Interpreting multi-wavelength observations of the Epoch of Reionization from next generation telescopes

M.Sc. Thesis

By

Sohini Dutta

DEPARTMENT OF ASTRONOMY, ASTROPHYSICS AND SPACE ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

MAY 2022

Interpreting multi-wavelength observations of the Epoch of Reionization from next generation telescopes

A THESIS

Submitted in partial fulfillment of the
requirements for the award of the degree
of
Master of Science

by

Sohini Dutta

DEPARTMENT OF ASTRONOMY, ASTROPHYSICS AND SPACE ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

MAY, 2022

INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled "**Interpreting multi-wavelength observations of the Epoch of Reionization from next-generation telescopes**" in the partial fulfilment of the requirements for the award of the degree of **MASTER OF SCIENCE** and submitted in the **DISCIPLINE OF ASTRONOMY, ASTROPHYSICS AND SPACE ENGINEERING, Indian Institute of Technology Indore**, is an authentic record of my own work carried out during the time period from August 2020 to May 2022 under the supervision of Dr Suman Majumdar, Assistant Professor, Department of Astronomy, Astrophysics and Space Engineering, Indian Institute of Technology, Indore.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

Dutta

Signature of the student with the date
Sohini Dutta

This is to certify that the above statement made by the candidate is correct to the best of my/our knowledge.

Suman Majumdar 20/05/2022

Signature(s) of the Supervisor(s) of

M.Sc. (with date)

Dr Suman Majumdar

Sohini Dutta has successfully given his/her M.Sc. Oral Examination was held on **05 May 2022**.

Manoneeta Chakraborty

Signature of PSPC Member #1
Date: 22/05/2022

Signature of PSPC Member #2
Date: 23/05/2022

Signature of PSPC Member #3
Date: 23/05/2022

Suman Majumder

Signature(s) of Supervisor(s).
Date: 20/05/2022

Signature of M.Sc. Co-Ordinator
Date: 25/05/2022

S. Das

Convener, DPGC
Date: 26/05/2022

Manoneeta Chakraborty

Signature of HoD (Officiating)
Date: 26/05/2022

Dedicated to Maa and Papa

ABSTRACT

The Epoch of Reionization (EoR) is one of the least understood periods in the history of the universe. In this project, our goal was to explore the potential of future multi-wavelength observations in understanding this epoch. We first explore the possibility of using the redshifted $158\,\mu m$ line from singly ionized carbon (CII) atoms for this purpose. The $158\,\mu m$ line emitted by the early sources of lights is expected to trace the star formation rate during the EoR. We have attempted to constrain the minimum dark matter halo mass, that hosts the early photon sources, using the CII power spectrum from the EoR using an Artificial Neural Networking (ANN) based signal emulator and Bayesian Inference via a Markov Chain Monte Carlo (MCMC) framework. We discovered that since the CII power spectrum is very featureless (the only feature it has is variation of its amplitude with the minimum host halo mass of the sources), one requires a very large training data set to train the emulator for it to be able to predict the signal power spectrum with an appreciable degree of accuracy. In addition, it was found that despite using a large training set, the emulator seems to have a tendency to produce degenerate data at certain points, especially towards the boundaries of our parameter space. This anomaly was discovered by using the MCMC algorithm in tandem with the emulator and comparing the obtained estimates of the minimum host halo mass with the same obtained from MCMC runs in tandem with the simulator.

We have further shown that the CII $\otimes$ 21cm cross-power spectrum has more features compared to the auto-power spectra of CII and thus would require a smaller training set to emulate the same. In addition, since this cross-power spectrum is dominated by the fluctuations in the HI 21cm field, this can potentially be used as a more robust statistic for constraining multiple EoR parameters. We plan to explore this thoroughly in future and would like to explore this statistics potential to constrain reionization history along with other EoR parameters.

ACKNOWLEDGEMENTS

I would like to express my immense gratitude to my supervisor, Dr Suman Majumdar for his guidance, encouragement, and valuable lessons. I could not have found the strength to complete this project had it not been for his timely interventions in times of stress and appropriate motivation in times of lethargy. I am grateful to my PSPC members Prof Abhirup Datta, Dr Manoneeta Chakroborty and Prof Subhendu Rakshit for highlighting the possible flaws in this project in the numerous evaluations over the past year. I would also like to thank Mr Chandra Shekhar Murmu and Mr Himanshu Tiwari for their constant guidance and assistance in helping me overcome the obstacles I faced during the course of this project. Without their limitless patience, my codes would still be bug-infested. I would like to thank my parents for their unconditional support in my ambitions, without which I can not even fathom being able to pursue my goals. Thank you for putting up with my tantrums, and for never giving up on me. And lastly, I would like to extend my thanks to my peers and friends, Saswata for keeping me sane (and for the excellent food), Gursharanjit for your help in understanding the enigma that is ANN, and Rashmi for being my pseudo-roommate. I would also like to thank my classmates for always inspiring me to do better. Thank you for making this project a lot more enjoyable!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The quest to understand the vast frontiers of space beyond our own little planet is perhaps the oldest endeavour of humankind. It is therefore quite surprising that after thousands of years of indagation and countless hours spent by humankind in order to study and demystify the universe, we still know strikingly little of its history and structure.

Though initiated as a purely philosophical and theological discipline, cosmology is now one of the cutting edges of science. It is the study of the universe and its evolution as a whole. The first firm step of cosmology from theology to science was, in my opinion, due in part to the invention of the telescope by Galileo Gallilei during the renaissance. The ability to see celestial objects with our own eyes firmly established the cosmos in the physical realm instead of the metaphysical one in the minds of most people.

Subsequently, the many breakthroughs in the field of science that came in the centuries since Galileo first trained his telescope skyward, paved the way for Einstein to completely revolutionize our understanding of how the universe potentially works.

Today, we stand on the shoulders of these great men and women as we attempt to once more do what thousands before us have done for millennia; look up and ponder where we came from.

1.1 A brief history of the universe

Combining Einstein's theory of general relativity and the standard model of particle physics gave rise to a theory of evolution of the universe that is popularly known as the Big Bang theory. It is a more generic form of the concordance model or the ΛCDM model (also known as LCDM model) of the universe.

The fundamental idea for such a model arose from the fact that if we rewind time according to Edwin Hubble's observations (Hubble, 2014), then intuitively, one can come to the conclusion that the universe must have been denser in the past. This is exactly what the ΛCDM model suggests. The universe and everything in it was born from a singularity of infinite density at an event called the Big-Bang.

The LCDM model, as the name suggests, consists of two key components that sets it apart from the steady-state models of the pre-Hubble era. It contains a cosmological constant (arising from Einstein's general theory of relativity and denoted by Λ) associated with dark energy and cold dark matter. The cosmological constant is the component of this model that contributes to the expansion of the universe and the cold dark matter aids in structure formation.

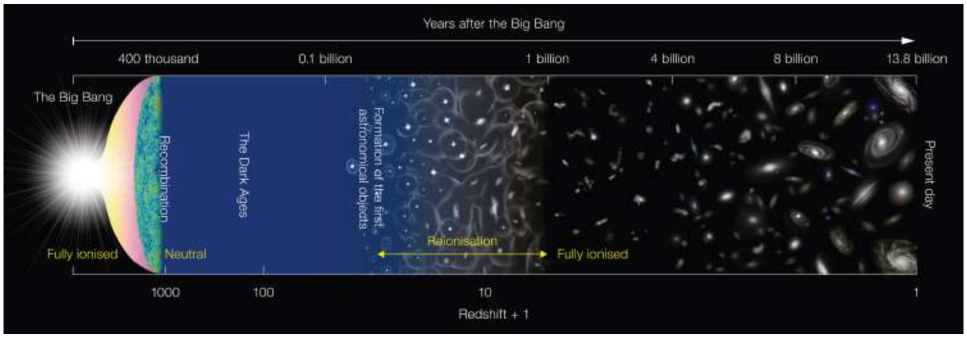


Figure 1.1: The evolution of the universe (Credit: NAOJ).

After the big bang, space-time expanded rapidly during a brief period called the inflation. In its infancy, the universe was a hot, uniform soup of subatomic particles and photons. Eventually, it cooled down to a temperature where the universe becomes too cold for nuclear fusion but remains too hot for neutral atoms to exist. So, the universe continued to cool for about another 18,000 years. This is when the recombination epoch begins. Although atoms had started forming, matter and radiation were still coupled, and were interacting during this period. Small perturbations in this plasma travelled as acoustic waves, producing over and under-densities in their path.

Eventually, after further cooling of the universe, radiation finally decoupled and streamed out freely as what we now know call the Cosmic Microwave Background. This happened about 400,000 years after the Big Bang. This epoch is also known as the epoch of last scattering. It is named as such because this is the epoch when an average CMB photon last scattered from an electron. This primordial radiation contains the imprints of the tiny fluctuations in the matter that existed right before it decoupled from matter. The CMB has an almost perfect black body spectrum. In fact, the CMB provides the best example of a blackbody spectrum we have found so far. Fluctuations in the CMB are of the order of a few hundredths of a kelvin, indicating a largely smooth, featureless universe.

In time, the expanding and cooling plasma allowed the baryons to recombine to form the first neutral hydrogen atoms. The radiation decoupled from baryons and fluctuations stopped propagating through acoustic means. This recombination of matter led to a period called as the dark ages. During this time period, larger structures began forming as cold dark matter started collapsing into halos and catalyzing structure formation.

As a result of this development, the first stars and galaxies materialised in the universe. The small perturbations that had previously existed in matter, started to grow, aided by the underlying cold dark matter. Large gravitational wells opened up that sucked in even more gas and grew in size. Thus, the first stars and galaxies were formed. This period is perhaps aptly named as the Cosmic Dawn. This happened about 250-350 million years after the Big Bang. These primordial entities became powerful sources of radiation as the stars went through nuclear fusion to sustain themselves.

The ultraviolet radiation from these sources started ionizing the neutral hydrogen that formed the intergalactic medium around the luminous structures. This happened in bubbles around the galaxies that had already formed. This curious period is known as the Epoch of Reionization. As a result of reionization, the universe became "transparent" to the rest of the electromagnetic spectrum once the neutral hydrogen was almost entirely ionized (Loeb and Furlanetto, 2013).

The cosmic dawn and epoch of reionization are particularly interesting since the first stars and galaxies were forming for the first time and developing around this time. We have some idea of what happened before reionization using our study of the cosmic microwave background. And what happened afterwards, can be studied across the electromagnetic spectrum. But what happened during this period of time is largely unknown, because it can be probed using very specific transition lines. These signals are incredibly faint and buried in bright foreground contaminants. We have been largely unsuccessful in observing them until recently due both in parts to our lack of understanding of this foreground and other technical challenges. The epoch of reionization is expected to give us unforeseen insight into the process of structure formation. Signals from this epoch are also being used to constrain cosmological constants as done by (McQuinn et al., 2005) during

Ω_Λ	0.685
Ω_m	0.315
h	0.674
σ_8	0.811

Table 1.1: Standard Cosmological parameters. Values taken from Lahav and Liddle, 2006.

relevant time of the universe to test our cosmological model.

Chapter 2

The Epoch of Reionization

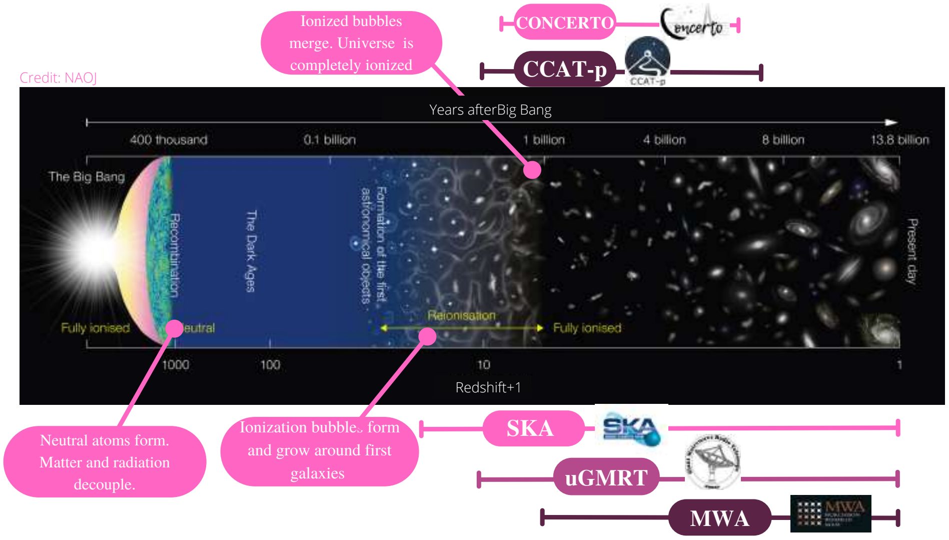


Figure 2.1: Various CII and 21cm experiments and their observational limits.

The CMB suggests that for the first time, hydrogen atoms formed about 400,000 thousand years after the big bang (corresponding to a redshift of 1100). However, a billion years later, UV radiation from newly formed stars and galaxies led to the same gas being ionized again into its constituent protons and electrons. This period was known as the epoch of reionization. It is estimated to have lasted from a redshift of about 15 to 6 (Choudhury and Ferrara, 2006).

2.1 Theory of Reionization

The course of reionization can be determined by counting photons from all galaxies as a function of time. To model reionization, one essentially needs to simulate the number of photons produced by each luminous source, or each group of luminous sources, and the effect such photons had on their local universe.

In this case, since the IGM was almost entirely made up of neutral hydrogen, we are primarily concerned with sources that are capable of producing ultraviolet radiation to ionise the hydrogen. There were two categories of such sources in the early universe; stellar radiation from galaxies and Quasars. However, during that time, galaxies were smaller and younger and hence, they hosted inconsequentially small black holes. This implies that the only sources we should account for are stars.

The ionized medium consists of two phases at this point: the highly ionized region and the ionization front that separates the ionized region from the neutral region.

The simplest model for reionization is the case of a single, isolated galaxy ionizing the IGM surrounding it. The formation of these ionized HII regions around individual galaxies, known as ionized bubbles, is what drives reionization.

Assuming a spherical, ionized volume V, separated from surrounding neutral hydrogen by an ionization front, and neglecting recombination events, the ionized proper volume V_p can be determined by:

$$\bar{n}_H V_p = Q_i \tag{2.1}$$

Here, $\bar{n}_H$ is the mean number density of hydrogen atoms and Q_i is the total number of ionizing photons produced by the source.

The size of the resulting HII regions depends on the source producing the ionizing radiation. Let us consider the total mass of the ionizing source halo to be M_h. We also assume that baryons are incorporated into the stars at an efficiency rate of $f_\star$. The escape fraction of the ionizing radiation is f_{esc} is the fraction of ionizing photons that escape the host galaxy without getting absorbed. Also, let N_{ion} be the number of ionizing photons per baryon. We also introduce a parameter A_{He} as a correction to convert the number of ionizing photons to the number of ionized hydrogen atoms (assuming that helium is singly ionized as well) given by:

$$A_{He} = \frac{4}{4 - 3Y_p} = 1.22 \tag{2.2}$$

where Y_p is the mass fraction of Helium.

This becomes necessary since the ionization potential of HeI is 24.4 eV which is sufficiently close to Hydrogen's 13.6 eV ionization potential to result in the simultaneous ionization of both species by the same radiation source.

Together, these factors can be used to determine the ionization efficiency in our simple model by:

$$\zeta = A_{He} f_\star f_{esc} N_{ion} \tag{2.3}$$

Neglecting events of recombination, the maximum comoving radius of the region that the halo mass M_h can ionize is:

$$r_{max} = (\frac{3}{4\pi}\frac{Q_i}{\bar{n}_H^0})^{1/3} = (\frac{3}{4\pi}\frac{\zeta}{\bar{n}_H^0}\frac{\Omega_b}{\Omega_m}\frac{M_h}{m_p})^{1/3} \tag{2.4}$$

(Loeb and Furlanetto, 2013) Here, $\bar{n}_H^0$ is the co-moving number density of hydrogen.

These parameters become instrumental later, when we try to make models of the Epoch of Reionization.

2.2 Probes of reionization

The Epoch of Reionization can be probed with some very specific transition lines that give us insight into the physical conditions of the epoch. To generate a holistic picture of this period in time, we primarily need a tracer for the intergalactic medium and a tracer for the sources. This general requirement is also required to model the progress of reionization.

