well as the traditional summary statistics considered in this work, and in Figure 5.10 we show the posterior constraints for the fiducial model for four different noise levels using nSansa as well as the TPS and TPS+TPDF. This comparison is for our fiducial thermal model and the Gaussian likelihood case for all the summaries. At all noise levels the constraints by nSansa are more precise and more accurate than the traditional summaries and the boost in precision increases gradually with the CNR. In particular, the nSansa FoM is 1.65 times larger than the TPS+TPDF case at $\mathrm{CNR}_6 = 20$ and 2.12 times larger at $\mathrm{CNR}_6 = 200$. At $\mathrm{CNR}_6 > 100$, the original architecture with a σ_p-query performs slightly better than its non-query counterpart in terms of precision for the bagged results of the ensemble. At all CNR the FoM distribution of the ensemble with the query generally centers around higher FoM than the non-query ensemble. All members in the query ensemble have a higher FoM than the TPS+TPDF combination, whereas for the non-query ensemble a tail extends to lower FoM, rendering it an inefficient architecture to be used standalone (without ensembles). It is also noteworthy that in our findings the best-performing member of a committee at one noise level is not necessarily so at another, further raising the issue of stability of the inference using a single member network. Nonetheless, the bagged results of the non-query version are especially interesting since, unlike its query counterpart or the traditional summaries, no information of the underlying noise level needs to be supplied whatsoever.

5.6 Conclusion and outlook

We extended the proof-of-concept analysis of N24 by incorporating certain real-world nuisance effects in the Lyα forest observations to make a deep learning field-level inference pipeline such as LyαNNA more reliable for actual data. Namely, we studied the impact of instrumental noise in the spectra on the precision and accuracy of the NN inference. We constructed a 1D ResNet convolutional architecture called nSansa to recover the power law TDR parameters (T_0, γ) of the IGM from medium resolution, noisy Lyα forest data akin to López et al. (2016). We trained this machinery with a large set of labeled mock spectra generated from hydrodynamical simulations, including a range of noise levels characteristic of targeted, individual spectroscopic observations. We performed an extensive

hyperparameter tuning based on Bayesian optimization using a completely disjoint validation set and thereafter trained a committee of 2200 networks with the best hyperparameter values for ensuring statistical stability of the results via ensemble learning. The resultant pipeline can be treated as an optimal compression of the Lyα transmission field into a summary statistic that is degenerate with the target parameters of interest.

We performed Bayesian inference with nSansa using a Gaussian likelihood (as is common practice) as well as without any assumptions of the form of the likelihood. In the latter case we estimated the joint density of the parameters (labels) and the NN-compressed summaries via a GMM surrogate for DELFI. Finally, we conducted a detailed investigation of this inference framework with a previously unseen realistic mock test dataset and characterized its behavior with varying noise levels on the spectra. When compared to traditional summary statistics of the field such as TPS and TPDF—each of them having been corrected for noise—our machinery exhibits enhanced precision and accuracy of inference, which suggests that it can extract useful non-Gaussian features of the transmission field that are not captured by the traditional summaries even when they are buried under noise.

We compared the Gaussian likelihood inference and DELFI against each other for nSansa to draw interesting conclusions about the NN-compressed summary statistic. The posterior precision for the two is comparable in the case of a GMM with 23 mixture components for DELFI and only mild deviations from a Gaussian posterior appear. In light of the computational resources required to build a 5D surrogate density and the eventual time of running MCMC with it, a Gaussian (distribution) can be regarded as a reasonable approximation of the form of likelihood for nSansa. We discourage a comparison of nSansa DELFI with the constraints of traditional summaries (TPS+TPDF) performed in this work at this point, since an equivalent likelihood-free inference framework for the latter is lacking as of yet, and we motivate future research in this direction to address the compatibility of simpler likelihood assumptions for TPS and TPDF.

It is instructive to note that the improvement in posterior precision over TPS+TPDF we found in this work (over the whole CNR_6 range studied) is not as large as in the pure, noise-free case of N24. This could partially stem from the fact that the traditional summary vectors are corrected for noise using known analytical properties (a subtraction of noise power for the TPS and a deconvolution of the noise PDF for the TPDF). On the other hand, supplying the information of σ_p in the nSansa workflow via a query only leads to a marginal gain in precision—only at high CNR—over an equivalent pipeline without the said query, signaling to the inability of the nSansa architecture to meaningfully utilize this auxiliary information of the noise level. Further investigation may potentially reveal alternative ways to account for this information into a field-level inference pipeline to yield a correction of the similar degree as for the human defined summaries.

Moreover, in this work we focused solely on two realistic nuisance effects, namely noise and spectral resolution. While relatively straightforward to handle with traditional summaries, they are two of the biggest concerns for spectrum-based inference from targeted spectroscopic observations of quasars and thus merit a methodological investigation such as this into the usefulness of inference with deep learning. However, they are accompanied by a set of other modeling and physical systematic effects, especially in large cosmological

survey datasets. These are, for instance, quasar continuum fitting uncertainties, metal absorption lines, damped Lyα absorbers (DLAs), etc. These must be adequately incorporated into the next generation field-level frameworks for robust and reliable inference with large real-world datasets using artificial intelligence.

Chapter 6

Conclusion and future directions

6.1 Summary

The unique absorption feature in the spectra of distant quasars, aptly named the Lyα forest, is one of the most powerful probes of cosmic physics at play over a wide range of scales. Not only is it a notable independent tracer of the BAO scale, it also provides valuable insights into the thermal state of the IGM and the cosmic reionization. Classically, analytically well-defined summary statistics, e.g., the 1D power spectrum of the Lyα transmission, have almost exclusively been used to constrain the values of cosmological and astrophysical parameters of interest from this signal. However, the transmission field has an inherently highly non-Gaussian PDF, rendering such statistics inefficient for encoding the full information. In the age of large spectroscopic surveys delivering humongous amounts of cosmological data, this has become a pressing problem.

This thesis, in some ways, is a systematic study of the lost information content while using traditional human-defined statistics and efficient ways to recover it responsibly. In other ways, it sets the stage for the upcoming generation of data analysis and inference techniques for the Lyα forest, namely the field-level inference or the simulation-based inference.

Artificial intelligence, in particular deep learning, was identified as a promising yet relatively unexplored avenue for field-level inference with the Lyα forest at the beginning of this research endeavor. It indeed offered a natural way of encoding highly stochastic non-Gaussian Lyα forest data with the aim of performing Bayesian inference, inspired by the recent, encouraging progress made in neighboring fields, e.g., weak gravitational lensing. We proposed to use labeled mock Lyα forest data sourced by giant cosmological hydrodynamic simulations to train those neural network architectures in a supervised fashion. It was thus imperative to ensure the accuracy and fidelity of those Lyα forest mocks before using them for training. In this thesis, I started with the task of mock data production and quality assurance. I then trained and built a deep learning framework for field-level inference from pure simulations as a proof of concept (PoC), and afterwards investigated the applicability and robustness of that framework in the presence of some unavoidable

data contaminants in real observations.

Mock data generation

In Chapter 3 I present a detailed analysis of the current state-of-the-art methods for creating mock Lyα forest datasets from cosmological hydrodynamic simulations. Popular tools for this purpose incorporate various assumptions into the workflow with the intention of reducing the computational cost. They lead to approximations that have not thus far been justified with adequate evidence. I dissected those approximations and methodically examined their applicability in different cosmological regimes. Based on my findings I recommended—and adopted myself—an optimal course through the multitude of approximate techniques for the production of accurate mock training datasets.

Proof-of-concept NN inference

It was instructive to first investigate the usefulness of deep learning for field-level inference and the improvement in inference outcomes by using such methods over the traditional statistics, in the scenario of pure, uncontaminated Lyα forest signals. This would provide a conceptual proof that neural networks may be employed to optimally extract relevant information from the underlying field that is not accessible to most commonly used traditional summaries. In Chapter 4, I reproduce the contents of the peer-reviewed research I led, Nayak et al. (2024), that demonstrated not only that the neural networks can extract information from the field beyond the 1D power spectrum and the PDF of the transmission, but also that for uncontaminated spectra the gain in inference precision can be factors-of-a-few over those summary statistics. In Chang et al. (2025) we found further evidence in support of those claims.

Increased realism in the data

In light of the extremely promising findings of the PoC analysis, the most natural step forward was to adapt this new framework for inference with real world datasets that contain a myriad of instrumental and physical systematic effects. Instrumental noise and resolution, quasar continuum estimation uncertainty, intergalactic metal absorption, and high column density (HCD) absorbers are some examples of those contaminating agents. They have the adverse effect of compromising the constraining power of the Lyα transmission field and introducing biases in our posterior constraints. In a particular instance, the impact of noise on the 1D power spectrum and the PDF of the Lyα transmission is analytically tractable and remediable, however, such understanding in the context of field-level inference is hitherto lacking. In Chapter 5, I reproduce the contents of another manuscript submitted for peer review, Nayak et al. (2025), that presents an analysis of our deep learning framework for noisy, medium resolution Lyα forest spectra and a characterization of the impact of the varying noise level in the data on the posterior constraints. We found that our machinery outperforms the traditional summary statistics under consideration even in the presence of noise, however, the gain in inference precision is not to the same degree as for

uncontaminated spectra. We substantiated these results by observing that no tractable correction is applied to the compressed summary vectors of our neural network as opposed to the 1D power spectrum or the PDF. We also performed likelihood-free inference (LFI) with our machinery via estimating the joint density of the NN summaries and parameter labels, similar to Alsing et al. (2018). We found that the posterior contours in the LFI case agree closely with those in a Gaussian likelihood inference scenario, indicating that the NN summary vectors are statistically well behaved. The findings of this study are indeed quite promising for the future of field-level inference with real-world Lyα forest data.

6.2 Outlook

Data systematics

This thesis paves a path to the next-generation inference techniques involving artificial intelligence with the Lyα forest from large cosmological surveys. As such it demonstrates the potential of neural networks to encode the relevant information of the field for inference, both in principle and practice, as an alternative to a lossy compression the traditional summaries offer. However, in order to be able to apply it to real observations, a handful of other systematic effects must also be accounted for in the training process. In the following I speculate potential ways to address a few of the most important systematics in a future work, and certain caveats to those.

- In Nayak et al. (2025) we addressed the noise and spectral resolution in the spectra. Although we varied the noise level, we kept the spectral resolution fixed, which is characteristic of a single spectroscopic instrument with known noise properties. The framework may not, however, be directly applied to data observed with a different instrument with a different resolution. One could account for a range of different resolutions similar to the noise level in the training data, and potentially marginalize over this uncertainty, or supply the known value of the resolving power $R \equiv \lambda/\Delta\lambda$ as a query to the network. In either scenario, the spectra must be sampled conforming to the Nyquist limit, which would then imply varying pixel and spectrum sizes in the training data and the network architecture must, in turn, be adapted for this.

