Studying Cosmic Evolution with 21 cm Intensity Mapping

by

Chris Anderson

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy
(Physics)

at the

UNIVERSITY OF WISCONSIN - MADISON
2017
This thesis describes early work in the developing field of 21-cm intensity mapping. The 21-cm line is a radio transition due to the hyperfine splitting of the ground state of neutral hydrogen (HI). Intensity mapping utilizes the aggregate redshifted 21-cm emission to map the three-dimensional distribution of HI on large scales. In principle, the 21-cm line can be utilized to map most of the volume of the observable Universe. But the signal is small, and dedicated instruments will be required to reach a high signal-to-noise ratio. Large spectrally smooth astrophysical foregrounds, which dwarf the 21-cm signal, present a significant challenge to the data analysis. This thesis describes the fundamental physics of the 21-cm line and derives the expected cosmological signal. I also provide an overview of the desired characteristics of a dedicated 21-cm instrument, and I list some instruments that are coming on-line in the next few years. I then describe the data analysis techniques and results for 21-cm intensity maps that were made with two existing radio telescopes, the Green Bank telescope (GBT) and the Parkes telescope. Both observations have detected the 21-cm HI signal by cross-correlating the 21-cm intensity maps with overlapping optical galaxy surveys. The GBT maps have been used to constrain the neutral hydrogen density at a mean redshift (z) of 0.8. The Parkes maps, at a mean redshift of 0.08, probe smaller scales. The Parkes 21-cm intensity maps reveal a lack of small-scale clustering when they are cross-correlated with 2dF optical galaxy maps. This lack of small-scale clustering is partially due to a scale-dependent and galaxy-color-dependent HI-galaxy cross-correlation coefficient. Lastly, I provide an overview of planned future analyses with the Parkes maps, with a proposed multi-beam receiver for the Green Bank telescope, and with simulations of systematic effects on foregrounds.
Contents

1 Introduction ........................................... 1
  1.1 Outline ........................................... 2
  1.2 My Contributions .................................. 3

2 Cosmology Overview .................................... 5
  2.1 The Expanding Universe ............................ 5
  2.2 The FLRW metric and the Friedmann Equations ..... 7
  2.3 The $\Lambda$CDM Model ............................. 12
    2.3.1 Quantifying Structure .......................... 13
    2.3.2 Epochs in the History of the Universe .......... 17
    2.3.3 The Need for Dark Matter and Dark Energy ..... 20
  2.4 The CMB Fluctuations and Structure Formation .... 26
    2.4.1 Matter Tracers .................................. 29

3 The 21-cm signal ...................................... 33
  3.1 Hyperfine Transition .............................. 33
  3.2 Radiative Transfer .................................. 44
  3.3 HI Spin Temperature ............................... 52
  3.4 Galaxy surveys using the 21-cm line ................ 59
  3.5 21-cm Intensity Mapping ........................... 61
    3.5.1 Tomographic mapping of matter using the 21-cm signal .... 61
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5.2 Foregrounds</td>
<td>65</td>
</tr>
<tr>
<td>4 Radio Telescopes and HI Astronomy</td>
<td>69</td>
</tr>
<tr>
<td>4.1 Fundamentals of Single Dish radio telescopes</td>
<td>70</td>
</tr>
<tr>
<td>4.1.1 Power measured in Temperature Units</td>
<td>71</td>
</tr>
<tr>
<td>4.1.2 The Collecting Area of an Antenna</td>
<td>72</td>
</tr>
<tr>
<td>4.1.3 Antenna Reciprocity Theorem, Impedance matching</td>
<td>73</td>
</tr>
<tr>
<td>4.1.4 Noise and the System Temperature</td>
<td>74</td>
</tr>
<tr>
<td>4.1.5 Map Calibration</td>
<td>77</td>
</tr>
<tr>
<td>4.1.6 Stokes Parameters</td>
<td>79</td>
</tr>
<tr>
<td>4.2 Observations with the Parkes L-band Receiver</td>
<td>80</td>
</tr>
<tr>
<td>4.3 Observations with the Green Bank Telescope (GBT) 800 MHz Receiver</td>
<td>82</td>
</tr>
<tr>
<td>4.4 Designing an 800 MHz Array for the GBT</td>
<td>85</td>
</tr>
<tr>
<td>4.5 Survey of Upcoming Low Redshift 21 cm Experiments</td>
<td>95</td>
</tr>
<tr>
<td>5 Making Maps From Parkes Radio Data</td>
<td>97</td>
</tr>
<tr>
<td>5.1 Observation Strategy and Data Format</td>
<td>98</td>
</tr>
<tr>
<td>5.2 RFI removal</td>
<td>100</td>
</tr>
<tr>
<td>5.3 The Mapmaker Equation</td>
<td>102</td>
</tr>
<tr>
<td>5.4 Parallelizing the Mapmaker Code</td>
<td>104</td>
</tr>
<tr>
<td>5.5 The Noise Covariance Model</td>
<td>106</td>
</tr>
<tr>
<td>5.6 Calibrating the Maps</td>
<td>108</td>
</tr>
<tr>
<td>5.6.1 Bandpass Calibration</td>
<td>108</td>
</tr>
<tr>
<td>5.6.2 Flux Calibration</td>
<td>110</td>
</tr>
<tr>
<td>5.7 On the Optimality of the Parkes Maps</td>
<td>115</td>
</tr>
<tr>
<td>6 Cutting through the Foregrounds</td>
<td>116</td>
</tr>
<tr>
<td>6.1 Identifying Foregrounds via Principal Component Analysis</td>
<td>117</td>
</tr>
<tr>
<td>6.2 Systematic Effects on Foregrounds</td>
<td>120</td>
</tr>
<tr>
<td>6.3 Parkes Foreground Removal</td>
<td>122</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>---------------------------------------------------------</td>
</tr>
<tr>
<td>6.4</td>
<td>Cross-Power Spectrum Estimation</td>
</tr>
<tr>
<td>6.4.1</td>
<td>Simulations</td>
</tr>
<tr>
<td>6.4.2</td>
<td>Cross-power spectrum</td>
</tr>
<tr>
<td>7</td>
<td>Results</td>
</tr>
<tr>
<td>7.1</td>
<td>GBT</td>
</tr>
<tr>
<td>7.2</td>
<td>Parkes</td>
</tr>
<tr>
<td>8</td>
<td>Continuing Work</td>
</tr>
<tr>
<td>8.1</td>
<td>Parkes Auto-Power Constraints</td>
</tr>
<tr>
<td>8.2</td>
<td>Simulating Instrumental Effects on Foreground Cleaning</td>
</tr>
<tr>
<td>8.3</td>
<td>Improving GBT Maps with the GBT-HIM Array</td>
</tr>
<tr>
<td>9</td>
<td>Conclusions</td>
</tr>
<tr>
<td>A</td>
<td>Calibrator Noise</td>
</tr>
<tr>
<td>A.1</td>
<td>MBCORR Backend</td>
</tr>
<tr>
<td>A.2</td>
<td>Extra Noise Induced</td>
</tr>
</tbody>
</table>
List of Figures

2-1 A diagram from the NASA/WMAP team of the history of the Universe in a $\Lambda$CDM model with inflation. Earlier times are on the left. The Universe starts with a brief period of exponential growth, which expands quantum fluctuations to large scales. This lasts for much less than a second, after which the inflaton field decays into the particles and radiation of the standard model. Expansion of the Universe continues at a lower rate, dominated first by radiation, then by matter. After 380,000 years, matter decouples from radiation, and the Universe becomes transparent. Expansion continues through the dark ages, until gravitational collapse forms the first stars, which reionize most of the matter in the Universe. Around 5 billion years ago, the rate of expansion begins to accelerate again as the energy density of matter drops below the energy density of the cosmological constant.

2-2 From [71]. Apparent magnitude is plotted as a function of redshift for observed type Ia supernovae. Apparent magnitude is a logarithmic measure of brightness, in which smaller numbers correspond to brighter objects. For the Type Ia supernova standard candle, apparent magnitude depends logarithmically on the distance to the object. The dashed line is the best-fit flat cosmology, in which $\Omega_m = 0.29$ and $\Omega_\Lambda = 0.71$. 

---

29 and $\Omega_\Lambda = 0.71$.
2-3 From the report of the dark energy task force [2]. The distance $D(z)$ vs. redshift relation and the growth factor $g(z)$ vs. redshift are the primary observables that can constrain the dark energy equation of state. The black curve is the $ΛCDM$ model. The green curve is a model with $Ω_m = 1$, which is strongly disfavored by current data. The red curve is a model fit to CMB data with $w = -0.9$. This model is consistent with current data but can be constrained by measuring $\sim 5\%$ deviations in the distance and growth factor in the redshift range $0 \leq z \leq 3$.

2-4 From [89]. The blue circle shows the comoving volume accessible in principle, using the 21-cm transition of neutral hydrogen. The red area depicts the galaxies detected by the Sloan Digital Sky Survey. The thick solid black line shows the surface of last scattering of the CMB, with the thickness shown to-scale. This solid black line is the farthest observable distance. The infinite redshift circle represents the outer radius that would be observable if the Universe were transparent over its full history.

3-1 The collisional de-excitation rate of HI from the triplet to the singlet spin state for the two dominant species, hydrogen atoms ($κ^{HH}_{10}$) and free electrons ($κ^{eH}_{10}$). The transition rate is the sum of these terms multiplied by their respective densities. The figure is from [31]. The $κ^{eH}_{10}$ curve is based on calculations by [30], and the $κ^{HH}_{10}$ curve is from calculations of [99] at $T < 300$ K and extrapolations to higher temperatures by K. Sigurdson, using the methods of [77].
3-2 The allowed Lyman alpha dipole transitions. The solid paths show transition chains that, if followed from the $s$ to $p$ orbital via absorption and then from $p$ back to $s$ via spontaneous emission, can couple the $s$ spin 0 and spin 1 states. The dashed lines show paths that cannot couple $s$ spin 0 and spin 1 states via Lyman alpha absorption and spontaneous emission. This figure is from [31].

3-3 The top panel is expected history of the HI kinetic temperature, the CMB temperature, and the spin temperature, using numerical simulations from the publicly available REFCAST code. The bottom panel is the resultant brightness temperature dip. This figure is from [31].

4-1 An antenna in series with a matched resistor through a narrow-band filter, in a black box at temperature $T$. This figure is from the excellent NRAO radio astronomy course: https://science.nrao.edu/opportunities/courses/era/. This figure is from [31].

4-2 The relative location and sizes of the 13 beams on the sky. Taken from [4].

4-3 The Green Bank telescope. When the 800 MHz receiver is operating, a boom extends out from the arm at the top left of the image and places the receiver at the prime focus of the paraboloid. This image is from the NRAO website: http://www.nrao.edu/pr/2013/GBTWVU/. The same off-axis design is used on roof-mounted satellite television dishes.

4-4 This photograph, taken from the feed arm of the GBT, shows the current 800 MHz corrugated horn feed at the area where it is stowed during high winds or to switch the telescope to a different receiver. The dimensions of this room provide an absolute length limit of 3.02 meters for any proposed GBT array.
This picture shows the SBA used for the 600 MHz GBT feed. All pieces painted white are conductors. The structural towers holding up the dipoles and smaller reflectors are dielectric materials. The back reflecter is actually an obtuse angled cone. The conical back reflector increases the bandwidth compared to a flat reflector. The antenna operates like a leaky resonant cavity.

Cross-section of the Winegard feed, from [72].

Cross-section of our corrugated SBA receiver. The bottom reflector is flat in the middle and then conical. The horizontal corrugation can be seen at the top of the cone, at approximately the same height as the dipole antennas. Directly above the horizontal corrugation is a vertical quarter wave corrugation, reminiscent of the Winegard feed. The outer edges of the feed are squared off at four places, which gives the feed a smallest diameter of 2.77 wavelengths or about 1.04 meters.

Simulated spill temperature for our corrugated SBA receiver across the 800 MHz frequency band. The spill temperature is calculated at a GBT elevation angle of 30, 60, and 90 degrees. The points for the 60 and 30 degree elevation angle spill temperatures are practically identical.

Structural drawing of the final hex + 1 7-element array. Small holes in the receivers help save weight and have minimal effect on the beam-pattern.

The beam 1 hitmap for half of the 2df1 field. Pixels where data was recorded are red, and blue pixels indicate positions on the sky that were never pointed at by the telescope. Lines with opposite slopes come from scans taken during rising and setting.
5-2 The left panel shows all of the data blocks for the XX polarization of beam 1 after RFI flagging and rebinning to 1 MHz frequency bands, and the right panel shows the same thing for the XX polarization of beam 10. The color scale shows the standard deviation divided by the mean ($\sigma/T_{\text{sys}}$ from equation 4.8) for each of the 64 frequency bins over the length of the block (typically 3 minutes). White spaces indicate masked frequency channels: this occurs at places where all 16 of the original 62.5 kHz frequency bins that contribute to a 1 MHz bin were masked by the RFI flagging algorithm. The vertical bands indicate variance present across the frequency band. Some of these are probably due to the beam passing through a bright point source, but the brightest ones are likely due to broadband RFI. There is also persistent narrow band RFI at the edges of the band, especially near 1290 MHz. The plot for beam 10 on the left illustrates the consistent high level of noise that led us to discard data from 3 of the 26 beams/polarizations.

5-3 This image matrix shows the first eigenvectors of the XX polarization sub-maps centered at 33 degrees right ascension, made from the 12 good XX beams. The plot at row $r$ and column $c$ represents the first eigenvector of beam $c$, calculated from the SVD of $C_{r,c}$, plotted across the full 64 MHz bandwidth. The eigenvectors are expected to represent the product of the bandpass shape and the power law of the diffuse emission present in the map. The column-to-column variation reflects the slightly different bandpass shapes of the various beams.

5-4 The final bandpass estimates for the 12 good XX beams, plotted over the full 64 MHz bandwidth. The ripples in the bandpass shapes appear to have a period near 5.7 MHz, and are probably caused by standing waves between the receiver cabin and the reflecting dish [13].
The two 21-cm sub-maps that overlap the 2dF SGP field are shown at band center. The three rows from top to bottom show: the maps before any foreground modes are removed; the maps after 10 modes are removed; and the inverse noise weights, which are roughly proportional to the time spent observing each pixel. The color scales, from left to right, refer to the maps from top to bottom. All beams have been combined, and the resolution has been degraded to 1.4 times the original beamsize. The point sources and diffuse galactic foregrounds have been strongly suppressed in the foreground cleaned maps, and the scale of the remaining fluctuations is consistent with thermal noise. It should be noted that the fluctuations on the right ascension edges of the cleaned maps saturate the scale, but their magnitude is consistent with thermal noise and sparse coverage. Some cross-hatched striping can be seen in the maps and weights due to the differing scan angles of the azimuthal scan strategy as the field rose and set. The noise implied by the weights is higher than thermal noise because the mapmaker’s noise estimation includes variance from the noise-cal measurement, residual RFI, and fluctuating foregrounds.

Results of the SVD calculation of foreground modes on a simulated 30 degree by 30 degree map with the same frequency range as our GBT maps; 700 to 900 MHz over 256 frequency channels. The simulated data was provided by Le Zhang, and it follows the Santos foreground model and the 21-cm model of [97]. The plot shows the square root of the eigenvalues $\Sigma$, normalized to 1 for the largest eigenvalue. The foregrounds are completely described by the three largest eigenvectors, which are smooth functions of frequency. The 21-cm power spectrum can be reconstructed with minimal signal loss after de-projecting these 3 frequency modes from every pixel of the map.
7-1 From [86]. The auto-power spectrum of the two GBT fields are shown in blue and green. The solid line shows the power spectrum before foreground removal. Blue and green dotted lines show the power spectrum of the thermal noise in both maps. The noise bias is avoided in the auto-power spectra by cross-correlating maps made with different parts of the data set. Error bars come from thermal noise plus sample variance. The dashed red line is the lower bound on the power spectrum set by the cross-power measurement.

7-2 From [86]. Confidence intervals on the constrained value of the power spectrum are shown. Dark gray shows the 68% confidence intervals, and light gray shows the 95% confidence intervals. The dotted red line is the lower bound set by the cross-power. The solid red line represents the 95% confidence interval that should be reachable based on thermal noise only, in the absence of any foreground residuals.
7-3 Left: Observed 1D cross-power averaged over the four Parkes fields and cleaned by removing 10 SVD modes. A circle denotes positive power and a × denotes negative power. The grey line is the mean of the simulations, for which we assume $b_{\text{HI}} = 0.85$ and $T_b = 0.064$ mK, given by the ALFALFA measurement of $\Omega_{\text{HI}} [47]$. The dashed black line is the corresponding dark matter power spectrum scaled as Eq 3.93. Plotted error bars are 1-$\sigma$, derived from the Monte Carlo simulations described in section 6.4.2. Right: The purple points are the average of the auto-power spectra of the 2dF galaxies in the regions that overlap our Parkes maps. Errors are the standard deviation of the mean over the four sub-map regions. The dashed black line is the simulated dark matter power spectrum. The solid grey line is the expected shot noise signal, simulated from 100 unclustered mock catalogs that follow the survey selection function. The green points are the 2dF auto-power data minus the simulated shot noise.

7-4 Similar to figure 7-3, but the 2dF galaxies have been split into red and blue populations. Left: Observed 1D cross-power spectrum between the HI maps and the red and blue 2dF galaxies, averaged over the four Parkes fields and cleaned by removing 10 SVD modes. The red cross-power points are slightly offset on the k-axis, for ease of reading. Right: Observed 1D 2dF red-blue galaxy cross-power spectrum averaged over the regions that overlap our Parkes maps. Errors are the standard deviation of the mean over the four sub-map regions.
8-1 From [85]. These plots show the SVD analysis of the frequency-frequency covariance matrix formed by correlating pairs of 21-cm intensity maps of the GBT wide field. The map pairs cover the same location on the sky but are made with different sub-seasons of the data, to avoid thermal noise bias. Left: The solid line shows the square root of the SVD singular values, ordered from largest to smallest and normalized such that the largest singular value is set to 1. The dashed line is an estimate of the SVD spectrum of the noise and variable systematic effects. It represents the SVD of the difference of sub-season maps. The singular values of the dashed noise estimate is also normalized to the largest value of the solid line SVD spectrum. Right: The frequency eigenvectors associated with the SVD spectrum of the GBT maps. The eigenvectors are ordered by their associated singular values, with the largest one on the top. The first vector resembles a smooth power law, but note that there is some non-smooth frequency structure. The second mode comes from mode mixing of angular structure into frequency structure due to the frequency-dependent main-beam and can be removed by a common-beam convolution, assuming a Gaussian beam-model. The projection of the third mode onto the angular structure of the map correlates strongly with polarized emission in the map. Its frequency structure also resembles the frequency structure of measured GBT polarization leakage patterns. Therefore, this mode is likely due to leakage of polarized intensity into the unpolarized maps. The origin of the rest of the modes is unknown. Approximately 30 foreground modes are removed in the analysis of 7.1 to reach the level of signal and thermal noise. The foreground modes that are removed do not resemble smooth polynomials.
8-2 The SVD spectrum of simulated maps with 21-cm signal, foregrounds, and point sources, convolved with a frequency-dependent Gaussian beam based on fits to the measured GBT beam with a Gaussian function. Two realizations of Gaussian thermal noise are added to the otherwise identical maps. Left: The square root of the SVD singular values, ordered from largest to smallest and normalized such that the largest singular value is set to 1. Right: The frequency eigenvectors associated with the SVD spectrum of the simulated maps. The first 5 smooth modes are due to foregrounds: the Gaussian beam convolution adds one smooth foreground mode (there are only 4 smooth foreground modes if no beam convolution is applied to the simulated maps). The eigenvectors are ordered by their associated singular values, with the largest one on the bottom (opposite convention of figure 8-1). The modes after the 5 largest are dominated by noise and 21-cm signal.