2.2.1 Lyman-α forest

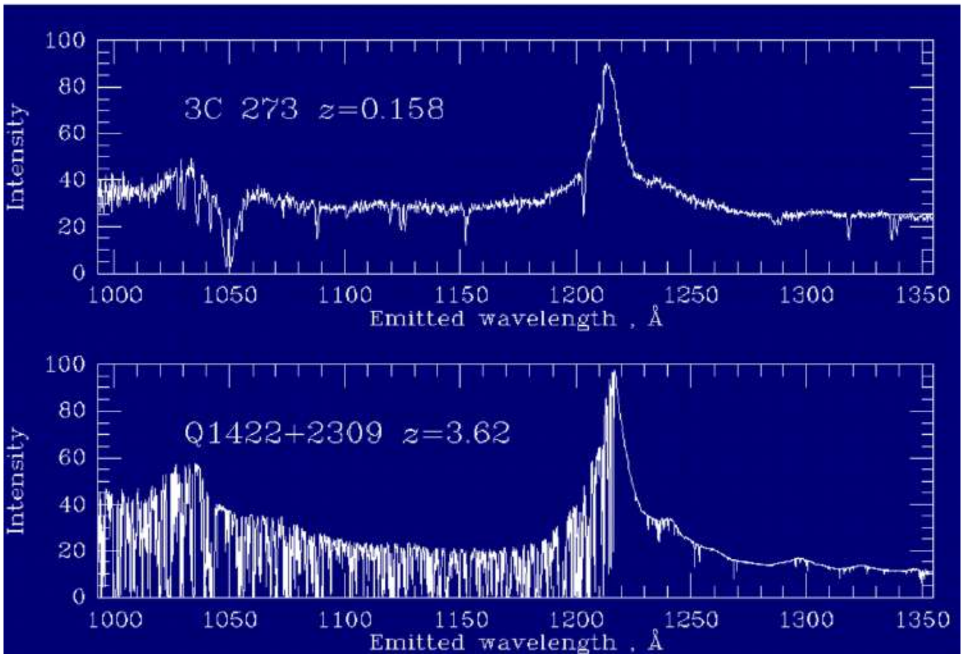


Figure 2.2: Lyman α forest in two quasars, 3C 273 at z=0.158 (top panel) and 1422+2309 at z=3.62 (bottom panel). Credits: "Active Galaxies and Quasars - the Lyman-alpha Forest", n.d.

At high redshift, all quasars have a large number of thin absorption lines that extend blue-ward from the wavelength of the quasar's own Lyman alpha emission line. These are Lyman alpha absorption from foreground structures, in which the quasar light probes a component of cosmic gas that is ordinarily invisible. Since we witness considerably more absorbers (be they clouds, filaments, or even crowding in velocity rather than space) approaching higher redshifts, this component evolves strongly with cosmic time. They haven't fully vanished, though. When the HST mission enabled the first accurate measurements of Lyman alpha at low redshifts, it was discovered that a few of these absorbers remained in the local Universe.They are related with galaxies in general but not precisely - for example, 3C 273 lies behind the Virgo cluster of galaxies and has a handful of absorbers in the cluster's redshift range, but they can't be unambiguously identified in position and redshift with specific surrounding galaxies. The history of galaxy formation may be intricately linked to the evolution of the Lyman-alpha forest.

(2.2) compares the radiated wavelengths of two quasars with very different redshifts, 3C 273 at z=0.158 and 1422+2309 at z=3.62, shifted to a similar scale. The powerful and broad emission peak in the high-redshift quasar is Lyman alpha, which is nearly halved by the commencement of the Lyman alpha forest. Only a few (but far more than zero) Lyman alpha absorbers are visible at low redshift in 3C 273, notably the powerful and broad absorption from its light intersecting the disc of a foreground spiral galaxy (ours). ("Active Galaxies and Quasars - the Lyman-alpha Forest", n.d.)

For absorption, the Cosmological radiative transfer equation gives:

$$I_\nu = I_{\nu_Q}(t_Q)\left(\frac{1}{1+Z_Q}\right)^3 e^{-\tau_\nu} \tag{2.5}$$

Here, I_ν is the observed specific intensity, $I_{\nu_Q}(t_Q)$ is the emitted specific intensity from the quasar, $\left(\frac{1}{1+Z_Q}\right)^3$ is the redshift dilution factor and τ is the optical depth. The optical depth τ is given by the following equation:

$$\tau_\nu \approx 10^5\left(\frac{n_{HI}}{n_H}\right) \tag{2.6}$$

This implies that the observed flux is essentially the emitted flux with a multiplicative factor of exp $(-10^5 X_{HI})$ where X_{HI} or the Neutral fraction is the ratio of mass density of neutral hydrogen to the density of total hydrogen.

This means that the observed flux will be very low, even if the amount of neutral hydrogen is very low. However, we can observe in 2.2 that the Lymann-α forest has certain emission patterns, signalling that the amount of neutral hydrogen fraction is less than 10^{-5}, indicating a highly ionised universe. This is the Gunn-Peterson effect (Gunn and Peterson, 1965), which is a phenomenon that occurs in the Lyman-α forest. As a result, Lymann-α can be used as a probe in EoR research. However, as we go higher in redshift (i.e. $z \geq 5.5$), this feature starts to vanish, and at even higher redshifts, it totally vanishes. As a result, we'll need more probes.

2.2.2 CMB Anisotropies

After last scattering, the radiated photons were streaming in all directions. These produced what we see as the Cosmic Microwave Background (CMB) today. However, when reionization occurred, the resulting free electrons began scattering these CMB photons. As a result, reionization left a permanent trace on these CMB photons, which could enable us to study EoR in principle. Now, the scattering process in question is Thomson scattering, which has an angular dependence. As a result, it redistributes the angular pattern of CMB photons, resulting in polarisation, as observed in the CMB temperature anisotropy.

To find the effect of reionization on the CMB, we study the optical depth from Thomson scattering:

$$\tau = \sigma_T c \int_t^t dt \overline{n_e} (1+z)^3 \tag{2.7}$$

As is evident, the optical depth is proportional to the number of free electrons i.e. $\overline{n_e}$ so, it can be a complementary probe to the Lymann-α transition and provide us previously unavailable information about the EoR.

2.2.3 21cm emission (HI line)

The 21cm line is one of the most important probes of the Epoch of Reionization. The idea is to trace the distribution of the neutral hydrogen, and therefore, trace the distribution of the IGM, since it is primarily composed of neutral hydrogen at this time. It can be used to distinguish ionized and neutral hydrogen quite effectively, and to study the fluctuations in the neutral hydrogen gas. The ionized regions will emit no 21cm radiation, while the neutral regions will. This transition was first suggested theoretically by Henrik van de Hulst in 1944 and first observed by Harold Ewen and Ed Purcell in 1951. (Loeb and Furlanetto, 2013)

The 21 cm radiation occurs from the hyper-fine splitting of the ground (n=1) state of hydrogen. The proton and electron both have independent spins or angular momenta and associated magnetic moments. In the ground state of hydrogen, there are two possibilities, either the proton and the electron will have parallel spins, i.e. their magnetic moments will be anti-parallel or they will have anti-parallel spins. There is a small energy difference of $2\hbar^2$ between these two possible states(Bradt, 2008). Therefore, using

$$\Delta E = h\Delta\nu \tag{2.8}$$

we get a corresponding transition frequency of $\nu \approx 1420.4MHz$ or a wavelength of 21.11 cm. However, observations of the 21cm line are made at wavelengths much longer than 21cm. This is because,while the rest-frame wavelength of this transition is 21cm, due to cosmological redshift, this wavelength gets stretched to longer wavelengths.

Hence, for a source situated at a redshift of z, the observed frequency will be:

$$\nu_o = \frac{1420}{1+z}MHz \tag{2.9}$$

and the observed wavelength will be:

$$\lambda_o = (1 + z) \times 21cm \tag{2.10}$$

Now, the radiative transfer equation for a spectral line is given by:

$$\frac{dI_\nu}{dl} = \frac{\phi(\nu)h\nu}{4\pi}[n_1 A_{10} - (n_1 B_{10})I_\nu] \tag{2.11}$$

here, dl is the proper path length element, $\phi(\nu)$ is the line profile function normalized by $\int \phi\nu d\nu = 1$, subscript 0 and 1 denote the upper and lower atomic levels, n_i denotes the number density of atoms at the different levels and A_{ij} and B_{ij} are Einstein coefficients for the transition between these levels, with i and j being the initial and final states respectively. (Loeb and Furlanetto, 2013)

Making use of the standard relation of atomic physics:

$$B_{10} = (\frac{g_0}{g_1})B_{01} \tag{2.12}$$

and

$$B_{10} = A_{10} = 2.85 \times \frac{c^2}{2h\nu^3} \tag{2.13}$$

and the relative populations of hydrogen atoms in the two states, we can define the spin temperature T_S as follows:

$$\left(\frac{n_1}{n_0}\right) = \left(\frac{g_1}{g_0}\right) exp\frac{-T_\star}{T_S} \tag{2.14}$$

where $T_\star = E_{10}/k_B = 68mK$ is equivalent to the transition energy E_{10}. In our regime of interest, $T_\star \ll T_S$ so all related exponentials can be expanded to leading order.

If we quantify the intensity I_ν by the brightness temperature T_b required of a blackbody radiator with spectrum B_ν such that $I_\nu = B_\nu(T_b)$, then, we can write the radiative transfer equation (eq 2.11) as:

$$T_b'(\nu) = T_S(1 - e^{-\tau_\nu}) + T_R'(\nu)e^{-\tau_\nu} \tag{2.15}$$

where τ_ν is the optical depth and (integral of absorption coefficient along the path of the ray) and T_R is the brightness temperature of the background radiation field (Loeb and Furlanetto, 2013).

Due to cosmological redshift, the emergent brightness $T_b'(\nu)$ creates an apparent brightness on earth as

$T_b(\nu_o) = T_b'(\nu)/(1+z)$

where ν_o is the observed frequency, given by $\nu_o = \nu/(1+z)$.

What makes 21cm radiation special is that essentially the entire neutral IGM is transparent to this radiation during the Cosmic Dawn era. This is because of the weakness of the transition which results in an absorption coefficient of only 1%. Additionally, since its transition energy is so low, it provides a sensitive thermometer of low temperature IGM and it can be seen across the entire IGM against the CMB as a low frequency radio emission.

Experiments like the Square Kilometer Array (SKA), the upgraded Giant Metrewave Telescope (uGMRT), Experiment to Detect the Global Epoch of Reionization Signal (EDGES) and Murchison Widefield Array (MWA) are some of the projects focused on detecting the 21cm signal from the Epoch of Reionization.

2.2.4 $158\mu m$ CII line

The $158\mu m$ CII line is an extremely viable, alternate probe of the Epoch of Reionization. It directly complements the purpose served by the 21cm line. Where 21cm line traces the distribution of the neutral hydrogen gas, and thus the state of reionization, the CII line serves as a tracer of the star formation rate (SFR). CII line therefore traces the distribution of the sources of reionization. This line originates from the fine structure transition between

$${}^2P_{3/2} \rightarrow^2 P_{1/2}$$

in singly ionized carbon atoms.

CII emission originates from

1. **The interstellar medium (ISM):** ISM is the gas permeating the space between stars. Due to the low ionization potential of CII ions, (11.3 eV), CII ions can exist in both neutral and ionized ISM. Hence, we can expect to see some CII emission from these regions. (Sutter et al., 2019)

2. **Photodissociation Regions:** These are dense and warm regions located between HII and regions and molecular clouds. Due to their proximity with energetic sources like O and B stars and AGNs, the physical and chemical properties of these regions are dominated by far ultraviolet radiation(FUV). The FUV radiation also causes transition from atomic to molecular hydrogen and from ionized carbon to carbon mono-oxide. (M. Silva et al., 2015)

3. **Ionized Regions (HII regions):** These are ionized regions of hydrogen gas in and around the galaxies. There has been some evidence that these regions may also produce CII emission. (Anderson et al., 2019)

4. **Cold atomic gas and CO dark molecular gas:** CII line is one of the most dominant coolants in neutral interstellar regions. Hence, it is expected that molecular clouds will also be a good source of CII emission. (Clark et al., 2019)

However, observations of the relative intensity of different intensity lines show that CII is emitted primarily from the photodissocation regions (PDRs).

The intensity of CII emission depends on the collisional rate, which relates it to the gas density and temperature, and therefore also on FUV strength. This intensity can be theoretically estimated as:

$$I_\nu = \frac{h\nu}{4\pi H(z)(1+z)^3} A_{ul} f_{CII}^{grd} n_{CII}(z) \times \frac{g_u}{g_1} exp(-T_{\star,ul}/T_{S,ul}) \times \left[1 - \frac{exp(T_{\star,ul}/T_{S,ul}) - 1}{(2h\nu^3/c^2 I_\nu)_{\nu ul}}\right] \tag{2.16}$$

where f_{CII}^{grd} is the fraction of CII ions at ground level, n_{CII} is the number density of singly ionized carbon atoms, T_S is the spin temperature and $T_\star \equiv h\nu_{ul}/k_B$ and ν_{ul} is the frequency of transition, g_u and g_1 are statistical weights values at 4 and 2 respectively. The Einstein spontaneous coefficient $A_{ul} = 2.36 \times 10^{-6} s^{-1}$. (M. Silva et al., 2015)

Since not all of the parameters involved in the above equation can be directly observed or estimated, the average intensity of a line can also be given by:

$$\bar{I}(z) = \int_{M_{min}}^{M_{max}} dM \frac{dn}{dM} \frac{L(M,z)}{4\pi D_L^2} y(Z) D_A^2 \tag{2.17}$$

where dn/dM is the halo mass function, M is the halo mass, M_{min} and M_{max} are minimum and maximum halo masses (in terms of solar mass $M_\odot$), D_L is the proper luminosity distance, D_A is the comoving angular diameter distance and $y(z) = d\chi/d\nu$ where χ is the comoving distance and ν is the observed frequency. (M. Silva et al., 2015)

Now CII emission can also be linked to other observable quantities. Since it is powered by FUV radiation, there is a link between the two that can be converted to a correlation between CII luminosity and far-infrared (FIR) luminosities given that in actively star forming galaxies, there is an established correlation between FUV and FIR.

CII luminosity can be linked to the FIR luminosity as follows:

$$L_{CII(M,z)}[L_\odot] = 0.003 \times L_{FIR} \tag{2.18}$$

where $R = \frac{L_{CII}}{L_{FIR}} = 0.003$ in nearby, late-type galaxies. (Boselli et al.,

2002) Furthermore, the Infrared (IR) and Far-infrared (FIR) luminosities are connected as:

$$L_{IR}(8-1000\mu m) = (1.89 \pm 0.26) L_{FIR}(40-120\mu m) \tag{2.19}$$

(Cardiel et al., 2003) Then, the IR luminosity is connected to star formation rate (SFR) denoted by ψ as:

$$L_{IR}(M,z)[L_{\odot}] = 5.8 \times 10^{9} \psi(M,z)[M_{\odot} yr^{-1}] \tag{2.20}$$

(Kennicutt, Jr., 1998) Using the above equations, a relation between CII luminosity and SFR can be established as:

$$L_{CII(M,z)}[L_{\odot}] = 9.22 \times 10^{6} \psi(M,z)[M_{\odot} yr^{-1}] \tag{2.21}$$

(M. Silva et al., 2015) To establish an upper and a lower bound on the values of L_{CII}, the L_{CII} and SFR relation can be parameterized as:

$$\log_{10}(L_{CII}[L_{\odot}]) = a_{L_{CII}} \times \log_{10}(\psi[M_{\odot}]) + b_{L_{CII}} \tag{2.22}$$

(M. Silva et al., 2015) Several models of CII luminosity can be developed and tested using the above relation by simply varying the parameters $a_{L_{CII}}$ and $b_{L_{CII}}$.

CII would thus be an excellent tracer of the SFR of the source. Since stars are the primary generators of ionizing photons in the epoch of reioinization, a line intensity mapping of CII luminosity in relevant redshift ranges would thus provide us with a reliable distribution of sources of reionization.

Some experiments like Carbon CII line in post-reionization and ReionizatiOn epoch (CONCERTO) and Cerro Chajnantor Atacama Telescope-prime (CCAT-p) are currently planned to specialise in performing line intensity mapping of the CII signal from the Epoch of Reionization.

2.3 Motivation behind present work

The epoch of reionization is a crucial and largely unexplored part of cosmic history. Studying it promises to provide us with answers regarding galaxy and star formation and the transition of the universe from the homogenous state pre-reionization to the clumpy state post-reionization that we see in the local universe.

As previously mentioned, we must find methods to understand the ongoing physical processes in order to have a complete history of the universe during this epoch. One such method is to use observed Fourier statistics in comparison with Bayesian Parameter estimation. It is especially useful since we do not have tomographic images from the epoch of reionization yet (although future telescopes are expected to be able to image this period effectively).

We would like to understand the astrophysics of the sources using the CII $158\,\mu m$ line intensity mapping. In addition, we would also like to study the correlation between the sources and the intergalactic medium (IGM). Such a cross-correlation is not only expected to give us more information about the evolutionary history of the universe, it will also enable us to estimate the three crucial parameter values, M_{hmin}, N_{ion} and R_{mfp}.