- Dealing with HCD systems such as the Damped Lyα absorbers (DLA) is relatively easier with neural networks, as it is with traditional summary statistics. For most cosmological applications, spectra containing HCDs may be identified with a reliable tool (e.g., Parks et al. 2018; Wang et al. 2022a) and simply be omitted during inference.

- Intergalactic metal lines often blend together with the Lyα forest and are practically impossible to separate, unlike HCDs. A field level machinery could potentially see this as an enhancement of the Lyα absorption along a line of sight and could produce biased inference results. With a model of intergalactic metal composition, this

process may also be mocked up in the training data. A neural network might learn to apply an offset to the compressed summaries such that the posterior distributions are unbiased. The reliability of this, however, rests in the fidelity of the model of intergalactic metal composition.

- A misestimation of the true quasar continuum generally introduces large scale (small k) envelope modulations to the overall amplitude of a Lyα forest transmission spectrum. While training a neural network, this process may be simulated as follows. A mock quasar continuum may first be applied to an individual transmission (F) spectrum and then this quasar-forest composite may be fed to a continuum prediction/fitting pipeline (such as Turner et al. 2024) that estimates a new continuum that may be different from the original one. For training a machine to properly recognize this systematic, it would be imperative to repeat this process randomly on the fly, by shuffling the true (mock) continua and the eventual continuum extraction scheme. However, this not only compounds the dimensionality of the problem, it also makes the training incredibly slow by the current best estimates.

A wider scope of application

In this thesis, I have exclusively focused on constraining the two parameters (T_0, γ) of the power-law temperature density relation of the IGM at $z \sim 2.2$. As such this neural network summary is uniquely sensitive to those parameters and is not useful for constraining other parameters of interest, as opposed to traditional summaries. Thus, such a field-level neural network, in principle, needs to be retrained for every new parameter we would like to infer with it. Heuristically, a ResNet architecture such as ours would learn the characteristic features of the data in the first few layers and then connect the relevant features to the parameter labels deeper on. This provides hope for the usefulness of techniques such as transfer learning and domain adaptation for downscaling the retraining problem.

Another potential issue while creating a NN for inference of cosmological parameters in particular is the sparsity of the training labels. In this thesis I used a post-processing technique to populate the (T_0, γ) parameter space using the same hydrodynamic simulation box. However, in order to populate the cosmological parameter space, new hydrodynamic simulations must be run with the labels sampled, which is many orders of magnitude more expensive than our post-processing. Existing suites such as Lyssa (Walther et al. 2025) populate the space of cosmological parameters too sparsely for a meaningful NN training. A potential remedy to this could in turn be generative AI models as lightweight field-level emulators of the hydrodynamic simulators. In this context, I refer the reader to Section 4.1, in particular the discussion of recent advances in AI applications in cosmology.

Interpretability

One of the biggest pitfalls of deep learning perhaps is its black box nature. Indeed, for this sole reason, faster and less expensive deep learning alternatives to traditional methods, including those of parameter inference, have not yet fully become the state of the art. This

healthy skepticism, however, keeps the booming research in AI applications to cosmology in check. It encourages a spirit of developing AI models that are interpretable at least up to a certain degree. In this thesis, I have not been able to touch upon this aspect of the framework. Besides its demonstrable prowess of optimal information extraction, it must also be made more interpretable if it is to shift the paradigm of Bayesian inference in cosmology. For this purpose, approaches such as saliency analysis may be adopted that attempt to learn the sensitivity of the extracted features (and eventual outputs) by the trained networks to small changes in the input data vectors.

Appendix A

An approximate lightcone model

This appendix corresponds to Chapter 4.

Due to unavailability of lightcone simulations, our skewers come entirely from a single cosmic epoch (snapshot) and do not capture any evolution of underlying hydrodynamic fields along the LOS. Consequently the spectra carry this modeling uncertainty. However, we implemented an approximate model of lightcone evolution for testing our machinery, by interpolating certain physical quantities among close-by snapshots (in z). For four snapshots at $z \sim 2.0$, 2.2, 2.4, and 2.6, we estimated the true underlying TDR parameters (see Figure A.1) and then linearly interpolated them to obtain $T_0(z)$ and $\gamma(z)$. The redshift span of our skewers is $\Delta z \sim 0.1$. Assuming the centers of our skewers at $z \sim 2.2$, we rescaled all the skewer temperatures according to the interpolated lightcone TDR for nonlightcone $\rho_\mathrm{b}/\bar{\rho}_\mathrm{b}$ (with the same procedure for sampling in the parameter space described in Section 4.2.1).

Additionally, we incorporated the overall evolution of the mean Lyα transmission through rescaling τ at the pixel level as follows. We first estimated $\tau_\mathrm{eff}(z) = -\log \bar{F}$ along our skewers from Becker et al. 2013 and then rescaled all the nonlightcone optical depth values as $\tau_0(z) \rightarrow \tau_\mathrm{lc}(z) = \tau_\mathrm{eff}(z)/\tau_\mathrm{eff}(z = 2.2) \cdot \tau_0(z)$ to eventually obtain mock Lyα transmission spectra. Figure A.2 shows the fractional variation in the TDR parameters and the Lyα transmission across our skewers according to this approximate lightcone

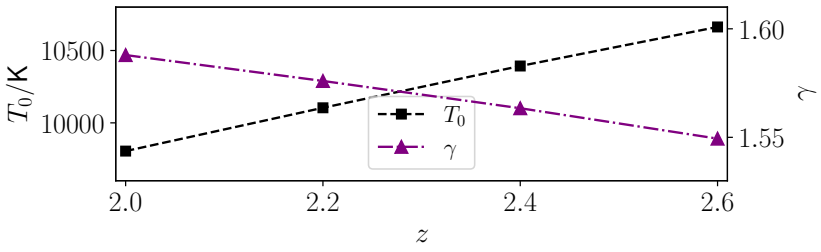

Figure A.1: TDR parameters at four different redshifts; a generally smooth, linear variation can be seen in both T_0 and γ.

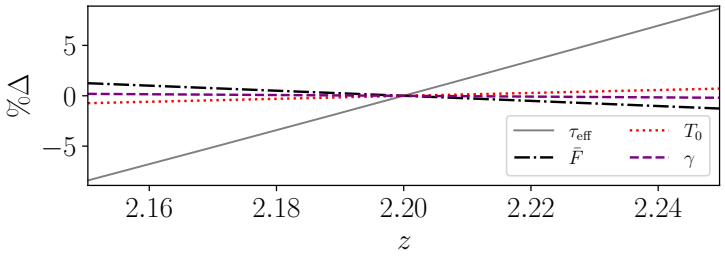

Figure A.2: Percentage variations in the TDR parameters and the Lyα transmission in our approximate lightcone model for the redshift span of our skewers.

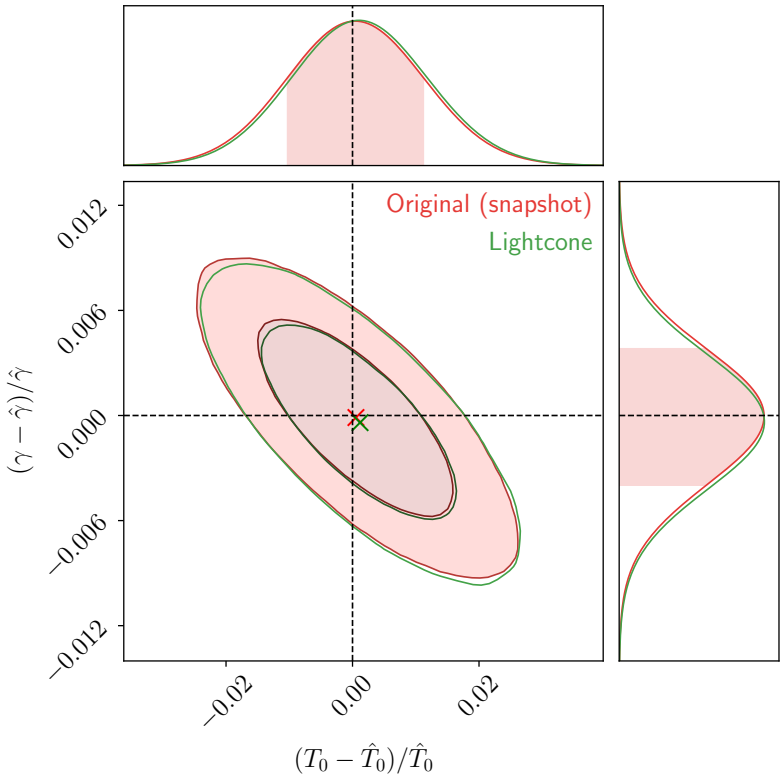

Figure A.3: Comparison of posterior contours between the original (snapshot) and approximate lightcone model test cases. Both carry information equivalent to 100 skewers. A mean (expectation) value of the posterior distributions are also shown with crosses of the corresponding colors. In both the cases, statistically inter-consistent posterior distributions are obtained, recovering the fiducial TDR of the snapshot.

Table A.1: Comparison metric values for the original (snapshot) and approximate lightcone test cases for Sansa (for information equivalent to 100 spectra).

Test set	$\delta\chi^2_{\mathrm{r}}$	FoM / FoM(orig.)
Snapshot (orig.)	0.002	- -
Lightcone	0.008	1.001

model. A maximum of $\lesssim 1.5\%$ deviation in T_0 and $\lesssim 0.4\%$ in γ can be seen. In $\bar{F}$, there is a maximum of $\lesssim 2.6\%$ variation across the redshift span of the skewers. We followed the above procedure for generating 4,000 lightcone spectra with the same 4,000 physical skewers as the "original" test case (Section 4.4).

With these 4,000 lightcone spectra, we performed inference with Sansa; see the posterior constraints in Figure A.3. We expect Sansa to be able to recover a "mean" TDR along the skewers, that is, the thermal parameters at the centers of the skewers, $\hat{T}_0 = 10104.15$ K and $\hat{\gamma} = 1.58$ (the fiducial values). The metrics for the lightcone and the original test cases are compared in Table A.1.

Appendix B

Orthogonal basis of the parameters

This appendix corresponds to Chapter 4.