8-3 The hypothetical Airy beam of a 100 meter circular aperture. This represents the beam that would be formed by an ideal receiver that uniformly illuminated the Green Bank Telescope. The beam-pattern as a function of angular distance from the center of the beam is displayed for three frequencies in the GBT bandwidth: 700 MHz (blue), 800 MHz (green), and 900 MHz (red). Each beam-pattern is normalized to 1.0 at the center for this plot. Significant mode-mixing can occur when bright structure is in the side-lobes of the beam due to the strong frequency dependence of the side-lobes.
The SVD spectrum of simulated maps with 21-cm signal, foregrounds, and point sources, convolved with a 100-meter Airy beam. Two realizations of Gaussian thermal noise are added to the otherwise identical maps. Left: The square root of the SVD singular values, ordered from largest to smallest and normalized such that the largest singular value is set to 1. Right: The frequency eigenvectors associated with the SVD spectrum of the simulated maps. The first 10 modes are due to foregrounds, modified by the systematic effect of the Airy beam convolution. The eigenvectors are ordered by their associated singular values, with the largest one on the bottom (opposite convention of figure 8.1). The first 10 modes are fairly smooth and are easily associated with the smooth foregrounds. The Airy beam is responsible for 5 of these modes: modulation of the smooth power law by the frequency-dependent side-lobes of the Airy beam is visible in the frequency structure of these modes. If a Gaussian beam is used instead, only the first 5 modes smooth modes are present, and the next 5 modes are not at all smooth, apparently dominated by the noise and 21-cm signal. The singular values of these 5 non-smooth modes are also lower with the Gaussian beam-model.

Forecasts of relative errors (a) and absolute errors (b) on the power spectrum at the location of the BAO wiggles. The purple curve is a hypothetical survey conducted with the current horn receiver, and the other colors show various times and angular coverages with the 7-element array. Only thermal noise is accounted for, so these errors should be viewed as an idealization. Foreground residuals may dominate over thermal noise.
List of Tables

3.1 Foreground parameters in the Santos model: from [73]. . . . . . . . 67

4.1 Estimate of $T_{sys}$ for both polarizations of a receiver in the proposed
7-element array. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 94
Chapter 1

Introduction

Cosmology is the study of the origin and history of the Universe. The past century has seen an incredibly rapid advance in this area of study. In that time, Cosmology has transformed from a mostly philosophical quandary to a precision science. The theoretical foundation for this transformation was the formulation of the theory of general relativity in 1915, which provided the theoretical underpinnings for an expanding Universe. Technologically, the advances have been driven by improved tools for observing the Universe. The development of radio receivers and space technology have allowed us to make the single most significant measurement in Cosmology: the detection of the relic cosmic microwave background (CMB) radiation from the very hot early Universe, and the measurement of the tiny fluctuations in this background radiation that were the seeds for the growth of galaxies. Combining the CMB maps with later time large scale structure (LSS) surveys of the clustering of matter is an extremely power tool for expanding our knowledge of the Universe [28]. This thesis focuses on a promising, relatively new, technique to map LSS, known as 21-cm intensity mapping. This technique uses the 21-cm line of neutral hydrogen (HI). Because hydrogen accounts for most of the visible matter in the Universe, and it existed even before stars ever formed, 21-cm intensity mapping has the potential to map the matter distribution of most of the observable Universe over almost its entire
history. Since the 21-cm line is a radio transition, these surveys can utilize powerful existing radio technology to map the LSS relatively quickly and cheaply. The technique is not without its challenges, however. The chief challenge is the existence of astrophysical foreground signals which must be disentangled from the HI signal. This thesis overviews some of the work I have been involved with on early 21-cm intensity mapping experiments.

1.1 Outline

Chapter 2 is an overview of the standard ΛCDM theory of cosmology and the methods for measuring the clustering of matter, focusing specifically on the power spectrum. I also overview how various dark energy models can be constrained by measuring the power spectrum at low redshifts. Chapter 3 derives the important fundamental physics of the 21-cm line, including the size of the expected signal and the observable power spectrum. It also reviews the large astrophysical foregrounds that present a challenge to detecting the 21-cm signal. Chapter 4 summarizes the fundamental properties of single-dish radio telescopes and the specific properties of receivers that my collaborators and I have used on the Green Bank and Parkes telescopes. It also summarizes our efforts in designing a multi-beam instrument for the Green Bank Telescope (GBT), and it includes a brief summary of near term future 21-cm surveys. Though it is not the main emphasis of this work, I also note the fast radio burst (FRB) that was discovered in our archived GBT data, and I describe the projected rate of FRB discoveries for our proposed multi-beam GBT array. Chapter 5 is a summary of the procedure for making 21-cm intensity maps from data taken with the Parkes telescope, including a sub-section describing how the code was parallelized. Chapter 6 reviews the methods that my collaborators and I have used to remove foregrounds from 21-cm intensity maps. Chapter 7 summarizes the results from 21-cm intensity maps made with the Green Bank and Parkes telescopes. Chapter 8 is a discussion of
my ongoing research, focusing on continuing analysis of the Parkes data and recent
efforts to simulate instrumental effects on the measured foreground signal. I conclude
in Chapter 9.

1.2 My Contributions

I joined a collaboration in 2011 that had already produced detections of cosmic struc-
ture by correlating 21-cm maps made with the Green Bank Telescope (GBT) with
the DEEP2 galaxy survey in 2010. At that time, the major contributors to the group
included Kevin Bandura (then at Carnegie Mellon University, now at West Virginia
University), Tzu-Ching Chang (then at the Academia Sinica Institute of Astronomy
Astrophysics, now at the Jet Propulsion Laboratory), Jeff Peterson (Carnegie Mel-
lon University), and Ue-Li Pen (University of Toronto, CITA). The group was then
designing an array of receivers for the GBT to increase its mapping speed. My task
was to simulate various feed antenna designs, in search of a small feed with a narrow
beam-pattern. The final receiver design was based on my model.

I moved principally to the data side of things around 2013. I’ve conducted many
observations with the GBT, contributing to our latest maps, and I authored several
of our successful proposals for GBT time. By this time, the map-making procedure
for GBT data was standardized to a pipeline developed by Kiyoshi Masui (now at the
University of British Columbia), a then graduate student of Ue-Li Pen’s. Roughly, the
steps in that pipeline are: RFI flagging (written principally by Liviu-Mihai Calin, then
an undergraduate student at the University of Toronto), calibration (written prin-
cipally by Tabitha Voytek, then at Carnegie Mellon University, now at the University
of KwaZulu-Natal), map-making (written by Kiyoshi Masui), foreground removal and
power spectrum estimation (written principally by Eric Switzer, now at NASA God-
dard Space Flight Center, and Yi-Chao Li, now at the University of KwaZulu-Natal).
Kiyo envisioned an upgrade to the map-making algorithm, which was becoming a
bottle-neck in computation time as the angular extent of the maps increased due to
new data. With help from Kiyo and Richard Shaw (at the University of Toronto,
CITA), I parallelized the map-making code, which relieved this bottle-neck and made
it possible to efficiently make larger maps. In 2014, our group was granted observing
time on the Parkes telescope. I traveled to Australia to operate the telescope with
Yi-Chao Li and Cheng-Yu Kuo (then at the Academia Sinica Institute of Astronomy
Astrophysics, now at the National Sun Yat-sen University). Upon our return,
Yi-Chao and I led the analysis of the Parkes data. Yi-Chao converted the data to
the same format as the GBT data and processed the data through the RFI flagger.
I computed the maps and performed absolute flux and bandpass calibration. I also
modified the foreground cleaning code for use with the Parkes maps and ran it on
the Parkes maps. Around this time Yi-Chao graduated. Nick Luciw (University of
Toronto, CITA) joined the collaboration, and he ran the power spectrum calculation
on the Parkes maps. I led the writing of a paper describing the results of our initial
analysis, which cross-correlates the Parkes maps with 2dF galaxy maps. This thesis
occasionally borrows language from that paper, which has been submitted to MN-
RAS. A future paper which will analyze constraints on shot noise with the Parkes
auto-power spectrum is in preparation.

I have also supervised UW Madison undergraduate students Aleks Cianciara in
simulations of cylinder telescopes for the upcoming Tianlai 21-cm intensity mapping
experiment and Ben Hoscheit in simulations of systematic effects on measured fore-
grounds in single-dish 21-cm experiments.
Chapter 2

Cosmology Overview

In this chapter, I give an overview of the development of today’s standard ΛCDM cosmological model.

2.1 The Expanding Universe

The modern big bang theory started with the discovery that distant galaxies appear to be receding away from us, no matter in what direction we look. The discovery was made possible by the work of the American astronomer Vesto Slipher. In 1912, Slipher was the first person to observe the shift in atomic spectral lines of distant galaxies that is now termed redshift. This spectral shift was interpreted as a relativistic Doppler shift, caused by the relative velocity of the distant galaxy to us. He found that some galaxies, such as Andromeda, were blueshifted, indicating that they were moving toward us. However, most galaxies he looked at were redshifted, meaning that they were moving away from us, some at very rapid rates.

In 1908, Henrietta Swan Leavitt discovered a relationship between the period and brightness of certain time-variable stars called Cepheids. She studied Cepheids in the large and small Magallanic clouds, two dwarf galaxies that orbit the Milky Way. She found that, when she narrowed her focus to just stars within a single dwarf galaxy,
she could accurately predict the luminosity of each Cepheid by simply measuring the period over which the Cepheid’s luminosity varied. Since stars within the same dwarf galaxy ought to be approximately the same distance from us, she was able to conclude that she had found a fundamental relationship between period and intrinsic luminosity in Cepheid stars. Her discovery, also published in 1912, made Cepheids the first standard candle of cosmology. The term standard candle refers to any object whose intrinsic luminosity can be determined even without knowing how far away it is. In the case of a Cepheid, the intrinsic luminosity is a simple function of its period. Knowing the intrinsic luminosity allows one to calculate the distance to the object by comparison to the measured luminosity, which decreases with the square of the distance.

In 1927 and 1929, respectively, a Belgian Catholic Priest and astronomer, Georges Lemaitre, and two American astronomers, Edwin Hubble and Milton Humason, independently discovered what has come to be known as Hubble’s law. They made this discovery by combining redshift measurements with distance measurements from Cepheid variable stars. They found that all distant galaxies are receding from us, and more distant galaxies are receding faster. In fact, the rate at which they recede appears to be proportional to their distance:

\[ v = H_0 D. \]  

The constant of proportionality, \( H_0 \), is known as Hubble’s constant (Hubble originally over-estimated it by nearly a factor of 10). We now know \( H_0 \) to be about \( 70 \frac{\text{km/s}}{\text{Mpc}} \) to within at least 10\% (there is currently some tension between different methods of measuring \( H_0 \)). Hubble’s Law is the observation that led to the idea of the big bang theory. It does not take much of a leap to imagine that if objects are moving away from us now, then they must have been closer in the past.
2.2 The FLRW metric and the Friedmann Equations

One of the consistent lessons of early astronomy is that we do not occupy a central celestial location. The Sun does not revolve around us. We are not at the center of the galaxy. In fact, there seem to be many galaxies. The Cosmological principle takes these lessons to the largest scales, asserting that we are not in a special place in the Universe. Of course this is not true on local scales, since the Earth is certainly very different from the nearly empty void of space. But if one zooms out to large enough scales, the Cosmological principle asserts that the Universe will look pretty much the same everywhere. The basis for the modern big bang theory comes from combining the Comological principle with the theory of general relativity, published by Einstein in 1915. In general relativity, the local geometry of space-time is described by a metric that relates the proper time of a moving particle, that is the time $d\tau$ passing in the particle’s rest frame, to the motion of that particle through space-time, as measured by an outside observer. The Cosmological principle implies the Friedmann-Lemaitre-Robertson-Walker (FLRW) metric for the Universe. In hyper-spherical coordinates, the FLRW metric implies that an observer at rest in the Universe will see a moving particle obey:

$$c^2 d\tau^2 = c^2 dt^2 - a(t)^2 \left[ dr^2 + S_k^2(r) d\Omega^2 \right] . \quad (2.2)$$

Let’s analyze what this means. As a simplification, the observer is free to choose the radial direction $\hat{r}$ of the FLRW coordinate system to coincide with the particle’s direction of motion, so that $d\Omega = 0$. For the case of this single moving particle, the only difference between this FLRW Universe and the flat Minkowski space of special relativity is that the FLRW Universe can expand or contract. That is, if distance markers were placed along the particle’s path in the $\hat{r}$ direction before the particle set out, those markers could get farther or closer to each other over time. That shrinking or expanding of the Universe is quantified by the dimensionless scale factor $a(t)$. The imaginary distance markers that get farther or closer apart as the
Universe expands or shrinks are what cosmologists refer to as comoving coordinates. Comoving coordinates do not measure the effect of the changing scale factor \( a(t) \). A solid ruler, on the other hand, would not expand with the Universe since its chemical bonds would easily resist any expansion effects. Coordinates measured by lining up many such rulers are referred to as proper coordinates: they will notice the effect of the scale factor \( a(t) \). During a proper time \( d\tau \) in the particle’s rest frame, the outside observer will have seen a particle with a velocity \( v \) travel a proper distance \( a(t)dr = vdt \) and a comoving distance \( dr = \frac{vdt}{a(t)} \) in a time \( dt \): and these are all related through equation 2.2. For a constant scale factor \( a(t) = 1 \), equation 2.2 implies the Lorentz transformation of special relativity. The only other difference between FLRW coordinates and Minkowski space can be observed by measuring the angular distance between multiple objects in different directions. The geometry of three dimensional space in an FLRW Universe is not generically Euclidean, or flat. There could also be either positive (sphere-like) or negative (saddle-like) curvature. Positive or negative curvature would alter the measured angles between distant objects. This is quantified by the function \( S_k(r) \), defined as

\[
S_k(r) = \begin{cases} 
\frac{\sin(\sqrt{k}r)}{\sqrt{k}}, & \text{if } k > 0 \\
r, & \text{if } k = 0 \\
\frac{\sinh(\sqrt{-k}r)}{\sqrt{-k}}, & \text{if } k < 0.
\end{cases}
\]

The factor \( k \) has units of inverse area and describes a possible Gaussian spatial curvature to the Universe. All current observations are consistent with a Euclidean flat \( k = 0 \) Universe. For the FLRW metric, the expansion of the Universe is completely described by the parameter \( a(t) \), which we know is increasing due to Hubble’s Law. By convention, \( a(t) \) is set to 1 at the present time, and \( a(t) \) is believed to have been smaller in the past. A far-away galaxy at some distance \( D(t) = a(t)r \) will be receding.
from us at a rate
\[ v(t) = \dot{a}(t)r = \frac{\dot{a}(t)}{a(t)}a(t)r = \frac{\dot{a}(t)}{a(t)}D(t). \]  
(2.4)

Therefore, the Hubble Constant must be
\[ H_0 \equiv \frac{\dot{a}}{a} \bigg|_{t=t_0}, \]  
(2.5)

where the scale-factor and its derivative are evaluated at the current time. The Hubble Parameter \( H(t) \) is a generalization of the Hubble constant, but the quantity is evaluated at time \( t \), which is not necessarily the current time. Since the only part of the metric that changes with time is the scale factor \( a(t) \), the spatial coordinates of a point at rest in the Universe are specified by the vector \( a(t)\vec{r} \). The spatial coordinates with the scale-factor divided out correspond to the comoving coordinates. Coordinates that include the expansion factor \( a(t) \) are proper coordinates. Since the scale-factor is 1 at present, comoving and proper coordinates agree only during the current epoch.

By combining the Cosmological principle with the Einstein Field equations, and assuming that the Universe is filled with a perfect fluid, one can derive the Friedmann equations, which govern how the scale factor changes due to the energy content of the Universe.
\[ \frac{\ddot{a}}{a} = -\frac{4\pi G}{3} \left( \rho + 3\frac{p}{c^2} \right) + \frac{\Lambda c^2}{3} \]  
(2.6)

and
\[ \left( \frac{\dot{a}}{a} \right)^2 = \frac{8\pi G\rho}{3} - \frac{k c^2}{a^2} + \frac{\Lambda c^2}{3}, \]  
(2.7)

where \( G \) is Newton’s gravitational constant, \( c \) is the speed of light, \( \rho \) is the density of the fluid, \( p \) is the pressure of the fluid, and \( \Lambda \) is a cosmological constant describing a hypothetical constant energy density component of the Universe. Equation 2.6 is often referred to as the acceleration equation, in which case equation 2.7 is exclusively referred to as the Friedmann equation. In general relativity, energy conservation is
replaced with covariant conservation of the local stress-energy tensor, which governs how the energy density of the fluid changes with the scale factor. This stress-energy conservation leads to the equation

$$\dot{\rho} = -3H \left( \rho + \frac{p}{c^2} \right).$$

This equation governs how the energy density of the various components of the Universe change as the Universe expands. Massive particles, such as protons and electrons, have much more energy in their mass \( E = mc^2 \) than they have due to the pressure of their motion. Thus, for massive particles, \( p \approx 0 \) in the above equation. Radiation, on the other hand, consists of mass-less photons, whose energy density is entirely due to momentum through the relation \( E = |P|c \) (even though radiation has no mass, \( \rho \) is defined via \( \rho = \frac{E}{c^2V} \)). Accounting for geometrical factors, this results in the relationship \( \rho_r = \frac{3p}{c^2} \) for radiation. Applying these results to equation 2.8, one finds:

$$\rho_m \propto \rho_{m0} a^{-3}$$
$$\rho_r \propto \rho_{r0} a^{-4}.$$  \hspace{1cm} (2.9)

Both of these dependencies have simple interpretations. For matter, the density decreases with the cube of the scale factor because the Universe becomes less dense as it expands. The density of radiation also decreases with the cube of the scale factor, but photons lose additional energy because their wavelength stretches with the scale-factor of the Universe. This is due to the gravitational redshift effect of general relativity.

Now that we know how the energy density of radiation and matter evolve with the scale factor of the Universe, we can combine this with the Friedmann equation to extrapolate the expansion history of the Universe back in time, based on its current
energy content.

\[
\left( \frac{\dot{a}}{a} \right)^2 = \frac{8\pi G \rho_m a^{-3} + \rho_r a^{-4} - \frac{k c^2}{a^2} + \frac{\Lambda c^2}{3}}{3}.
\] (2.10)

In order to remove some of the constants in this equation, we define the critical density as the energy density necessary to reproduce the current observed Hubble constant \( H_0 \approx 70 \frac{\text{km/s}}{\text{Mpc}} \). We define parameters \( \Omega_x = \frac{8\pi G \rho_x}{3H_0^2} \) as the contribution to the critical density from component \( x \) today. With these definitions, the above equation becomes

\[
\frac{\dot{a}}{a} = H_0 \sqrt{\Omega_m a^{-3} + \Omega_r a^{-4} + \Omega_k a^{-2} + \Omega_\Lambda}.
\] (2.11)

The above equation provides the expansion history of the Universe, so long as energy is not transferred from matter to radiation, or vice-versa. The fact that universal expansion also increases a photon’s wavelength gives us the most important tool of cosmology. If a photon is emitted at a wavelength \( \lambda_e \) when the Universe has a scale factor \( a \), then we will measure that photon’s wavelength now, when the scale factor is 1, as

\[
\lambda_o = \frac{\lambda_e}{a} \equiv (1 + z) \lambda_e,
\] (2.12)

where I have defined redshift \( z \equiv \frac{1}{a} - 1 \). As a photon travels through the Universe at the speed of light, \( c dt = a(t) dr \) (choosing our coordinate system so that the motion is purely radial). Applying this result and a coordinate transformation to equation 2.11, one finds that the comoving distance transversed by the photon (and the proper distance to the source of the photon) is given by

\[
D(z) = \frac{c}{H_0} \int_0^z \frac{dz}{\sqrt{\Omega_m (z+1)^3 + \Omega_r (z+1)^4 + \Omega_k (z+1)^2 + \Omega_\Lambda}}.
\] (2.13)

With this equation, we now have a method for mapping the distance to far-off objects by measuring redshift. Since atoms emit specific patterns of spectral lines, we can look for redshifted emission of these patterns, and use this equation to calculate the
distance to the object. This equation forms the basis of spectroscopic galaxy surveys and also 21-cm intensity mapping, which is the subject of this thesis. Redshift is also used to measure the distance to the cosmic microwave background (CMB) radiation, which will be discussed in section 2.3.2.