Our specific aim with this project is to conceptualize a method to use Bayesian Inference techniques to constrain the aforementioned parameters. In addition, we also want to investigate whether Artificial Neural Network based emulators would be a computationally cheaper alternative to semi-analytical simulations when it comes to performing such parameter estimations.

Our expectation is that these methods can be used in combination with observed statistics to give us better constraints on the values of these physical parameters.

Chapter 3

Simulating the Epoch of Reionization

In order for us to calculate the power spectra of the 21cm and CII lines, we first need to simulate a 21cm field and a CII field at our redshifts of interest. For this purpose, we first simulated a dark matter distribution using an N-body code. An FoF algorithm (Mondal et al., 2015) was then used to identify the halos from this dark matter distribution.

3.1 Introduction to simulation techniques

3.1.1 N-body Simulation:

N-body simulations are simulations of dynamical systems in physics and astronomy under the influence of physical forces. In cosmology, they are used to simulate linear structure formation such as formation of galactic halos due to the influence of dark matter.

In general, N-body simulations involve a large number of particles (denoted by N). The motion of each of these particles is governed by 6N number of differential equations. Solving these can prove to be extremely computationally expensive. Hence, several calculation optimization methods can be used to decrease computation time and reduce the loss of accuracy.

In this case, the N-body simulation by Mondal et al., 2015 uses the particle-mesh formalism for this purpose. This method is detailed below:

Particle-mesh method:

In this method, space is divided into a grid or a "mesh" system. Particles are assumed to be distributed on the vertices of such a mesh. In order to calculate the gravitational potential on each particle, one only needs to use the Poisson equation for gravitational potential given by:

$$\Delta^2\Phi = 4\pi G\rho \tag{3.1}$$

here G is the universal gravitational constant and ρ is the particle density at each mesh point.

3.1.2 Friends-of-Friends algorithm:

Friends-of-Friends (FoF for short) is a popular tool used in simulations to identify groups of particles. In cosmological simulations as the simulation used in this work, such an algorithm is used to identify dark matter halos in the simulation volume.

In practice, it is done with the help of a property known as "linking length". The algorithm demands that any particle found inside the radius of this linking length be considered linked and part of a group or cluster of particles. Hence, a particle is directly linked to all the particles in its vicinity within the linking length (its "friends" so to speak) and indirectly linked to all the particles that are in the vicinity and within range of its friends (friends-of-friends) ("Friends-Of-Friends Algorithm — SWIFT: SPH With Inter-dependent Fine-grained Tasking 0.9.0 documentation", 2014).

3.2 Simulating HI and CII maps

We generate HI and CII maps from this halo distribution. Our simulation volume is $215^3 cMpc^3$ with 3072^3 grids and 1536^3 particles as in Murmu et al., 2021. It results in a grid separation of 0.07 cMpc and a particle-mass resolution of $\approx 10^8 M_\odot$ (here, $cMpc^3$ stands for "comoving Mpc^3").

The linking length used for the FoF algorithm is 0.2 times the mean interparticle separation in the simulation and the resulting halo mass resolution is $10^9 M_\odot$ (Murmu et al., 2021).

3.2.1 Simulating HI maps

The semi-numerical ReionYuga code (Choudhury et al., 2009;Majumdar et al., 2014;Mondal et al., 2015) was used to simulate the neutral hydrogen maps. It is based on an excursion set formalism (Furlanetto et al., 2004). The neutral hydrogen distribution follows the underlying dark matter distribution. Independent parameters that determine the reionization history are the minimum mass of dark matter halos contributing to reionization (M_{h-min}), ionizing photon emission efficiency N_{ion}, and ionizing photon mean-free path (R_{mfp}). The minimum halo mass sets the lower limit for halo mass, below which no halos are considered to contribute to the reionization of the IGM. The ionizing photon emission frequency is a proportionality constant between the number of ionizing photons being emitted by a source and the halo mass of the halo hosting them. The mean free path of ionizing photons is the maximum radius to smooth the hydrogen and photon density and determine the ionization condition at grid points (Murmu et al., 2021).

By tuning these three parameters, several reionization histories can be generated. The simulation starts with a smoothing radius and ends up at R_{mfp} to check whether at any radius, the smoothed photon density exceeds the smoothed neutral hydrogen density at a given grid point. If this condition is met, the grid point is considered to be ionized (Murmu et al., 2021).

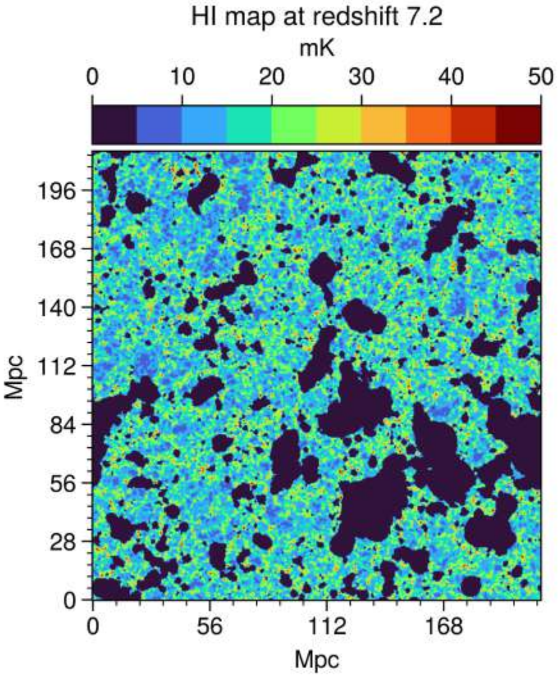


Figure 3.1: HI map at z=7.2. The black bubbles indicate ionized bubbles of HII and the coloured regions are regions containing HI gas.

The reionization history determines how the fluctuations in the HI 21-cm signal will evolve. For example, in a reionization scenario where the R_{mfp} is higher, ionized regions will evolve faster. Similarly, in case of a lower minimum halo mass, reionization will be boosted due to the contribution to the process from smaller halos. Thus one can study various reionization scenarios by simply varying these three parameters.

3.2.2 Simulating CII maps

The CII maps are generated by simply painting the CII luminosity (L_{CII}) onto the dark matter halos identified from the N-body simulation. The CII luminosity L_{CII} depends on the star formation rate which in turn, depends on the mass of the dark matter halo.

Therefore, to estimate CII luminosity, we use the following relation between SFR and halo mass:

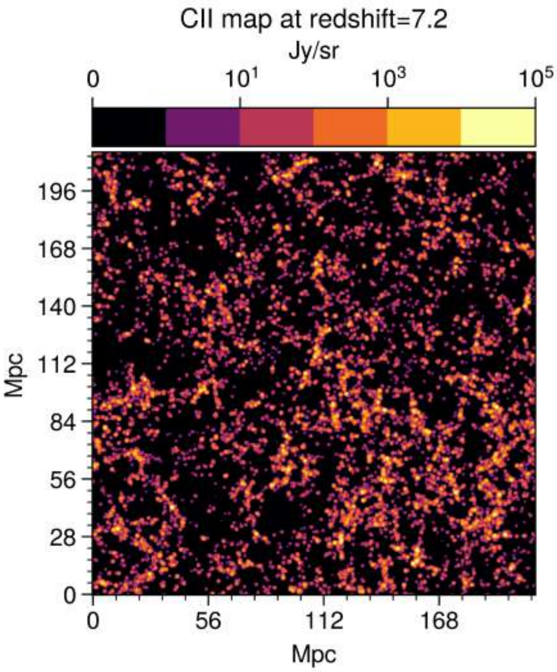


Figure 3.2: CII intensity map at z=7.2. Each point on the map is an average of intensity from a cluster of several galaxies.

$$\begin{aligned} \frac{SFR}{M_\odot yr^{-1}} = & \\ & 2.25 \times 10^{-26}(1 + (z-7)7.5 \times 10^{-2}) \\ & \times M^a \left(1 + \frac{M}{c_1}\right)^b \left(1 + \frac{M}{c_2}\right)^d \left(1 + \frac{M}{c_3}\right)^e \end{aligned} \tag{3.2}$$

with $a = 2.59, b = -0.62, d = -0.4, e = -2.25, c_1 = 8 \times 10^8 M_\odot, c_2 = 7 \times 10^8 M_\odot, c_3 = 10^{11} M_\odot$ adopted from Silva et al 2013 (M. B. Silva et al., 2013). Then the SFR can be related to the CII luminosity as given in Silva et al 2015:

$$\log_{10}\left(\frac{L_{CII}}{L_\odot}\right) = a_{L_{CII}} \times log_{10}\left(\frac{SFR}{M_\odot yr^{-1}}\right) + b_{L_{CII}} \tag{3.3}$$

(M. Silva et al., 2015)

Based on this relation, the CII luminosity distribution is generated from our previously generated halo distribution. Next, using cloud-in-cell method, the CII luminosity distribution is mapped to a coarse grid map of CII intensity using the equation:

$$I_{CII} = \frac{1}{\Delta V} \sum_i \frac{L_{CII}(M_i, z)}{4\pi H(z)} \tag{3.4}$$

(Murmu et al., 2021). This finally gives a CII intensity map as shown in (Fig. 3.2). The choice of parameters in this work corresponds to the m_1 model of Silva et al. 2015 as $a_{L_{CII}} = -0.8475, b_{L_{CII}} = 7.2203$ (M. Silva et al., 2015). This leads to the brightest value of average CII intensity amongst the four models discussed in the paper (Murmu et al., 2021) and corresponds to the fit on high redshift galaxies by De Looze et al., 2014.

3.3 Results of N-body simulations

We look at 4 redshifts with sufficiently spaced out neutral fractions to study how each of these statistical measures vary with reionization state.

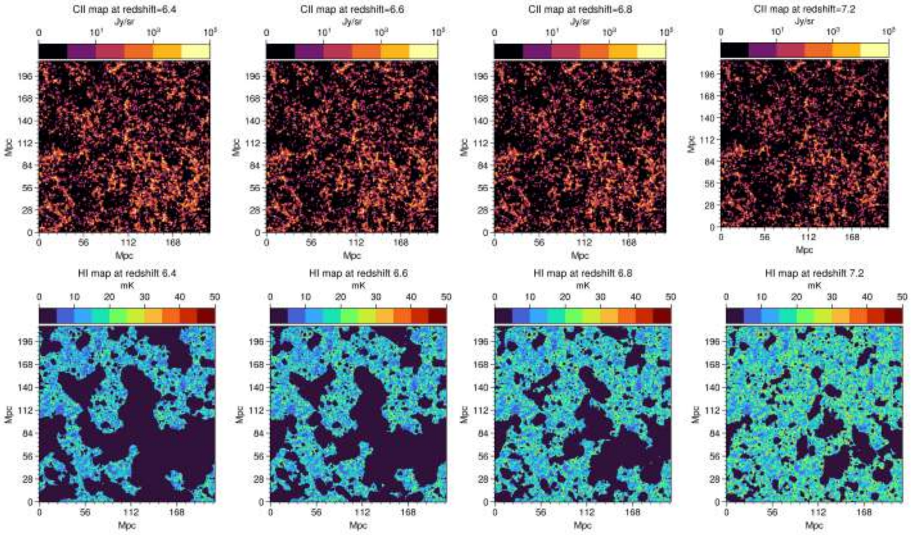


Figure 3.3: The top panel displays CII intensity maps and bottom panel displays HI maps. Each point on the CII maps corresponds to a cluster of sources. The black bubbles in the HI maps are ionized regions and the coloured regions are regions containing neutral hydrogen.

As can be seen from Fig 3.3, the ionized bubbles start appearing first around the location of the brightest ionization sources. This implies, that at larger length scales (smaller k-modes) we expect the CII and HI signals to be strongly anti-correlated. Also, with decreasing redshift, the bubbles in the HI maps start merging and get larger. This is an indication of more and more neutral hydrogen getting ionized, leading to an increase in the size of the bubbles. Conversely, in the CII maps, the number of sources increase with decreasing redshift. This is also intuitive, because one would expect more and more luminous structures such as stars and galaxies to form as the universe got older.

Chapter 4

Statistics

Statistical analysis of data obtained from the Epoch of Reionization is a crucial part of the process of extracting information from them. Additionally, constructing detailed images of 21cm and CII lines from high redshift becomes impossible due to their faintness. This is more applicable to the 21cm radiation. While 21cm is contaminated by bright foregrounds (Datta et al., 2010), CII signals are contaminated by CO emission from galaxies lying between the observer and the source (M. Silva et al., 2015). Regardless, both signals benefit tremendously from statistical analysis methods. In addition, often the instrumental noise is comparable to, and sometimes exceeds the signal (Nasirudin et al., 2020). In such cases especially, statistical quantities become lifesavers .

There are two widely used statistics that one should be familiar with: power spectrum and cross-power spectrum. These are used to study the signals received for fluctuations or perturbations. They are essential tools for understanding how the universe appeared at high redshifts and how signals of interest correlate with each other.

4.1 Power Spectrum

The usage of power spectrum for these signals is motivated by its success in analysis of the CMB signal and in galaxy surveys for constraining the cosmological parameters.

Consider the case of a fluctuation in brightness temperature due to a switch from an HII region to an HI region. This boundary is quite sharp and the brightness temperature drops drastically almost instantaneously. Now, if the fluctuations in the brightness temperature between the two regions 2 and 1 is given by:

$$\delta_{21}(\mathbf{x}) \equiv \frac{[\delta T_b(\mathbf{x}) - \delta\bar{T}_b]}{\delta\bar{T}_b} \tag{4.1}$$

then the power spectrum can be defined as:

$$\langle\tilde{\delta}_{21}(\mathbf{k}_1)\tilde{\delta}_{21}(\mathbf{k}_2)\rangle \equiv (2\pi)^3\delta_D(\mathbf{k}_1 + \mathbf{k}_2)P_{21}(k_1) \tag{4.2}$$

where δ_D is the Dirac Delta function and $\langle ... \rangle$ indicate an ensemble average. (Furlanetto et al., 2004)

Power spectrum in general is the three dimensional Fourier transform of the corresponding two point correlation function and parameterizes the correlations present in the appropriate field. As such, it can also be derived as follows:

Let us denote matter density fluctuations as $\delta(x)$. If we assume that we are simply looking at two different points of the field at the same cosmic time, then, the Fourier transform of this field is given as:

$$\Delta(\mathbf{k}) = \int d^3x\delta(\mathbf{x})e^{(-i\mathbf{k}.\mathbf{x})} \tag{4.3}$$

Then,

$$\langle\Delta(\mathbf{k})\Delta^*(\mathbf{k'})\rangle = \int d^3x \int d^3y\langle\delta(\mathbf{x})\delta(\mathbf{y})\rangle e^{-i(\mathbf{k}.\mathbf{x}-\mathbf{k'}.\mathbf{y})} \tag{4.4}$$

if we make a shift in the origin by **a** then,

$$\langle\Delta(\mathbf{k})\Delta^*(\mathbf{k'})\rangle = \int d^3x \int d^3y\langle\delta(\mathbf{x})\delta(\mathbf{y})\rangle e^{-i(\mathbf{k}.\mathbf{x}-\mathbf{k'}.\mathbf{y})}e^{-i(\mathbf{k}-\mathbf{k'}).\mathbf{a}} \tag{4.5}$$

As $\langle\Delta(\mathbf{k})\Delta^*(\mathbf{k}')\rangle$ is supposed to be homogeneous and isotropic (following the cosmological principle), eq. 4.4 and eq. 4.5 are equivalent and the integrals should not depend on **a**.

$\therefore\ e^{-i(\mathbf{k}-\mathbf{k}').\mathbf{a}} = 1$

$$\begin{aligned}
\langle\Delta(\mathbf{k})\Delta^*(\mathbf{k}')\rangle = & \\
& \int d^3x \int d^3y \langle\delta(\mathbf{x})\delta(\mathbf{y})\rangle e^{-i(\mathbf{k}.\mathbf{x}-\mathbf{k}'.\mathbf{y})-i(\mathbf{k}-\mathbf{k}').\mathbf{a}} \\
& = (2\pi)^3\delta^D(\mathbf{k}-\mathbf{k}')P(\mathbf{k})
\end{aligned} \tag{4.6}$$

Here, P(**k**) is the power spectrum. Additionally, since the universe is homogeneous and isotropic, the direction of k should not control the value of the power spectrum. Hence,

$$P(\mathbf{k}) = P(|k|) = P(k)$$

It is related to the two point correlation function as:

$$P(k) = \int d^3x e^{-i\mathbf{k}.\mathbf{x}}\eta(x) \tag{4.7}$$

Where $\eta(x)$ is the two point correlation function.