Heuristically, the training of the network is most efficient when our training sample captures the most characteristic variations in the data with respect to the two parameters of interest, T_0 and γ. Indeed, as found by many previous analyses (e.g., Walther et al. 2019), there appears to be an axis of degeneracy in the said parameter space given by the orientation of the elongated posterior contours. This presented us with an alternative parametrization of the space accessible via an orthogonalization of a parameter covariance matrix. By doing a mock likelihood analysis with a linear interpolation emulator of the TPS on a preexisting grid of thermal models, we first obtained a (rescaled) parameter covariance matrix $\mathbf{C}$ and then diagonalized that such that $\mathbf{\Lambda} = \mathbf{V}^T\mathbf{C}\mathbf{V}$ is a 2×2 diagonal matrix ($\mathbf{V}$ symmetric). An "orthogonal" representation of the parameters was then found by a change of basis,

$$\mathbf{A}^T = \mathbf{V}\boldsymbol{\pi}^T, \tag{B.1}$$

where $\mathbf{A} = (\alpha, \beta)$ and $\boldsymbol{\pi} = (T_0, \gamma)$. In the above definition, β represents the degeneracy direction in the $\boldsymbol{\pi}$ parameters and α corresponds to the axis of the most characteristic deviation. Thence, we sampled the parameter space for training with an 11×11 regular grid in the orthogonal parameter space.

Appendix C

Biases due to a limited prior range

This appendix corresponds to Chapter 4.

After a cursory network training we observed that when the true labels fall close to the edges of our prior range set by the training sample as shown in Figure 4.4, the mean network predictions are biased toward the center of that space. We also found that in the central region of our prior space, all the network predictions are roughly Gaussian distributed but closer to the edges of that prior they are more skewed, that is, the mode of the distribution is closer to the ground truth than the median. To regularize this, we sampled an extended, sparser grid of thermal models along the degeneracy direction β (in which SANSA provides weaker constraints), to augment our training dataset as shown in Figure C.1. After retraining SANSA on this extended grid, we observed that a (quantifiable) bias still exists (Figure C.2) but the predictions have mostly Gaussianized. The mean point predictions for each of the thermal models on the original grid along a given orthogonal parameter axis fall on approximately a straight line, and hence we can perform a linear transformation of all the raw network predictions such that they satisfy our expectation, $\tilde{\boldsymbol{\pi}} = \hat{\boldsymbol{\pi}}$. This tractable transformation can be represented as follows. In the orthogonal parameters,

$$\tilde{\mathbf{A}}_{\mathrm{f}}^{T} = \mathbf{M}\tilde{\mathbf{A}}_{\mathrm{i}}^{T} + \mathbf{c}^{T}, \tag{C.1}$$

where $\tilde{\mathbf{A}} = (\tilde{\alpha}, \tilde{\beta})$ is a point prediction vector in the orthogonal basis, $\mathbf{M}$ is a diagonal matrix and the subscripts "i" and "f" denote the original and transformed states of the vector, respectively. This linear transformation is also shown for each parameter α and β independently in Figure C.2. We note that since the change of basis is also a linear operation, the overall transformation in $\boldsymbol{\pi}$ parameters is linear and preserves the Gaussianity of the point predictions. Following Eq. (B.1), the transformation applied to the actual parameters $\tilde{\boldsymbol{\pi}}$ looks like

$$\tilde{\boldsymbol{\pi}}_{\mathrm{f}}^{T} = \mathbf{W}\tilde{\boldsymbol{\pi}}_{\mathrm{i}}^{T} + \mathbf{d}^{T}, \tag{C.2}$$

where $\mathbf{W} = \mathbf{V^{-1}MV}$ and $\mathbf{d} = \mathbf{cV^{-1}}$ due to the symmetric $\mathbf{V}$. All the network covariance estimates $\tilde{\mathbf{C}}$ can also be linearly mapped with a matrix operation $\tilde{\mathbf{C}} \rightarrow \mathbf{W}\tilde{\mathbf{C}}\mathbf{W}^{T}$. We used our validation set to fit the linear transformation parameters, $\mathbf{W}$ and $\mathbf{d}$, for each trained

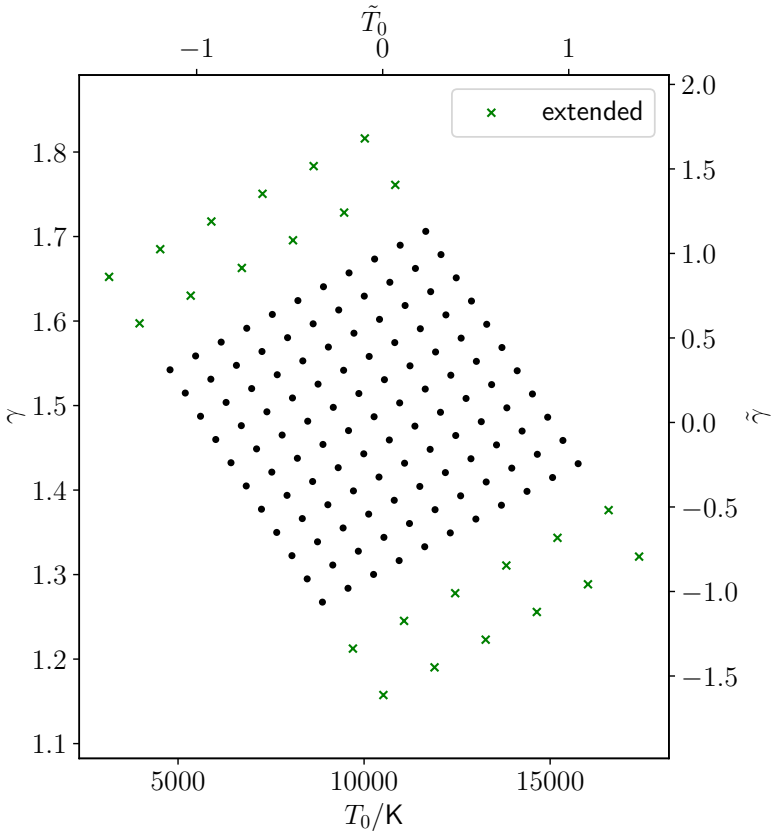

Figure C.1: Extended grid of thermal models for regularizing the point predictions of SANSA. The extension is sparser than the actual training grid as it is only used for rectifying the skewed network predictions for models on the original grid by implicitly allowing a larger prior volume during training. No further test (or network validation) is performed for the models on the extended grid.

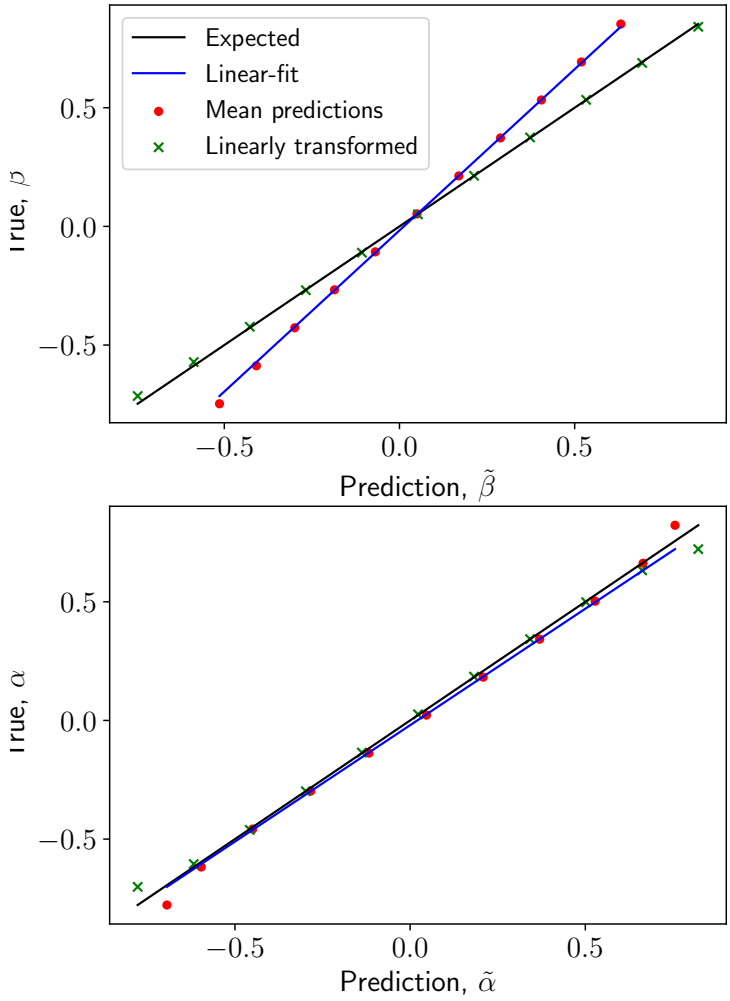

Figure C.2: Mean point predictions on the validation set from a network trained on an extended grid as shown in Figure C.1 and a straight line fit of each independent orthogonal parameter (averaged over the other). The mean predictions after applying the linear transformation can be seen to follow the expected $y = x$ behavior.

neural network in the committee individually (the two parameter combinations (true labels) closest to the prior boundaries in each orthogonal parameter were not considered for this fitting). The full neural network SANSA presented in this paper has this tractable linear transformation incorporated as a final, unbiased layer.

Appendix D

Hyperparameter optimization I

This appendix corresponds to Chapter 4.

As in every deep learning implementation, our algorithm is defined by a large set of hyperparameters that must be tuned in order to arrive at the best possible location on the bias-variance trade-off. Our hyperparameters include the dropout rate, the amplitude of the kernel regularization (l_2), number of residual blocks, number of filters and kernel size in each residual block, weight initialization, amount of downsampling per pooling layer, size of the batches of training data, learning rate, Adam β_1 parameter, etc. Two strategies were adopted to explore this vast hyperparameter space. Hyperparameters having a finite number of discrete possible values (e.g., architecture in the residual parts) were manually tuned with informed heuristic choices of values to try. The rest of the hyperparameters having continuous spectra; namely kernel regularization, dropout, learning rate, and Adam β_1; were tuned with a Bayesian optimization algorithm — based on tree-structured Parzen estimators — for an informed search of the space and an economical use of the resources. This was performed using the predefined routines of the python package Optuna (Akiba et al. 2019).

Appendix E

Training progress

This appendix corresponds to Chapter 4.