The Friedmann equation is a valid solution to Einstein’s equations, but we do not live in a perfectly homogeneous Universe. It seems reasonable that the Universe is homogeneous on large scales, but is there any proof? Very nearly. It was proven [84] that, if general relativity is correct, and if the cosmic microwave background radiation is nearly uniform for all comoving observers in the local Universe, then the local Universe is very nearly homogeneous and isotropic; and thus described by the FLRW metric and the Friedmann equations. The CMB is uniform to a part in 100,000 as we observer it on Earth. This fact, combined with the Copernican principal: the idea that we occupy no special observational frame, implies that the Friedmann equation and FLRW metric is correct, so long as general relativity is believed.

2.3 The ΛCDM Model

The current standard model of cosmology is the ΛCDM model: Λ stands for a cosmological constant, and CDM stands for cold dark matter. The ΛCDM model posits that we are in an expanding Universe that started with a big bang and whose current energy density is approximately 30% matter and 70% an unknown component, called dark energy, whose properties are consistent with a cosmological constant. Only 15% of the matter, roughly 5% of the total energy density of the Universe, is believed to be ordinary visible matter. The rest of the matter has been inferred to exist through its gravitational effects but does not appear to otherwise interact with light or with ordinary matter. This extra matter is referred to as dark matter. Dark matter is cold, in the sense that the dark matter particles have a large enough mass that their pressure is negligible, compared to the energy of their mass.
The main pillars of the big bang theory are the observation of the expanding Universe, the match of observed helium and hydrogen abundances to the predictions of big bang nucleosynthesis, and the existence of the cosmic microwave background radiation. Dark matter and dark energy are later additions to the theory.

### 2.3.1 Quantifying Structure

Much of what we know about cosmology comes from measurements of structure in the Universe: galaxies, clusters, and groups. In this section, I briefly divert from cosmology to build up the mathematical tools used to quantify measurements of large scale structure (LSS).

The matter density field is the primary quantity of interest in LSS studies. It is useful to define the fractional density perturbation:

\[
\delta(\vec{x}) = \frac{\rho(\vec{x}) - \bar{\rho}}{\bar{\rho}},
\]

(2.14)

where \(\bar{\rho}\) is the average density of the Universe. Cosmological models cannot predict where dense structure will form, but they can predict how matter tends to cluster. One way to quantify the clustering of matter is to measure correlations between the density perturbations as a function of their separation. The most important function for describing and measuring this is the two-point correlation function:

\[
\xi(r) = \langle \delta(\vec{x})\delta(\vec{x} + \vec{r}) \rangle
\]

(2.15)

Theoretically, the average is over multiple realizations of the Universe, drawn from the statistical model given by the cosmology (this might sound odd, but it makes more sense if one keeps in mind the idea that structure was seeded by random quantum fluctuations, which could have been different). The double bracket symbol \(\langle \rangle\) indicates an average over multiple realizations of the Universe. The over-bar notation (as in \(\bar{\rho}\)) represents a spatial average over a single realization of the Universe. The assumption
of homogeneity on large scales means that the correlation function depends only on the separation between two points, not their absolute positions, which is why there is no \( \vec{x} \) dependence of \( \xi(r) \). Similarly, the assumption of isotropy means that the directions of the separations are also irrelevant, which is why \( r \) is shown as a scalar.

In reality, we have access to only one Universe, so how is one to measure \( \xi(r) \)? Imagine that a large survey of the Universe has been conducted, and a pixelized density perturbation map has been created from this survey. One can exploit isotropy and homogeneity to estimate \( \xi(r) \) by averaging over all pairs of pixels in the map that are separated by a distance of approximate length \( r \). In doing this, one replaces the average over multiple realizations of the Universe with an average over different locations of our single observed Universe. This procedures assumes the principle of ergodicity. The idea is that the density fluctuations at different pairs of points located throughout our Universe make a fair sample of the clustering that would have developed between a fixed pair of points if the Universe had reformed many times.

Cosmological models are most predictive on very large scales, where the evolution of the perturbations is well described by linear perturbation theory. Therefore, it is natural to make predictions and report measurements in Fourier space, where linear modes evolve independently of each other. The Fourier conventions used here and most commonly in the literature are:

\[
\delta(\vec{k}) = \mathcal{F} \left[ \delta(\vec{x}) \right] = \int \delta(\vec{x}) e^{i\vec{k} \cdot \vec{x}} d^3\vec{x} \tag{2.16}
\]

\[
\delta(\vec{x}) = \mathcal{F}^{-1} \left[ \delta(\vec{k}) \right] = \frac{1}{(2\pi)^3} \int \delta(\vec{k}) e^{-i\vec{k} \cdot \vec{x}} d^3\vec{k} \tag{2.17}
\]

The Fourier transform of the two point correlation function is called the power spectrum.

\[
P(k) = \mathcal{F} \left[ \xi(r) \right] = \int \xi(r) e^{i\vec{k} \cdot \vec{r}} d^3\vec{r} \tag{2.18}
\]
The $k$ in $P(k)$ is written as a scalar because the assumption of isotropy implies that the power spectrum is equal for all $\vec{k}$ vectors of the same magnitude. $P(k)$ is related to the Fourier transformed density perturbations by

$$\langle \delta(\vec{k})\delta(\vec{k}') \rangle = (2\pi)^3 \delta_D^3(\vec{k} + \vec{k}') P(k)$$

(2.19)

Here $\delta_D^3(\vec{k})$ is the three dimensional Dirac delta function. Equation 2.19 can be demonstrated as follows. One starts by expressing each $\delta(\vec{k})$ as the Fourier transform of $\delta(\vec{x})$.

$$\langle \delta(\vec{k})\delta(\vec{k}') \rangle = \left\langle \int d^3\vec{x} e^{i\vec{k} \cdot \vec{x}} \int d^3\vec{r} e^{i\vec{k} \cdot \vec{r}} \right\rangle$$

$$\langle \delta(\vec{k})\delta(\vec{k}') \rangle = \int d^3\vec{x} d^3\vec{r} e^{i(\vec{k} + \vec{k'}) \cdot \vec{x}} \langle \delta(\vec{x})\delta(\vec{x} + \vec{r}) \rangle e^{i\vec{k'} \cdot \vec{r}}$$

$$\langle \delta(\vec{k})\delta(\vec{k}') \rangle = \int d^3\vec{x} e^{i(\vec{k} + \vec{k'}) \cdot \vec{x}} \int d^3\vec{r} \xi(r) e^{i\vec{k'} \cdot \vec{r}}$$

$$\langle \delta(\vec{k})\delta(\vec{k}') \rangle = (2\pi)^3 \delta_D^3(\vec{k} + \vec{k}') P(k)$$

$$\langle \delta(\vec{k})\delta^*(\vec{k}') \rangle = (2\pi)^3 \delta_D^3(\vec{k} - \vec{k}') P(k)$$

(2.20)

In the third line, we used the definition of the two-point correlation function (which encapsulates the assumptions of homogeneity and isotropy). In the fourth line, we used the definition of the power spectrum as the Fourier transform of the two-point correlation function and the Fourier transform property that $\int d^3\vec{x} e^{i(\vec{k} + \vec{k'}) \cdot \vec{x}} = (2\pi)^3 \delta_D^3(\vec{k} + \vec{k'})$. In the last line, we exploited the facts that $\delta(\vec{r})$ is real, which implies that $\delta(-\vec{k}) = \delta^*(\vec{k})$.

Any real survey will only be able to observe a finite but hopefully very large volume $V$ of the Universe. The calculation of $\delta(k)$ must therefore be computed, not as a Fourier transform over all space, but as Fourier series expansion over the volume $V$.

$$\delta(\vec{k}) = \int_V \delta(\vec{x}) e^{i\vec{k} \cdot \vec{x}} d^3\vec{x}$$

(2.21)
For a Fourier series, the identity \( \int d^3 \vec{x} e^{i(\vec{k} + \vec{k}')} \cdot \vec{x} \) becomes \( V \delta_{\vec{k},-\vec{k}'} \), where \( \delta_{\vec{k},-\vec{k}'} \) is the Kronecker delta. Therefore, 2.20 for a finite volume survey becomes

\[
\frac{\langle \delta(\vec{k}) \delta^*(\vec{k}') \rangle}{V} = \delta_{\vec{k},\vec{k}'} P(k). \tag{2.22}
\]

From equation 2.21, one can see that the Fourier coefficients of the over-density, let’s call them \( a_{\vec{k}} \), are given by \( a_{\vec{k}} = \frac{\delta(\vec{k})}{V} \). Therefore, equation 2.22 actually says that the power spectrum is the variance of those Fourier coefficients multiplied by the volume of the survey. Since the power spectrum is a property of the full density field, it should tend towards a constant number, regardless of the volume of the survey (so long as it is sufficiently large to be representative of the Universe as a whole). Therefore, the variance of the Fourier coefficients scale like: \( \langle a_{\vec{k}} a_{\vec{k}}^* \rangle \propto \frac{1}{V} \). It may seem odd that a larger volume sample will have smaller amplitude Fourier coefficients, but the reason is that the density of the k-space sampling is proportional to the volume, so larger volumes allow one to resolve the power into finer bins in k-space.

In this thesis and in many cosmology papers, plots of the power spectrum do not display the quantity \( P(k) \). Instead, they show the quantity \( \Delta^2(k) \), which is often referred to as the dimensionless power spectrum. It is defined as

\[
\Delta^2(k) = \frac{k^3 P(k)}{2\pi^2}. \tag{2.23}
\]

The dimensionless power spectrum has a fairly straight-forward physical interpretation, which can be stated as follows. Imagine measuring the density in a sphere of radius \( R \) that is randomly placed in the Universe. On average, the matter contained in that sphere will have a density equal to the average density in the Universe. Therefore, the average over-density in that randomly placed sphere is zero. But the variance of the over-density contained within the randomly placed sphere will not be zero, and (making the reasonable assumption that \( \Delta^2(k) \) increases at higher \( k \)) it will
roughly be

\[ \sigma^2(R) \approx \Delta^2 \left( \frac{1}{R} \right). \] (2.24)

The power spectrum has been the primary data reduction tool in studies of cosmic structure. If the fluctuations have a Gaussian distribution, which the CMB fluctuations do to high certainty, then the 2-point correlation function or the power spectrum contains all the information. However, many theories predict some degree of deviation from Gaussianity, and non-linear gravitational collapse also causes deviations from Gaussianity to develop over time. Higher order statistical tools such as the 3-point correlation function or the bi-spectrum have been developed to characterize non-Gaussianity. However, since this thesis is on early pioneering work in 21-cm intensity mapping, we will not use these tools and will instead focus on attempts to accurately measure the power spectrum.

2.3.2 Epochs in the History of the Universe

Returning to our picture of an expanding Universe, if we extrapolate the expansion of the Universe back in time, it must have been denser and hotter in the past. In the far past, the Universe was hot enough that most atoms were ionized. During that early epoch, the high density of photons and ions kept the photons coupled to the free electrons and protons through Thompson scattering. Radiation and baryonic matter acted like a single fluid, and they were in thermal equilibrium. As the Universe expanded and cooled, the electrons and protons combined to form neutral atoms (as set by earlier big bang nucleosynthesis, the atoms were roughly 75% hydrogen by mass and 25% helium by mass, with trace amounts of some heavier elements). Once the photons’ temperature dropped below the energy needed to excite the atoms from their ground states, the chance of photons scattering with a neutral atom became negligible. This marked a transition in the Universe, when it rather abruptly went from opaque to transparent, and radiation thermally decoupled from matter. Since that
point, photons have free-streamed through the nearly transparent Universe. Photons from this ‘last scattering surface’ are still visible today. They have an almost perfect thermal black-body distribution, which has been redshifted by the subsequent expansion of the Universe from their original temperature of 3,000 Kelvin at the time of decoupling. This cosmic microwave background (CMB) radiation was first discovered in 1964 by Arno Penzias and Robert Wilson. The observed temperature of the CMB is 2.7 Kelvin, which implies that the Universe today is roughly 1100 times larger than it was at the time of decoupling. The extreme homogeneity of the CMB temperature, which is the same in every direction to within 1 part in 100,000, justifies the FLRW metric’s assumption of isotropy.

To many cosmologists, the isotropy of the CMB also represents a philosophical challenge. If the early expansion history was dominated solely by radiation and matter, then the speed of light would have been too small for the entire surface of last scattering to be causally connected. Therefore, it should not have reached thermal equilibrium. This philosophical quandary, and others, are answered by a theory known as inflation, which is accepted by most cosmologists today but not confirmed. In inflationary theory, the very early Universe was dominated by some unknown scalar field. This scalar field drove an early phase of exponential growth, which quickly expanded patches of the Universe that were previously in causal contact to huge sizes that were out of causal contact. Quantum fluctuations in this field were also expanded to large scales, and they account for the tiny temperature fluctuations in the CMB. Figure 2-1 shows the ΛCDM model with an initial inflationary period.

After the inflationary period, the dominant components in the Friedmann equation were first radiation, and later matter. The Universe continued to expand, but the gravitational effect of these components slowed the rate of expansion. By the time of decoupling of matter and radiation at redshift \( z \sim 1100 \), the Universe’s energy density was dominated by matter. The expansion continued, with no stars or galaxies yet formed. However, the small over-densities, initially only different by 1 part in 100,000,
Figure 2-1: A diagram from the NASA/WMAP team of the history of the Universe in a $\Lambda$CDM model with inflation. Earlier times are on the left. The Universe starts with a brief period of exponential growth, which expands quantum fluctuations to large scales. This lasts for much less than a second, after which the inflaton field decays into the particles and radiation of the standard model. Expansion of the Universe continues at a lower rate, dominated first by radiation, then by matter. After 380,000 years, matter decouples from radiation, and the Universe becomes transparent. Expansion continues through the dark ages, until gravitational collapse forms the first stars, which reionize most of the matter in the Universe. Around 5 billion years ago, the rate of expansion begins to accelerate again as the energy density of matter drops below the energy density of the cosmological constant.
slowly collapsed into denser structures through gravitational attraction. Some time around $z \sim 30$, the first stars began to form from these over-densities. The time between matter-radiation decoupling and star formation is known as the Dark Ages.

Once stars began to form, the nuclear processes from the stars produced high-energy photons that mostly reionized the neutral hydrogen and helium in the Universe. The stars also created the heavier elements. By a redshift of $z \sim 6$, the majority of the Universe is believed to have reionized, but a small $\sim 1\%$ fraction of hydrogen was dense enough to form clouds of neutral hydrogen (HI) that are self-shielded from ionizing radiation. These HI clouds occurred at the location of collapsed matter overdensities. They were crucial to star and galaxy formation, since stars are unlikely to have gravitationally collapsed from warm ionized gas and are instead expected to have evolved from cold neutral (HI) clouds, to molecular ($\text{H}_2$) clouds, to stars [68].

Evidence from type Ia supernovae has shown that the expansion rate of the Universe has begun to accelerate again, indicating that a dark energy component is beginning to dominate the energy density of the Universe.

2.3.3 The Need for Dark Matter and Dark Energy

Dark matter and dark energy are rather unnatural elements that have been added to an otherwise simple theory. Therefore, I briefly review the reasons for introducing them.

There are several reasons for introducing dark matter. One piece of evidence comes from measurements of the rotation curves of galaxies. When a galaxy is viewed edge-on, astronomers can measure the velocity of the stars and neutral hydrogen, which rotate about the center of mass of the galaxy, by observing a Doppler shift in their spectral lines. A rotation curve is a plot of this velocity as a function of radial distance from the center of the galaxy. These curves show surprisingly high velocities, and the velocities do not decrease as expected at the outer edge of the galaxies. The most popular explanation is that most of the matter of galaxies is a form of dark
matter that extends far beyond the visible extent of the galaxy in what is known as a dark matter halo. A similar type of study looks at the velocities of individual galaxies in gravitationally bound galaxy clusters. The virial theorem relates the velocities of orbiting objects in stable gravitationally bound structures to the mass of the entire gravitating system. The observed velocities of the galaxies again imply a much higher mass for the clusters than what is visible in the luminous matter of the galaxies. Lastly, the physics of the 1 part in 100,000 fluctuations of the CMB provide an excellent test. In the early Universe, regular baryonic matter and photons were coupled through Thompson scattering, but any non-baryonic dark matter would have no way to influence photons or ordinary matter except through gravitation. The presence of significant amounts of dark matter imparts several features in the power spectrum of the CMB fluctuations that are sensitive to the ratio of dark matter to ordinary baryonic matter. These signatures of dark matter have been found in the CMB. All of these results can be consistently explained if approximately 85% of matter in the Universe is dark matter.

Dark energy is the most recent addition to the standard model of cosmology. The strongest evidence for its existence comes from another standard candle, the Type Ia supernova. Like the Cepheid variable star, the intrinsic luminosity of type Ia supernovae can be inferred from how its luminosity changes over time. Type Ia supernovae are due to exploding white dwarf stars. They are much brighter than Cepheid stars and can thus be detected out to much higher distances and redshifts. Measurements of redshift versus inferred distance to type Ia supernovae (see figure 2-2) have revealed that the Universe is currently accelerating and has changed from negative to positive acceleration in the past 5 billion years [71, 70, 61]. From the acceleration equation (2.6), an accelerating Universe is only possible if there is either a cosmological constant or some other mysterious component with a negative pressure. Whatever is responsible for the recent acceleration is referred to as dark energy. The current standard model explains the acceleration behavior as due to a cosmological constant,
Figure 2-2: From [71]. Apparent magnitude is plotted as a function of redshift for observed type Ia supernovae. Apparent magnitude is a logarithmic measure of brightness, in which smaller numbers correspond to brighter objects. For the Type Ia supernova standard candle, apparent magnitude depends logarithmically on the distance to the object. The dashed line is the best-fit flat cosmology, in which $\Omega_m = 0.29$ and $\Omega_\Lambda = 0.71$.

whose energy density only recently became noticeable as the matter density dropped due to the expanding Universe. This $\Lambda$CDM concordance model is also supported by measurements of $\Omega_m$ from the x-ray emission of large clusters [3] and from the number density of galaxy clusters, which is sensitive to dark energy’s retarding effects on structure growth [14].