We however use the dimensionless power spectrum

$$\Delta^2(\mathbf{k}) = \left(\frac{k^3}{2\pi^2}\right)P(\mathbf{k}) \tag{4.8}$$

which roughly quantifies the variance when the field is smoothed on the scale of $x = 2\pi/k$.

4.2 Cross-Power Spectrum

One of the biggest challenges in working with the power spectra of 21cm and CII lines is the fact that they are heavily contaminated by foreground, particularly in case of 21cm radiation. In case of 21cm, the faint, redshifted signal is buried in bright foreground that is several order of magnitudes higher. This makes the extraction of the signal truly challenging. Although the foreground contamination in CII observations is not quite as strong, it too is contaminated by bright CO emissions from various galaxies. These effects affect the auto-power spectrum calculations significantly.

Cross-power spectrum on the other hand, becomes greatly useful in this regard. It is expected that the 21cm and CII foregrounds will be uncorrelated at large scales (i.e., our scales of interest). Cross-correlating these two uncorrelated fields thus automatically eliminates their effect in the final product.

Analogous to the auto power spectrum above, the cross power spectrum for two signals S1 and S2 is given as:

$$\langle \tilde{\delta}^{*}_{S1}(\mathbf{k}')\tilde{\delta}_{S2}(\mathbf{k})\rangle = (2\pi)^3\delta^D(\mathbf{k}-\mathbf{k}')P_{S1\times S2}(\mathbf{k}) \tag{4.9}$$

And the cross-correlation coefficient for these two signals is given as:

$$r_{S1\times S2}(k) = \frac{P_{S1\times S2}(k)}{\sqrt{P_{S1}(k)P_{S2}(k)}} \tag{4.10}$$

Here $P_{S1\times S2}$ is the cross power spectrum of signals S1 and S2 and P_{S1} and P_{S2} are the auto-power spectra of the signals S1 and S2 (Murmu et al., 2021).

The cross-correlation coefficient indicates the degree and nature of correlation between the two signals. A highly negative coefficient implies that the signals are strongly anti-correlated. A highly positive coefficient implies that the signals are strongly correlated. And a coefficient with values approaching 0 shows uncorrelated signals. Also, the values of the correlation coefficient should always lie between +1 and -1 since it is essentially the cosine of the phase difference of the two fields being studied. The following formalism clarifies this further:

Assume that there are two fields *S1* and *S2*. In general, they can be described by the following $F_{S1} = A(k)e^{i\theta}$ and $F_{S2} = B(k)e^{i\phi}$ in the Fourier space.

From 4.9, the power spectra of these two fields can be written as:

$P_{S1\times S2} =< F_{S1}F^{*}_{S2} >=< Ae^{i\theta}Be^{-i\phi} >= |A||B| < e^{i(\theta-\phi)} >= |A||B|cos(\theta - \phi)$

Similarly, $P_{S1} = |A|^2 cos\theta$ and $P_{S2} = |B|^2 cos\phi$

Also, from 4.10, the corresponding cross-correlation coefficient becomes:

$r_{S1\times S2} = \frac{P_{S1\times S2}(k)}{\sqrt{P_{S1}(k)P_{S2}(k)}} = cos(\theta - \phi)$

4.3 Statistical Analysis from simulated CII and HI maps

In this project, we have computed the HI power spectrum, the CII power spectrum and the $CII \otimes 21cm$ cross-power spectrum from HI and CII maps that have been generated by using the simulations discussed above in sections 3.2.2 and 3.2.1.

4.3.1 CII power spectrum

We have calculated the CII power spectrum (Murmu et al., 2021) for coeval cubes simulated at redshifts of z= 6.4,6.6,6.8 and 7.2. Coeval cubes simply mean that each and every point in a given cosmological box are at the same cosmic time or redshift. The power spectrum has been compared with Murmu et al., 2021 for consistency check.

We observe that the nature and the magnitude of the CII power spectrum does not evolve significantly with redshift. Since the CII line is a tracer of reionization sources, this is to be expected from such a plot. The only observable change is the slight increase in magnitude of the power spectrum.

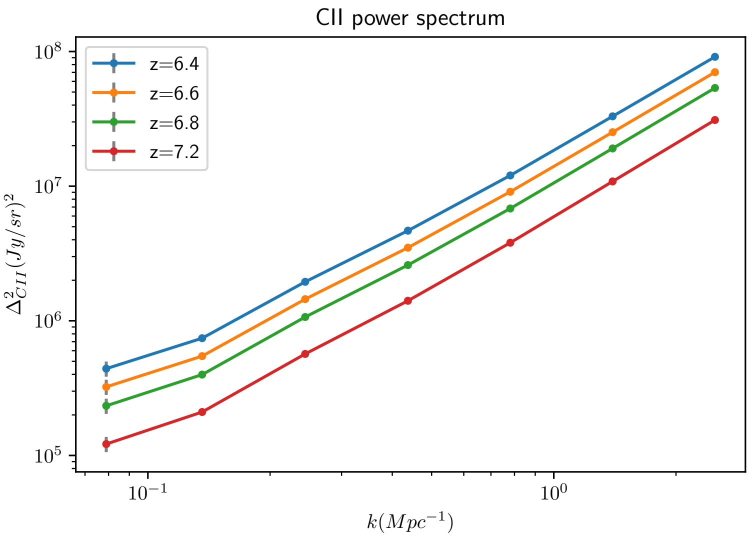


Figure 4.1: CII power spectrum at the four redshifts.

4.4 HI 21cm power spectrum

We have also calculated the HI 21cm power spectrum at the aforementioned four redshifts to compare the evolution of this statistic as reionization progresses. The reionization history that has been used is plotted as a relation between neutral fraction and redshift in (Fig. 4.2).

As is clearly visible from (Fig. 4.3) here, the power spectrum of neutral hydrogen changes significantly with changing redshift. This is a clear indication of the fact that the state of the IGM is constantly changing with changing redshift (and thus neutral fraction).

4.5 $CII \otimes 21cm$ cross-power spectrum

The CII and 21cm power maps were used to also calculate their cross-power spectrum (Eq 4.9) at respective redshifts. Hence, the cross power spectrum has contributions both from the HI field and the CII field. The correlation coefficient (Eq 4.10) shows the nature and degree of correlation between the two fields.

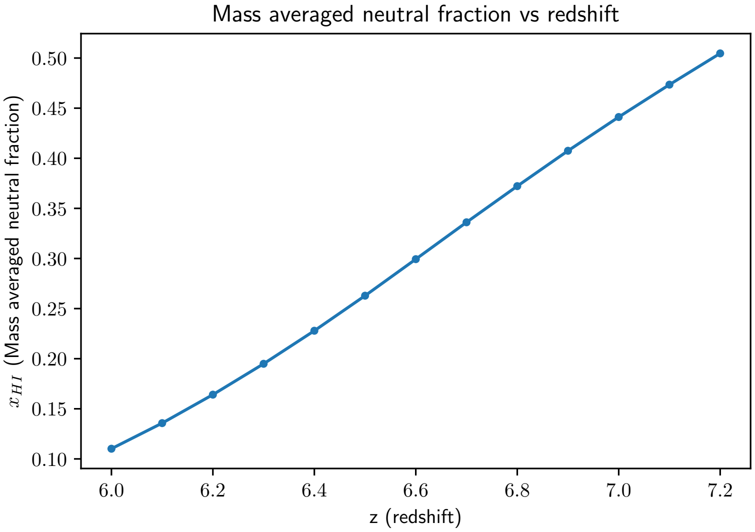


Figure 4.2: Mass averaged neutral fraction vs redshift. Plots like these essentially lay out how the state of reionization progressed in the model being used. This is a visualization of how fast the reionization progressed in our model and what path it took.

As is evident from (Fig. 4.5), the correlation coefficient approaches a value of -1.0 at smaller k-modes or large length scales. This implies that the HI and the CII fields are strongly anti-correlated at such length scales. On the contrary, at larger k-modes or smaller length scales, the correlation coefficient approaches a value of 0, indicating that the two fields are uncorrelated at these k-modes.

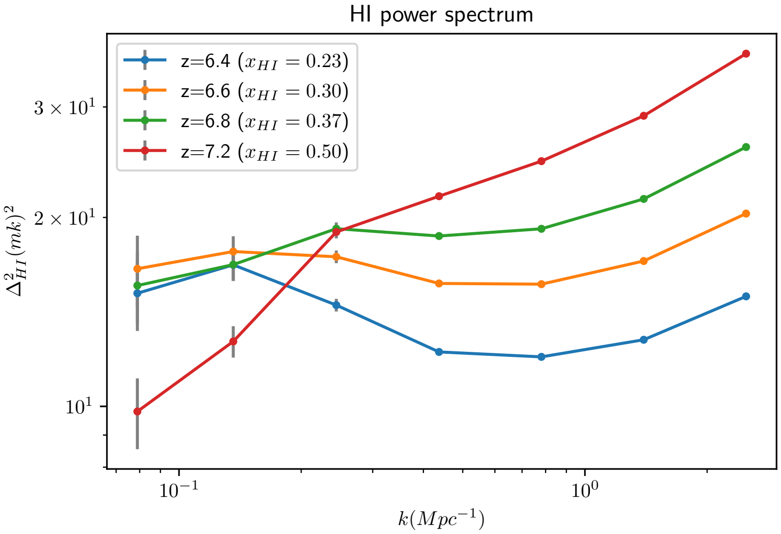


Figure 4.3: The HI power spectrum for all four redshifts.

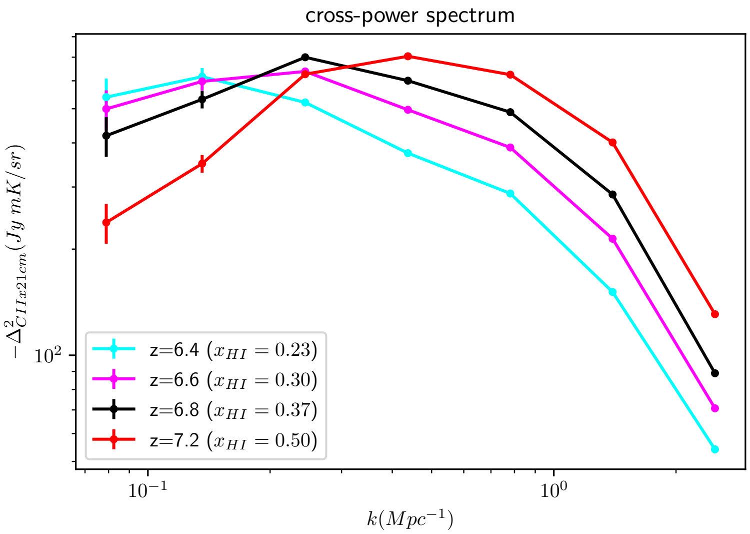


Figure 4.4: The $CII \otimes 21cm$ cross-power spectrum.

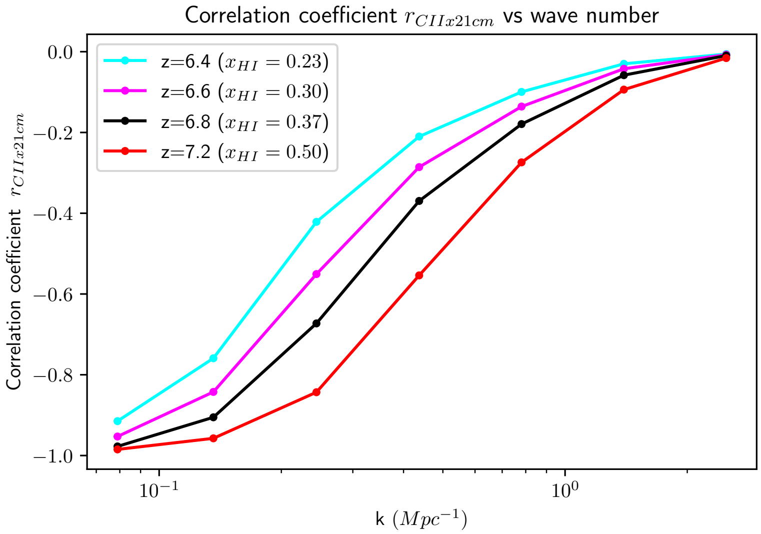


Figure 4.5: Correlation coefficient for $CII \otimes 21cm$ cross-power spectrum.

Chapter 5

Emulating the power spectra

5.1 Introduction

In order to perform parameter estimation using a Markov Chain Monte Carlo formalism, one needs a model that can serve as the theoretical basis for our Bayesian estimator. One option for such an estimator is a semi-numerical simulation. The second, and more computationally efficient one is an emulator.

One of the main goals of this project was to test whether such an emulator can serve as a more effective alternative for parameter estimation purposes. In this work, make use of a simple Artificial Neural Networking structure (ANN) to emulate power spectra.

We have used this code to simulate the CII power spectrum at a redshift of 7.0 and CII$\otimes$21cm cross-power spectrum at the same redshift.

For this project we have implemented the ANN using Python. The training data-set has been prepared by using a semi-numerical formalism similar to the one used in Murmu et al., 2021 and detailed in 3.2.2.

5.2 Introduction to Artificial Neural Networking

Artificial neural networks are computing systems inspired by biological neural networks that exist in higher animal brains (Goodfellow et al., 2016). Any ANN consists of interconnected nodes called artificial neurons that form the basis of the ANN architecture.

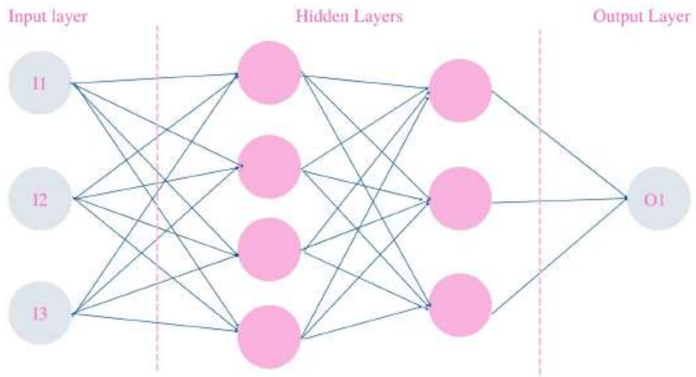


Figure 5.1: A basic ANN architecture.

An ANN primarily consists of three types of “layers”: the input layer, the hidden layers and the output layer.

The number of nodes in the input and output layers directly depend on the nature of the data that the ANN needs to be trained on and are explicitly defined as such.

The number of nodes in the input layer, as the name suggests, corresponds to the number of parameters the data depends on. Its job is to bring data into the ANN system for further processing by the subsequent layers of artificial neurons.

The number of nodes in the output layer on the other hand depends on the dimensions of the output data. The output layer coalesces and concretely produces the end result.

The last kinds of layers in a traditional ANN are the hidden layers. There may be one or more hidden layers in an ANN but most rudimentary systems contain only one or two hidden layers. The number of hidden layers and the number of nodes in each layer is determined by the nature and the complexity of the data. The best combination for these numbers (i.e, the number of layers and the number of nodes in each layer) can be determined by several tuning methods.

Each of these layers are connected to the subsequent layer with the help of an activation function. This activation function determines whether a neuron should be activated or not by calculating the weighted sum and adding the bias with it.

5.2.1 Activation Functions

The purpose of an activation function is to introduce non-linearity within a neural network. They enable back-propagation within an ANN since they provide both the gradient of the loss function and the error, in order to update the weights and biases. Without this non linearity, an ANN will be a simple linear regression model, which is not productive.

Some of the most commonly used activation functions are discussed below.

Commonly used activation functions

Sigmoid function: The sigmoid function is a non-linear activation function that is primarily used in feed-forward neural networks. It is differentiable and real and is defined for real input values containing positive derivatives everywhere. It is represented by:

$$f(x) = \frac{1}{1+e^{-x}} \tag{5.1}$$

The major disadvantages of using sigmoid function are gradient saturation, slow convergence, sharp damp gradients during back-propagation from within deeper hidden layers to the input layers, and non-zero centered output

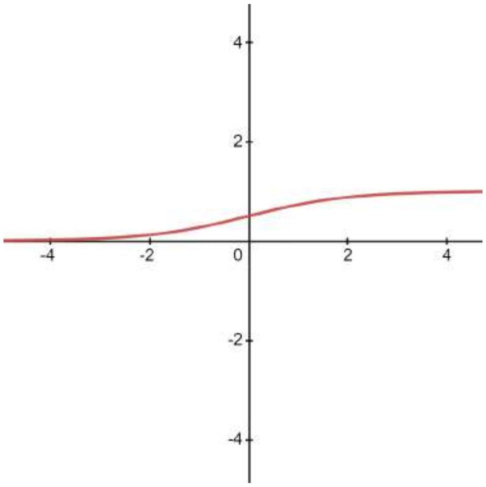


Figure 5.2: The sigmoid function.

that causes the gradient updates to propagate in varying directions.(Nwankpa et al., 2018)

Hyeprbolic tan (tanh): Another type of AF is the hyperbolic tangent function, often known as the tanh function. It's a more rounded, zero-centered function with a range of -1 to 1.