We show the learning curves for SANSA in Figure E.1 containing the four network metrics — MSE, χ^2, $\log|\tilde{\mathbf{C}}|$, and the NL3 loss — for the best-performing network in our committee. Following our expectation, the values of the loss, MSE, and $\log|\tilde{\mathbf{C}}|$ decreased for the training set and also for the validation set (albeit somewhat stochastically) over the initial epochs and eventually the validation loss stopped improving. On the other hand, χ^2 for training converged to ~ 2 and that for validation fluctuated on the higher side, occasionally coming close to 2, with a slow overall gain over the epochs. We restored the network to its state at an epoch j^* at which the loss value was minimal while $|\chi^2 - 2| < 0.05$ for the validation set. This helped us make sure that the network predictions are generalized enough and regularized under the Gaussian likelihood cost function. For this network $j^* = 531$. The same qualitative behavior was observed for all the networks in our committee.

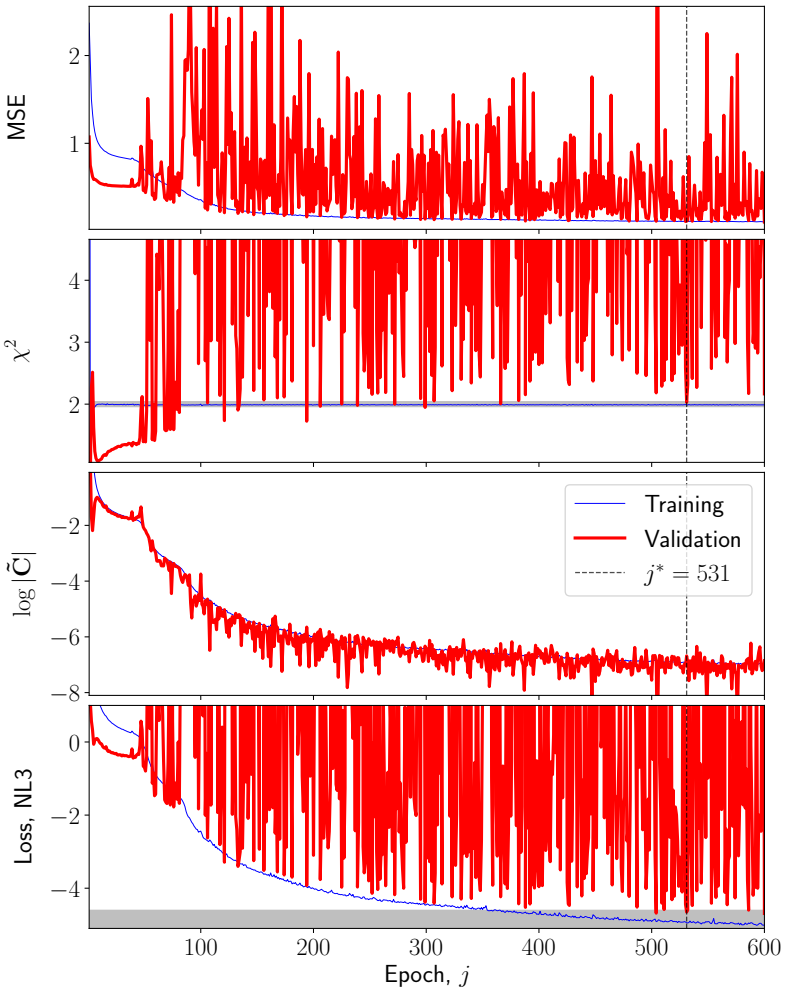

Figure E.1: Learning curves of the best-performing network in the committee of SANSA. This network is trained for > 600 epochs. The best state of the network that can be possibly achieved is determined by the minimal value of NL3 while $|\chi^2 - 2| < 0.05$ for the validation set and that occurs at epoch $j^* = 531$ during training.

Appendix F

Single network versus committee

This appendix corresponds to Chapter 4.

As mentioned in Section 4.3.4, combining outputs of multiple similar neural networks is shown to improve the eventual outcome. Motivated by this, we employed a committee of 20 networks with the same architecture but having different initial weights, and training with different random batches and augmentation on the fly. In practice, the likelihood model of inference described in Section 4.3.5 can be built for each individual network in the committee in the same way as for the committee itself since the procedure relies purely on the predictions of the network(s) from the validation set. Hence, it was possible to compare the posterior constraints of an individual network with that of the committee. We present a comparison of the committee with the best-performing member network (defined as the one leading to the posterior constraints with the highest FoM of all the individual networks) in Table F.1 and Figure F.1 for our "original" test set of spectra from the fiducial thermal model. The constraints by the committee are $\sim 4\%$ tighter than the best-performing network in FoM and they are slightly more accurate as well. Even though $N_{\textsc{Sansa}} = 20$ is statistically a small number of sample members in a committee, the aggregate results of the ensemble are still a little better than the best-performing network, conforming to the popular findings.

Table F.1: Comparison metric values of the committee and its best-performing member for our original test case for Sansa (for information equivalent to 1,000 spectra).

	$\delta\chi^2_{\rm r}$	FoM / FoM(committee)
Committee	0.015	- -
Best network	0.045	1.035

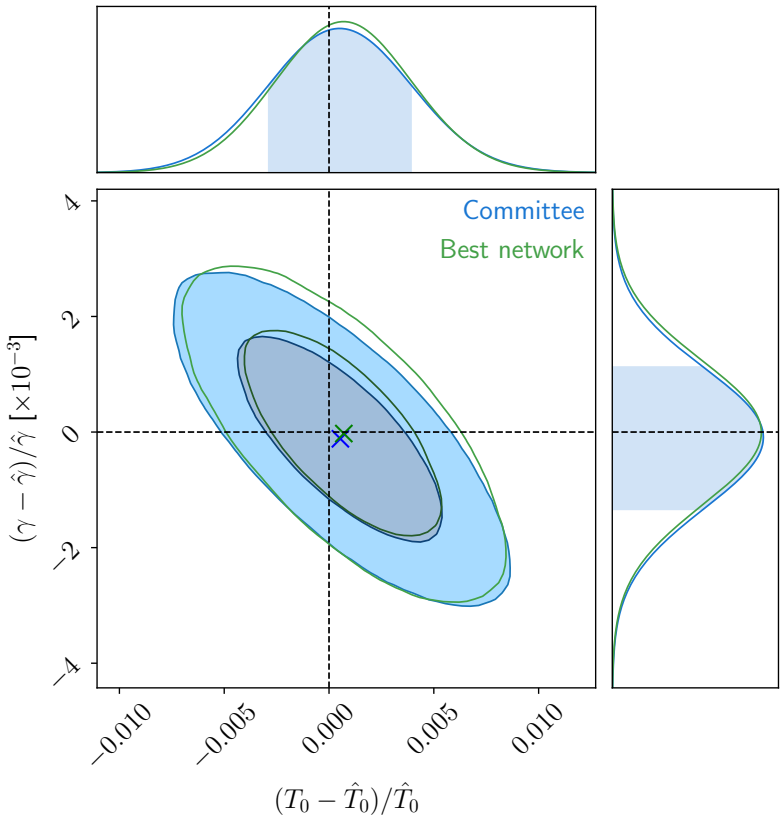

Figure F.1: Comparison of the posterior contours obtained by the committee with those by the best-performing network in the committee for the fiducial thermal model, with information equivalent to 1,000 spectra. The committee constraints are slightly more accurate and ∼4% tighter than those of the best member network.

Appendix G

Parameter space sampling

This appendix corresponds to Chapter 5.

We sampled our training, validation, and test labels in the parameter space using Sobol' sequences for an optimal prior volume filling. In the orthogonal space (α, β) of the TDR parameters (T_0, γ) as discussed in Appendix B of N24, we first created a sample of 256 labels of our training set in the range $-1.1 < \alpha < 1.1$ and $-1.5 < \beta < 1.5$. We then applied rescalings of the labels outside our prior, $-1 \leq \alpha, \beta \leq 1$, as shown in Figure 5.1 (gray region), such that the 2D density of labels in that region is 1/4th of the original. This is motivated by our observations in Appendix C of N24. (It is noteworthy that, unlike N24, we did not apply a linear transformation to the predictions by nSansa after training and used the raw predictions throughout all our analyses. Any biases due to the limited range in sampling are automatically taken into account while creating an emulator for the likelihood analysis and a GMM surrogate for DELFI.) We sampled the validation and test sets both entirely within our prior range.

Appendix H

Hyperparameter optimization II

This appendix corresponds to Chapter 5.

After manually optimizing the overall skeleton of the nSansa architecture, we proceeded to tune some of the more refined hyperparameters of the system. Our objective entailed finding values of hyperparameters θ that minimize $\mathcal{L}^*_{\rm val}(\theta)$, i.e., $\theta_{\rm opt} = {\rm argmin}_\theta \, \mathcal{L}^*_{\rm val}(\theta)$. We used the open-source Python package Optuna for tuning the hyperparameters β_1 (Adam), dropout rates and the l_2 amplitudes of regularization of convolutional kernels and MLP layers, and the number of nodes in the hidden layer of the MLP. Our search space priors are listed in Table H.1. We employed Bayesian optimization based on Gaussian processes (GP) with the `GPSampler` class of Optuna. We initially ran 128 trials with hyperparameters sampled with a quasi Monte Carlo (QMC) strategy based on Sobol' sequences (`QMCSampler`) and then built and refined our GP surrogate model to sample hyperparameter sets for further trials. We distinguish the two phases as the QMC (initial phase) and PSC (post surrogate construction phase). Optuna uses log of expected improvement as the acquisition function by default. We refer interested readers to the Optuna documentation for more details (https://optuna.readthedocs.io).