The discovery of dark energy was completely unexpected. Understanding its nature is one of the major goals of modern cosmology. There is no theory that naturally explains its energy density, or why it is so close to that of matter at the present epoch,
but much smaller than the energy density of matter in the past, when the Universe was smaller and matter was denser. Dark energy has only been known about for 20 years, and its effects are only significant in the ‘recent’ Universe, meaning the last 5 billion years. The available data cannot yet nail down the dark energy equation of state, which relates its energy density, $\rho c^2$, to its pressure, $p$, through the parameter $w$:

$$\rho c^2 = wp.$$  \hspace{1cm} (2.25)

Dark energy is clearly close to a cosmological constant, which has $w = -1$, but it may have some slight deviation. A general dark energy component modifies the Friedmann equation and the redshift-distance relation as

$$D(z) = \frac{c}{H_0} \int_0^z \frac{dz}{\sqrt{\Omega_m(z+1)^3 + \Omega_r(z+1)^4 + \Omega_k(z+1)^2 + \Omega_{DE}(z+1)^3(1+w)}}. \hspace{1cm} (2.26)$$

For a cosmological constant, $w = -1$. However, $w = -0.9$ for instance, is also supported by the data. In fact, $w$ may change as a function of time. The dark energy task force [2] outlines comprehensive strategies for constraining the dark energy equation of state. All of these strategies involve measuring large scale structure (LSS) in the redshift range $0 \leq z \leq 3$. Large scale structure (LSS) surveys can constrain dark energy by measuring $D(z)$ through the baryon acoustic oscillation (BAO) standard ruler (explained further in the next section) and by measuring the growth factor $g(z)$. The growth factor quantifies the growth of small perturbations in the matter density, which slowly collapse over time due to gravity. If one defines $\delta(r) = \frac{\rho(r)-\bar{\rho}}{\bar{\rho}}$ as the local fluctuation in the matter density relative to the average matter density of the Universe and $\delta(\vec{k})$ as its Fourier transform in comoving coordinates (see section 2.3.1 for explicit definitions), then, according to linear perturbation theory, fluctuations with a wavelength that is much smaller than the observable Universe will grow as the
Universe evolves via the equation

\[ \ddot{\delta}(\vec{k}) + 2H \dot{\delta}(\vec{k}) - 4\pi G \bar{\rho} \delta = 0. \]  
(2.27)

Since \( \vec{k} \) does not enter into this equation, small fluctuations in the linear regime all grow at the same rate. The growth rate is sensitive to dark energy because it depends on the energy content of the Universe through the Hubble parameter \( H \) and the mean matter density \( \rho \). We can define a universal growth factor \( g(z) \), which relates the matter over-density at the time of decoupling \( (z \approx 1100) \) to the matter over-density at redshift \( z \).

\[ \delta(\vec{k}, z) = g(z)\delta(\vec{k}, z \approx 1100). \]  
(2.28)

We cannot observe a particular matter over-density \( \delta(\vec{k}, z) \) at more than one redshift, but the observed power spectrum of the CMB should equal the power spectrum that was present at all locations in the Universe at \( z \approx 1100 \). Therefore,

\[ P(\vec{k}, z) = g^2(z)P(\vec{k}, z \approx 1100). \]  
(2.29)

So, one can directly probe \( g(z) \) by measuring the power spectrum at redshift \( z \) and comparing it to the CMB power spectrum. However, this is complicated in practice by the fact that LSS is most easily measured by observing light-emitting objects like galaxies. Galaxies are biased tracers of the underlying matter distribution, which means that the galaxy power spectrum is multiplied by the square of the galaxy bias:

\[ P_{gg}(\vec{k}, z) = b_g^2 g^2(z)P(\vec{k}, z \approx 1100). \]  
(2.30)

This unknown galaxy bias is degenerate with the growth factor. The situation is somewhat more complicated for HI LSS surveys. However, both types of LSS surveys can remove this degeneracy by measuring redshift space distortions (RSDs).
Figure 2-3: From the report of the dark energy task force [2]. The distance $D(z)$ vs. redshift relation and the growth factor $g(z)$ vs. redshift are the primary observables that can constrain the dark energy equation of state. The black curve is the $\Lambda$CDM model. The green curve is a model with $\Omega_m = 1$, which is strongly disfavored by current data. The red curve is a model fit to CMB data with $w = -0.9$. This model is consistent with current data but can be constrained by measuring $\sim 5\%$ deviations in the distance and growth factor in the redshift range $0 \leq z \leq 3$. 
2.4 The CMB Fluctuations and Structure Formation

The power spectrum of the CMB fluctuations is well-understood theoretically, and the predicted power spectrum is a strong function of the energy make-up of the Universe. The CMB fluctuations have been measured to high precision by three generations of satellite missions: COBE, WMAP, and Planck. These are the most important data sets of modern physical cosmology. Their analysis has allowed for unprecedented precision constraints on cosmological parameters. As an example, one of the most important features in the CMB fluctuations are the baryon acoustic oscillation (BAO) peaks. These peaks represent ripples in the baryon-photon fluid. Their scale is determined by well-understood physics that governed the speed of sound in the baryon-photon fluid and the time between the big bang and the decoupling of matter and radiation. The angular scale at which we see these peaks tells us that the Universe is close to flat, and the relative strengths of the harmonics of these peaks constrain the ratio of baryonic to dark matter. Although the CMB is the most important data set of modern cosmology, most of the data available from the CMB has already been extracted by these experiments. In fact, even if all instrumental noise could somehow be removed from these measurements, the constraining power of the CMB is limited by the fact that it is only a two-dimensional probe of the specific time of matter-radiation decoupling in the history of the Universe. The CMB only measures fluctuations from a thin spherical shell whose radius is set by the distance photons have traveled between the time of radiation-matter decoupling and today. The finite volume of the Universe probed by the CMB is a limitation because the exact configuration of the initial fluctuations of the Universe cannot be predicted: theories like inflation only predict the average properties of these fluctuations. These average properties could be more accurately measured if one could probe a larger volume of the Universe. Errors on measured parameters due to sampling a finite volume of the Universe are referred to as sample variance.

There is another way to measure the fluctuations imprinted in the Universe. The
small fluctuations seen in the CMB temperature are believed to also be the seeds for
the growth of modern day galaxies and clusters. If we can detect these structures and
measure their redshift, then we can use equation 2.13 to calculate the distance to the
clusters along the line-of-sight. It should then be possible to make a 3-dimensional
tomographic map of the large scale structure (LSS) in the Universe. The density
fluctuations in such a map will be much like the CMB fluctuations, but they differ
in two important ways. First, a map of matter in the Universe can cover a much
larger total volume than a CMB map, and therefore it can potentially put a much
better constraint on the statistical properties of the initial fluctuations of the Universe.
Second, such a LSS map will probe the clustering in the Universe at different times
in its history for each redshift that is observed, due to the finite speed of light.
The fluctuations that were present at the time of matter-radiation decoupling, which
are seen in the CMB, will have evolved over billions of years, through the effect of
gravitation. A tomographic redshift map would allow us to observe the history of
this gravitational evolution (if we assume ergodicity: that every large enough volume
is a fair sampling of the Universe). The evolution of the density fluctuations with
time present both an advantage and a difficulty. On large scales, the evolution of the
fluctuations can be calculated from linear perturbation theory, and the increase in
fluctuations is quantified through the growth factor $g(z)$. The growth factor depends
on the energy content of the Universe, so measuring it accurately at low redshifts
will constrain dark energy models. However, once an object becomes about 2.7 times
as dense as the rest of the Universe, its evolution will be strongly non-linear and
it will quickly collapse into a very dense object. This cannot currently be modeled
to high accuracy. In the cold dark matter picture, this non-linear collapse happens
first with the smallest objects. Therefore, the power spectrum of the LSS will not
be as useful on small scales, particularly at recent times. The most useful part of
the power spectrum will be mainly in the largest scales, where the predictions of
theory can be easily compared with the observations. Another advantage to the
different redshifts visible in the LSS is that the baryon acoustic oscillations (BAOs) will provide a standard ruler. The scale of the BAO feature is precisely known from CMB studies. After matter-radiation decoupling, the BAO feature will simply expand with the Universe, maintaining a constant comoving size. Observing the BAO feature in the angular direction will constrain $D(z)$, and observing the BAO feature in the redshift direction will constrain the local value of $H(z)$. Both of these measurements will be sensitive to the equation of state of dark energy at $0 \leq z \leq 3$.

It should be noted that observations of smaller collapsed objects can in fact be used, but the best tool is not the power spectrum. $N$-body simulations reveal a simple functional form for the dark matter halo mass function $\frac{dn(M,z)}{dM}$ [65], which quantifies the number density of collapsed dark matter halos of mass $\approx M$, as a function of redshift $z$. This mass function depends sensitively on the energy content of the Universe at redshifts $0 \leq z \leq 3$. Measurements of the number density of high mass clusters can therefore be used to constrain various dark energy models [6]. Such measurements, however, are difficult to perform with the 21-cm intensity mapping technique that is the subject of this thesis. I will therefore focus on the power spectrum of the density fluctuations.

To roughly quantify the statistical power of LSS compared to the CMB, let’s estimate the number of independent Fourier modes that could be analyzed in these two types of surveys. For the CMB, let’s assume that Silk damping sets the maximum usable mode roughly to $l \sim 1000$. Then there are about $10^6$ observable Fourier modes in the CMB. Now consider a hypothetical LSS survey that maps the full Universe from $z = 0$ to $z = 3$. This maps a comoving volume with a radius of about 6 Gpc. Conservatively assuming that Fourier modes corresponding to scales smaller than 10 Mpc are not useful because the non-linear evolution cannot be modeled on those scales, a map of this size still gives us access to roughly $10^9$ linear or near-linear Fourier modes. Thus, LSS can theoretically achieve much lower sample variance than the CMB on most scales.
2.4.1 Matter Tracers

Measuring LSS is complicated by the fact that most matter apparently does not strongly interact with light through any force except the rather weak force of gravity. We must therefore rely on visible bayronic matter as tracers of the total underlying matter distribution. The most common method for tracing matter is to identify galaxies through the optical or near infrared emission of their stars. Optical galaxy surveys usually consist of two stages. The first stage identifies objects in the sky that are above a certain brightness threshold. Each identified object must then be classified as a galaxy or a star. This is usually achieved by assuming that extended structure corresponds to galaxies, and point-like objects are stars. A second observation is then conducted on these galaxy locations, using a spectograph. Each galaxy’s redshift is determined by detecting a shift in atomic spectral lines. Spectroscopic galaxy surveys have been the main tool for measuring LSS in the low-redshift Universe. Spectroscopic surveys provide very accurate redshift measurements, but they require long observation times. Since these surveys identify individual objects, high-redshift detections are difficult due to the inverse square law. Corrections must also be made for extinction of light due to intervening dust. The largest galaxy redshift surveys to date have been the 2dF survey [21] and the Sloan Digital Sky Survey [95]. These two surveys have detected spectra for around 230,000 and 3 million galaxies, respectively. The power spectrum of LSS galaxy surveys is computed by assuming that optical galaxies are a biased stochastic tracer of the underlying dark matter distribution. Define the local galaxy over-density \( \delta_g(r) \) as

\[
\delta_g(r) = \frac{n(r) - \bar{n}}{\bar{n}},
\]  

where \( n(r) \) is the local galaxy number density, and \( \bar{n} \) is the average galaxy number density. Note that \( n(r) \) will be either 0 or \( \frac{1}{V} \), if the volume is chosen to be small enough. Then, if \( \delta_{DM}(r) \) is the underlying dark matter over-density, this model
asserts that the probability $p(r)$ of finding a galaxy at position $r$ will be

$$p(r) = \bar{n}V \left(1 + b_g \delta_{DM}(r)\right),$$

(2.32)

where we have defined $b_g$ as the linear galaxy bias. The probability accounts for unknown baryonic stellar physics. The galaxy power spectrum, $P_{gg}$, will look like the dark matter power spectrum, $P_{DM}$, but it will me multiplied by the square of the linear galaxy bias:

$$P_{gg} = b_g^2 P_{DM}.$$  

(2.33)

The 21-cm line of neutral hydrogen (HI) provides a different way to trace matter. Hydrogen composes roughly 75% of the baryonic matter in the Universe, so neutral hydrogen will be an abundant tracer, even at late times when the neutral fraction is $\sim 1\%$. The 21-cm line of neutral hydrogen is the longest wavelength strong transition in the spectrum of neutral hydrogen [62]. Line confusion is not expected to be an issue, so that emission lines with wavelengths longer than 21-cm can be unambiguously identified as redshifted 21-cm emission. Extinction is also not an issue, since the size of cosmic dust is much smaller than the 21-cm wavelength, which lies in the radio spectrum. The HIPASS and ALFALFA surveys [52, 47] have already used the 21-cm emission line to survey low-redshift galaxies with large HI masses, using the Parkes and Arecibo radio telescopes. Such surveys that identify individual galaxies share the problem with optical galaxy surveys that the signal falls off as the distance squared, and therefore they cannot extend to high redshifts.

In this thesis, we advocate the technique of 21-cm intensity mapping, which does not attempt to identify individual galaxies but instead collects the emission from many galaxies that occupy large voxels. Since individual galaxies do not need to be detected, intensity mapping surveys can efficiently make maps of LSS out to very high redshift [45, 93, 63, 16]. In principle, 21-cm intensity mapping can in fact map most of the volume of the Universe, even extending into the dark ages, where traditional
Figure 2-4: From [89]. The blue circle shows the comoving volume accessible in principle, using the 21-cm transition of neutral hydrogen. The red area depicts the galaxies detected by the Sloan Digital Sky Survey. The thick solid black line shows the surface of last scattering of the CMB, with the thickness shown to-scale. This solid black line is the farthest observable distance. The infinite redshift circle represents the outer radius that would be observable if the Universe were transparent over its full history.

Galaxy surveys are not possible because stars have not yet formed. Figure 2-4, from [89], shows a slice of the comoving volume of the Universe accessible by measuring the 21-cm line.

Although 21-cm intensity mapping can extend to high redshifts, it has a few complications which make it difficult in practice. First, since individual objects are not detected, the size of the local over-density is not computed by the number density of objects but by the brightness temperature of the 21-cm emission. There will be noise on this measurement for any real radio telescope. It is therefore important to
observe for long periods of time, with multiple receivers, and with as little extra noise added from the receivers as possible. To reach a signal-to-noise ratio greater than 1 on each voxel of a 21-cm intensity map will require dedicated instruments observing for long periods of time. The path-finding projects in this thesis, for instance, all have a signal-to-noise ratio less than 1 on each voxel. A more fundamental challenge is the existence of spectrally smooth free-free and synchrotron foregrounds at $\sim 10^4$ times the level of the signal. In principle, these can be disentangled from the 21-cm signal at all but the longest wave-lengths, since the 21-cm signal will follow the clumpy distribution of matter in the redshift direction. However, physical effects like imperfect bandpass calibration, frequency-dependent beams, and polarization leakage make foreground removal difficult in practice. Lastly, the 21-cm power spectrum, $P_{TT}$, will be normalized by both an HI bias and a mean brightness temperature $T$, which depends on the average HI density of the Universe:

$$P_{TT} = T^2 b_{HI}^2 P_{DM}. \quad (2.34)$$

It is extremely difficult to measure the mean brightness temperature $T$ of the 21-cm signal, which depends on the density of HI. This means that it will be difficult in practice to disentangle the redshift evolution of the growth factor from the evolution of the comoving HI fraction and HI bias. However, it should be possible to disentangle these effects if one can accurately measure redshift space distortions (RSDs). This is discussed further in section 3.5.1.
Chapter 3

The 21-cm signal

3.1 Hyperfine Transition

In this section, I briefly review the hyperfine splitting of the $n = 1$ hydrogen energy level. The focus will be to derive the wavelength and the transition amplitude. From these, I will calculate the Einstein A and B coefficients. These will in turn allow me to calculate the expected magnitude of the 21-cm signal from radiative transfer arguments in the next section.

The nuclear spin $\vec{S}_p$ of the proton produces a magnetic moment

$$\vec{\mu}_p = \frac{g_p e}{2m_p} \vec{S}_p$$

(3.1)

where $g_p \approx 5.59$ is the g-factor of the proton. The proton’s magnetic moment will generate a magnetic field

$$\vec{B} = \frac{\mu_0}{4\pi} \left[ \frac{3\vec{r} (\vec{r} \cdot \vec{\mu}_p) - r^2 \vec{\mu}_p}{r^5} + \frac{8\pi}{3} \vec{\mu}_p \delta^D (\vec{r}) \right]$$

(3.2)

The delta function describes the extremely high field in the center of a dipole when it is compressed down to infinitesimal size. The magnetic field interacts with the effective magnetic moment of the electron, which has contributions from both the
orbital and spin angular momentum, and is given by

\[ \vec{\mu}_e = -\frac{e}{2m_e} \vec{L} - \frac{g_e e}{2m_e} \vec{S}_e \]  

(3.3)

where \( g_e \approx 2.002 \) is the very precisely known g-factor of the electron. The perturbing spin interaction potential energy of \(-\vec{\mu}_e \cdot \vec{B}\) is therefore (using \( \vec{\mu}_p = \frac{g_p e}{2m_p} \vec{S}_p \))

\[
V = \frac{\mu_0 g_p e^2}{16\pi m_e m_p} \frac{3}{r^5} \left( \vec{\hat{r}} \cdot \vec{S}_p \right) \vec{\hat{r}} \cdot \left( \vec{L} + g_e \vec{S}_e \right) - r^2 \vec{S}_p \cdot \left( \vec{\hat{L}} + g_e \vec{S}_e \right)
\]

\[+ \frac{\mu_0 g_p e^2}{6m_e m_p a_0^3} \vec{S}_p \cdot \left( \vec{L} + g_e \vec{S}_e \right) \delta^D (\vec{\hat{r}}) \]  

(3.4)

The zeroth order perturbation to the energy level of \(|\Psi_n\rangle\) is given by the amplitude \(\langle \Psi_n | V | \Psi_n \rangle\). For the \(n = 1\) state, the orbital angular momentum is zero, and \(\langle \vec{\hat{x}} | \Psi_0 \rangle = e^{-1/2} e^{-3/2} e^{-r/a_0}\), where \(a_0\) is the Bohr radius. The energy correction to the \(n = 1\) state will thus be given by the integral

\[
\Delta E_{n=1} = \frac{\mu_0 g_e g_p e^2}{16\pi^2 m_e m_p a_0^3} \int \int \int \int r^{-2} e^{-r/a_0} \left[ 3 \left( \vec{\hat{r}} \cdot \vec{S}_p \right) \left( \vec{\hat{r}} \cdot \vec{S}_e \right) - \vec{S}_p \cdot \vec{S}_e \right] dr \sin(\theta) \, d\theta \, d\phi
\]

\[+ \frac{\mu_0 g_e g_p e^2}{6\pi m_e m_p a_0^3} \int \int \int e^{-2r/a_0} \vec{S}_p \cdot \vec{S}_e \delta^D (\vec{\hat{r}}) \, r^2 dr \sin(\theta) \, d\theta \, d\phi \]  

(3.5)

One can show that the angular portion of the integral in the first term of equation 3.5 is exactly zero due to symmetry and cancellations of terms. Although the radial integral diverges, this occurs only due to the portion near \(r \sim 0\), where the energy should really only be determined by the second term, with the delta function. It is therefore argued that the first term is unambiguously zero, and the only contribution to the energy splitting comes from the second term of equation 3.5. This term is easy to calculate due to the delta function.

\[
\Delta E_{n=1} = \frac{\mu_0 g_e g_p e^2}{6\pi m_e m_p a_0^3} \vec{S}_p \cdot \vec{S}_e 
\]  

(3.6)
If we define $\vec{S}$ as the total spin angular momentum of the atom $\vec{S} = \vec{S}_p + \vec{S}_e$, then we can then use the standard trick, $\vec{S}_p \cdot \vec{S}_e = \frac{1}{2} \left[ S^2 - S_p^2 - S_e^2 \right]$, to express the energy perturbation as
\[
\Delta E_{n=1} = \frac{\mu_0 g_e g_p e^2}{6\pi m_e m_p a_0^3} \frac{1}{2} \left[ S^2 - S_p^2 - S_e^2 \right] \tag{3.7}
\]
For the Hydrogen case, where both the proton and the electron have spin $1/2$, the total atomic spin will either be in the triplet $s = 1$ or the singlet $s = 0$ state. Using the fact that $\hat{S}^2 |S, S_z\rangle = \hbar^2 s(s + 1) |S, S_z\rangle$, we see that
\[
\begin{align*}
  s = 1 \quad & \hat{S}_p \cdot \hat{S}_e = \hbar^2 \left( 1 - \frac{3}{4} \right) \\
  s = 0 \quad & \hat{S}_p \cdot \hat{S}_e = -\hbar^2 \frac{3}{4}
\end{align*}
\]
The spin 1 and zero states thus have energy perturbations
\[
\begin{align*}
  s = 1 \quad & \Delta E_{n=1} = \frac{\mu_0 g_e g_p e^2 \hbar^2}{24\pi m_e m_p a_0^3} \\
  s = 0 \quad & \Delta E_{n=1} = -3\frac{\mu_0 g_e g_p e^2 \hbar^2}{24\pi m_e m_p a_0^3}
\end{align*}
\]
The energy difference between the two spin states is
\[
E_{10} = \frac{\mu_0 g_e g_p e^2 \hbar^2}{6\pi m_e m_p a_0^3} = \frac{2\alpha^4 m_e^2 c^2 g_e g_p}{3m_p} \tag{3.8}
\]
Plugging in the constants, one finds
\[
\begin{align*}
  E_{10} &= 5.8743 \mu eV, \\
  \nu_{10} &= 1.4204 \text{ MHz}, \\
  \lambda_{10} &= 21.106 \text{ cm}, \\
  T_* &= \frac{E_{10}}{k} \approx 68 \text{ mK}.
\end{align*}
\tag{3.9}
\]
The energy gap is very small, resulting in a long wavelength transition in the radio spectrum. On the last line, we have followed Field [29] in defining $T_*$ as the equivalent
temperature of the 21-cm transition. In practically all cases the spin temperature \( T_s \), defined via the Boltzmann factor \( \frac{N_1}{N_0} = \frac{g_1 e^{-T_s}}{g_0} \), will be much higher than \( T_* \). Therefore, \( \frac{N_1}{N_0} \approx \frac{g_1}{g_0} \left( 1 - \frac{T_s}{T_*} \right) \). This fact is critical in making HI a three dimensional tracer of matter. Stimulated emission nearly cancels the effect of absorption, and this keeps the HI optically thin.