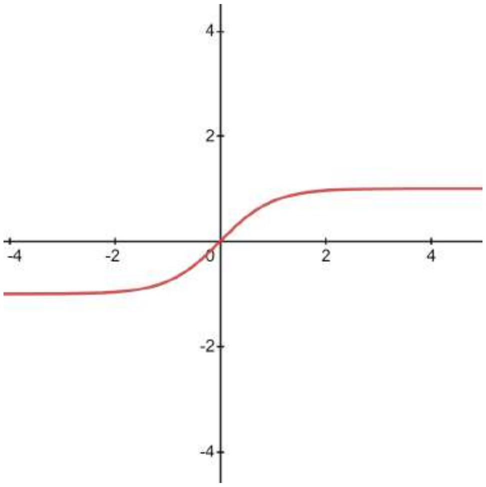


Figure 5.3: Tan hyperbolic function.

Because it provides higher training performance for multilayer neural networks, the tanh function is considerably more widely utilised than the sigmoid function. The tanh function's primary advantage is that it gives a zero-centered output, which helps with back-propagation. The tanh function has mostly been applied to natural language processing and speech recognition applications in recurrent neural networks.

However, the tanh function, like the sigmoid function, has a limitation: it cannot address the vanishing gradient problem. Furthermore, when the input value is 0, the tanh function can only achieve a gradient of 1. (x is zero). As a result, throughout the computing process, the function may produce some dead neurons.(Nwankpa et al., 2018)

Rectified Linear Unit (ReLU): The rectified linear unit is a very commonly used activation function in neural networks. It was first proposed by **nair2010**. The ReLU has a faster learning rate and provides greater generalization and success rate when compared to sigmoid and tanh functions. It has a nearly linear feature which preserves some of the linearity in linear models. These are then easier to optimize with gradient descent methods.

Each input element is subjected to a threshold operation by the ReLU activation function, with values less than zero being set to zero. The ReLU activation function is thus given by:

$$f(x) = \begin{cases} x_i & \text{if } x_i \geq 0 \\ 0 & \text{if } x_i < 0 \end{cases} \tag{5.2}$$

This function overcomes the vanishing gradient problem seen in earlier forms of activation functions by correcting the values of the inputs less than zero and setting them to zero (sigmoid and tanh).

The most major benefit of utilising the ReLU function in calculation is that it ensures faster computation by obviating the need to compute exponentials and divisions, resulting in faster overall computation. The ReLU function also introduces sparsity in the hidden units by squishing the values between zero and maximum.(Nwankpa et al., 2018)

Exponential Linear Units (ELU): The exponential linear units (ELUs) function is a type of AF that can be used to speed up neural network training

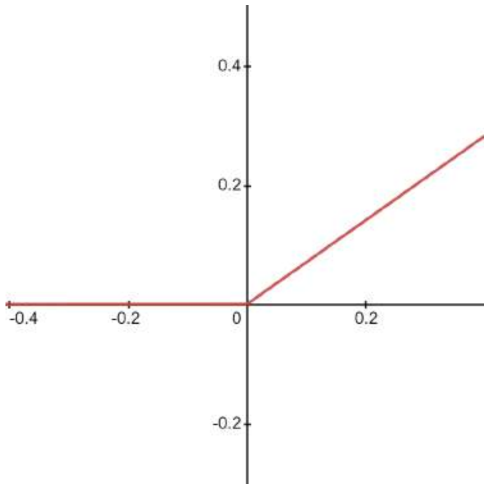


Figure 5.4: ReLU activation function.

(just like ReLU function). The ELU function’s major advantage is that it can solve the vanishing gradient problem by employing identity for positive values and boosting the model’s learning properties.

Negative ELU values drive the mean unit activation closer to zero, lowering computational complexity and increasing learning speed. The ELU is a great alternative to the ReLU since it reduces bias changes during training by pushing mean activation towards zero.

The ELU function is given by:

$$f(x) = \begin{pmatrix} x, & \text{if } x > 0 \\ \alpha e^{x} - 1, & \text{if } x \leq 0 \end{pmatrix} \tag{5.3}$$

The major disadvantage that ELU however has is that it is not a zero centered function. (Nwankpa et al., 2018)

The exponential linear units (ELUs) function is a type of AF that can be used to speed up neural network training (just like ReLU function). The ELU function’s major advantage is that it can solve the vanishing gradient problem by employing identity for positive values and boosting the model’s learning properties.

Negative ELU values drive the mean unit activation closer to zero, low-

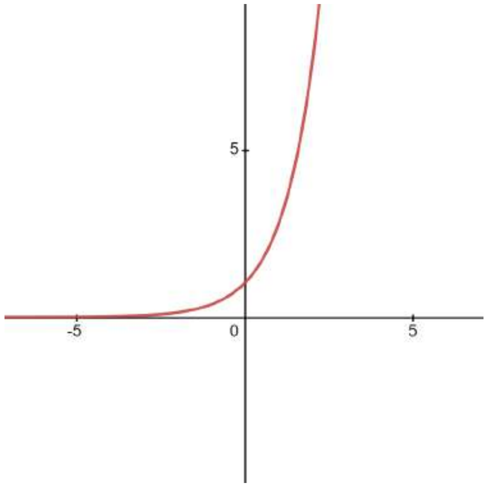


Figure 5.5: ELU activation function.

ering computational complexity and increasing learning speed. The ELU is a great alternative to the ReLU since it reduces bias changes during training by pushing mean activation towards zero.

5.2.2 Loss and loss functions

The cost or loss function has a critical role to play because it must correctly condense all features of the model down to a single number, with improvements in that number indicating a stronger model. A loss function must be chosen for calculating the model's error throughout the optimization phase.

This can be a difficult challenge to solve because the function must capture the problem's attributes and be motivated by project and stakeholder concerns. (Goodfellow et al., 2016)

Maximum Likelihood

To estimate the inaccuracy of a set of weights in a neural network, a variety of functions can be utilised.

We prefer a function that maps the space of candidate solutions onto a smooth (but high-dimensional) landscape that the optimization method may reasonably travel using iterative model weight updates.

Maximum likelihood estimation, or MLE, is a framework for inference that aims to discover the best statistical estimates of parameters using historical training data, which is exactly what the neural network is attempting to achieve.

We have a training dataset with one or more input variables, and we need a model to estimate model weight parameters that translate examples of the inputs to the output or target variable as accurately as possible.

The model attempts to create predictions that fit the data distribution of the target variable given input. A loss function measures how closely the distribution of predictions provided by a model fits the distribution of target variables in the training data when maximum likelihood is used.

The use of maximum likelihood as a framework for estimating model parameters (weights) for neural networks and machine learning in general has the advantage that the estimate of the model parameters increases as the number of examples in the training data set grows. This is referred to as the "consistency" attribute. (Goodfellow et al., 2016)

Mean-Squared error loss

The average of the squared differences between the anticipated and actual values is determined as the Mean Squared Error loss, or MSE for short.

Regardless of the sign of the predicted and actual numbers, the result is always positive, and a perfect value is 0.0. Although it can be employed in a maximum optimization process by making the score negative, the loss value is minimised. (Goodfellow et al., 2016)

5.3 Structure of the CII emulator

The CII emulator developed in the course of this project was built using the *keras* package (Chollet et al., 2015) built into python. It has two densely connected layers, each with 25 neurons. Both layers use the ELU activation functions. The loss function used was Mean Squared Error.

To determine the above "hyperparameters" such as the number of layers, the number of neurons etc, a "hyperparameter tuning" was performed. This was achieved by using the *keras Tuner* package (O'Malley et al., 2019) available for python. Searching for the best parameters can be done by using various techniques such as grid search and random search, but we used the Bayesian Optimization method for our purposes. These values control the over fitting and underfitting of a learning algorithm. To find the best set of hyperparameters, generally, the model is evaluated for each set of proposed hyperparameters.

Grid search selects a grid of hyperparameter values and compares them all. The min and max values for each hyperparameter must be determined by guesswork. A random sample of points on the grid is evaluated using a random search. It performs better than grid search. Smart hyperparameter tuning selects a few hyperparameter settings, assesses the validation matrices, makes hyperparameter adjustments, and re-evaluates the validation matrices.

Bayesian optimizaton is a sequential design strategy developed to optimize black box functions. Black box functions are functions that do not assume any explicit form. This is therefore ideal for our use-case. In this case, the black box function is the ANN based emulator (whose functional form is not known).

The Bayesian strategy is to regard the target function as a random function and apply a prior to it. The prior encapsulates assumptions about the function's behaviour. The prior is updated to produce the posterior distribution over the objective function after accumulating the function evaluations, which are handled as data. The posterior distribution is then used to create an acquisition function (also known as infill sampling criterion) that defines the query point for the next query (Bergstra et al., n.d.).

The data set to train and test the emulator was generated by using the simulations discussed in Chapter 3 (Mondal et al., 2015,Murmu et al., 2021) using the relationship between L_{CII} and dark matter halo mass M as described in Eq. 3.3. In this equation, the following SFR model was implemented:

$$log_{SFR} = \alpha(logM_{halo} - 10) + \gamma log(\frac{1+z}{6}) + \delta \tag{5.4}$$

(Ma et al., 2018)

We used these two models in tandem to simulate power-spectra. This forms the data set that was used to train and test the emulator. Two training sets were created on two different occassions. One consisted of 236 individual samples and the second, much larger set consisted of 2167 samples.

After the whole data-sets were generated, random points were picked from them to serve as the 'training set'. The training set refers to the set that is used to train the emulator and makes up about 90% of the whole simulated data-set. The rest of the data points constitute the testing set. This set is used to verify the trained model's accuracy while predicting unseen data.

The data was scaled using the *Standard Scaler* method available in the Scikit-learn python library (Pedregosa et al., 2011). It performs the following operation on the data:

If x is the point that is being scaled,

$$z = \frac{x - \mu}{\sigma} \tag{5.5}$$

here, μ is the mean of the data and σ is the standard deviation of our data. z is the scaled data point. It standardizes the whole data set by removing the mean and scaling to unit variance. This standardization process is an important pre-processing step in a machine learning algorithm since it ensures that no individual group(s) of data dominates the emulator model (Pedregosa et al., 2011).

Following the standardization process, the simulated data is provided to the emulator for training. Once the training is done, the emulator stores the model generated in a file that can be called at the user's convenience

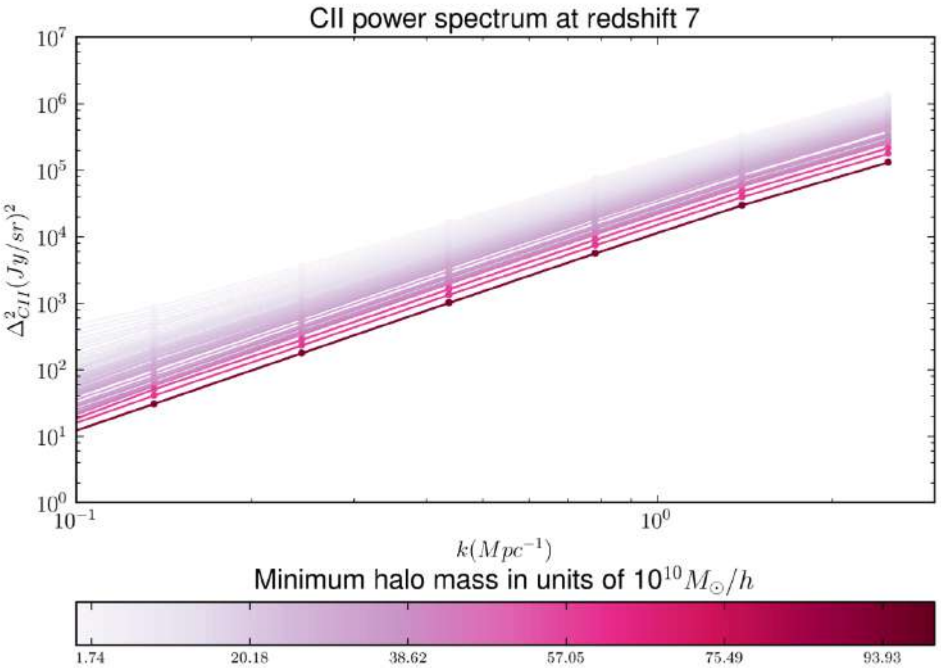


Figure 5.6: Data-set generated using simulations (Murmu et al., 2021,Chapter 3,Sub-section 4.3.1). There are 236 individual sets of CII power spectra. Each of these correspond to a different value of M_{hmin}.

to predict or emulate the power-spectrum for any parameter value that is provided to it. The predicted data from the emulator also needs to undergo an inverse transform process to be physically relevant to our science case.

5.4 Emulation results

At first, a total data set of 236 points was used to train and test this emulator. Out of these, 212 points were randomly picked to form the training data set and 24 points were randomly picked to form the test data set. The value of M_{hmin} was varied linearly within the range of $0.073 \times 10^{10} M_\odot h^{-1}$ and $158.293 \times 10^{10} M_\odot h^{-1}$. For each data point, the power spectrum for the CII signal was calculated based on the corresponding value of M_{hmin}. The M_{hmin} determines the smallest mass of dark matter halo that we considered contributes to the CII emission process for each case.

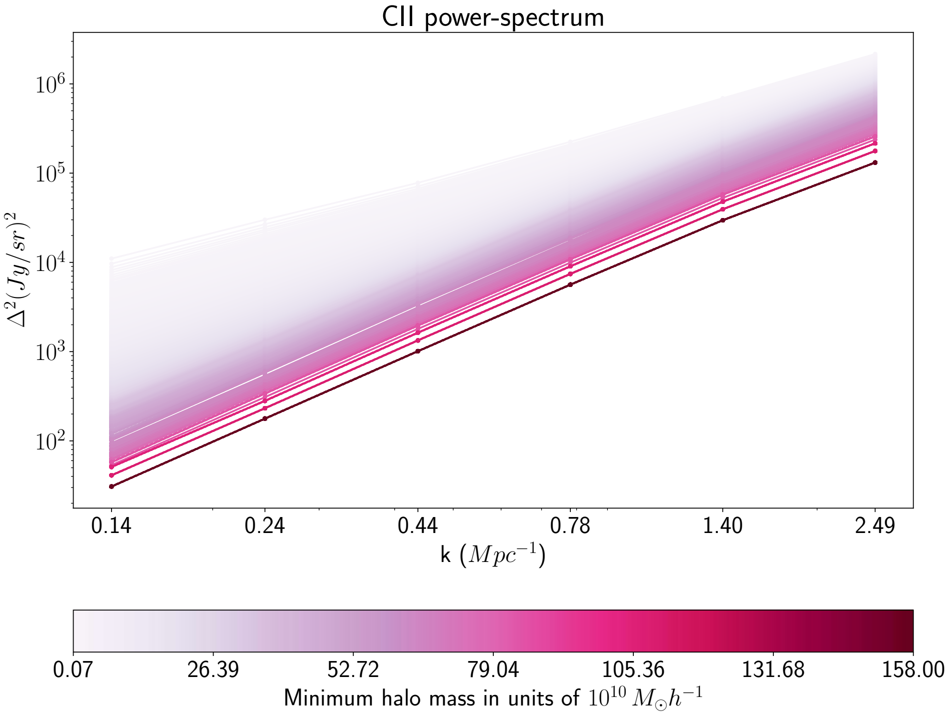


Figure 5.7: Data-set generated using simulations (Murmu et al., 2021,Chapter 3,Sub-section 4.3.1). There are 2167 individual sets of CII power spectra. Each of these correspond to a different value of M_{hmin}.

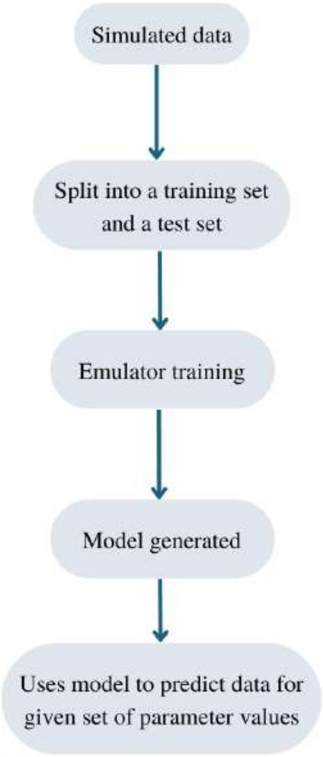

Simulated data
Split into a training set and a test set
Emulator training
Model generated
Uses model to predict data for given set of parameter values

Figure 5.8: Flowchart laying out the basic steps in the emulator.