Table H.1: The search space for our hyperparameter optimization

Hyperparameter	Range	Sampling
Adam β_1	$[0.88, 0.92]$	(unif. in linear space)
Kernel decay, $l_2^{\rm conv}$	$[10^{-8}, 10^{-4}]$	(unif. in log space)
Weight decay, $l_2^{\rm dense}$	$[10^{-8}, 10^{-4}]$	(unif. in log space)
Dropout (conv.), p_1	$[0.0, 0.2]$	(unif. in linear space)
Dropout (MLP), p_2	$[0.0, 0.1]$	(unif. in linear space)
# hidden nodes, $N_{\rm d}$	$[75, 100]$	(unif. over integers)

Figure H.1 shows a scatter of the minimum validation loss obtained, $\mathcal{L}^*_{\rm val}$, as a function of trial number as well as histograms of $\mathcal{L}^*_{\rm val}$. The best value of $\mathcal{L}^*_{\rm val}$ is found in the PSC phase at trial 165+1. The distribution of $\mathcal{L}^*_{\rm val}$ leans only *slightly* toward smaller values in the

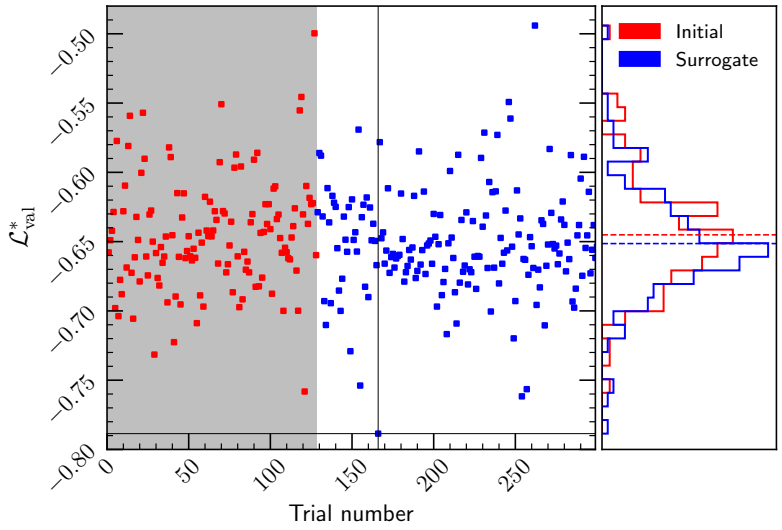

Figure H.1: Left: Minimal validation loss against trial number for our Bayesian hyperparameter tuning. The gray shaded region indicates the QMC phase to build the surrogate out of. **Right**: Normalized histograms of the minimal validation loss across trials in our hyperparameter optimization. We show the QMC and the PSC phases separately. The dashed vertical lines show the mean loss values in both the phases. The PSC phase leans slightly toward smaller loss values than the QMC phase.

PSC phase than in QMC contrary to our expectation of an informed Bayesian optimization algorithm, hinting at the inefficacy of the acquisition function and the surrogate generation internal to Optuna. Nevertheless, the best $\mathcal{L}^*_{\text{val}}$ is a $> 2\sigma$ outlier in both the distributions.

Bibliography

Abdul Karim, M., J. Aguilar, S. Ahlen, et al. (Oct. 2025). "DESI DR2 results. I. Baryon acoustic oscillations from the Lyman alpha forest." In: Phys. Rev. D 112.8, 083514, p. 083514. DOI: 10.1103/2wwn-xjm5.

Akiba, T., S. Sano, T. Yanase, et al. (2019). "Optuna: A Next-generation Hyperparameter Optimization Framework." In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Almgren, A. S., J. B. Bell, M. J. Lijewski, et al. (Mar. 2013). "Nyx: A Massively Parallel AMR Code for Computational Cosmology." In: ApJ 765.1, 39, p. 39. DOI: 10.1088/0004-637X/765/1/39.

Alpher, R. A., H. Bethe, and G. Gamow (Apr. 1948). "The Origin of Chemical Elements." In: *Phys. Rev.* 73 (7), pp. 803–804. DOI: 10.1103/PhysRev.73.803. URL: https://link.aps.org/doi/10.1103/PhysRev.73.803.

Alsing, J., B. Wandelt, and S. Feeney (July 2018). "Massive optimal data compression and density estimation for scalable, likelihood-free inference in cosmology." In: MNRAS 477.3, pp. 2874–2885. DOI: 10.1093/mnras/sty819.

Armengaud, E., N. Palanque-Delabrouille, C. Yèche, et al. (Nov. 2017). "Constraining the mass of light bosonic dark matter using SDSS Lyman-α forest." In: MNRAS 471.4, pp. 4606–4614. DOI: 10.1093/mnras/stx1870.

Bahcall, J. N. and E. E. Salpeter (Nov. 1965). "On the Interaction of Radiation from Distant Sources with the Intervening Medium." In: ApJ 142, pp. 1677–1680. DOI: 10.1086/148460.

Becker, G. D., J. S. Bolton, M. G. Haehnelt, et al. (Jan. 2011). "Detection of extended He II reionization in the temperature evolution of the intergalactic medium." In: MNRAS 410.2, pp. 1096–1112. DOI: 10.1111/j.1365-2966.2010.17507.x.

Becker, G. D., J. S. Bolton, P. Madau, et al. (Mar. 2015). "Evidence of patchy hydrogen reionization from an extreme Lyα trough below redshift six." In: MNRAS 447.4, pp. 3402–3419. DOI: 10.1093/mnras/stu2646.

Becker, G. D., P. C. Hewett, G. Worseck, et al. (Apr. 2013). "A refined measurement of the mean transmitted flux in the Lyα forest over $2 < z < 5$ using composite quasar spectra." In: MNRAS 430.3, pp. 2067–2081. DOI: 10.1093/mnras/stt031.

Blanton, M. R., M. A. Bershady, B. Abolfathi, et al. (July 2017). "Sloan Digital Sky Survey IV: Mapping the Milky Way, Nearby Galaxies, and the Distant Universe." In: AJ 154.1, 28, p. 28. DOI: 10.3847/1538-3881/aa7567.

Boera, E., G. D. Becker, J. S. Bolton, et al. (Feb. 2019). "Revealing Reionization with the Thermal History of the Intergalactic Medium: New Constraints from the Lyα Flux Power Spectrum." In: ApJ 872.1, 101, p. 101. DOI: 10.3847/1538-4357/aafee4.

Boera, E., M. T. Murphy, G. D. Becker, et al. (July 2014). "The thermal history of the intergalactic medium down to redshift z = 1.5: a new curvature measurement." In: MNRAS 441.3, pp. 1916–1933. DOI: 10.1093/mnras/stu660.

Bolton, J. S., M. Viel, T. .-. Kim, et al. (May 2008). "Possible evidence for an inverted temperature-density relation in the intergalactic medium from the flux distribution of the Lyα forest." In: MNRAS 386.2, pp. 1131–1144. DOI: 10.1111/j.1365-2966.2008.13114.x.

Bolton, J. S., G. D. Becker, M. G. Haehnelt, et al. (Mar. 2014). "A consistent determination of the temperature of the intergalactic medium at redshift z = 2.4." In: MNRAS 438.3, pp. 2499–2507. DOI: 10.1093/mnras/stt2374.

Bondi, H. and T. Gold (Jan. 1948). "The Steady-State Theory of the Expanding Universe." In: MNRAS 108, p. 252. DOI: 10.1093/mnras/108.3.252.

Boonkongkird, C., G. Lavaux, S. Peirani, et al. (Mar. 2023). "LyAl-Net: A high-efficiency Lyman-α forest simulation with a neural network." In: *arXiv e-prints*, arXiv:2303.17939, arXiv:2303.17939. DOI: 10.48550/arXiv.2303.17939.

Bosman, S. E. I., F. B. Davies, G. D. Becker, et al. (July 2022). "Hydrogen reionization ends by z = 5.3: Lyman-α optical depth measured by the XQR-30 sample." In: MNRAS 514.1, pp. 55–76. DOI: 10.1093/mnras/stac1046.

Breiman, L. (2004). "Bagging predictors." In: *Machine Learning* 24, pp. 123–140. URL: https://api.semanticscholar.org/CorpusID:47328136.

Busca, N. G., T. Delubac, J. Rich, et al. (Apr. 2013). "Baryon acoustic oscillations in the Lyα forest of BOSS quasars." In: A&A 552, A96, A96. DOI: 10.1051/0004-6361/201220724.

Busca, N. and C. Balland (Aug. 2018). "QuasarNET: Human-level spectral classification and redshifting with Deep Neural Networks." In: *Submitted to MNRAS*, arXiv:1808.09955, arXiv:1808.09955. DOI: 10.48550/arXiv.1808.09955.

Chabanier, S., J. D. Emberson, Z. Lukić, et al. (Jan. 2023). "Modelling the Lyman-α forest with Eulerian and SPH hydrodynamical methods." In: MNRAS 518.3, pp. 3754–3776. DOI: 10.1093/mnras/stac3294.

Chabanier, S., N. Palanque-Delabrouille, C. Yèche, et al. (July 2019). "The one-dimensional power spectrum from the SDSS DR14 Lyα forests." In: J. Cosmology Astropart. Phys. 2019.7, 017, p. 017. DOI: 10.1088/1475-7516/2019/07/017.

Chang, S., P. Nayak, M. Walther, et al. (Aug. 2025). "Human vs. machine – 1:3. Joint analysis of classical and ML-based summary statistics of the Lyman- forest." In: *arXiv e-prints*, arXiv:2508.03264, arXiv:2508.03264. DOI: 10.48550/arXiv.2508.03264.

Chollet, F. et al. (2015). *Keras*. https://keras.io.

Croft, R. A. C., D. H. Weinberg, N. Katz, et al. (Mar. 1998). "Recovery of the Power Spectrum of Mass Fluctuations from Observations of the Lyα Forest." In: ApJ 495.1, pp. 44–62. DOI: 10.1086/305289.

Cuceu, A., A. Font-Ribera, S. Nadathur, et al. (May 2023). "Constraints on the Cosmic Expansion Rate at Redshift 2.3 from the Lyman-α Forest." In: Phys. Rev. Lett. 130.19, 191003, p. 191003. DOI: 10.1103/PhysRevLett.130.191003.

Dawson, K. S., D. J. Schlegel, C. P. Ahn, et al. (Jan. 2013). "The Baryon Oscillation Spectroscopic Survey of SDSS-III." In: AJ 145.1, 10, p. 10. DOI: 10.1088/0004-6256/145/1/10.

DESI Collaboration, B. Abareshi, J. Aguilar, et al. (Nov. 2022). "Overview of the Instrumentation for the Dark Energy Spectroscopic Instrument." In: AJ 164.5, 207, p. 207. DOI: 10.3847/1538-3881/ac882b.

Dietterich, T. G. (2000). "Ensemble Methods in Machine Learning." In: *Multiple Classifier Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–15.

Draine, B. T. (2011). *Physics of the Interstellar and Intergalactic Medium*.

du Mas des Bourboux, H., J. Rich, A. Font-Ribera, et al. (Oct. 2020). "The Completed SDSS-IV Extended Baryon Oscillation Spectroscopic Survey: Baryon Acoustic Oscillations with Lyα Forests." In: ApJ 901.2, 153, p. 153. DOI: 10.3847/1538-4357/abb085.

Ďurovčíková, D., H. Katz, S. E. I. Bosman, et al. (Apr. 2020). "Reionization history constraints from neural network based predictions of high-redshift quasar continua." In: MNRAS 493.3, pp. 4256–4275. DOI: 10.1093/mnras/staa505.

Einstein, A. (Jan. 1915). "Zur allgemeinen Relativitätstheorie." In: *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften*, pp. 778–786.