In order to know the magnitude of the 21-cm signal, we must know the absorption and emission coefficients. So we now turn to calculating the Einstein A and B coefficients. We will first calculate the absorption rate \( B_{01} \) under the influence of a perturbing radiation field at the resonant frequency \( \nu_{10} \), using time dependent perturbation theory. Following Einstein, we will use detailed balance arguments to derive the spontaneous emission \( (A_{10}) \) and stimulated emission \( (B_{10}) \) rates from the absorption rate. One would expect that, because the spin up and spin down states with \( l = 0 \) both have even parity, the 21-cm transition should be electric dipole forbidden. One might also expect that the spins should naturally interact with the magnetic field. The following analysis will indeed find that the magnetic dipole term is the highest order non-vanishing transition matrix element, and it dominates the absorption rate.

Using minimal coupling to account for the Lorentz force, the perturbed hydrogen Hamiltonian can be written as

\[
H = \frac{1}{2m_e} \left( \vec{p} + e\vec{A} \right)^2 - e\phi + \Phi(r) \tag{3.10}
\]

where \( \Phi(r) \) is the atomic potential, including the electric potential due to the proton and all fine and hyperfine corrections. The perturbing field is represented by the vector and scalar potentials \( \vec{A} \) and \( \phi \). Expanding this,

\[
H = \frac{1}{2m_e} \left( p^2 + e\vec{A} \cdot \vec{p} + e\vec{p} \cdot \vec{A} + e^2 A^2 \right) - e\phi + \Phi(r) \tag{3.11}
\]

This expression can be simplified by choosing to represent the perturbing field in the
Coulomb gauge ($\nabla \cdot A = 0$), so that $\vec{A}$ and $\vec{p}$ commute, and by dropping the $\frac{e^2}{2m_e}A^2$ term, which is second order in the small perturbing potential. Then, noting that $\phi = 0$ for a plane wave perturbation, we have

\[
H = H_0 + H_1
\]

\[
H_0 = \frac{p^2}{2m_e} + \Phi(r)
\]

\[
H_1 = \frac{e\vec{A} \cdot \vec{p}}{m_e}
\]

(3.12)

Here $H_0$ is the full hydrogen Hamiltonian, and the $H_1$ term accounts for interactions of the point charges with a small perturbing radiation field. However, $H_1$ as written is incomplete because it ignores the energy of the atomic spin coupling with the radiation field. We will have to add this in by hand, with a $H_s = -\vec{\mu} \cdot \vec{B}$ term, which we will work out later. First, let us specify the perturbing potential. In the Coulomb gauge, the electric and magnetic fields are related to the potential via

\[
\vec{E} = -\frac{\partial \vec{A}}{\partial t} - \nabla \phi
\]

\[
\vec{B} = \nabla \times \vec{A}
\]

(3.13)

We can represent a plane wave with electric field polarization $\hat{\epsilon}$ and propagation direction $\hat{n}$ by setting $\phi = 0$ and

\[
\vec{A} = 2A_0 \hat{\epsilon} \cos \left( \frac{\omega}{c} \hat{n} \cdot \vec{x} - \omega t \right) = A_0 \hat{\epsilon} \left[ e^{i \left( \frac{\omega}{c} \hat{n} \cdot \vec{x} - \omega t \right)} + e^{-i \left( \frac{\omega}{c} \hat{n} \cdot \vec{x} - \omega t \right)} \right]
\]

(3.14)

Plugging this into equation 3.13, one finds

\[
\vec{E} = 2\omega A_0 \hat{\epsilon} \sin \left( \frac{\omega}{c} \hat{n} \cdot \vec{x} - \omega t \right)
\]

\[
\vec{B} = 2\frac{\omega}{c} A_0 (\hat{n} \times \hat{\epsilon}) \sin \left( \frac{\omega}{c} \hat{n} \cdot \vec{x} - \omega t \right)
\]

(3.15)

which perfectly represents a plane wave with electric field amplitude $E_0 = 2\omega A_0$ and
Using equations 3.12 and 3.14, the time-dependent perturbing Hamiltonian can be written as

\[ H_1(t) = \frac{eA_0 \hat{\epsilon} \cdot \hat{p}}{m_e} \left[ e^{i(\hat{\epsilon} \cdot \hat{n} \cdot \vec{x} - \omega t)} + e^{-i(\hat{\epsilon} \cdot \hat{n} \cdot \vec{x} - \omega t)} \right] + H_s \]  

(3.16)

The 21-cm perturbing wavelength is much larger than the extent of the electron’s wavefunction, so we shall expand the spatial exponential terms. Keeping only the terms up to linear order in \( \frac{\omega}{c} \),

\[ H_1(t) = \frac{eA_0 \hat{\epsilon} \cdot \hat{p}}{m_e} \left[ \left( 1 + i \frac{\omega}{c} \hat{n} \cdot \vec{x} \right) e^{-i \omega t} + \left( 1 - i \frac{\omega}{c} \hat{n} \cdot \vec{x} \right) e^{i \omega t} \right] + H_s \]  

(3.17)

The zeroth order term is known as the electric dipole term. Labeling the electric dipole term \( H_{ed} \), we can write the perturbing Hamiltonian as

\[ H_1(t) = H_{ed} + \frac{i \omega}{c} \frac{eA_0}{m_e} (\hat{\epsilon} \cdot \hat{p}) (\hat{n} \cdot \vec{x}) \left[ e^{-i \omega t} - e^{i \omega t} \right] + H_s \]  

(3.18)

Note that \( \hat{\epsilon} \cdot \hat{p} \) and \( \hat{n} \cdot \vec{x} \) commute due to the orthogonality of \( \hat{\epsilon} \) and \( \hat{n} \). Using the vector identity \((\vec{a} \times \vec{b}) \cdot (\vec{c} \times \vec{d}) = (\vec{a} \cdot \vec{c})(\vec{b} \cdot \vec{d}) - (\vec{b} \cdot \vec{c})(\vec{a} \cdot \vec{d})\), one finds that

\[ (\hat{\epsilon} \cdot \hat{p})(\hat{n} \cdot \vec{x}) = (\hat{n} \times \hat{\epsilon}) \cdot (\vec{x} \times \hat{p}) + (\hat{\epsilon} \cdot \vec{x})(\hat{n} \cdot \hat{p}) \]  

(3.19)

Note that, since the magnetic field direction \( \hat{b} \) is \( \hat{n} \times \hat{\epsilon} \), the first term looks like \( \hat{b} \cdot \vec{L} \).

One can also show, for the basic unperturbed Hydrogen Hamiltonian \( H = \frac{\vec{p}^2}{2m} + \Phi(\vec{x}) \), that

\[ (\hat{\epsilon} \cdot \hat{p})(\hat{n} \cdot \vec{x}) = \frac{im_e}{\hbar} [H, (\hat{\epsilon} \cdot \vec{r})(\hat{n} \cdot \vec{x})] - (\hat{\epsilon} \cdot \vec{x})(\hat{n} \cdot \hat{p}) \]  

(3.20)

This follows from the canonical commutation relation and the fact that \([x_i, H] = ih\frac{\partial H}{\partial p_i}\). Combining equations 3.19 and 3.20,

\[ (\hat{\epsilon} \cdot \hat{p})(\hat{n} \cdot \vec{x}) = \frac{1}{2} \hat{b} \cdot \vec{L} + \frac{im_e}{2\hbar} [H, (\hat{\epsilon} \cdot \vec{r})(\hat{n} \cdot \vec{x})] \]  

(3.21)
The second term represents the traceless electric quadrupole moment, which we shall label $H_{eq}$.

$$H_1(t) = H_{ed} + \frac{i\omega}{2c} \frac{eA_0}{m_e} \hat{b} \cdot \vec{L} \left[ e^{-i\omega t} - e^{i\omega t} \right] + H_{eq} + H_s$$  \hspace{1cm} (3.22)

Let us now consider the contribution to the perturbing Hamiltonian from the interaction of the atom’s spin with the magnetic field of the radiation. This term, including both the electron and proton spin contribution, will be of the form $H_s = -\mu_e \cdot \vec{B} - \mu_p \cdot \vec{B}$. From equation 3.15,

$$H_s = \frac{i\omega}{c} A_0 (\mu_e + \mu_e) \cdot (\hat{n} \times \hat{\epsilon}) \left[ e^{i(\omega \hat{n} \cdot \vec{x} - \omega t)} - e^{-i(\omega \hat{n} \cdot \vec{x} - \omega t)} \right]$$  \hspace{1cm} (3.23)

Again, we can expand the spatial component of the exponentials. There is already one factor of $\frac{\omega}{c}$ in the prefactor, so to linear order in $\frac{\omega}{c}$,

$$H_s = \frac{i\omega}{c} A_0 (\mu_e + \mu_e) \cdot (\hat{n} \times \hat{\epsilon}) \left[ e^{i(\omega \hat{n} \cdot \vec{x} - \omega t)} - e^{-i(\omega \hat{n} \cdot \vec{x} - \omega t)} \right]$$  \hspace{1cm} (3.24)

where we have identified the unit vector $\hat{n} \times \hat{\epsilon}$ as the direction of the magnetic field, labeled as $\hat{b}$ . Our full perturbing Hamiltonian, to linear order in $\frac{\omega}{c}$, is

$$H_1(t) = H_{ed} + H_{eq} + \frac{\omega}{c} \frac{eA_0}{m_e} \left( \hat{b} \cdot \vec{L} - g_e S_e \cdot \hat{b} + \frac{g_p m_e}{m_p} S_p \cdot \hat{b} \right) \left[ e^{-i\omega t} - e^{i\omega t} \right]$$  \hspace{1cm} (3.25)

where we have utilized the fact that $\sin(\theta) = \frac{i}{2} [e^{-i\theta} - e^{i\theta}]$. In the interaction picture, the perturbing time-dependent potential causes the eigenstates to mix according to

$$i\hbar \frac{dc_n(t)}{dt} = \sum_m \langle n | H_1(t) | m \rangle e^{i\omega_{nm}(t-t_0)} c_m(t)$$  \hspace{1cm} (3.26)

where $\omega_{nm} = \frac{E_n - E_m}{\hbar}$. This provides a set of coupled differential equations between the amplitudes of eigenstates of the original Hamiltonian. We are interested in how the perturbing radiation mixes the $|n_1, s_0\rangle$ and $|n_1, s_1\rangle$ states.
In fact, the electric dipole, electric quadrupole, and the $\hat{b} \cdot \vec{L}$ term all contribute zero to the transition amplitude $\langle n_1, s_0 | H_1(t) | n_1, s_1 \rangle$. The arguments are as follows.

- For the electric dipole term, $\langle n_1, s_0 | H_{ed}(t) | n_1, s_1 \rangle \propto \langle n_1, s_0 | \vec{\epsilon} \cdot \vec{p} | n_1, s_1 \rangle \propto \int \epsilon_i r_i e^{-2r/a_0} dV = 0$. The odd parity of the $\epsilon_i r_i$ term ensures that the angular integral between any two states of equal angular momentum is zero.

- The $\hat{b} \cdot \vec{L}$ term is zero since $\langle n_1, s_0 | \vec{L} | n_1, s_1 \rangle = 0$; both states have zero angular momentum.

- Finally, the electric quadrupole term is zero since $\langle n_1, s_0 | [H, (\hat{\epsilon} \cdot \vec{x})(\hat{n} \cdot \vec{x})] | n_1, s_1 \rangle \propto E_{10} e \int x_i x_j e^{-2r/a_0} dV = 0$. The integral is zero due to the spherical symmetry of the wavefunction and the fact that $\hat{\epsilon}$ is perpendicular to $\hat{n}$.

Therefore, the critical matrix elements for this transition will be

$$\langle m | H_1(t) | n \rangle = \frac{\omega e A_0}{c m} \langle m | -g_e \vec{S}_e \cdot \hat{b} + \frac{g_p m_p}{m} \vec{S}_p \cdot \hat{b} | n \rangle \sin(\omega t) \quad (3.27)$$

The matrix element from the spin zero state $|0, 0\rangle$ to each of the triplet spin 1 states, $|1, 1\rangle$, $|1, -1\rangle$, and $|1, 0\rangle$ must be calculated. To do so, we recall that the total spin states $|S, S_z\rangle$ can be written in terms of the z-projection of the electron and proton spin states $|s_{ze}, s_{zp}\rangle$ as follows.

$$
\begin{align*}
|1, 1\rangle &= |\frac{1}{2}, \frac{1}{2}\rangle \\
|1, 0\rangle &= \frac{1}{\sqrt{2}} \left(|\frac{1}{2}, \frac{1}{2}\rangle + |\frac{1}{2}, -\frac{1}{2}\rangle\right) \\
|1, -1\rangle &= |\frac{-1}{2}, \frac{1}{2}\rangle \\
|0, 0\rangle &= \frac{1}{\sqrt{2}} \left(|\frac{-1}{2}, \frac{1}{2}\rangle - |\frac{1}{2}, -\frac{1}{2}\rangle\right)
\end{align*}
$$
Utilizing the Pauli matrices for the electron and proton spins, one finds that

\[ \langle 1, 1 | H_1(t) | 0, 0 \rangle = -\frac{\omega e A_0}{c m_e} \frac{\hbar}{2\sqrt{2}} \left( g_e + \frac{g_p m_e}{m_p} \right) \hat{b} \cdot (\hat{x} + i\hat{y}) \sin(\omega t) \]

\[ \langle 1, -1 | H_1(t) | 0, 0 \rangle = \frac{\omega e A_0}{c m_e} \frac{\hbar}{2\sqrt{2}} \left( g_e + \frac{g_p m_e}{m_p} \right) \hat{b} \cdot (\hat{x} - i\hat{y}) \sin(\omega t) \]

\[ \langle 1, 0 | H_1(t) | 0, 0 \rangle = \frac{\omega e A_0 \hbar}{c 2m_e} \left( g_e + \frac{g_p m_e}{m_p} \right) \hat{b} \cdot \hat{z} \sin(\omega t) \]  

(3.28)

We can now calculate the absorption probability. Consider an HI atom, in the \( n=1 \) energy level, that starts in the \( |0, 0\rangle \) singlet spin state. At time \( t=0 \), a plane wave of radiation, at frequency \( \omega \), turns on. To first order in perturbation theory, the amplitude of the \( |1, 1\rangle \) spin state is given by

\[ c_{1,1}^{(1)}(t) = \int_0^t \frac{-i}{\hbar} H_{1,1,0,0} \sin(\omega t') e^{i\omega t'} dt', \]

\[ c_{1,1}^{(1)}(t) = \frac{H_{1,1,0,0}}{2\hbar} \left[ \frac{1 - e^{i(\omega_{10}+\omega)t}}{\omega_{10} + \omega} + \frac{1 - e^{i(\omega_{10}-\omega)t}}{\omega_{10} - \omega} \right]. \]  

(3.29)

Here I have split the time dependent component, \( \sin(\omega t) \), out of the definition of the transition amplitude \( H_{1,1,0,0} \). So, \( H_{1,1,0,0} = \frac{\omega e A_0}{c m_e} \frac{\hbar}{2\sqrt{2}} \left( g_e - \frac{g_p m_e}{m_p} \right) \hat{b} \cdot (\hat{x} - i\hat{y}) \). The first term in \( c_{1,1}^{(1)}(t) \) has a resonance when the frequency of the radiation is equal to \( \frac{E_f - E_i}{\hbar} \), and so it is relevant only for stimulated emission. Since we are calculating the probability of going to a higher energy state, we are interested in the resonance of the second term, which is responsible for absorption. This term is resonant as \( \omega \) approaches \( \omega_{10} \). As \( \omega \) approaches \( \omega_{10} \), only the second term in equation 3.29 is significant.

\[ c_{1,1}^{(1)}(t) \approx \frac{H_{1,1,0,0}}{2\hbar} \left[ \frac{1 - e^{i(\omega_{10}-\omega)t}}{\omega_{10} - \omega} \right]. \]  

(3.30)

The probability of finding the atom in the \( |1, 1\rangle \) state is given by

\[ P_{1,1}^{(1)}(t) = \frac{|H_{1,1,0,0}|^2 t^2}{2\hbar^2} \sin^2 \left( \frac{(\omega_{10} - \omega)}{2} t \right). \]  

(3.31)
In order to complete the derivation of Fermi’s Golden rule for radiation mediated atomic transitions, we must account for the fact that equation 3.31 unrealistically assumes that the incoming radiation is a perfectly monochromatic plane wave. All natural radiation has a finite bandwidth, so let us instead consider the case of incident radiation with a finite but narrow bandwidth and a constant energy density per unit bandwidth \( u \), related to the electric field and vector potential amplitudes (making use of equation 3.15) as follows

\[
U = \int_{\omega - \Delta \omega}^{\omega + \Delta \omega} \frac{u}{2\pi} d\omega = \frac{\epsilon_0}{2} E_0^2 = 2\epsilon_0 \omega^2 A_0^2 \bigg|_{\omega = \omega_{10}}.
\]

Equation 3.31 must now be replaced by an integral, effectively counting the probability of an atomic transition from all frequencies near resonance.

\[
P_{1,1}^{(1)}(t) = \frac{|H_{1,1,0,0}|^2 t^2}{4h^2} \frac{u}{4\pi\epsilon_0 \omega_{10}^2 A_0^2} \int_{\omega - \Delta \omega}^{\omega + \Delta \omega} \text{sinc}^2 \left( \frac{\left( \frac{\omega_{10} - \omega}{2} \right) t}{\Delta \omega} \right) d\omega
\]

(3.33)

This integral is performed over probability, not amplitude, because the different frequencies are assumed to have uncorrelated phases. Although the bandwidth of the resonance is very narrow, in the long time limit the integral covers nearly all of the area under the squared sinc curve, and the integral converges to \( \frac{2\pi}{\tau} \). Therefore,

\[
P_{1,1}^{(1)}(t) = \frac{|H_{1,1,0,0}|^2 ut}{8\epsilon_0 \omega_{10}^2 A_0^2 h^2}.
\]

(3.34)

Plugging in the matrix element \( H_{1,1,0,0} \), one finds

\[
P_{1,1}^{(1)}(t) = \frac{e^2 ut}{64\epsilon_0 c^2 m_e^2} \left( g_e + \frac{g_p m_e}{m_p} \right)^2 \left( \hat{b} \cdot \hat{x} \right)^2 + (\hat{b} \cdot \hat{y})^2.
\]

(3.35)