This emulator terminates the training process if the validation accuracy does not improve after 20 training epochs. In addition, a model is stored at every instance of accuracy improvement. This model is written over the following time such an instance occurs and so on and so forth. These two processes together ensure that a) the data isn't overfitted and b) only the highest accuracy model is stored for final use.

This produced a model with a prediction accuracy of 95.3%. This implies, that it was able to predict the correct values for 95.3% of the unseen test parameters we were providing.

Hence, we picked nine random points from our test set of parameters and used the trained model to predict the CII power spectrum and compare it with the existing simulated power spectrum.

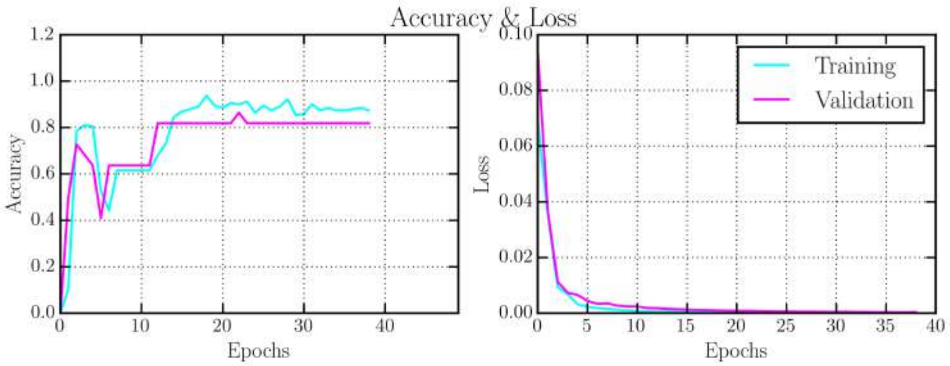


Figure 5.9: This graph shows how the accuracy and loss function of the emulator progressed while training it using the dataset containing 236 samples.

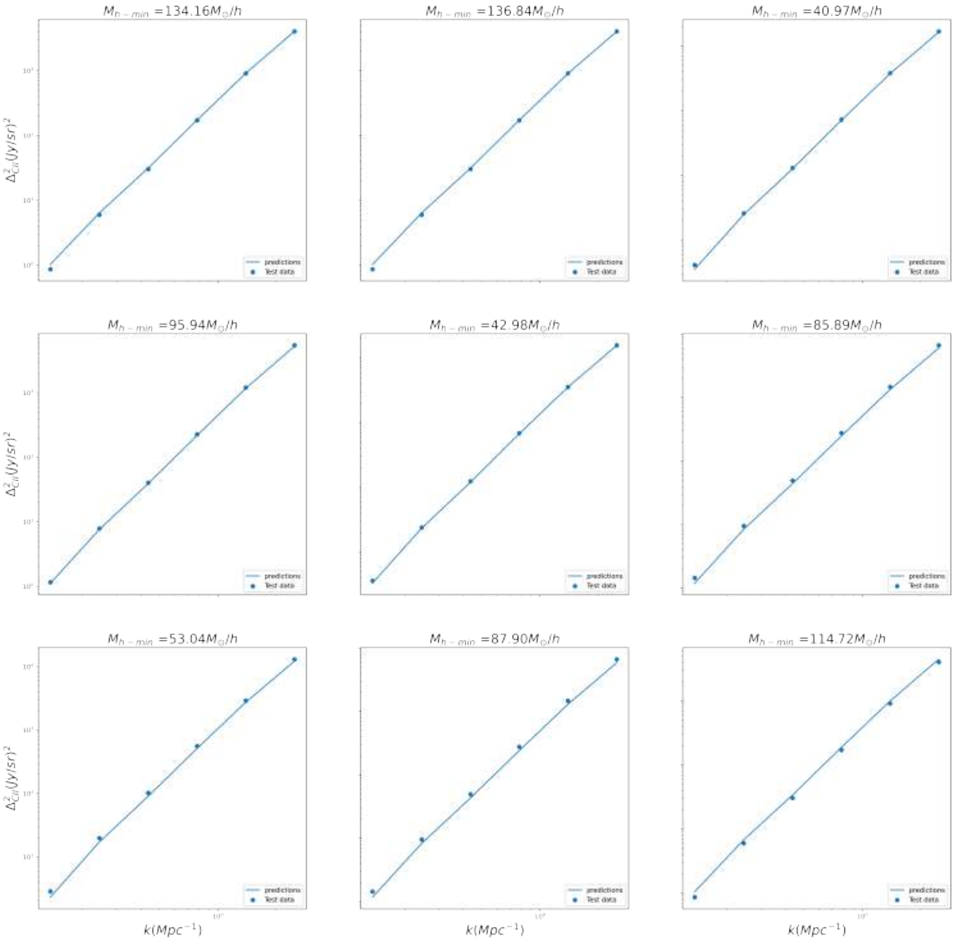


Figure 5.10: This panel compares power spectra as they were simulated and then were predicted by using the ANN after training it with 212 data points. The dots are ANN predictions and the lines are the simulated power spectra.

However, we discovered after running our parameter estimation algorithm, that this version of the emulator was not sufficiently accurate for our use case (discussed in the next chapter). It essentially produced degenerate data for two distinct values of parameters. This led to a biased estimation from the MCMC algorithm.

Hence, we decided to generate a second, much larger data set to train the same emulator. This data-set has 2167 samples and the smallest halo contains 10 dark matter particles while the largest halo contains about 21,000 dark matter particles.

A prediction accuracy of 96.8% was achieved after training with the larger data set. This indicates that out of all the unseen test points the model has attempted to predict, it was able to correctly predict 96.8% points. The predicted power-spectra were also checked against simulated power-spectra.

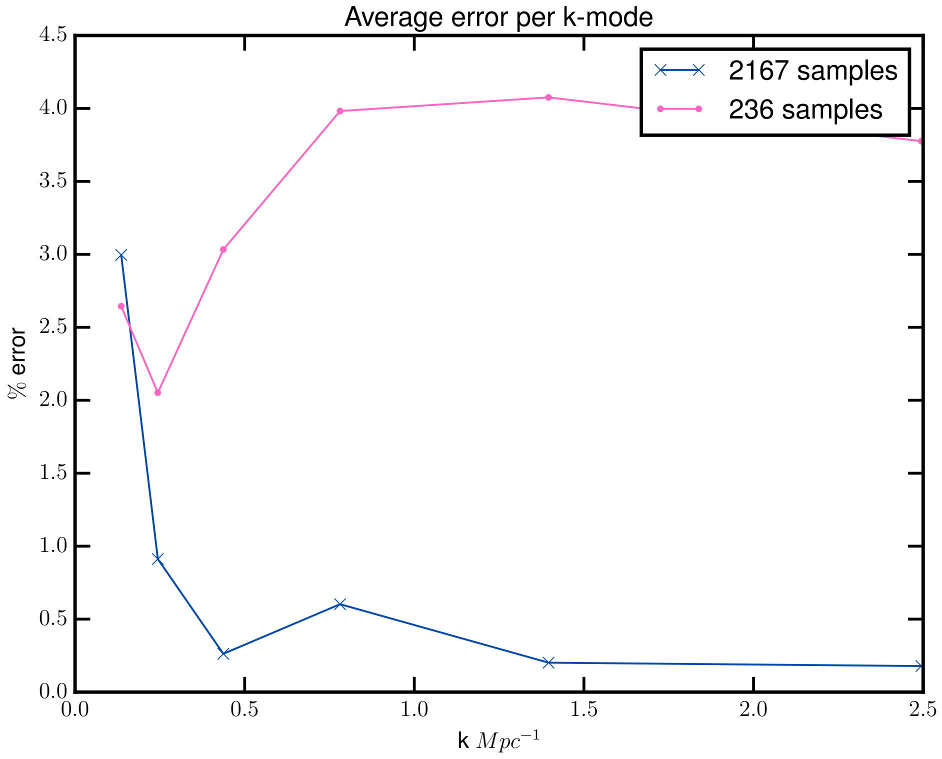


Figure 5.11: Average error comparison between the two models.

We found that this model is significantly more robust at emulating the

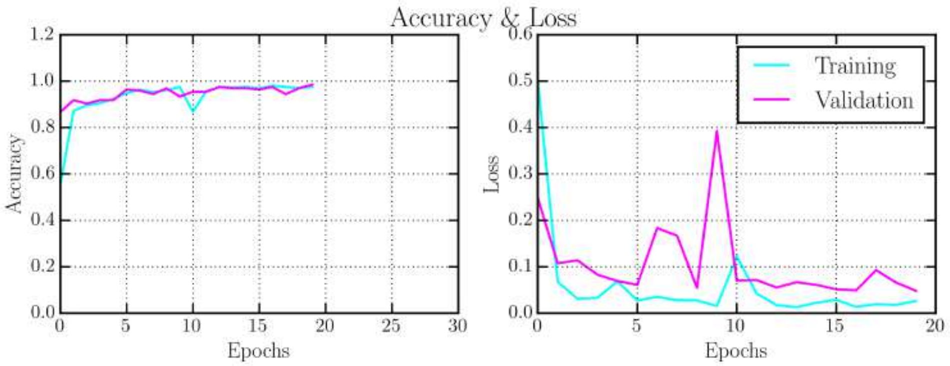


Figure 5.12: The left panel shows the development of accuracy during the training process of the emulator. The right panel shows the progress of the loss function during the same process.

CII power spectrum for any unseen parameter value, as can be inferred from Fig. 5.11. Despite the significant improvement, it was found that this better, more detailed model too is incapable of completely eliminating the tendency to emulate degenerate power spectrum values, especially towards the boundaries of the parameter set provided for training. This anomalous behaviour appears to be due to the featureless nature of the CII power spectrum as a statistic. The only feature it possesses is the amplitude value. Hence, even a slight deviation in this single feature value leads to a large deviation in the parameter it corresponds to and this mismatch is clearly captured by the MCMC algorithm.

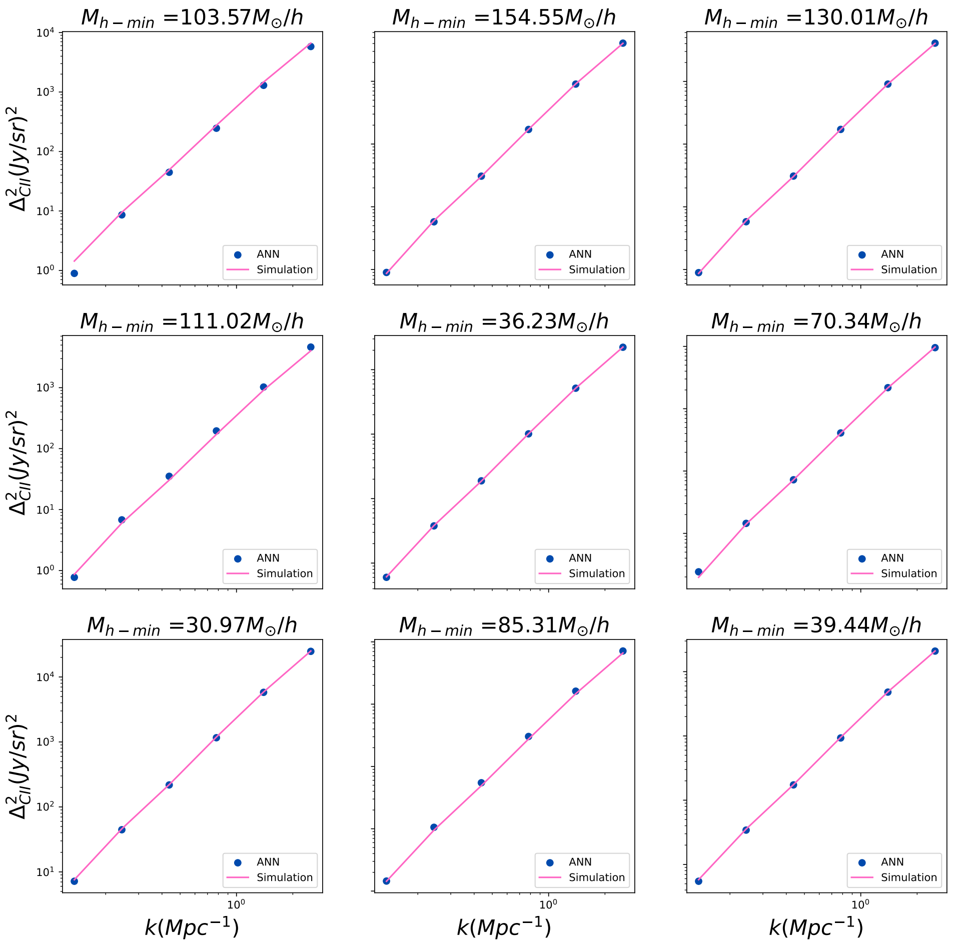
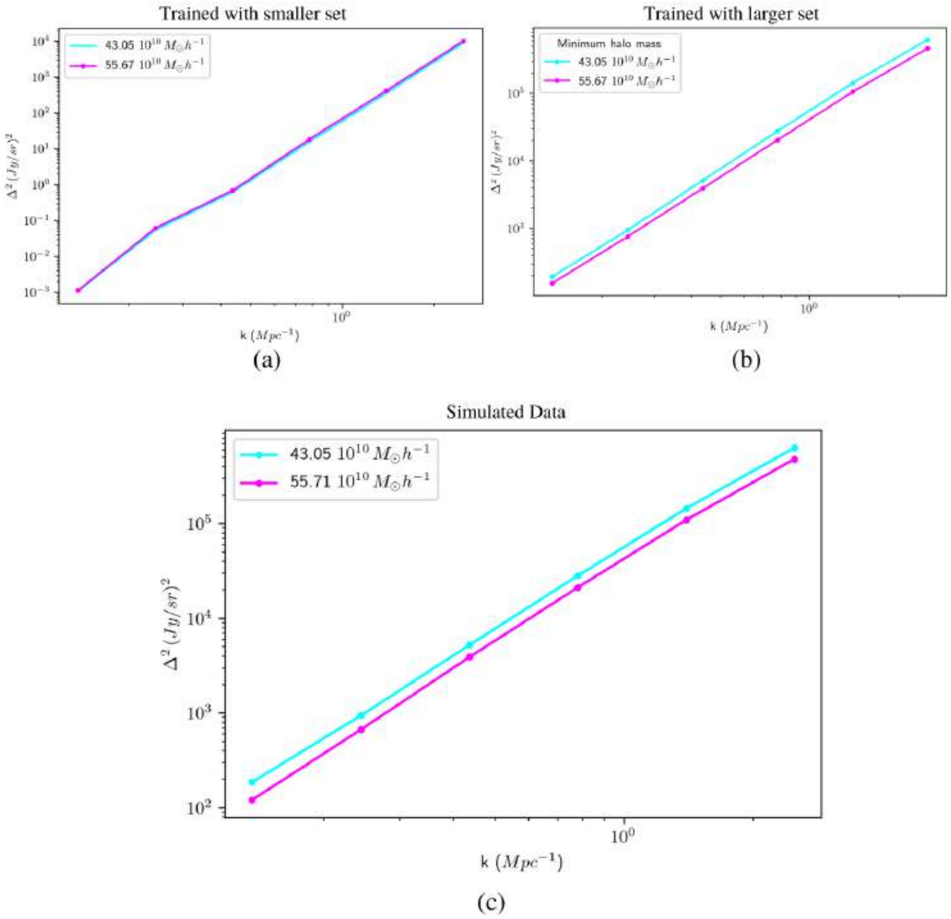


Figure 5.13: This panel compares power spectra as they were simulated and then were predicted by using the model generated after training the ANN with 2167 samples. The dots are ANN predictions and the lines are the simulated power spectra. These nine M_{hmin} values were randomly picked from the available values of M_{hmin} within the unseen test set. These were then passed to the emulator to obtain corresponding predictions.



(a)

(b)

(c)

Figure 5.14: The degeneracy can be seen in these plots very clearly. We have compared the case for two M_{hmin} values (in units of $10^{10}\,M_{\odot}h^{-1}$). First panel a) on the left in shows the power spectrum as emulated by the emulator after it was trained with the smaller data set. The panel b) on the right shows the power spectrum emulated after the emulator was trained with the larger data set and the bottom panel c) above shows the comparison between the power spectrum values as simulated by the simulator.

Chapter 6

Parameter estimation using Bayesian Inference

While there are several parameter estimation techniques that have been developed, like probability plotting, rank regression, maximum likelihood estimation, Bayesian parameter estimation etc, we opt to use the Bayesian technique for our problem. Given a set of data or observations, as is our case, Bayesian statistics seeks to update the posterior probability as one explores the parameter space.