— (Jan. 1917). "Kosmologische Betrachtungen zur allgemeinen Relativitätstheorie." In: *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften*, pp. 142–152.

Fluri, J., T. Kacprzak, A. Lucchi, et al. (Sept. 2019). "Cosmological constraints with deep learning from KiDS-450 weak lensing maps." In: Phys. Rev. D 100.6, 063514, p. 063514. DOI: 10.1103/PhysRevD.100.063514.

Fluri, J., T. Kacprzak, A. Lucchi, et al. (Apr. 2022). "Full w CDM analysis of KiDS-1000 weak lensing maps using deep learning." In: Phys. Rev. D 105.8, 083518, p. 083518. DOI: 10.1103/PhysRevD.105.083518.

Fluri, J., T. Kacprzak, A. Refregier, et al. (Dec. 2018). "Cosmological constraints from noisy convergence maps through deep learning." In: Phys. Rev. D 98.12, 123518, p. 123518. DOI: 10.1103/PhysRevD.98.123518.

Gaikwad, P., F. B. Davies, and M. G. Haehnelt (Mar. 2025). "Measuring photo-ionization rate and mean free path of HeII ionizing photons at $2.5 \leq z \leq 3.6$: Evidence for late and rapid HeII reionization Part-II." In: *arXiv e-prints*, arXiv:2503.04893, arXiv:2503.04893. DOI: 10.48550/arXiv.2503.04893.

Gaikwad, P., M. Rauch, M. G. Haehnelt, et al. (June 2020). "Probing the thermal state of the intergalactic medium at $z > 5$ with the transmission spikes in high-resolution Ly α forest spectra." In: MNRAS 494.4, pp. 5091–5109. DOI: 10.1093/mnras/staa907.

Gaikwad, P., R. Srianand, M. G. Haehnelt, et al. (Sept. 2021). "A consistent and robust measurement of the thermal state of the IGM at $2 \leq z \leq 4$ from a large sample of Ly α forest spectra: evidence for late and rapid He II reionization." In: MNRAS 506.3, pp. 4389–4412. DOI: 10.1093/mnras/stab2017.

Garzilli, A., J. S. Bolton, T. .-. Kim, et al. (Aug. 2012). "The intergalactic medium thermal history at redshift z = 1.7-3.2 from the Lyα forest: a comparison of measurements using wavelets and the flux distribution." In: MNRAS 424.3, pp. 1723–1736. DOI: 10.1111/j.1365-2966.2012.21223.x.

Glorot, X. and Y. Bengio (May 2010). "Understanding the difficulty of training deep feedforward neural networks." In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Y. W. Teh and M. Titterington. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR, pp. 249–256. URL: https://proceedings.mlr.press/v9/glorot10a.html.

Goobar, A., S. Perlmutter, G. Aldering, et al. (Jan. 2000). "The Acceleration of the Universe: Measurements of Cosmological Parameters from Type Ia Supernovae." In: *Physica Scripta Volume T* 85, p. 47. DOI: 10.1238/Physica.Topical.085a00047.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press.

Gordon, C., A. Cuceu, J. Chaves-Montero, et al. (Nov. 2023). "3D correlations in the Lyman-α forest from early DESI data." In: J. Cosmology Astropart. Phys. 2023.11, 045, p. 045. DOI: 10.1088/1475-7516/2023/11/045.

Grazian, C. and Y. Fan (Sept. 2019). "A review of Approximate Bayesian Computation methods via density estimation: inference for simulator-models." In: *arXiv e-prints*, arXiv:1909.02736, arXiv:1909.02736. DOI: 10.48550/arXiv.1909.02736.

Gunn, J. E. and B. A. Peterson (Nov. 1965). "On the Density of Neutral Hydrogen in Intergalactic Space." In: ApJ 142, pp. 1633–1636. DOI: 10.1086/148444.

Gupta, A., J. M. Zorrilla Matilla, D. Hsu, et al. (May 2018). "Non-Gaussian information from weak lensing data via deep learning." In: Phys. Rev. D 97.10, 103515, p. 103515. DOI: 10.1103/PhysRevD.97.103515.

Guth, A. H. (Jan. 1981). "Inflationary universe: A possible solution to the horizon and flatness problems." In: Phys. Rev. D 23.2, pp. 347–356. DOI: 10.1103/PhysRevD.23.347.

Harrington, P., M. Mustafa, M. Dornfest, et al. (Apr. 2022). "Fast, High-fidelity Lyα Forests with Convolutional Neural Networks." In: ApJ 929.2, 160, p. 160. DOI: 10.3847/1538-4357/ac5faa.

He, K., X. Zhang, S. Ren, et al. (Dec. 2015a). "Deep Residual Learning for Image Recognition." In: *arXiv e-prints*, arXiv:1512.03385, arXiv:1512.03385. DOI: 10.48550/arXiv.1512.03385.

— (Feb. 2015b). "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification." In: *arXiv e-prints*, arXiv:1502.01852, arXiv:1502.01852. DOI: 10.48550/arXiv.1502.01852.

Hiss, H., M. Walther, J. Oñorbe, et al. (May 2019). "A Novel Statistical Method for Measuring the Temperature-Density Relation in the IGM Using the b - N $_{HI}$ Distribution of Absorbers in the Lyα Forest." In: ApJ 876.1, 71, p. 71. DOI: 10.3847/1538-4357/ab1418.

Hu, T., V. Khaire, J. F. Hennawi, et al. (Sept. 2022). "Measuring the thermal and ionization state of the low-z IGM using likelihood free inference." In: MNRAS 515.2, pp. 2188–2207. DOI: 10.1093/mnras/stac1865.

Huang, L., R. A. C. Croft, and H. Arora (Oct. 2021). "Deep forest: Neural network reconstruction of the Lyman-α forest." In: MNRAS 506.4, pp. 5212–5222. DOI: 10.1093/mnras/stab2041.

Hubble, E. (1929). "A relation between distance and radial velocity among extra-galactic nebulae." In: *Proceedings of the National Academy of Sciences* 15.3, pp. 168–173. DOI: 10.1073/pnas.15.3.168. URL: https://www.pnas.org/doi/abs/10.1073/pnas.15.3.168.

Hui, L. and N. Y. Gnedin (Nov. 1997). "Equation of state of the photoionized intergalactic medium." In: MNRAS 292.1, pp. 27–42. DOI: 10.1093/mnras/292.1.27.

Iršič, V., M. Viel, T. A. M. Berg, et al. (Apr. 2017a). "The Lyman α forest power spectrum from the XQ-100 Legacy Survey." In: MNRAS 466.4, pp. 4332–4345. DOI: 10.1093/mnras/stw3372.

Iršič, V., M. Viel, M. G. Haehnelt, et al. (July 2017b). "First Constraints on Fuzzy Dark Matter from Lyman-α Forest Data and Hydrodynamical Simulations." In: Phys. Rev. Lett. 119.3, 031302, p. 031302. DOI: 10.1103/PhysRevLett.119.031302.

Ivanov, M. M., M. W. Toomey, and N. G. Karaçaylı (Mar. 2025). "Fundamental Physics with the Lyman-Alpha Forest: Constraints on the Growth of Structure and Neutrino Masses from SDSS with Effective Field Theory." In: Phys. Rev. Lett. 134.9, 091001, p. 091001. DOI: 10.1103/PhysRevLett.134.091001.

Jacobus, C., P. Harrington, and Z. Lukić (Nov. 2023). "Reconstructing Lyα Fields from Low-resolution Hydrodynamical Simulations with Deep Learning." In: ApJ 958.1, 21, p. 21. DOI: 10.3847/1538-4357/acfcb5.

Jeffrey, N., L. Whiteway, M. Gatti, et al. (Jan. 2025). "Dark energy survey year 3 results: likelihood-free, simulation-based wCDM inference with neural compression of weak-lensing map statistics." In: MNRAS 536.2, pp. 1303–1322. DOI: 10.1093/mnras/stae2629.

Kacprzak, T. and J. Fluri (July 2022). "DeepLSS: Breaking Parameter Degeneracies in Large-Scale Structure with Deep-Learning Analysis of Combined Probes." In: *Physical Review X* 12.3, 031029, p. 031029. DOI: 10.1103/PhysRevX.12.031029.

Karaçaylı, N. G., C. Ravoux, P. Martini, et al. (Sept. 2025a). "DESI DR1 Lyα 1D power spectrum: Validation of estimators." In: *arXiv e-prints*, arXiv:2509.13593, arXiv:2509.13593. DOI: 10.48550/arXiv.2509.13593.

Karaçaylı, N. G., P. Martini, J. Aguilar, et al. (Oct. 2025b). "DESI DR1 Lyα 1D power spectrum: the optimal estimator measurement." In: J. Cosmology Astropart. Phys. 2025.10, 004, p. 004. DOI: 10.1088/1475-7516/2025/10/004.

Karaçaylı, N. G., P. Martini, J. Guy, et al. (Mar. 2024). "Optimal 1D Lyα forest power spectrum estimation - III. DESI early data." In: MNRAS 528.3, pp. 3941–3963. DOI: 10.1093/mnras/stae171.

Kingma, D. P. and J. Ba (Dec. 2014). "Adam: A Method for Stochastic Optimization." In: *arXiv e-prints*, arXiv:1412.6980, arXiv:1412.6980. DOI: 10.48550/arXiv.1412.6980.

Kobayashi, T., R. Murgia, A. De Simone, et al. (Dec. 2017). "Lyman-α constraints on ultralight scalar dark matter: Implications for the early and late universe." In: Phys. Rev. D 96.12, 123514, p. 123514. DOI: 10.1103/PhysRevD.96.123514.

Lee, K.-G., J. F. Hennawi, D. N. Spergel, et al. (Feb. 2015). "IGM Constraints from the SDSS-III/BOSS DR9 Lyα Forest Transmission Probability Distribution Function." In: ApJ 799.2, 196, p. 196. DOI: 10.1088/0004-637X/799/2/196.

Lemaître, G. (Mar. 1931). "The expanding universe." In: MNRAS 91, pp. 490–501. DOI: 10.1093/mnras/91.5.490.

Liang, Y., P. Melchior, S. Lu, et al. (Aug. 2023). "Autoencoding Galaxy Spectra. II. Redshift Invariance and Outlier Detection." In: AJ 166.2, 75, p. 75. DOI: 10.3847/1538-3881/ace100.

Lidz, A., C.-A. Faucher-Giguère, A. Dall'Aglio, et al. (July 2010). "A Measurement of Small-scale Structure in the 2.2 <= z <= 4.2 Lyα Forest." In: ApJ 718.1, pp. 199–230. DOI: 10.1088/0004-637X/718/1/199.