Similarly,

\[
P_{1,-1}^{(1)}(t) = \frac{e^2 ut}{64\epsilon_0 c^2 m_e^2} \left( g_e + \frac{g_p m_e}{m_p} \right)^2 \left( \hat{b} \cdot \hat{x} \right)^2 + (\hat{b} \cdot \hat{y})^2.
\]

(3.36)
and

\[ P_{1,0}^{(1)}(t) = \frac{e^2 u t}{32 \epsilon_0 c^2 m_e^2} \left( g_e + \frac{g_p m_e}{m_p} \right)^2 (\hat{b} \cdot \hat{z})^2. \]  

(3.37)

Let \( \hat{n} \) be the direction of propagation of the radiation field. Since we are considering the atom to initially be spin zero, we are free to choose the orientation of the z-axis to be in the \( \hat{n} \) direction. The direction \( \hat{b} \) of the magnetic field must be perpendicular to the propagation direction, but otherwise we are free to choose that direction as well. Let us choose \( \hat{b} = \hat{x} \). Then, let us define \( P_{1}^{(1)}(t) \) as the first order probability that a spin 0 HI atom has gone to any of the spin 1 states because of absorption of radiation.

\[ P_{1}^{(1)}(t) = P_{1,1}^{(1)}(t) + P_{1,-1}^{(1)}(t) = \frac{e^2 u(\nu_{10}) t}{32 \epsilon_0 c^2 m_e^2} \left( g_e + \frac{g_p m_e}{m_p} \right)^2. \]  

(3.38)

Note that \( P_{1}^{(1)}(t) \) is independent of the direction of the incoming radiation. The time derivative of \( P_{1}^{(1)}(t) \) yields the time rate of increase of probability of the atom going spin 0 to spin 1 due to the incoming radiation. Dividing this derivative by the radiation energy density per unit frequency yields the Einstein absorption coefficient \( B_{01} \).

\[ B_{01} = \frac{1}{u(\nu_{10})} \frac{dP_{1}^{(1)}(t)}{dt} = \frac{e^2}{32 \epsilon_0 c^2 m_e^2} \left( g_e + \frac{g_p m_e}{m_p} \right)^2. \]  

(3.39)

Now, we can use the principle of detailed balance to derive the spontaneous emission coefficient \( A_{10} \) and the stimulated emission coefficient \( B_{10} \). If we imagine a group of HI atoms in thermal equilibrium with a blackbody cavity, both at temperature \( T \), then the net rate at which the atoms change their energy level must be zero. Inside the thermal enclosure, the blackbody will emit radiation with specific intensity given by the Planck spectrum. Integrating over \( 4\pi \) steradians and dividing by the speed of light converts this power per unit area per unit frequency per steradian to an energy density per unit frequency per unit frequency: \( u(\nu) = \frac{4\pi}{c} B_{\nu}(T) = \frac{8\pi \hbar \nu^3}{e^2} \left[ e^{\frac{\hbar \nu}{kT}} - 1 \right]^{-1} \). Detailed balance
requires that

\[ N_1A_{10} = N_0u(\nu)B_{01} - N_1u(\nu)B_{10}. \]  

(3.40)

At equilibrium, the relative number of atoms in the two states is given by the Boltzmann factor:

\[ \frac{N_1}{N_0} = \frac{g_1}{g_0} e^{-\frac{E_{10}}{kT}} = 3e^{-\frac{E_{10}}{kT}}. \]  

(3.41)

So,

\[ A_{10} = \left[ \frac{g_0}{g_1} e^{\frac{E_{10}}{kT}} B_{01} - B_{10} \right] \frac{8\pi h\nu^3}{c^3} \left[ e^{\frac{h\nu}{kT}} - 1 \right]^{-1}. \]  

(3.42)

The only way for the above to be satisfied at \( h\nu = E_{10} \) and at all temperatures is if

\[ B_{10} = \frac{g_0}{g_1} B_{01}, \]

\[ A_{10} = \frac{8\pi h\nu^3}{c^3} B_{10}. \]  

(3.43)

Utilizing these equations and equation 3.39,

\[ A_{10} = \frac{\pi h e^2 \nu_{10}^3}{12\epsilon_0 c^5 m_e^2} \left( g_e + \frac{g_p m_e}{m_p} \right)^2. \]  

(3.44)

\[ 3.2 \text{ Radiative Transfer} \]

The unavoidable background radiation source everywhere in the universe is the blackbody spectrum of the Cosmic Microwave Background (CMB), whose current temperature is \( T_{CMB}(z = 0) \approx 2.7K \) and whose temperature in the past was \( T_{CMB}(z) = (1 + z)T_{CMB}(z = 0) \). For a clump of neutral hydrogen at redshift \( z \) to be detectable through its 21-cm line, its spin states must be out of thermal equilibrium with the CMB. This will be quantified by the spin temperature, \( T_s \), of the HI gas, and we will see that the HI contributes a net emission of 21-cm photons if \( T_s > T_{CMB}(z) \) or a net absorption of 21-cm photons if \( T_s < T_{CMB}(z) \). As the Universe expands, this emission or absorption feature will be redshifted to \( (z + 1) \times 21 \text{ cm} \). In principle,
the redshift allows us to make three dimensional tomographic maps of the HI in the Universe. Let us now derive how the intensity of CMB photons changes as it passes through regions of HI gas.

The observable that we can measure is the specific intensity, \( I_{\nu}(\nu, \hat{n}) \), which is the power per unit frequency per unit area per unit solid angle coming from direction \( \hat{n} \) and measured at frequency \( \nu \).

\[
I_{\nu}(\nu, \hat{n}) = \frac{dE}{d\nu dA d\Omega} \quad (3.45)
\]

A common unit for specific intensity at radio wavelengths is Janskys per steradian \( (1 \text{ Jy/sr} = 10^{-26} \text{ Wm}^{-2}\text{Hzsr}^{-1}) \). In the absence of scattering, emission, absorption, or gravitational redshift (due to Hubble expansion, for example), \( I_{\nu}(\nu, \hat{n}) \) is a conserved quantity. This is because the radiated power per unit area from a distant source falls off like \( 1/r^2 \), but the solid angle that the source subtends also falls off as \( 1/r^2 \), and these two terms cancel in the definition of \( I_{\nu}(\nu, \hat{n}) \). We will therefore set up a differential equation describing how \( I_{\nu}(\nu, \hat{n}) \) changes only as it passes through matter that can absorb and emit radiation. This is known as the radiative transfer equation.

\[
\frac{dI_{\nu}(\nu, \hat{n})}{ds} = -\kappa_{\nu}(\nu, n_{HI}, T_s) I_{\nu}(\nu, \hat{n}) + \epsilon_{\nu}(\nu, n_{HI}, T_s) \quad (3.46)
\]

where \( ds \) is the proper length of the differential path length element. Equation 3.46 must be integrated along the path from the source (CMB last scattering surface) to the observer (us). The first term accounts for absorption of CMB photons by the HI, where \( \kappa_{\nu} \) is the absorption coefficient, which depends on the local properties of the HI gas. We will see later that this term is proportional to the local HI density and inversely proportional to the local spin temperature. The second term is due to 21-cm emission, and \( \epsilon_{\nu} \) is the emissivity of the HI gas. There is no term for scattering, since the scattering cross section of photons with neutral hydrogen is essentially zero. Equation 3.46 also does not include energy loss of the photons due
to Hubble expansion. Hubble expansion causes the photons to redshift so that photons we observe at frequency $\nu$ that were emitted at redshift $z$ were originally emitted at a higher frequency, $\nu(1+z)$. Hubble expansion also decreases the density of the photons by a multiplicative factor of $(z+1)^{-3}$, and it decreases each photon’s energy like $(1+z)^{-1}$, but it also bunches the photons tighter in frequency space so that photons now seen in a bandwidth $\Delta\nu$ were emitted in a bandwidth $\Delta\nu(1+z)$. Combining all these factors, one finds that Hubble expansion alters the specific intensity like

$$I^o_\nu(\nu) = \frac{I^e_\nu(\nu(1+z))}{(1+z)^3},$$

where $I^o_\nu$ and $I^e_\nu$ are the observed and emitted specific intensities, respectively. If one plugs this in for a blackbody spectrum, one finds that the effective temperature of a blackbody spectrum falls off like $(1+z)^{-1}$ as the Universe expands. Therefore, we will change our quantity of interest from specific intensity to brightness temperature $T_b$, and then we just have to remember that $T_b$ falls of like $(1+z)^{-1}$.

Before continuing, one other simplification is necessary. At late redshifts, long after reionization, the HI clumps are expected to have a fairly uniform spin temperature, $T_s$, defined by the Boltzmann factor.

$$\frac{n_1}{n_0} = \frac{g_1}{g_0} e^{-\frac{E_{10}}{kT_s}} = 3 e^{-\frac{68mK}{T_s}}$$

Here, $n_1$ is the number density of HI in the spin 1 triplet state, $n_0$ is the number density of HI in the spin 0 state, $g_i$ are the statistical weights for the two spin states, and $E_{10}$ is the energy of the 21-cm spin flip transition. In temperature units, $E_{10}/k = T_* = 68mK$. The HI gas is in local thermodynamic equilibrium, and therefore Kirchhoff’s law of thermal radiation applies. The argument is as follows: if one imagines placing the HI gas into thermal contact with a blackbody cavity at the same temperature as the gas’s spin temperature, with a narrow band filter between the two systems, then by the second law of thermodynamics there must be no net exchange of energy between
the two systems. To achieve this, the ratio between the emission and absorption coefficient of the HI gas must be fixed, and that ratio must be the blackbody spectrum.

\[ \epsilon_\nu (\nu, n_{HI}, T_s) = \kappa_\nu(\nu, n_{HI}, T_s) B_\nu(T_s) \] (3.49)

With Kirchhoff’s Law, equation 3.46 becomes:

\[ \frac{dI_\nu}{\kappa_\nu(\nu, n_{HI}, T_s) ds} = -I_\nu + B_\nu(T_s) \] (3.50)

I have dropped the frequency and direction dependence of \( I_\nu \) for notational simplicity. As one final notational convenience, I define the dimensionless optical depth \( \tau_\nu \).

\[ \tau_\nu = \int_{\text{source}}^{\text{observer}} \kappa_\nu(\nu, n_{HI}, T_s) ds \]
\[ d\tau_\nu = -\kappa_\nu(\nu, n_{HI}, T_s) ds \] (3.51)

In the absence of emission, a signal at optical depth \( \tau_\nu \) would decrease due to absorption by a factor of \( e^{-\tau_\nu} \) before reaching the observer. The negative sign in the bottom equation of 3.51 indicates that the optical depth decreases going from source to observer. The radiative transfer equation can now be written as:

\[ -\frac{dI_\nu}{d\tau_\nu} = -I_\nu + B_\nu(T_s) \] (3.52)

This equation can be solved by multiplying both sides by \( e^{-\tau_\nu} d\tau_\nu \) and integrating by parts.

\[ \int_0^{\tau_\nu} e^{-\tau_\nu} I_\nu d\tau_\nu' + \int_0^{\tau_\nu} e^{-\tau_\nu} I_\nu d\tau_\nu' = \int_0^{\tau_\nu} [I_\nu - B_\nu(T_s)] e^{-\tau_\nu} d\tau_\nu' \]

\[ e^{-\tau_\nu} I_\nu \bigg|_0^{\tau_\nu} + \int_0^{\tau_\nu} e^{-\tau_\nu} I_\nu d\tau_\nu' - \int_0^{\tau_\nu} e^{-\tau_\nu} I_\nu d\tau_\nu' - B_\nu(T_s) \int_0^{\tau_\nu} e^{-\tau_\nu} d\tau_\nu' \]

\[ I_\nu e^{-\tau_\nu} - I_\nu = B_\nu(T_s) (e^{-\tau_\nu} - 1) \]

\[ I_\nu = I_{\nu\text{CMB}} e^{-\tau_\nu} + B_\nu(T_s) (1 - e^{-\tau_\nu}) \] (3.53)
In the second line, we have assumed a uniform spin temperature in the HI cloud, allowing us to take $B_\nu(T_s)$ outside the integral. Let us now recast the equation in terms of brightness temperature. A specific intensity $I_\nu$ has a brightness temperature, $T_b$, defined as the blackbody temperature that would produce an equivalent specific intensity: $I_\nu = B_\nu(T_b)$. In the Rayleigh-Jeans limit, $B_\nu = \frac{2\nu^2kT}{e^{\nu/kT}-1}$, so $T_b = \frac{\nu^2}{2k}$. In terms of brightness temperatures, equation 3.53 becomes

$$T_b = T_{CMB}(z)e^{-\tau_\nu} + T_S(1-e^{-\tau_\nu}) \quad (3.54)$$

This is the brightness temperature that would have been seen by an observer at redshift $z$ situated directly behind the cloud of HI, which is also at redshift $z$. But as the photons travel to us at redshift 0, the energy density of the photons continues to drop due to the expansion of the Universe. The apparent temperature of an expanding and redshifting blackbody spectrum falls off like $T(z = 0) = T(z)/(z + 1)$. Therefore,

$$T_b = \frac{T_{CMB}(z)}{z + 1} e^{-\tau_\nu} + \frac{T_S}{z + 1}(1-e^{-\tau_\nu}) \quad (3.55)$$

We must also be careful to evaluate the optical depth $\tau_\nu$ not at the original frequency $\nu$, but at the redshifted frequency $\nu(z + 1)$. The difference in brightness temperature from the unaltered CMB brightness temperature can be expressed in the very simple form

$$\Delta T_b = \frac{T_s - T_{CMB}(z)}{z + 1}(1-e^{-\tau_\nu}) \quad (3.56)$$

If one assumes that $\tau_\nu \ll 1$ (which we will later see is a good assumption), then

$$\Delta T_b \approx \frac{T_s - T_{CMB}(z)}{z + 1} - \tau_\nu \quad (3.57)$$

To complete this picture, we must compute the optical depth $\tau_\nu$, which we know from equation 3.51 is the absorption coefficient integrated along the proper length of the photons’ path. The important physical processes in determining $\kappa_\nu$ are the
absorption rate of 21-cm photons, determined by $B_{01}$, and the stimulated emission of 21-cm photons, determined by $B_{10}$. Unlike the more familiar optical cases, the 21-cm line is such a long wavelength that all theoretical spin temperature models have $kT_s \gg \hbar c/\lambda_{21}$. This means that for all the HI in the intergalactic medium, almost (but slightly fewer than) 3/4 will be in the higher energy spin triplet state. Therefore we must be careful to include stimulated emission, which will nearly cancel absorption and greatly diminish the absorption coefficient. If we consider a radiation field of specific intensity $I_\nu$, the rate of increase of spin 1 atoms per unit time per unit area per unit path length per steradian will be given by (we are ignoring spontaneous emission, since we are calculating the absorption coefficient only)

$$\frac{dN_1}{dtdAdsd\Omega} = \frac{n_{HI}I_\nu}{c} \left[ \frac{N_1}{N_0 + N_1} B_{10} - \frac{N_0}{N_0 + N_1} B_{01} \right]$$  \hspace{1cm} (3.58)

where $n_{HI}$ is the neutral hydrogen number density in proper coordinates. This increase in spin 1 HI atoms will result in a decrease in specific intensity, given by

$$\frac{dI}{ds} = -h\nu\phi(\nu) \frac{dN_1}{dtdAdsd\Omega}$$  \hspace{1cm} (3.59)

where $\phi(\nu)$ is the line profile of the 21-cm transition, normalized such that $\int \phi(\nu)d\nu = 1$. This equation, along with equation 3.43, allows us to identify the absorption coefficient and recast it in terms of the emission rate $A_{10}$ as

$$\kappa_\nu = \frac{\lambda_{21}^2}{8\pi} n_{HI} A_{10} \left[ \frac{N_1}{N_0 + N_1} - \frac{g_1}{g_0} \frac{N_0}{N_1 + N_0} \right] \phi(\nu)$$  \hspace{1cm} (3.60)

The fractions in the above equation can be filled in using the Boltzmann factor: $\frac{N_1}{N_0} = \frac{g_1}{g_0} e^{-\frac{E_{10}}{kT_s}}$, with $\frac{g_1}{g_0} = 3$ in our case. If we make the further reasonable assumption that $kT_s \gg E_{10}$, then the above expression can be Taylor expanded about $e^{-\frac{E_{10}}{kT_s}} \approx 1$
in the small parameter \(1 - e^{-\frac{E_{10}}{kT_s}}\). That expansion results in the expression

\[
\kappa_{\nu} = \frac{\lambda_{21}^2}{8\pi} \left( 3 \frac{A_{10}}{4} n_{HI} A_{10} (1 - e^{-\frac{E_{10}}{kT_s}}) \phi(\nu) \right)
\]

(3.61)

for \(\kappa_{\nu}\). The normalized line profile \(\phi(\nu)\) in principal will include natural line broadening from the uncertainty principle, temperature broadening, and pressure broadening. In all cases we expect that \(kT_s \gg E_{10}\), so that

\[
\kappa_{\nu} \approx \frac{\lambda_{21}^2}{8\pi} \left( 3 \frac{A_{10}}{4} E_{10} \frac{n_{HI} A_{10}}{kT_s} \phi(\nu) \right) = \frac{3A_{10}E_{10}\lambda_{21}^2 n_{HI}}{32\pi} \frac{n_{HI}}{kT_s} \phi(\nu)
\]

(3.62)

This must be integrated along the line of sight to obtain the optical depth.

\[
\tau_{\nu} = \int_{\text{source}}^{\text{observer}} \kappa_{\nu} ds \approx \int \frac{3A_{10}E_{10}\lambda_{21}^2 n_{HI}}{32\pi} \frac{n_{HI}}{kT_s} \phi(\nu) ds
\]

(3.63)

Let us consider the mean optical depth caused by the HI in the intergalactic medium (IGM), which we will assume is an isotropic distribution of HI. To calculate \(\tau_{\nu}\), we need to determine the range of proper lengths over which to integrate. If we assume that the IGM is following the Hubble flow exactly, then we can apply the Friedmann equation to the photon null geodesic:

\[
ds = adr = cdt = c \frac{a}{da/dt} \frac{da}{a} = - \frac{cdz}{(1 + z)H(z)}
\]

(3.64)

This allows us to replace the integral over proper distance with an integral over redshift.

\[
\tau_{\nu} \approx \int \frac{3cA_{10}E_{10}\lambda_{21}^2 n_{HI}}{32\pi} \frac{n_{HI}}{kT_s(1 + z)H(z)} \phi(\nu(z)) dz
\]

(3.65)

The notation on the line profile emphasizes the fact that we will need to change variables on the line profile from \(\nu\) to \(z\) to carry out the integral. As stated earlier, the line profile accounts for the intrinsic, pressure, and thermal broadening. If this broadening were the same at all proper distances, then the optical depth as a function
of redshift would be the convolution of \( \frac{n_{HI}}{T_s(1+z)H(z)} \) with the line profile. The real situation is a bit more complicated, since the line profile will in general be different at different proper distances. However, we will ignore all complications by assuming that all line profiles are narrow enough that \( \frac{n_{HI}}{T_s(1+z)H(z)} \) does not evolve significantly over the entire width of the profile. In that case, we can replace \( \phi(\nu) \) with \( \delta(\nu - \nu_{10}) \), where \( \nu_{10} \) is the rest frequency of the 21-cm line (we are calculating the 0 redshift case initially). Changing coordinates, \( \delta(\nu - \nu_{10}) = \delta(\nu_{10}z + 1 - \nu_{10}) = (z + 1)^2 \delta(z) \).

\[
\tau_\nu \approx \int \frac{3cA_{10}E_{10}\lambda_{21}^2 n_{HI}(z+1)}{32\pi \nu_{10} kT_S H(z)} \delta(z)dz
\]

\[
\tau_\nu \approx \frac{3A_{10}E_{10}\lambda_{21}^3 n_{HI}}{32\pi} \frac{H(z)}{kT_S H_0}
\] (3.66)

Now let us consider the redshift dependence of the terms in equation 3.66. Let us define the comoving HI density \( \Omega_{HI} \) as the HI energy density at redshift \( z \), in comoving coordinates, divided by the total energy density today. By this definition, \( n_{HI} = (1+z)^3\Omega_{HI} \frac{\rho_{tot}}{\rho_{HI}} = (1+z)^3\Omega_{HI} \frac{3H_0^2}{8\pi G m_H} \), where \( \rho_{tot} \) is the mass needed to supply the current total energy density of the Universe and \( m_H \) is the Hydrogen mass. The quantities determining the Hubble parameter also evolve with redshift. The energy in radiation is negligible for all redshifts of interest, so \( H(z) = H_0 [\Omega_m (1+z)^3 + \Omega_\Lambda]^{1/2} \).