6.1 Bayesian Statistics

The seeds of Bayesian Statistics were sown by Rev Thomas Bayes who solved Bernoulli's problem of updating the probability of an event given new information. This led to the formulation of the now famous Bayes' theorem. In 1820, Pierre-Simon de-Laplace applied Bayes' theorem to celestial mechanics and medical statistics. Though these ideas of Bayes' and Laplace experienced a drop in popularity in the subsequent decades, they saw a resurgence in the 1920's with the Neo-Bayesians. The idea was that probability is essentially a lack of information and must be updated as more information becomes available. This began with John Maynard Keynes, an economist, and was carried forward by several others. Today, Bayesian Statistics is used widely in a range of fields, from astrophysics and cosmology to commercial applications via machine learning.

The foundational idea of Bayesian statistics is that probability is a degree of belief in a proposition allocated by the observer given the available information with uncertainty arising from incomplete data or noise. This is radically different from the frequentist approach that demands the prior knowledge of an ensemble in order to determine probability. Bayesian approach hence allows us to deal with situations where an ensemble can not be imagined, as is the case in cosmology.

6.1.1 Bayesian Inference

The famous Bayes' Theorem is given by:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \tag{6.1}$$

Here, $p(\theta|D)$ is the posterior, $p(D|\theta)$ is the likelihood, $p(\theta)$ is the prior and $p(D)$ is the evidence. Here θ is the given parameter(s) and D is the data. The posterior probability represents the knowledge of the system in light of the data given. The likelihood indicates the likelihood of the data given a hypothesis (about a parameter θ), the prior contains all the information we know about the parameter before the experiment and the evidence is essentially the probability of the data given a particular model. In most use cases (including ours) this can be ignored. Evidence becomes crucial only in model selection problems. Generally, what we would like to compute is:

$$p(\theta|D) \propto p(D|\theta)p(\theta) \tag{6.2}$$

Given a parameter estimation problem, we are generally interested in the posterior distribution for a certain quantity. But for this purpose, one needs to sample or explore the posterior distribution.

The idea behind sampling any probability distribution is that a collection of sufficiently large and preferably uncorrelated samples is a good approximation of the whole distribution. Hence any quantity estimated based on this sampled collection will correspond to the same quantity for the distribution the samples have been drawn from.

This generates a great need for generic sampling algorithms that will not only effectively and efficiently sample any given probability distribution, but will also post process the sampled set in order to calculate statistically meaningful quantities such as mean, variance, co-variance etc.

6.2 Iterative Sampling Methods

Ideally, we would like to draw samples that are independent from each other but in general, it becomes impractical to sample any arbitrary distribution in this manner without using a sampling method that updates its sampling scheme as it explores the parameter space and learns about the distribution. The ideal sampling method that we are looking for, therefore, needs to be iterative in nature.

These algorithms generate a sequence of samples in which at each stage of sampling, some of the information obtained from generating the previous step of the sample is used to decide which part of the parameter space to explore next. The result is that these algorithms, while very powerful and general, produce samples of the target density that are correlated, non-uniformly weighted or both.

There are several such iterative sampling methods available, however we opted for the Markov Chain Monte Carlo method for our use case. The following section discusses this sampling method in greater detail.

6.2.1 Markov Chain Monte Carlo (MCMC) method

Markov chain Monte Carlo (MCMC) methods are a class of algorithms for sampling from a probability distribution used in statistics. By recording states from a Markov chain that has the desired distribution as its equilibrium distribution, a sample of the desired distribution can be obtained. The more steps there are, the more closely the sample's distribution resembles the target distribution. The Metropolis–Hastings method is one of many chain-building algorithms available.

A markov chain is a sequence of points $x_s = x_1, x_2, x_3 ... x_N s$ where each s'th (x_s) point depends explicitly and solely on the s-1'th (x_{s-1}) point. The s'th sample, therefore, is drawn from the probability distribution of the form $P(x_s|x_{s-1})$.

The algorithm

There is no specific formula that dictates how to choose the first sample x_1. This is because if a sufficiently large number of samples are chosen from the target distribution and the MCMC algorithm is designed to explore this space, then the results should be independent of the starting point. In practicality, to implement this, the starting point is picked at random and multiple, independent chains are run parallely to ensure that the results are truly robust.

Having established that, the following is the general algorithm that is followed to pick the s'th (x_s) element once the chain has been started:

1. Draw a trial point from the proposal distribution $P(x_{trial}|x_{s-1}, trial)$, the form of which can be chosen to increase the efficiency of the algorithm.

2. Accept the trial point with the probability $P(accept|x_{trial}, x_{s-1})$.

3. If the trial point is accepted, then set $x_s = x_{trial}$. Otherwise set $x_s =$

x_{s-1}.

This process is repeated at least $\approx 10^3$ times building up a large set of samples.

Stationary distribution and detailed balance: For an MCMC algorithm to be useful, the main requirement is that at least in the high s limits, it draws from *p(**x*** which is the target density. If s-1'th element was drawn from *p(**x***, then s'th element will be drawn from the distribution:

$$\begin{aligned} Pr(\mathbf{x}_s) = \int p(x'_{s-1}) Pr(x_s|x'_{s-1}) dx'_{s-1} \\ = p(x_s) + \int [p(x') Pr(x_s|x', trial) Pr(accept|x_s, x') \\ -p(x_s) Pr(x'|x_s, trial) Pr(accept|x', x_s)] dx' \end{aligned} \tag{6.3}$$

Here, the s'th point will be drawn from the target density if the above integrand is equal to 0. Hence, this criterion will be satisfied if for any pair of points x_1 and x_2:

$$\begin{aligned} p(x_1) Pr(x_2|x_1, trial) Pr(accept|x_1, x_2) = \\ p(x_2) Pr(x_1|x_2, trial) Pr(accept|x_2, x_1) \end{aligned} \tag{6.4}$$

this implies that the two points must satisfy:

$$\frac{Pr(x_2|x_1, trial) Pr(accept|x_1, x_2}{Pr(x_1|x_2, trial) Pr(accept|x_2, x_1)} = \frac{p(x_2)}{p(x_1)} \tag{6.5}$$

This is known as the equation of *detailed balance*. Since this relationship is defined in terms of a *ratio* of the density ratio between two points, there is no need for the target density to be normalized.

There are several algorithms available to satisfy the detailed balance equation. The Metropolis algorithm (also known as Metropolis Hastings) is the most basic of such algorithms. In this case, the proposal distribution is taken to be independent of target density, so that $Pr(x_2|x_1, trial) = Pr(x_1|x_2, trial)$. Then the acceptance probability must include the necessary dependence on $p(x)$.

Burn-in: It is completely valid and indeed preferred to begin sampling from that place in the parameter space if the target distribution is known to be unimodal and the approximate location of the peak is also known. Repeat chains can be started from the same place if the goal density is known to be unimodal; if it is multimodal, the problem becomes substantially more difficult.

Unfortunately, it is not usually known where the target values will be acceptable but one of the key features of MCMC algorithms is that they can be run with an arbitrary starting point. The resultant chain will then tend to propagate towards the peak of the density. This process is called "burn-in". These first elements in the chain, however, can have arbitrarily small densities that would most likely not have been sampled in a chain of plausible length. A very effective (but crude) solution to deal with this challenge is to simply remove these initial points. No algorithm exists for this process, since it is simply a process of chopping up the first few (usually the first $\approx 10 - 15\%$) points.

6.2.2 The Metropolis Hastings algorithm

The Metropolis Hastings algorithm is a two step process used to generate the next link in the chain. Given a previous point x_{s-1}, the next point is given by the following process:

1. Draw a point from the proposal distribution $Pr(x_{trial}|x_{s-1})$, the form of which can be chosen to increase the efficiency of the algorithm.

2. Accept the trial with the probability

   $$Pr(accept|x_{trial}, x_{s-1}) = min\,[\frac{p(x_{trial}}{p(x_{s-1})}, 1] = min[e^{ln[p(x_{trial}]-ln[p(x_{s-1})],1}]$$

   which is unity if the trial point is more probable than the previous point and 0 only if $p(x_trial) = 0$.

3. If the trial point is accepted, then set $x_S = x_{trial}$; otherwise set $x_s = x_{s-1}$.

Independent of the nature of the proposal distribution, the MH algorithm satisfies the detailed balance equation which implies that the proposal distribution will be ***p(x)***. However, if the proposal distribution is too concentrated relative to the scales on which the target density varies, then $p(x_{trial}) \approx p(x)$ for all proposed points x_{trial}. The acceptance probability is then

$$Pr(accept|x_{trial}, x) \approx 1$$

and almost all points are accepted. This results in a very slow exploration of the distribution and the algorithm takes a large number of steps to converge.

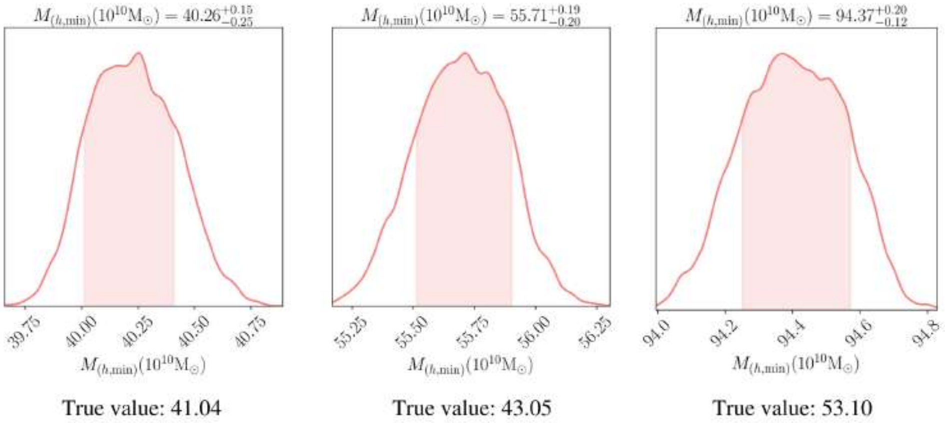
6.3 Implementation of the MCMC algorithm for parameter estimation

In this work, the CosmoHammer package (Akeret et al., 2012) was used to impliment the MCMC algorithm. The MCMC sampler essentially explores the parameter space in a random fashion (known as a random walk). The probability of the point is evaluated at every step relative to the probability at the previous step. The new step (or the trial) is accepted or rejected based on this comparison (Sub-section 6.2.2).

This particular algorithm uses 4 random walkers to explore the parameter space. Each walker takes a 1000 steps, out of which 100 steps are rejected as burn-in iterations.

The algorithm takes in a simulated power-spectrum data (ideally, a power-spectrum from observational data but since we do not have observational data, we shall use the simulated data as our true data) and walks through the parameter space to provide the most likely value of the parameter. In our case, the parameter is minimum halo mass M_{hmin}. This is because we have established a one to one correspondence between the halo mass and the CII luminosity (Sub-section 3.2.2). Since in this case we know the true value of M_{hmin} corresponding to the power-spectrum data we provide to our MCMC algorithm, we can compare and contrast the accuracy of both our estimation process and our emulator.

We first ran this algorithm using the first emulator model (let's call this model A). The estimation produced by this combination was severely biased. Some of the estimations are presented below:



True value: 41.04

True value: 43.05

True value: 53.10

Figure 6.1: Parameter estimation using the first emulator. All the values are quoted in units of $10^{10}\ M_\odot h^{-1}$.

True value ($10^{10}\ M_{hmin} h^{-1}$)	Estimated value ($10^{10}\ M_{hmin} h^{-1}$)	Percentage error in estimation
41.04	40.26	1.9
43.05	55.71	29.4
53.10	94.376	77.7

Table 6.1: Summary of results obtained from parameter estimation using MCMC and the emulator model trained with the smaller data set (model A).

To investigate the cause of this bias, we decided to emulate the power

True value ($10^{1}0\, M_{hmin}h^{-1}$)	Estimated value ($10^{1}0\, M_{hmin}h^{-1}$)	Percentage error in estimation
14.90	11.27	24.36
66.25	65.90	0.52
70.33	70.396	0.09
81.51	81.158	0.43
84.43	81.17	3.8

Table 6.2: Summary of results obtained from parameter estimation using MCMC and the emulator model trained with large data set (model B).

spectra using this model A of the emulator once for the true value, and once for the estimated value. We randomly picked the point at 43.05 $\times 10^{10}\, M_{\odot}h^{-1}$ for this investigation and the results have been reported in Fig. 5.14.

Hence, in order to make the emulator more functional, we decided to train it with a much larger data set and generated a second model (let's call this model B). Thereafter, using this model, we performed parameter estimation at 4 different points at random roughly spanning the available parameter space and attempted to perform a parameter estimation on them using the MCMC algorithm. The following results were obtained:

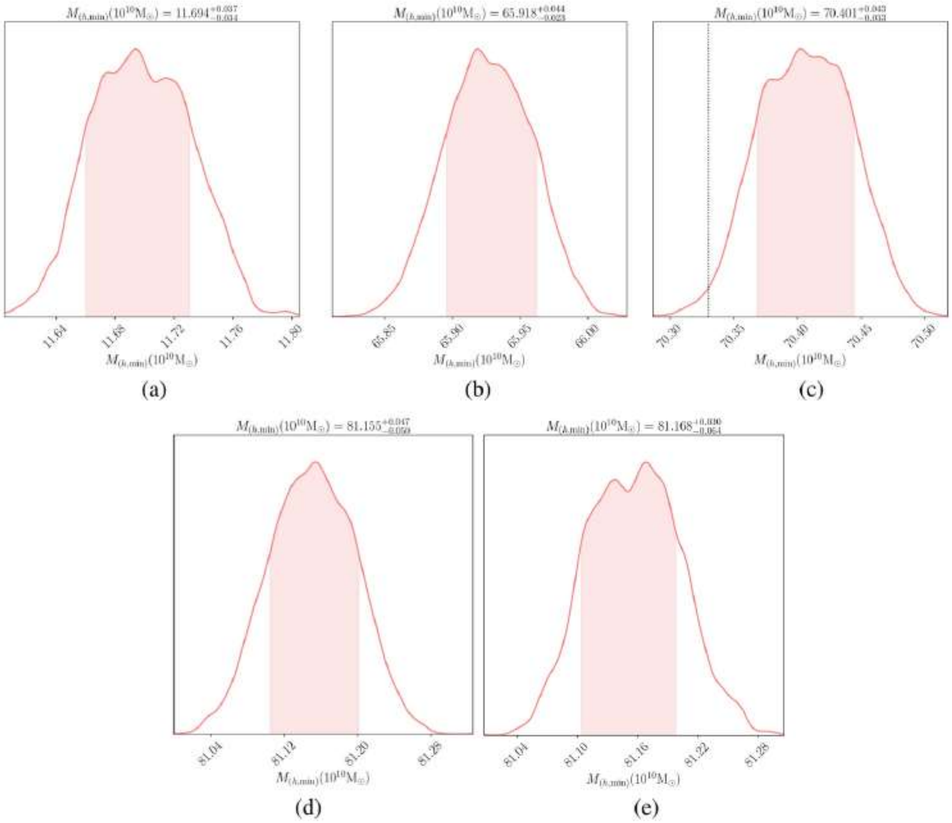


Figure 6.2: Plots showing estimations performed by our MCMC algorithm using the ANN based emulator as the underlying model. The values indicated in these plots are the estimates. The true values are (a) 14.90 (b) 66.25 (c) 70.33 (d) 81.51 and (e) 84.43 all in units of $10^{10}\ M_{\odot}h^{-1}$.

As is clear, the bias has been significantly reduced. However, there *still* remains some bias. To concretely determine whether this was caused by the emulator or the MCMC algorithm itself, we decided to run the MCMC algorithm yet again, but this time using the simulator as our model. This took significantly longer than the process with the emulator (≈ 18 hours as compared to ≈ 30 minutes with the emulator). However, this turned out to be more accurate.