López, S., V. D'Odorico, S. L. Ellison, et al. (Oct. 2016). "XQ-100: A legacy survey of one hundred $3.5 \lesssim z \lesssim 4.5$ quasars observed with VLT/X-shooter." In: A&A 594, A91, A91. DOI: 10.1051/0004-6361/201628161.

Lu, T., Z. Haiman, and J. M. Zorrilla Matilla (Mar. 2022). "Simultaneously constraining cosmology and baryonic physics via deep learning from weak lensing." In: MNRAS 511.1, pp. 1518–1528. DOI: 10.1093/mnras/stac161.

Lucas Makinen, T., A. Heavens, N. Porqueres, et al. (Jan. 2025). "Hybrid summary statistics: neural weak lensing inference beyond the power spectrum." In: J. Cosmology Astropart. Phys. 2025.1, 095, p. 095. DOI: 10.1088/1475-7516/2025/01/095.

Lukić, Z., C. W. Stark, P. Nugent, et al. (Feb. 2015). "The Lyman α forest in optically thin hydrodynamical simulations." In: MNRAS 446.4, pp. 3697–3724. DOI: 10.1093/mnras/stu2377.

Lynds, R. (Mar. 1971). "The Absorption-Line Spectrum of 4c 05.34." In: ApJ 164, p. L73. DOI: 10.1086/180695.

Maitra, S., S. Cristiani, M. Viel, et al. (Oct. 2024). "Parameter estimation from the Lyα forest in the Fourier space using an information-maximizing neural network." In: A&A 690, A154, A154. DOI: 10.1051/0004-6361/202450259.

Martizzi, D., M. Vogelsberger, M. C. Artale, et al. (July 2019). "Baryons in the Cosmic Web of IllustrisTNG - I: gas in knots, filaments, sheets, and voids." In: MNRAS 486.3, pp. 3766–3787. DOI: 10.1093/mnras/stz1106.

McDonald, P., J. Miralda-Escudé, M. Rauch, et al. (Nov. 2000). "The Observed Probability Distribution Function, Power Spectrum, and Correlation Function of the Transmitted Flux in the Lyα Forest." In: ApJ 543.1, pp. 1–23. DOI: 10.1086/317079.

McGreer, I. D., A. Mesinger, and V. D'Odorico (Feb. 2015). "Model-independent evidence in favour of an end to reionization by $z \approx 6$." In: MNRAS 447.1, pp. 499–505. DOI: 10.1093/mnras/stu2449.

McQuinn, M. and P. R. Upton Sanderbeck (Feb. 2016). "On the intergalactic temperature-density relation." In: MNRAS 456.1, pp. 47–54. DOI: 10.1093/mnras/stv2675.

Meiksin, A. (May 2000). "Spectral analysis of the Lyα forest using wavelets." In: MNRAS 314.3, pp. 566–572. DOI: 10.1046/j.1365-8711.2000.03315.x.

Melchior, P., Y. Liang, C. Hahn, et al. (Aug. 2023). "Autoencoding Galaxy Spectra. I. Architecture." In: AJ 166.2, 74, p. 74. DOI: 10.3847/1538-3881/ace0ff.

Miralda-Escudé, J., M. Haehnelt, and M. J. Rees (Feb. 2000). "Reionization of the Inhomogeneous Universe." In: ApJ 530.1, pp. 1–16. DOI: 10.1086/308330.

Moriwaki, K., T. Nishimichi, and N. Yoshida (July 2023). "Machine learning for observational cosmology." In: *Reports on Progress in Physics* 86.7, 076901, p. 076901. DOI: 10.1088/1361-6633/acd2ea.

Mukhanov, V. F. and G. V. Chibisov (Aug. 1982). "Vacuum energy and large-scale structure of the Universe." In: *Soviet Journal of Experimental and Theoretical Physics* 56.2, p. 258.

Murphy, M. T., G. G. Kacprzak, G. A. D. Savorgnan, et al. (Jan. 2019). "The UVES Spectral Quasar Absorption Database (SQUAD) data release 1: the first 10 million seconds." In: MNRAS 482.3, pp. 3458–3479. DOI: 10.1093/mnras/sty2834.

Nasir, F., P. Gaikwad, F. B. Davies, et al. (Oct. 2024). "Deep learning the intergalactic medium using Lyman-alpha forest at $4 \leq z \leq 5$." In: MNRAS 534.2, pp. 1299–1316. DOI: 10.1093/mnras/stae2153.

Nayak, P., M. Walther, and D. Gruen (Oct. 2025). "LyαNNA II: Field-level inference with noisy Lyα forest spectra." In: *arXiv e-prints*, arXiv:2510.19899, arXiv:2510.19899. DOI: 10.48550/arXiv.2510.19899.

Nayak, P., M. Walther, D. Gruen, et al. (Sept. 2024). "LYαNNA: A deep learning field-level inference machine for the Lyman-α forest." In: A&A 689, A153, A153. DOI: 10.1051/0004-6361/202348485.

O'Meara, J. M., N. Lehner, J. C. Howk, et al. (Sept. 2017). "The Second Data Release of the KODIAQ Survey." In: AJ 154.3, 114, p. 114. DOI: 10.3847/1538-3881/aa82b8.

Oñorbe, J., F. B. Davies, Lukić, et al. (July 2019). "Inhomogeneous reionization models in cosmological hydrodynamical simulations." In: MNRAS 486.3, pp. 4075–4097. DOI: 10.1093/mnras/stz984.

Oñorbe, J., J. F. Hennawi, and Z. Lukić (Mar. 2017). "Self-consistent Modeling of Reionization in Cosmological Hydrodynamical Simulations." In: ApJ 837.2, 106, p. 106. DOI: 10.3847/1538-4357/aa6031.

Palanque-Delabrouille, N., C. Yèche, J. Baur, et al. (Nov. 2015). "Neutrino masses and cosmology with Lyman-alpha forest power spectrum." In: J. Cosmology Astropart. Phys. 2015.11, pp. 011–011. DOI: 10.1088/1475-7516/2015/11/011.

Palanque-Delabrouille, N., C. Yèche, N. Schöneberg, et al. (Apr. 2020). "Hints, neutrino bounds, and WDM constraints from SDSS DR14 Lyman-α and *Planck* full-survey data." In: J. Cosmology Astropart. Phys. 2020.4, 038, p. 038. DOI: 10.1088/1475-7516/2020/04/038.

Parks, D., J. X. Prochaska, S. Dong, et al. (May 2018). "Deep learning of quasar spectra to discover and characterize damped Lyα systems." In: MNRAS 476.1, pp. 1151–1168. DOI: 10.1093/mnras/sty196.

Penzias, A. A. and R. W. Wilson (July 1965). "A Measurement of Excess Antenna Temperature at 4080 Mc/s." In: ApJ 142, pp. 419–421. DOI: 10.1086/148307.

Planck Collaboration, N. Aghanim, Y. Akrami, et al. (Sept. 2020). "Planck 2018 results. VI. Cosmological parameters." In: A&A 641, A6, A6. DOI: 10.1051/0004-6361/201833910.

Ravoux, C., M. L. Abdul Karim, E. Armengaud, et al. (Dec. 2023). "The Dark Energy Spectroscopic Instrument: one-dimensional power spectrum from first Ly α forest samples with Fast Fourier Transform." In: MNRAS 526.4, pp. 5118–5140. DOI: 10.1093/mnras/stad3008.

Ravoux, C., M.-L. Abdul-Karim, J.-M. Le Goff, et al. (May 2025). "DESI DR1 Lyα 1D power spectrum: The Fast Fourier Transform estimator measurement." In: *arXiv e-prints*, arXiv:2505.09493, arXiv:2505.09493. DOI: 10.48550/arXiv.2505.09493.

Ribli, D., B. Á. Pataki, J. M. Zorrilla Matilla, et al. (Dec. 2019). "Weak lensing cosmology with convolutional neural networks on noisy data." In: MNRAS 490.2, pp. 1843–1860. DOI: 10.1093/mnras/stz2610.

Riess, A. G., A. V. Filippenko, P. Challis, et al. (Sept. 1998). "Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant." In: AJ 116.3, pp. 1009–1038. DOI: 10.1086/300499.

Rogers, K. K. and H. V. Peiris (Feb. 2021). "Strong Bound on Canonical Ultralight Axion Dark Matter from the Lyman-Alpha Forest." In: Phys. Rev. Lett. 126.7, 071302, p. 071302. DOI: 10.1103/PhysRevLett.126.071302.

Rubin, V. C. (1983). "The Rotation of Spiral Galaxies." In: *Science* 220.4604, pp. 1339–1344. DOI: 10.1126/science.220.4604.1339. URL: https://www.science.org/doi/abs/10.1126/science.220.4604.1339.

Sarkar, A. K. and S. K. Sethi (Sept. 2024). "Weighing massive neutrinos with Lyman-α observations." In: J. Cosmology Astropart. Phys. 2024.9, 003, p. 003. DOI: 10.1088/1475-7516/2024/09/003.

Schaye, J., T. Theuns, M. Rauch, et al. (Nov. 2000). "The thermal history of the intergalactic medium*." In: MNRAS 318.3, pp. 817–826. DOI: 10.1046/j.1365-8711.2000.03815.x.

Schöneberg, N., G. F. Abellán, A. P. Sánchez, et al. (Oct. 2022). "The H_0 Olympics: A fair ranking of proposed models." In: Phys. Rep. 984, pp. 1–55. DOI: 10.1016/j.physrep.2022.07.001.

Seljak, U., A. Slosar, and P. McDonald (Oct. 2006). "Cosmological parameters from combining the Lyman-α forest with CMB, galaxy clustering and SN constraints." In: J. Cosmology Astropart. Phys. 2006.10, 014, p. 014. DOI: 10.1088/1475-7516/2006/10/014.

Slosar, A., V. Iršič, D. Kirkby, et al. (Apr. 2013). "Measurement of baryon acoustic oscillations in the Lyman-α forest fluctuations in BOSS data release 9." In: J. Cosmology Astropart. Phys. 2013.4, 026, p. 026. DOI: 10.1088/1475-7516/2013/04/026.

Sobol', I. (1967). "On the distribution of points in a cube and the approximate evaluation of integrals." In: *USSR Computational Mathematics and Mathematical Physics* 7.4, pp. 86–112. DOI: https://doi.org/10.1016/0041-5553(67)90144-9. URL: https://www.sciencedirect.com/science/article/pii/0041555367901449.