Combining these, we have

\[
\tau_\nu \approx \frac{9}{256\pi^2} \frac{H_0 A_{10} \lambda_{21}^3 E_{10}}{G m_H} \frac{\Omega_{HI}(1+z)^{3/2}}{kT_S [\Omega_m + \Omega_\Lambda(1+z)^{-3}]^{1/2}}
\] (3.67)

All the constants are well known except for \( H_0 \) and the spin temperature \( T_s \). Keeping three significant figures of precision, \( G = 6.67 \cdot 10^{-11} m^3 kg^{-1}s^{-2} \) and \( m_H = 1.67 \cdot 10^{-27} kg \). The decay rate of the 21-cm line, \( A_{10} \) is \( 2.85 \cdot 10^{-15} s^{-1} \), and \( \lambda_{21} \approx 0.211 m \).

We will use the usual parametrization of the less well known Hubble constant: \( H_0 = h \cdot 100 kmps^{-1} Mpc^{-1} \). In SI units, \( H_0 = h \cdot 3.24 \cdot 10^{-18} s^{-1} \). To three digits, \( E_{10} / k = 68.2 mK \).
Plugging in all these constants, we find

$$\tau_\nu \approx h \frac{189 mK}{T_s} \frac{\Omega_{HI}(1 + z)^{3/2}}{[\Omega_m + \Omega_\Lambda(1 + z)^{-3}]^{1/2}}$$

(3.68)

$\Omega_{HI}$ is much less than 1 due to the dominance of dark matter and dark energy and due to the high ionization fraction of hydrogen after reionization. All models for the spin temperature have $T_s \gg 189 mK$, and $h$ is known to be about 0.7. These facts mean that the $\tau_\nu \ll 1$ approximation is very well justified for the IGM. We can therefore plug equation 3.68 directly into equation 3.57 to derive the mean brightness temperature increase caused by the IGM.

$$\Delta T_b \approx 189 mK \frac{T_s - T_{CMB}(z)}{T_s} \frac{h\Omega_{HI}(1 + z)^{1/2}}{[\Omega_m + \Omega_\Lambda(1 + z)^{-3}]^{1/2}} \approx 189 mK \frac{h\Omega_{HI}(1 + z)^{1/2}}{[\Omega_m + \Omega_\Lambda(1 + z)^{-3}]^{1/2}}$$

(3.69)

### 3.3 HI Spin Temperature

The following analysis closely follows that of Field [29], supplemented by the comprehensive review of Furlanetto, Oh, and Briggs [31]. For a system in thermal equilibrium, the relative occupation of two differing energy states is a simple function of their energy difference, degeneracies, and the temperature of the system.

$$\frac{N_1}{N_0} = \frac{g_1}{g_0} e^{-\frac{\Delta E}{kT}}$$

(3.70)

However, even when a system is not in thermal equilibrium, one can still define a temperature using the same formula. For HI atoms in the spin 1 and spin 0 states, this is called the spin temperature $T_s$.

$$\frac{N_1}{N_0} = \frac{g_1}{g_0} e^{-\frac{\Delta E}{kT_s}} = 3e^{-\frac{T_s}{T_s}}$$

(3.71)
In a closed system in thermal equilibrium, the population of all states would be determined by a single temperature, and so the spin temperature \( T_s \) would be equal to \( T \), which would also control the relative populations of all other states. But in most astrophysical situations equilibrium is not achieved, and various processes, with different effective temperatures, can couple together in complicated ways.

The spin temperature of HI is generally driven by three competing factors: collisions couple \( T_s \) to the gas kinetic temperature \( T_K \), absorption and stimulated emission couple \( T_s \) to the temperature of the ambient 21-cm radiation \( T_R \) (mostly due to the CMB), and spin flipping high energy transitions (mainly the Lyman alpha transition) couple \( T_s \) to high energy radiation from stars, characterized by the Lyman temperature \( T_L \).

The general expression for the spin temperature can be derived from the basic idea that a stable system must have an equal rate of spin 0 to spin 1 transitions as it has spin 1 to spin 0 transitions. Writing the rate of spin 0 to 1 transitions on the left, and including the three competing factors mentioned above, we have

\[
n_0 \left( B_{01} u_R(\nu_{10}) + P_{01}^L + P_{01}^C \right) = n_1 \left( A_{10} + u_R(\nu_{10})B_{10} + P_{10}^L + P_{10}^C \right)
\]

where \( P_{01}^L \), for example, is the probability rate of that an HI atom in the spin 0 state will transition to a spin 1 state due to Lyman radiation. Some critical physics is being hidden by these definitions, but we shall elucidate it as we go forward. For now, let us use the definition of spin temperature from equation 3.71 to write \( n_1 \) as a function of \( n_0 \) and \( T_s \). We shall assume that \( T_s \gg T_* \) so that \( n_1 \approx 3n_0 \left( 1 - \frac{T_*}{T_s} \right) \). Plugging this in, and dividing both sides of eq. 3.72 by \( n_0 \), we find that

\[
\left( B_{01} u_R(\nu_{10}) + P_{01}^L + P_{01}^C \right) = 3 \left( 1 - \frac{T_*}{T_s} \right) \left( A_{10} + u_R(\nu_{10})B_{10} + P_{10}^L + P_{10}^C \right).
\]

If we use eq. 3.43 to rewrite all the Einstein coefficients in terms of \( A_{10} \), and the fact that \( u(\nu_{10}) = \frac{8\pi\hbar c}{e^2} \left[ e^{T_R/T_s} - 1 \right]^{-1} \approx \frac{8\pi\hbar c}{e^2} \frac{T_R}{T_s} \) (assuming that \( T_R \gg T_* \)), then 3.73 can
be rearranged to show

\[ 3 \left(1 - \frac{T_s}{T_s}\right) = \frac{3 \frac{T_R}{T_s} A_{10} + P_{01}^L + P_{01}^C}{A_{10} \left(1 + \frac{T_R}{T_s}\right)} + P_{10}^L + P_{10}^C \]  

(3.74)

We can simplify this equation by reducing the number of terms. First, let us examine \( P_{01}^C \), the probability rate that a spin 0 HI atom transitions to spin 1 due to collisions. For any colliding particle (HI or otherwise) at the same kinetic temperature as the HI, this rate ought to be proportional to the number density of the colliding particle times some function of the kinetic temperature \( T_k \). The rate \( P_{10}^C \) at which a spin 0 atom transitions to spin 1 due to collisions ought also to be a function of the number density times some function of the kinetic temperature. Therefore, the ratio of the two rates must be a function of kinetic temperature only, and since it depends only on kinetic temperature, we can derive it for the case of thermal equilibrium and generalize to all other cases. In thermal equilibrium, detailed balance and the Boltzmann factor mean that

\[ n_0 P_{01}^C = n_1 P_{10}^C = \frac{g_1}{g_0} P_{10}^C n_0 e^{-\frac{T_s}{T_k}} \approx 3 n_0 P_{10}^C \left(1 - \frac{T_s}{T_k}\right) \]  

(3.75)

Therefore, we have \( P_{01}^C = 3 P_{10}^C \left(1 - \frac{T_s}{T_k}\right) \). For Lyman radiation, one can also argue that there exists an effective Lyman temperature \( T_L \) such that \( P_{01}^L = 3 P_{10}^L \left(1 - \frac{T_s}{T_L}\right) \).

Using these, eq. 3.74 becomes

\[ 1 - \frac{T_s}{T_s} = \frac{\frac{T_R}{T_s} A_{10} + \left(1 - \frac{T_s}{T_L}\right) P_{10}^L + \left(1 - \frac{T_s}{T_k}\right) P_{10}^C}{A_{10} \left(1 + \frac{T_R}{T_s}\right) + P_{10}^L + P_{10}^C} \]  

(3.76)

Solving for \( T_s \), I find

\[ T_s = \frac{T_R + T_s + y_C T_k + y_L T_L}{1 + y_C + y_L} \approx \frac{T_R + y_C T_k + y_L T_L}{1 + y_C + y_L} \]  

(3.77)
where

\[ y_C = \frac{T_* P_{10}^C}{T_C A_{10}} \] (3.78)

and

\[ y_L = \frac{T_* P_{10}^L}{T_L A_{10}}. \] (3.79)

Now, let us turn to calculating the unknown transition rates \( P_{10}^C \) and \( P_{10}^L \). First, we will consider \( P_{10}^C \), the rate at which a spin 1 atom will transition to spin 0 due to particle collisions. The dominant particles present are expected to be neutral hydrogen (HI) atoms and ionized hydrogen in the form of protons and electrons. To a lesser degree, there will be collisions from various other species, such as Helium and heavier elements. There will also be stripped electrons and nuclei from these various other elements. The full collision coefficient will sum over contributions from each of these particles. Borrowing notation from [31], we write this as

\[ P_{10}^C = \Sigma_i n_i \kappa_{10}^i, \] (3.80)

where \( n_i \) is the number density of each species in cm\(^{-3}\) and \( \kappa_{10}^i \) is the collisional de-excitation rate out of the triplet state due to that species, in units of cm\(^3\)s\(^{-1}\). The de-excitation rate \( \kappa_{10}^i \) will depend on the temperature through the scattering cross-section and the velocity distribution. For now, let us ignore high energy collisions that could transition the atom above the n=1 state and instead focus on lower energy collisions that can only cause spin flips, since orbital angular momentum is inaccessible at n=1. Collisional spin flips can occur through magnetic interactions or due to the electrostatic interaction forcing the original electron out of the atom and replacing it with another electron of opposite spin. The particle exchange term dominates, since the magnetic interaction is suppressed relative to the electrostatic interaction by a factor of \( v^2/c^2 \), which is small both for the bound electron and for free electrons at realistic temperatures. The average cross-section for HI-HI electron exchange as a function of the energy can be worked out by a partial wave scattering expansion. By
Figure 3-1: The collisional de-excitation rate of HI from the triplet to the singlet spin state for the two dominant species, hydrogen atoms ($\kappa_{10}^{HH}$) and free electrons ($\kappa_{10}^{eH}$). The transition rate is the sum of these terms multiplied by their respective densities. The figure is from [31]. The $\kappa_{10}^{eH}$ curve is based on calculations by [30], and the $\kappa_{10}^{HH}$ curve is from calculations of [99] at $T < 300$ K and extrapolations to higher temperatures by K. Sigurdson, using the methods of [77].

ensuring that the z-projection of the total spin is unchanged, one can track the allowed transitions from the 3 spin 1 triplet states to the spin 0 singlet state. Averaging the cross-section over the energies of the thermal distribution allows one to compute $\kappa_{10}^{HH}$ at various kinetic temperatures. Calculating $\kappa_{10}^{eH}$, the scattering cross-section due to free electrons, involves a similar partial wave expansion. A plot summarizing these de-excitation rates is found in figure 3-1.

Although figure 3-1 shows $\kappa_{10}^{eH} \gg \kappa_{10}^{HH}$ at all temperatures, the low ionization fraction at low temperatures means that hydrogen-hydrogen collisions still dominate the total collisional de-excitation probability until the Universe becomes significantly
Figure 3-2: The allowed Lyman alpha dipole transitions. The solid paths show transition chains that, if followed from the $s$ to $p$ orbital via absorption and then from $p$ back to $s$ via spontaneous emission, can couple the $s$ spin 0 and spin 1 states. The dashed lines show paths that cannot couple $s$ spin 0 and spin 1 states via Lyman alpha absorption and spontaneous emission. This figure is from [31].

heated and ionized. At $T_k > 6200 \text{K}$, hydrogen-electron collisions begin to excite the $n=2$ level [30]. The $n=2$ transitions will in turn emit Lyman radiation, which will be absorbed by other hydrogen atoms in the cloud and come to dominate the spin temperature via the Wouthuysen-Field effect, which is the effect determining $P_{10}^L$.

Therefore, let us turn to calculating $P_{10}^L$. The absorption of Lyman alpha radiation is dominated by the dipole term and is therefore governed by dipole selection rules, which constrain the possible transitions. If we call $F$ the total spin of the atom, then the selection rules for a transition from a spin 1 photon are $\Delta F = 0, 1$, but $\Delta F \neq 0$ if $F = 0$, so there cannot be a spin 0 to spin 0 transition. This creates a web of possible transitions, illustrated in figure 3-2.

Skipping over the details of the calculation, I will simply note that HI that is optically thick to Lyman radiation will scatter Lyman photons many times, which will couple the Lyman temperature to the kinetic temperature: $T_L \approx T_K$.

We now turn to the evolution of the spin temperature. The detailed evolution of
the spin temperature is a complicated subject, but it can be broken into phases. After recombination, the CMB temperature will fall like $(1 + z)$. In the early dark ages, Compton scattering between CMB photons and the small number of free electrons couples the kinetic temperature of the gas to the CMB temperature. The coupling becomes inefficient by $z \approx 150$, at which point the kinetic temperature of the HI scales like $(1 + z)^2$, since it is a monatomic non-relativistic ideal gas. During the dark ages, there is no Lyman radiation, so $T_s$ is determined by the relative strengths of coupling to kinetic and CMB temperatures. A simulated temperature history before reionization is shown in figure 3-3. The result is a dip in spin temperature, creating an observable absorption feature in the dark ages.

The relevant phase for low-redshift 21-cm intensity mapping is the phase after stars reionize the Universe. Although most of the Universe is ionized, HI will be present
in dense clouds that are self-shielded from ionizing radiation. Stellar radiation drives
the kinetic temperature of neutral clouds up, and the spin temperature couples to
the kinetic temperature via the Wouthuysen-Field effect, so that \( T_s \gg T_{CMB} \) at late
times.

### 3.4 Galaxy surveys using the 21-cm line

Although I focused on deriving the magnitude of the mean 21-cm signal in section 3.2,
the vast majority of HI after reionization is expected to be found in small self-shielding
clumps (the largest of which are DLAs), usually associated with galaxies. Therefore,
the HI is expected to trace matter similarly to galaxies: rather than smoothly tracing
the dark matter over-densities, we expect to find the HI in discrete massive clumps,
which have a higher probability of occurring where the dark matter is densest. At
very low redshifts, we expect a single pointing of a large radio telescope to usually
contain one or fewer HI galaxies. Consider the case that a galaxy at redshift \( z \) has a
total HI mass \( M_{HI} \). We can use equation 3.63 with \( \int n_{HI} ds = \frac{M_{HI}}{m_H A_g} \),
where \( m_H \) is the mass of a hydrogen atom, \( M_{HI} \) is the total HI mass in the galaxy,
and \( A_g \) is the cross-sectional area of the galaxy, to derive the average optical depth in the direction
of that galaxy.

\[
\tau_\nu = \frac{3 A_{10} E_{10} A_{21}^2}{32 \pi} \frac{M_{HI}}{m_H A_g k T_S} \phi(\nu).
\] (3.81)

Assuming \( T_s \gg T_{CMB}(z) \) and \( \tau_\nu \ll 1 \), and remembering to evaluate the optical depth
at \( \nu(z + 1) \), the increase in brightness temperature in the direction of the galaxy will be

\[
\Delta T_b = \frac{T_s}{z + 1} \tau_\nu = \frac{3 A_{10} E_{10} A_{21}^2}{32 \pi m_H k} \frac{M_{HI}}{(z + 1) A_g} \phi(\nu(z + 1)).
\] (3.82)

At any significant redshift, however, the angular size of the galaxy will be smaller
than even a large focusing dish like Arecibo can resolve. The measurable quantity
is not the increased brightness temperature \( \Delta T_b \), but the increase in the system
temperature of the antenna that will occur when an antenna is pointed at a galaxy. This will scale like

$$\Delta T_{\text{sys}} = \frac{\Omega_g}{\Omega_A} \Delta T_b = \frac{A_g(z + 1)^2}{R(z)^2 \Omega_A} \Delta T_b = \frac{3A_{10}E_{10}\lambda_{21}^2}{32\pi m_H k} \frac{M_{HI}(z + 1)}{R(z)^2 \Omega_A} \phi(\nu(z + 1)), \quad (3.83)$$

where $\Omega_A$ is the effective size of the telescope beam in steradians and $\Omega_g = \frac{A_g(z+1)^2}{R(z)^2}$ is the angular size that the galaxy appears to subtend in steradians. The factor of $(z + 1)^2$ in the numerator of the angular size of the galaxy appears because a distant gravitationally bound (and thus constant size) galaxy would appear larger than it would in a non-expanding Universe, due to the fact that it was closer to us when the light was emitted. As we’ll show in chapter 4, the power collected by a radiometer over a certain bandwidth is $P = \int kT_{\text{sys}}(\nu) d\nu$. Since $\int \phi(\nu(z + 1)) d\nu = (1 + z)^{-1}$, we find that

$$\Delta P = \int k \Delta T_{\text{sys}}(\nu) d\nu = \frac{3A_{10}E_{10}\lambda_{21}^2}{32\pi m_H} \frac{M_{HI}}{R(z)^2 \Omega_A}. \quad (3.84)$$

Thus, the measurable signal falls off, as one might expect, with the square of the distance to the galaxy. In fact it falls off even more quickly, since the size of the telescope beam is proportional to the square of the wavelength divided by the collecting area, $\Omega_A \propto \frac{\lambda^2}{A}$. The wavelength gets stretched by a factor of $1 + z$, so $\Omega_A \propto \frac{\lambda_{21}^2 (1 + z)^2}{A}$. We also see from this and from equation 3.84 that the size of the signal from a small source is proportional to the collecting area of the antenna. Though we have derived this for the single dish case, the same is true of an interferometer; the size of the signal from a point source scales with the total collecting area of all the elements of the interferometer.

In order to catalog the 21-cm galaxies in the sky, the signal produced by the source must be high enough to stand out from ordinary thermal noise (see equation ??) without excessively long integration times. Even using radio telescopes with large collecting areas (such as the 300 meter diameter Aricebo telescope or the 64 meter diameter Parkes telescope), cataloging galaxies works only out to very modest redshifts.
before the factor of the distance squared makes blind detection impossible. The HI Parkes All Sky Survey (HIPASS) [52] and Arecibo Legacy Fast ALFA (ALFALFA) [35] survey used this technique to detect individual galaxies out to $z \sim 0.04$ and $z \sim 0.06$ respectively. Higher redshift detections of individual galaxies, out to $z \sim 0.2$, can still be made by pointing telescopes at known galaxy positions, but this requires long integration times to overcome thermal noise. The sky coverage of this method can be extended if one abandons the requirement of detecting individual galaxies. Instead, one can measure the 21-cm signal from the known locations and redshifts of many galaxies and boost the signal to noise by co-adding them. This technique is known as galaxy stacking, and it can yield high signal to noise measurements of the typical HI mass of optically selected galaxies. It was used to infer the comoving HI density at $z < 0.13$ [24] and at $z = 0.24$ [41].

### 3.5 21-cm Intensity Mapping

As one attempts to look farther back in distance and time/redshift, detection of individual galaxies becomes impractical, not only because of equation 3.84, but also because Universal expansion means that galaxies were denser in the past and because the angular extent of the telescope covers larger physical areas. This means that at higher redshifts, every voxel one can observe will be expected to contain one or more galaxies. At these redshifts, one can no longer hope to distinguish individual galaxies, but it is still possible to measure large scale fluctuations in the mean density of HI galaxies.