The following result was obtained:

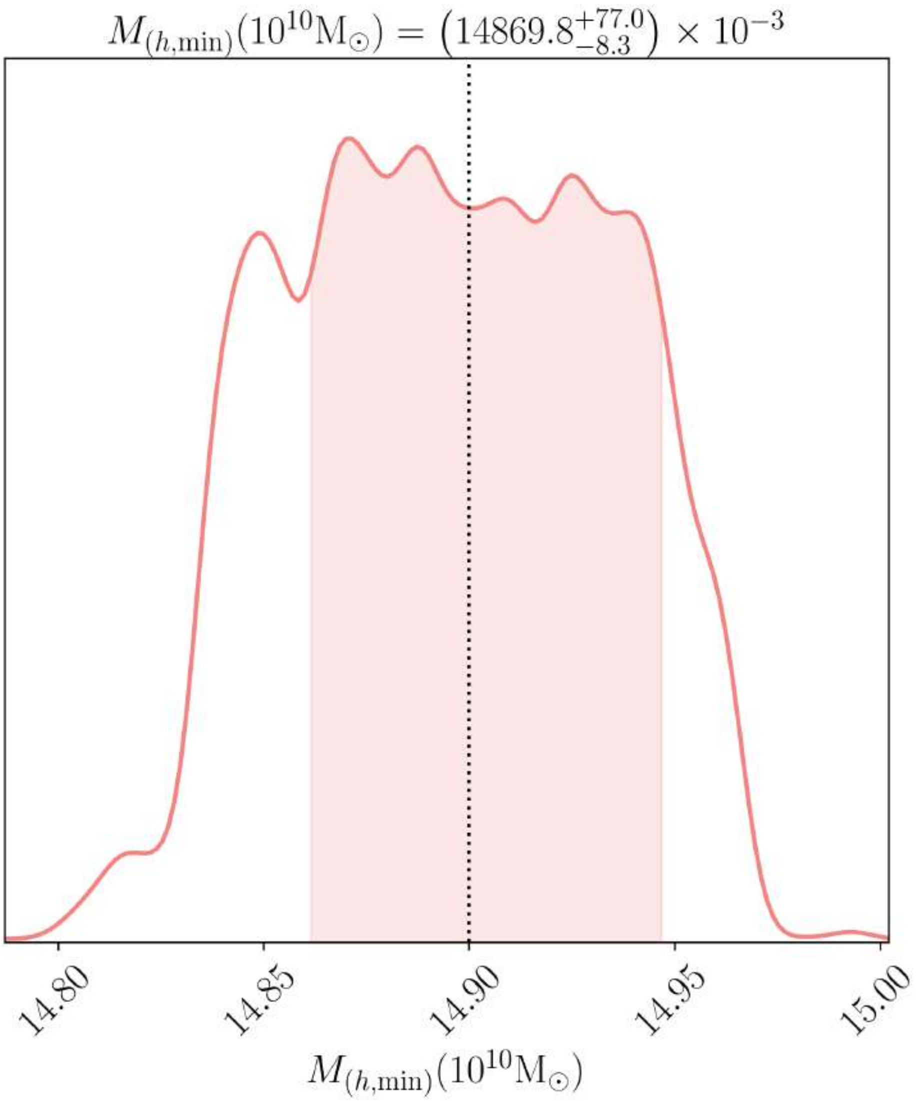


Figure 6.3: Estimation obtained by running the MCMC algorithm in tandem with the simulator. This time, the estimate falls within the 1 σ limit as should be the ideal case after the execution of an MCMC algorithm. The true value is $14.90 \times 10^{10}\ M_\odot h^{-1}$.

Hence, a clear conclusion can be drawn that this discrepancy or bias seen in earlier runs of the MCMC algorithm originates from the emulator models. As stated in Sec. 5.4, this appears to be due to the nature of the CII power spectrum. Since it does not have any features beyond its amplitude, even a slight deviation in the emulation of this amplitude leads to the power spectrum corresponding to a completely different parameter value.

Chapter 7

Discussion and future scope

The goal of this project was to conceptualize a feasible method of performing parameter estimation using Fourier statistics of line emissions from EoR that can be calculated from observations to be made by next generations telescopes and interferometers. The three parameters that are necessary and sufficient to model the epoch of reionizatin are minimum halo mass of dark matter halos (M_{hmin}), ionization efficiency (N_{ion}), and mean free path of ionizing photons (R_{mfp}).

Since these values define the history of our universe, we can observe their effects on the observed statistics. Hence, it is reasonable to expect that they can be estimated from these statistics as well. For this purpose, we decided to use an MCMC algorithm that would explore the parameter space (in this case, simply M_{hmin}) and give us the posterior probability of the parameter. But to do this, one has two options at hand.

1. Use a simulator to simulate the power spectrum at every step the random walker takes

2. Use an emulator to emulate the power spectrum at every step the random walker takes

While the first method is the more traditional way of doing this, it is usually extremely computationally expensive and time consuming. Hence, the second method with an emulator was chosen.

The following steps were taken to achieve this goal:

- N-body simulations were used to generate CII and HI 21cm maps.
  - An FoF algorithm was used to locate the halos in the dark matter field generated by the N-body code.
  - Every halo was assigned a CII luminosity associated with it, based on its halo mass, using the formalism given in Eq. 2.22. These luminosities were then painted on top of the dark matter halos
  - The HI maps were created using an excursion set formalism on the halo catalogue generated by the friends of friends algorithm with the assumption that baryonic matter traces dark matter distribution.
- These maps were then used to calculate power-spectra for CII and HI, and cross-power spectra for $CII \otimes 21cm$ signals.
- An ANN based emulator was built with the help of the Keras package (Chollet et al., 2015) to emulate the above signals and Bayesian Inference (Sec. 6.3) was performed.
  - The first emulator was trained and tested with 236 data points. (Fig. 5.6)
    - ∗ This emulator was found to have 86% accuracy of predictions.
    - ∗ Upon running the MCMC algorithm that was implemented with the help of the Cosmo Hammer package (Akeret et al., 2012,Sec. 6.3), in tandem with this emulator it was found to be prone to producing degenerate data points (Fig. 5.14,Sec. 5.4) leading to erroneous estimation.

  - Emulator trained the second time with a larger data set containing 2167 samples.

    * This emulator was found to have 96.8% accuracy of predictions when provided with unseen parameter values.

    * Upon running the MCMC in tandem with this emulator, it was found to have much better functionality. However, degeneracy remained, especially as one approaches the boundaries of the parameter space.

- To determine the cause of the problem, the MCMC algorithm was run again using the simulator (Chapter 3,Sub-section 3.2.2,Sub-section 4.3.1). This produced estimations within 1σ limit of the true value.

While this simulated run took approximately 20 times longer to execute and was much more computationally expensive, as one would expect, it provided much more accurate estimates of the minimum halo mass.

So to conclude, it appears that due to the featureless nature of the CII power spectrum, using an emulator for performing such parameter estimation exercises may not be the desirable path to walk on. Since the only feature that defines the CII power spectrum is its amplitude, even a slight shift in amplitude results in the power spectrum corresponding to a different parameter value.

Therefore, it is actually computationally cheaper to use our simulator to perform this task. This is aided by the fact that simulation of CII maps and generating power spectra from such maps can be done fairly quickly (as far as simulators are concerned). Although using the emulator is ≈ 20 times faster, the significant improvement in accuracy while using the simulator is non-trivial.

7.1 Further scope

This work can be further expanded into several different projects to further probe the epoch of reionization using statistics and machine learning techniques.

A few of these avenues are briefly discussed below:

7.1.1 Parameter estimation using cross-power spectrum:

A similar kind of treatment can be done by using cross-power spectrum, for example, $CII \otimes 21cm$ cross-power spectrum. As has been already shown by (Tiwari et al., 2021), the feature rich 21cm power spectrum can be successfully emulated by using an ANN based emulator. Hence, it can be reasonably expected that the $CII \otimes 21cm$ cross-power spectrum, which is dominated by the features of the 21cm power spectrum can be successfully emulated as well. In addition, this will enable us to estimate the other two parameters associated with the epoch of reionization, i.e. N_{ion} and R_{mfp} since these affect the HI 21cm field prominently.

We intend to extend the present project to explore this avenue and implement techniques described in this project on the cross-power spectrum.

7.1.2 Using realistic observational scenarios:

Our model is based on the ideal CII signal and only uses Gaussian noise in all of the operations. An important extension of this work would be to introduce realistic noise into the signal and use such a noisy signal for this kind of parameter estimation problems. Such an exercise would be able to more accurately display the feasibility of this manner of estimation with realistic data.

There can be several sources of realistic noise, such as:

1. Imperfect foreground subtraction

2. Ionospheric fluctuations with time

3. Array configuration

This kind of analysis using Bayesian statistics to estimate values of physically relevant parameters can be extremely useful.

Developing techniques in machine learning, particularly deep learning show great promise in being new and lucrative tools of exploring several physical processes and bringing down the computation cost of expensive processes. We hope this thesis and our future work generates sufficient interest in this direction.

Bibliography

Active Galaxies and Quasars - the Lyman-alpha Forest. (n.d.). Retrieved April 29, 2022, from http://pages.astronomy.ua.edu/keel/agn/forest.html

Akeret, J., Seehars, S., Amara, A., Refregier, A., & Csillaghy, A. (2012). CosmoHammer: Cosmological parameter estimation with the MCMC Hammer. *Astronomy and Computing*, *2*, 27–39. https://doi.org/10.1016/j.ascom.2013.06.003

Anderson, L. D., Makai, Z., Luisi, M., Andersen, M., Russeil, D., Samal, M. R., Schneider, N., Tremblin, P., Zavagno, A., Kirsanova, M. S., Ossenkopf-Okada, V., & Sobolev, A. M. (2019). The Origin of [C II] 158 μm Emission toward the H II Region Complex S235. *APJ*, *882*(1), Article 11, 11. https://doi.org/10.3847/1538-4357/ab1c59

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (n.d.). Algorithms for Hyper-Parameter Optimization.

Boselli, A., Gavazzi, G., Lequeux, J., & Pierini, D. (2002). [CII] at 158 μm as a star formation tracer in late-type galaxies. *Astronomy and Astrophysics*, *385*(2), 454–463. https://doi.org/10.1051/0004-6361:20020156

Bradt, H. (2008). Astrophysics processes: The physics of astronomical phenomena. *Astrophysics Processes: The Physics of Astronomical Phenomena*, 1–504. https://doi.org/10.1017/CBO9780511802249

Cardiel, N., Elbaz, D., Schiavon, R. P., Willmer, C. N. A., Koo, D. C., Phillips, A. C., & Gallego, J. (2003). A Multiwavelength Approach to the Star Formation Rate Estimation in Galaxies at Intermediate Redshifts. *The Astrophysical Journal*, *584*(1), 76–99. https://doi.org/10.1086/345594/FULLTEXT/

Chollet, F. et al. (2015). *Keras*.

Choudhury, T. R., & Ferrara, A. (2006). Updating reionization scenarios after recent data. *Monthly Notices of the Royal Astronomical Society: Letters*, *371*(1), L55–L59. https://doi.org/10.1111/J.1745-3933.2006.00207.X

Choudhury, T. R., Haehnelt, M. G., & Regan, J. (2009). Inside-out or outside-in: The topology of reionization in the photon-starved regime suggested by Lyα forest data. *Monthly Notices of the Royal Astronomical Society*, *394*(2), 960–977. https://doi.org/10.1111/J.1365-2966.2008.14383.X/2/MNRAS0394-0960-F9.JPEG

Clark, P. C., Glover, S. C. O., Ragan, S. E., & Duarte-Cabral, A. (2019). Tracing the formation of molecular clouds via [CII], [CI] and CO emission. *Mon. Not. R. Astron. Soc*, *000*(0000), 0–000.

Datta, A., Bowman, J., & Carilli, C. (2010). Bright source subtraction requirements for redshifted 21 cm measurements. *The Astrophysical Journal*, *724*(1), 526.

De Looze, I., Cormier, D., Lebouteiller, V., Madden, S., Baes, M., Bendo, G. J., Boquien, M., Boselli, A., Clements, D. L., Cortese, L., Cooray, A., Galametz, M., Galliano, F., Graciá-Carpio, J., Isaak, K., Karczewski, O. Ł., Parkin, T. J., Pellegrini, E. W., Rémy-Ruyer, A., … Sturm, E. (2014). The applicability of far-infrared fine-structure lines as star formation rate tracers over wide ranges of metallicities and galaxy types. *aap*, *568*, Article A62, A62. https://doi.org/10.1051/0004-6361/201322489

Friends-Of-Friends Algorithm — SWIFT: SPH With Inter-dependent Fine-grained Tasking 0.9.0 documentation. (2014). Retrieved April 16, 2022, from https://swift.dur.ac.uk/docs/FriendsOfFriends/algorithm_description.html

Furlanetto, S. R., Zaldarriaga, M., & Hernquist, L. (2004). The Growth of H ii Regions During Reionization. *The Astrophysical Journal*, *613*(1), 1–15. https://doi.org/10.1086/423025/FULLTEXT/

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. https://books.google.co.in/books?id=Np9SDQAAQBAJ

Gunn, J. E., & Peterson, B. A. (1965). On the density of neutral hydrogen in intergalactic space. *The Astrophysical Journal*, *142*. https://doi.org/10.1086/148444

Hubble, E. P. (2014). 106. A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae. *A Source Book in Astronomy and Astrophysics, 1900–1975*, 725–728. https://doi.org/10.4159/harvard.9780674366688.c114

Kennicutt, Jr., R. C. (1998). The Global Schmidt Law in Star□forming Galaxies. *The Astrophysical Journal*, *498*(2), 541–552. https://doi.org/10.1086/305588

Lahav, O., & Liddle, A. R. (2006). The Cosmological Parameters 2006. (September 2021), 1–18. https://doi.org/10.1088/0954-3899/33/1/001

Loeb, A., & Furlanetto, S. R. (2013). The first galaxies in the universe. *The First Galaxies in the Universe*, 1–540. https://doi.org/10.5860/choice.50-6740

Ma, X., Hopkins, P. F., Garrison-Kimmel, S., Faucher-Giguère, C.-A., Quataert, E., Boylan-Kolchin, M., Hayward, C. C., Feldmann, R., & Kereš, D. (2018). Simulating galaxies in the reionization era with FIRE-2: galaxy scaling relations, stellar mass functions, and luminosity functions. *MNRAS*, *478*(2), 1694–1715. https://doi.org/10.1093/mnras/sty1024

Majumdar, S., Mellema, G., Datta, K. K., Jensen, H., Choudhury, T. R., Bharadwaj, S., & Friedrich, M. M. (2014). On the use of seminumerical simulations in predicting the 21-cm signal from the epoch of reionization. *Monthly Notices of the Royal Astronomical Society*, *443*(4), 2843–2861. https://doi.org/10.1093/MNRAS/STU1342

McQuinn, M., Zahn, O., Zaldarriaga, M., Hernquist, L., & Furlanetto, S. R. (2005). Cosmological Parameter Estimation Using 21 cm Radiation from the Epoch of Reionization. *The Astrophysical Journal*, *653*(2), 815–834. https://doi.org/10.1086/505167

Mondal, R., Bharadwaj, S., Majumdar, S., Bera, A., & Acharyya, A. (2015). The effect of non-Gaussianity on error predictions for the Epoch of Reionization (EoR) 21-cm power spectrum. *Monthly Notices of the Royal Astronomical Society: Letters*, *449*(1), L41–L45. https://doi.org/10.1093/MNRASL/SLV015

Murmu, C. S., Majumdar, S., & Datta, K. K. (2021). C ii and H i 21-cm line intensity mapping from the EoR: Impact of the light-cone effect on auto and cross-power spectra. *Monthly Notices of the Royal Astronomical Society*, *507*(2), 2500–2509. https://doi.org/10.1093/mnras/stab2347

Nasirudin, A., Murray, S. G., Trott, C. M., Greig, B., Joseph, R. C., & Power, C. (2020). The Impact of Realistic Foreground and Instrument Models on 21 cm Epoch of Reionization Experiments. *The Astrophysical Journal*, *893*(2), 118. https://doi.org/10.3847/1538-4357/AB8003

Nwankpa, C., Ijomah, W., Gachagan, A., & Marshall, S. (2018). Activation Functions: Comparison of trends in Practice and Research for Deep Learning. *arXiv e-prints*, Article arXiv:1811.03378, arXiv:1811.03378.

O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., Invernizzi, L., et al. (2019). Keras Tuner.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Silva, M., Santos, M. G., Cooray, A., & Gong, Y. (2015). Prospects for Detecting C Ii Emission During the Epoch of Reionization. *Astrophysical Journal*, *806*(2), 209. https://doi.org/10.1088/0004-637X/806/2/209

Silva, M. B., Santos, M. G., Gong, Y., Cooray, A., & Bock, J. (2013). INTENSITY MAPPING OF Lyα EMISSION DURING THE EPOCH OF REIONIZATION. *The Astrophysical Journal*, *763*(2), 132. https://doi.org/10.1088/0004-637X/763/2/132

Sutter, J., Dale, D. A., Croxall, K. V., Pelligrini, E. W., Smith, J. D. T., Appleton, P. N., Beirão, P., Bolatto, A. D., Calzetti, D., Crocker, A., De Looze, I., Draine, B., Galametz, M., Groves, B. A., Helou, G., Herrera-Camus, R., Hunt, L. K., Kennicutt, R. C., Roussel, H., & Wolfire, M. G. (2019). USING [CII] 158 μm EMISSION FROM ISOLATED ISM PHASES AS A STAR-FORMATION RATE INDICATOR.

Tiwari, H., Shaw, A. K., Majumdar, S., Kamran, M., & Chaudhury, M. (2021). Improving constraints on the reionization parameters using 21-cm bispectrum. http://www.gmrt.ncra.tifr.res.in