Srivastava, N., G. Hinton, A. Krizhevsky, et al. (Jan. 2014). "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." In: *J. Mach. Learn. Res.* 15.1, pp. 1929–1958.

Telikova, K. N., P. S. Shternin, and S. A. Balashev (Dec. 2019). "Thermal State of the Intergalactic Medium at z $\sim$ 2-4." In: ApJ 887.2, 205, p. 205. DOI: 10.3847/1538-4357/ab52fe.

Theuns, T., J. Schaye, and M. G. Haehnelt (July 2000). "Broadening of QSO Lyα forest absorbers." In: MNRAS 315.3, pp. 600–610. DOI: 10.1046/j.1365-8711.2000.03423.x.

Theuns, T. and S. Zaroubi (Oct. 2000). "A wavelet analysis of the spectra of quasi-stellar objects." In: MNRAS 317.4, pp. 989–995. DOI: 10.1046/j.1365-8711.2000.03729.x.

Tohfa, H. M., S. Bird, M.-F. Ho, et al. (June 2024). "Forecast Cosmological Constraints with the 1D Wavelet Scattering Transform and the Lyman-α Forest." In: Phys. Rev. Lett. 132.23, 231002, p. 231002. DOI: 10.1103/PhysRevLett.132.231002.

Turner, W., P. Martini, N. G. Karaçaylı, et al. (Nov. 2024). "New Measurements of the Lyα Forest Continuum and Effective Optical Depth with LyCAN and DESI Y1 Data." In: ApJ 976.1, 143, p. 143. DOI: 10.3847/1538-4357/ad8239.

Upton Sanderbeck, P. and S. Bird (Aug. 2020). "Inhomogeneous He II reionization in hydrodynamic simulations." In: MNRAS 496.4, pp. 4372–4382. DOI: 10.1093/mnras/staa1850.

Viel, M., G. D. Becker, J. S. Bolton, et al. (Aug. 2013). "Warm dark matter as a solution to the small scale crisis: New constraints from high redshift Lyman-α forest data." In: Phys. Rev. D 88.4, 043502, p. 043502. DOI: 10.1103/PhysRevD.88.043502.

Viel, M., J. S. Bolton, and M. G. Haehnelt (Oct. 2009). "Cosmological and astrophysical constraints from the Lyman α forest flux probability distribution function." In: MNRAS 399.1, pp. L39–L43. DOI: 10.1111/j.1745-3933.2009.00720.x.

Viel, M., J. Lesgourgues, M. G. Haehnelt, et al. (Mar. 2005). "Constraining warm dark matter candidates including sterile neutrinos and light gravitinos with WMAP and the Lyman-α forest." In: Phys. Rev. D 71.6, 063534, p. 063534. DOI: 10.1103/PhysRevD.71.063534.

Walther, M., E. Armengaud, C. Ravoux, et al. (Apr. 2021). "Simulating intergalactic gas for DESI-like small scale Lymanα forest observations." In: J. Cosmology Astropart. Phys. 2021.4, 059, p. 059. DOI: 10.1088/1475-7516/2021/04/059.

Walther, M., J. F. Hennawi, H. Hiss, et al. (Jan. 2018). "A New Precision Measurement of the Small-scale Line-of-sight Power Spectrum of the Lyα Forest." In: ApJ 852.1, 22, p. 22. DOI: 10.3847/1538-4357/aa9c81.

Walther, M., J. Oñorbe, J. F. Hennawi, et al. (Feb. 2019). "New Constraints on IGM Thermal Evolution from the Lyα Forest Power Spectrum." In: ApJ 872.1, 13, p. 13. DOI: 10.3847/1538-4357/aafad1.

Walther, M., N. Schöneberg, S. Chabanier, et al. (May 2025). "Emulating the Lyman-Alpha forest 1D power spectrum from cosmological simulations: new models and constraints from the eBOSS measurement." In: J. Cosmology Astropart. Phys. 2025.5, 099, p. 099. DOI: 10.1088/1475-7516/2025/05/099.

Wang, B., J. Zou, Z. Cai, et al. (Mar. 2022a). "Deep Learning of Dark Energy Spectroscopic Instrument Mock Spectra to Find Damped Lyα Systems." In: ApJS 259.1, 28, p. 28. DOI: 10.3847/1538-4365/ac4504.

Wang, R., R. A. C. Croft, and P. Shaw (Sept. 2022b). "Deep forest: neural network reconstruction of intergalactic medium temperature." In: MNRAS 515.2, pp. 1568–1579. DOI: 10.1093/mnras/stac1786.

Weinberg, S. (2008). *Cosmology*. Cosmology. OUP Oxford.

Wiener, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*. First published edition, based on a classified 1942 wartime report. Cambridge, MA: The Technology Press of The Massachusetts Institute of Technology.

Wolfson, M., J. F. Hennawi, F. B. Davies, et al. (Dec. 2021). "Improving IGM temperature constraints using wavelet analysis on high-redshift quasars." In: MNRAS 508.4, pp. 5493–5513. DOI: 10.1093/mnras/stab2920.

Yèche, C., N. Palanque-Delabrouille, J. Baur, et al. (June 2017). "Constraints on neutrino masses from Lyman-alpha forest power spectrum with BOSS and XQ-100." In: J. Cosmology Astropart. Phys. 2017.6, 047, p. 047. DOI: 10.1088/1475-7516/2017/06/047.

Zaldarriaga, M. (Jan. 2002). "Searching for Fluctuations in the Intergalactic Medium Temperature Using the Lyα Forest." In: ApJ 564.1, pp. 153–161. DOI: 10.1086/324212.

Acknowledgments

I begin by expressing my sincere gratitude toward the German Academic Exchange Service (DAAD) and the German government for providing me with a prestigious scholarship out of public funds for carrying out this doctorate at the LMU Munich. Without it, I would not have been able to move to Germany and carry out this research. I am thankful to the Faculty of Physics of the LMU Munich and the Leibniz Rechenzentrum (LRZ) of the Munich Scientific Network for making invaluable computing resources available for this dissertation work.

To Prof. Daniel Grün, my doctoral advisor, I am forever indebted. From day one until now, you have mentored and supported me unwaveringly. I thank you: for identifying the potential in me; for helping and guiding me remotely with my research proposal for the DAAD scholarship in the time of the COVID-19 pandemic; making sure my move to Germany and LMU was smoother than I could imagine; for creating a fun and enriching research environment; for bringing together an even more fun and supportive ACAI group; for putting me in collaboration with Michael; for providing financial reinforcement for my sustenance in Munich and also for a multitude of conferences, academic visits, and observing shifts; for an opportunity to develop teaching skills. But most of all, I thank you for being my teacher. Not only have you been a remarkable research advisor, you have also taught me numerous life lessons and have been a constant source of inspiration. You have pushed me to be a good scientist and a good human being. You have been incredibly patient with me, especially when I made mistakes and the going got tough. I hope I have deserved your unrelenting belief in me.

To Dr. Michael Walther, my scientific mentor and collaborator, I cannot imagine the last four years of my PhD without you by my side. You led me on a stellar journey into the microcosm of the Lyα forest and reionization. With you, I have enjoyed doing research more than I ever thought possible. You have been a fantastic teacher and friend. When I most needed it, you held my hand and showed me the path, and when the times were right, you allowed me to be independent and figure things out on my own. Thank you for making me comfortable enough to ask the stupidest questions and learn by making the silliest errors. I hope to keep our friendship alive forever.

Together to both Daniel and Michael, you have been the dream team of teachers in my doctoral expedition. A Sanskrit *shloka* says,

गुरुर्ब्रह्मा गुरुर्विष्णुः गुरुर्देवो महेश्वरः ।
गुरुः साक्षात् परब्रह्म तस्मै श्री गुरवे नमः ॥

I paraphrase it as follows: "my *Guru* is the living embodiment of the divine spirit whom I revere." You both are such *Gurus* to me.

My special thanks to Patrick and Miriam Gebhardt for welcoming and accommodating me, practically a stranger at the time, in your home when I moved to Germany and initially had no place to live. And since then I have always felt at home with you, like you are my second family in Germany.

I am deeply grateful to the University Observatory (USM) community for an exceptional environment of astronomical research and education. It has been the best place for a passionate astronomer like me to pursue a doctorate at. My love of telescopes has been amply fulfilled here by access not only to the state-of-the-art 2m telescope at Mt. Wendelstein, but also to a historical 200-year-old Fraunhofer refractor right in the USM backyard. To Arno, Raphael, and Luis, thank you for an unforgettable Wendelstein experience! Special words of appreciation to Tadziu for his expert IT support throughout my time at USM.

My sincere gratitude to Drs. Oliver Friedrich and Luca Tortorelli for their inspiring roles of vigorous scientists and senior colleagues. Our spirited discussions on numerous occasions, scientific and otherwise, have been a rare medium of personal growth for me. My thanks to Dr. Céline Péroux for providing an outside perspective with my research from time to time.

I am indebted to three very special friends during my time at the LMU and in Germany: Julian, Bhashin, and Surya. You have been extraordinary companions through thick and thin. You have been there in my brightest moments and my darkest hours. You have reproached and counseled me as much as you have laughed and celebrated with me. I shall cherish those times as long as I live.

In this journey I have met and known many great people at USM with whom I have interacted about science and life. My heartfelt thanks to Berni, Elisa, Yun-Hsin, Jamie, Justin, Dylan, Jed, Anik, Laurence, Rintaro, Johanna, Elsa, Leon, Marco, Jaydeep, Malte, Nikolas, Varun, Zekang, Yunhe, Naman, Sarah, Julius, Alina, Sreyas. My apologies in case I have failed to mention any names. To Sookyung and Uday, I am proud and happy to have been a part of your Masters theses and to have helped you in any small way to achieve milestones in your careers. I wish you both immense success in your future endeavors.

I reserve these final words for my parents. મમ્મી, પપ્પા, thank you so very much for your unfaltering faith in me and your tremendous support that propelled me to keep working hard toward my aspirations. You never, for once, questioned the choices I made in my career and life. Instead of taking an orthodox approach like countless of your peers, you chose to see the world with a progressive lens like me. Without it, I could have never even imagined moving to another country in pursuit of higher education. You never once complained that I did not visit you even a single time in the last four years; I understand that it has been tough, but please know that I have spent every one of those days working toward my PhD and in turn making you proud. Thank you for your understanding even when you didn't understand what I was doing, for your love and support when I didn't have any, and for raising me to always pursue my curiosity—the very trait that makes me a scientist.

॥ विद्यायाः न अंतः अस्ति ॥