#### 3.5.1 Tomographic mapping of matter using the 21-cm signal

The relevant equation is still 3.69, but now we are interested not only in the mean signal produced by all the HI, but in the fluctuations in that signal. If one assumes that HI is simply a biased tracer of the matter distribution, then the brightness
temperature increase due to HI 21-cm emission is

\[ T_b = h189mK(1 + b_{HI}\delta_{DM})\frac{\Omega_{HI}(1 + z)^{1/2}}{[\Omega_m + \Omega_\Lambda(1 + z)^{-3}]^{1/2}}. \]  

(3.85)

We have dropped the \( \Delta T_b \) notation used previously for clarity, but the expression in 3.85 is still the brightness temperature increase relative to the brightness temperature of the CMB. The symbol \( \delta_{DM} \) refers to the local dark matter over-density. Let’s relabel the average brightness temperature increase due to 21-cm radiation as \( \bar{T}_b \). Then

\[ T_b = \bar{T}_b(1 + b_{HI}\delta_{DM}) \]  

(3.86)

It is actually very difficult to measure the average 21-cm brightness temperature \( \bar{T}_b \) because of the presence of bright foregrounds. However, it is possible to measure the fluctuations in the brightness temperature.

\[ \delta T_b = \bar{T}_b b_{HI}\delta_{DM}. \]  

(3.87)

If we take the Fourier transform of the temperature fluctuation map, we can calculate its power spectrum \( P_{TT} \). In this linear bias model, the brightness temperature power spectrum is related to the dark matter power spectrum \( P_{\delta\delta} \) via

\[ P_{TT} = \bar{T}_b^2b_{HI}^2P_{\delta\delta}. \]  

(3.88)

Thus, even though individual objects cannot be detected, the large scale dark matter power spectrum is still measurable, up to the normalization constant \( \bar{T}_b^2b_{HI}^2 \). The sensitivity of a telescope to this signal is not determined by the collecting area of the telescope but instead by its system temperature and by the number of receivers (see section ??). The only requirement on the collecting area is that it be large enough that the telescope beam is not too large: a large beam will wash out the scales of interest in the measured power spectrum. By dividing \( P_{TT} \) by the dark matter
power spectrum, which can be accurately simulated on large scales, one can obtain a measurement of the average brightness temperature multiplied by the HI bias.

When we attempt to reconstruct the 3D power spectrum, we must convert redshifts to distances in the line-of-sight direction, assuming some fiducial cosmology. But the redshift space power spectrum is distorted by peculiar velocities, an effect known as redshift space distortion (RSD). On large linear scales, the Kaiser effect [40] enhances the power of modes along the line-of-sight due to gravitational infall. On smaller non-linear scales, the random velocities of virialized collapsed matter will wash out power in the line-of-sight direction. One way to parametrize the non-linear effect is with a velocity dispersion $\sigma_v$ [58], which can be fit to the data. Using this RSD model, the power spectrum is given by

$$P_{TT}(k\parallel, k\perp, z) = \bar{T}^2 b^2 H_I \frac{(1 + \beta_{HI} \mu^2)^2}{1 + (k\mu \sigma_v / H(z))^2} P_{\delta\delta}(k, z),$$

(3.89)

where $k\parallel$ refers to k-vectors along the line-of-sight, $k\perp$ refers to vectors perpendicular to the line-of-sight, $k = \sqrt{k^2 + k^2_\perp}$, and $\mu$ is the cosine of the angle between the line-of-sight and the direction of the full k-vector. The parameter $\beta_{HI} = f(z) / b_{HI}$, where $f(z)$ is the dimensionless linear growth rate. The numerator accounts for enhancement from the linear Kaiser effect, and the denominator accounts for the decrease in power at smaller scales due to the velocity dispersion of collapsed objects. On large linear scales, the numerator dominates. Expanding the numerator, we find that

$$P_{TT}^{\text{lin}}(k\parallel, k\perp, z) = \bar{T}^2 b^2_{HI} (1 + 2\beta_{HI} \mu^2 + \beta^2_{HI} \mu^4) P_{\delta\delta}^{\text{lin}}(k, z).$$

(3.90)

The different $\mu$ dependencies can be teased out by performing a multipole expansion. Note that because of the $b_{HI}^{-1}$ dependence of $\beta_{HI}$, each multipole depends on bias and the linear growth rate differently (the final term actually has no bias dependence). As noted by [51], this provides an opportunity to break the degeneracy between $b_{HI}$ and $\bar{T}_b$ that is present in the undistorted power spectrum, if we assume a $\Lambda$CDM form for
the linear growth rate. A measurement of \( \tilde{T}_b \) in turn allows the comoving HI fraction \( \Omega_{HI} \) to be measured, through equation 3.69.

Future surveys will have much lower thermal noise than the early path-finder experiments discussed in this thesis. They will want to constrain dark energy models, which they can do if they measure the power spectrum near the BAO wiggles as a function of redshift. Their constraining power will be even greater if they can also measure the evolution of the growth factor \( g(z) \) (rather than assuming a \( \Lambda \)CDM form for structure growth). Measuring the BAO scale is fairly straightforward, but it is more difficult to measure the growth factor, since a redshift-dependence of \( \tilde{T}_b \) or \( b_{HI} \) can masquerade as structure growth. However, redshift space distortions offer a possible solution. Let us write equation 3.90 in terms of the CMB power spectrum and the growth factor \( g(z) \), explicitly including possible redshift dependence of \( \tilde{T}_b \) and \( b_{HI} \):

\[
P_{TT}(k\parallel, k\perp, z) = \tilde{T}_b(z)^2 b_{HI}(z)^2 g(z)^2 \left[ 1 + 2 \beta_{HI}(z) \mu^2 + \beta_{HI}(z)^2 \mu^4 \right] P_{\delta\delta}^{\text{CMB}}(k)
= \tilde{T}_b(z)^2 g(z)^2 \left[ b_{HI}(z)^2 + b_{HI}(z) f(z) \mu^2 + f(z)^2 \mu^4 \right] P_{\delta\delta}^{\text{CMB}}(k). \tag{3.91}
\]

Since \( f(z) \) is the logarithmic derivative of the growth factor, any dark energy model will predict both \( f(z) \) and \( g(z) \). Therefore, if a particular dark energy model is assumed and the CMB power spectrum is used, then measurements of the first, second, and fourth moment of the power spectrum in \( \mu \) will directly yield measurements of \( \tilde{T}_b(z)^2 b_{HI}(z)^2 \), \( \tilde{T}_b(z)^2 b_{HI}(z) \), and \( \tilde{T}_b(z)^2 \) respectively. These three measurements of two unknown functions can then be checked for consistency. The dark energy model that yields the most consistent result should be favored.

Foregrounds, to be discussed in more detail in the next section, are much brighter than the 21-cm signal. Even if the foregrounds are mostly removed, their residuals can still bias the 21-cm power spectrum. Thus, an important technique at low redshifts is to cross-correlate optical galaxy surveys with 21-cm intensity maps. Cross-correlation requires less perfect foreground removal, because the residual foregrounds should not
correlate with the galaxy maps. The HI-galaxy cross-power spectrum will be

\[ P_{Tg} = \bar{T}_b r b_{HI} b_g P_{\delta\delta}, \quad (3.92) \]

where \( b_g \) is the galaxy bias and \( r \) is the galaxy-HI correlation coefficient. The galaxy-HI correlation coefficient obeys the Schwarz inequality: \(-1 \leq r \leq 1\). Galaxies, like HI clouds, are biased tracers of the underlying dark matter distribution that have an increased likelihood of occurring near dark matter over-densities. Galaxy stellar luminosity is not directly related to HI mass, so there will be some lack of overlap between these two matter tracers. On large scales, however, \( r \) is expected to be constant and close to 1. In redshift space, the HI-galaxy cross-power spectrum is

\[ P_{Tg}(k_{||}, k_{\perp}, z) = \bar{T}_b r b_{HI} b_g \frac{(1 + \beta_{HI} \mu^2)(1 + \beta_g \mu^2)}{1 + (k \mu \sigma_{v}/H(z))^2} P_{\delta\delta}(k, z). \quad (3.93) \]

Consistency between the auto and cross-power spectra will be an important test that must be passed before we can trust 21-cm auto-power spectra at higher redshifts, where there are no galaxy surveys to cross-correlate with.

At low redshifts \((0 < z < 2.5)\), the major scientific goal of 21-cm intensity mapping is to measure large scale structure accurately enough to detect BAO wiggles during the accelerated expansion phase of the Universe. A comparison of redshift versus angular diameter distance can constrain dark energy models. Some forecasts of constrains on BAO wiggles from single dish 21-cm intensity mapping are included in section 8.3.

At high redshifts \((6 < z < 20)\), the major scientific goal is to detect the epoch of reionization, which can, among other things, tighten constraints on the Thompson scattering optical depth to the CMB.

### 3.5.2 Foregrounds

At frequencies below the 1420 MHz rest frequency of the HI 21-cm transition, the dominant source of power in the radio sky is synchrotron radiation and free-free
emission from the Milky Way, from bright point sources, and from a sea of unresolved point sources. For 21-cm intensity mapping, these contaminating sources of power are referred to as foregrounds. The brightness temperature of each of these sources is described to high accuracy by a power law frequency spectrum with a negative index $\alpha$ between 2 and 3.

$$T_b^{fg} \propto \nu^{-\alpha}. \quad (3.94)$$

The brightness temperature of the foregrounds will be on the order of 1 Kelvin at 1420 MHz and will quickly increase at lower frequencies due to the power law. Even at the highest frequencies, where the foregrounds are smallest, they are about 3 orders of magnitude brighter than the 21-cm signal. It is only the smoothness of the foreground spectra that makes 21-cm intensity mapping possible. The 21-cm signal will trace matter perturbations in both the angular direction and the frequency direction. In the frequency direction, the matter perturbations will vary much more quickly than the smooth foregrounds, so that most of the 21-cm signal can be easily separated from the foreground signal. However, this procedure assumes a hypothetical ideal radio antenna that sees the unmodified brightness temperature of the sky. A real antenna will see the effects of noise, imperfect bandpass calibration, mode mixing via convolution of the sky with the telescope beam, and leakage of polarized foregrounds into unpolarized intensity. In chapter 6, I discuss the ways that these real effects complicate this simple picture and make foreground removal a significant challenge.

Since the foregrounds are so much brighter than the signal, it is critical that they be smooth to a high degree of accuracy. One might imagine that small-scale wiggles in a source’s electron energy distribution could lead to wiggles in the emission spectrum, causing it to deviate from a power law. But the broad frequency emission properties of single electron free-free and synchrotron radiation tend to smooth out any possible wiggles [74]. The resulting free-free or synchrotron spectrum is very well approximated by a smooth power law, even at the milli-Kelvin level of the 21-cm signal. However, the power law index $\alpha$ will vary between free-free and synchrotron
sources and will also depend on the local properties of the electron energy distribution of the source. Each line of sight will contain several sources with different electron energy distributions, and the net emission will be the sum of these sources, which will have some spread in individual power law indexes. The resultant emission will be near a power law, but will decohere from a perfect power law over large frequency separations.

The average power of the foregrounds will contribute to the system temperature of the telescope and thus the noise in the map (see section 4.1.4), but it will not be an issue for producing a 21-cm intensity map, since the average power in each frequency slice is usually thrown away when a 21-cm intensity map is made. The fluctuations in the foregrounds, however, will survive. A model for these fluctuations is [73]

\[
C_f(\ell, \nu, \nu') = A \left( \frac{1000}{\ell} \right)^\beta \left( \frac{\nu^2}{\nu'\nu} \right)^{2\alpha} \exp \left( -\frac{\ln^2(\nu/\nu')}{2\xi^2} \right),
\]

(3.95)

where \( \nu_f \) is 130 MHz and \( \xi \) describes the decoherence from a perfect power law. A table of the parameters for the various sources in the Santos model is reproduced in 3.1. There is of course also the constant 2.7 Kelvin black-body emission from the CMB and its small angular fluctuations. These fluctuations will be very small compared to the other sources and will be flat functions of frequency in brightness temperature units.

Finally I’d like to note that, if they are seen over relatively narrow bandwidths,
power law functions of frequency will look basically like straight lines, and power laws with different indexes will look like lines with different slopes. Because of this, the combined foregrounds will be a nearly separable function of angle and frequency. It is this separability that makes the SVD/PCA foreground identification and removal technique, described in chapter 6, attractive.
Chapter 4

Radio Telescopes and HI Astronomy

A radio telescope typically measures incident radio waves coherently, retaining both amplitude and phase information. A typical radio telescope consists of a focusing dish, a receiver, a coherent (often cryogenically cooled) amplifier, filters to remove signal outside the desired bandwidth, local oscillators which downshift the frequency of the measured radiation, and finally a digitizer and digital spectrometer which digitally processes the signal. It is typical for radio telescopes to measure both linear polarizations of incident radiation with orthogonal dipole antennas in the receiver.

Because they retain phase information, radio telescopes can be used as either single receiver instruments, in which the voltage stream from the receiver is correlated with itself, or as interferometers, in which the signals from receivers at different locations, using different reflecting dishes, are cross-correlated. These cross-correlated signals, known as visibilities, are sensitive to sky-patterns that look approximately like double-slit interference patterns, where the distance between the two receivers corresponds to the slit distance. The visibilities thus produce a signal that is roughly equivalent to a Fourier mode on a patch of the sky; the size of the patch is dictated by the size of each element’s primary beam, and the frequency and orientation of the Fourier mode is determined by the length and direction of the separation between the receivers. By combining many pairs of differently placed receivers, arrays of
radio telescopes can map the sky with a resolution comparable to that of a single dish instrument with a size equal to the largest separation between receivers. Large baselines, coupled with large unfocused primary beams, allow for the possibility of making fairly high resolution maps of large swaths of the sky without the need to ever re-point the instruments, since the sky will drift past every day. With the decrease in price and increase in computing power of digital electronics, interferometers have become very attractive for future HI surveys. However, for all of these advantages, interferometers are much more complicated than single dish instruments. An \( N \) element interferometer will correlate \( \frac{N(N-1)}{2} \) pairs of voltages per polarization, and each voltage stream must be consistently calibrated in both gain and phase.

On the other hand, several single-dish radio telescopes with excellent staff support and sufficient resolution for BAO studies already exist. In the next sections, I will outline the basic properties of single dish radio telescopes and summarize 21 cm intensity mapping experiments that have utilized existing single dish radio telescopes on the Green Bank and Parkes telescopes. I will then survey upcoming 21 cm intensity mapping experiments, many of which will use interferometers specially designed for 21 cm intensity mapping.

### 4.1 Fundamentals of Single Dish radio telescopes

Many of the most important aspects of radio telescopes can be derived from a simple thermodynamic thought experiment. Consider an ideal loss-less antenna connected to a matched resistor through a narrow-band filter, all enclosed in a black box (albedo of zero) kept at temperature \( T \). The antenna can be as simple as a short dipole antenna or as complicated as the 100 meter Green Bank Telescope. The situation is illustrated in figure 4-1.
4.1.1 Power measured in Temperature Units

Since all elements are at temperature $T$, the resistor will generate a mean square voltage due to Johnson noise equal to $\langle V^2 \rangle = 4kTR\Delta\nu$, where $R$ is its resistance. The antenna and the resistor are matched, which means that half the voltage will drop across the resistor and the other half across the antenna. Therefore, the resistor will dissipate a power $P = kT\Delta\nu$, and an equal amount of power will be transmitted to the antenna. The loss-less antenna will dissipate this power by radiating it to the walls of the box. At the same time, the box will radiate power which the antenna will collect and transmit to the resistor. By the second law of thermodynamics, the power the antenna collects and transmits to the resistor must be equal to the power that the resistor transmitted to the antenna (otherwise the resistor would not stay at temperature $T$):

$$P = kT\Delta\nu.$$  \hspace{1cm} (4.1)

This means that the power an antenna collects is proportional to the temperature of the object it sees, provided that the object has a black-body spectrum ($I_\nu = B_\nu = \frac{2\nu^2}{c^2}kT = \frac{2kT}{\lambda^2}$ in the Rayleigh-Jeans limit) and fills the primary beam of the antenna.
This idea is used to define the brightness temperature $T_b$:

$$T_b = \frac{c^2 I_\nu}{2k\nu^2} = \frac{\lambda^2 I_\nu}{2k}. \quad (4.2)$$

An object filling the primary beam of an antenna will generate a power $kT_b\Delta \nu$ at the terminals of the antenna. If the object has a thermal spectrum, then $T_b$ is equal to the physical temperature $T$ of the object.

### 4.1.2 The Collecting Area of an Antenna

We can also use this thought experiment to deduce the total collecting area of the antenna. The antenna will see, from the walls of the box, a black-body spectrum at temperature $T$ in all directions. In the Rayleigh-Jeans limit, the specific intensity it sees is $B_\nu = \frac{2\nu^2}{c^2} kT = \frac{2kT}{\lambda^2}$. If we define $A_e(\theta, \phi)$ as the effective collecting area of the antenna in the $(\theta, \phi)$ direction then, since the total power collected must be $kT\Delta \nu$,

$$P = kT\Delta \nu = \int \frac{1}{2} B_\nu(\nu, T)A_e(\theta, \phi)d\Omega\Delta \nu = \frac{kT}{\lambda^2} \int A_e(\theta, \phi)d\Omega\Delta \nu, \quad (4.3)$$

where the factor of $\frac{1}{2}$ comes from the fact that a single antenna is only sensitive to one polarization of radiation. The average effective collecting area of the antenna must then be

$$\langle A_e(\theta, \phi) \rangle = \frac{A_e(\theta, \phi)d\Omega}{4\pi} = \frac{\lambda^2}{4\pi}. \quad (4.4)$$

This means that, for any particular wavelength, the effective collecting area of any loss-less antenna, averaged over all directions, is identical. The difference between a simple dipole antenna and a highly focused antenna like the GBT is that the GBT has nearly all of its collecting area focused on a small angular patch of the sky in the forward direction, while the dipole antenna spreads out its sensitivity almost equally in all directions.
4.1.3 Antenna Reciprocity Theorem, Impedance matching

For radio astronomy, we are much more interested in an antenna’s reception pattern $A_e(\theta, \phi)$ than its transmission pattern. However, the transmission pattern is much simpler to simulate, because one can calculate it with a single simulation, in which the terminals of the antenna are excited with some power and the resulting far-field radiation pattern is monitored in all directions. Fortunately, the antenna reciprocity theorem states that the transmission and reception pattern are proportional to each other.

The transmission pattern is typically quantified by the antenna gain $G(\theta, \phi)$, defined as follows. If the antenna radiates a total power $P$ (equal to $kT\Delta\nu$ in the previous thought experiment), then $G(\theta, \phi)$ is the ratio of the power radiated in the $(\theta, \phi)$ direction to that of a hypothetical isotropic antenna (which would radiate $\frac{P}{4\pi}$ in all directions). The antenna reciprocity theorem says that the gain and the beam-pattern are simply related:

$$G(\theta, \phi) = \frac{4\pi A_e(\theta, \phi)}{\lambda^2}.$$ (4.5)

Thus, one can simulate the antenna’s transmission pattern and easily deduce the reception pattern. The antenna reciprocity theorem can be derived from the Lorentz reciprocity theorem, which states that, for fields generated by current densities $J_1$ and $J_2$,

$$\int_V [J_1 \cdot E_2 - J_2 \cdot E_1]dV = \int_S [E_1 \times H_2 - E_2 \times H_1] \cdot dS = 0,$$ (4.6)

where the surface integral is zero if the integral is performed far away from the antennas, where the fields look like plane waves.

One of the caveats in the thought experiment at the beginning of this chapter is that the resistor and the antenna are assumed to be matched. If they are matched, then any voltage wave traveling down the transmission line from the resistor to the
antenna will not be reflected back but will instead be fully radiated by the antenna. The converse is also true: any power collected by the antenna will be fully transmitted to the resistor (or, in the case of a radio antenna, it will be transmitted to an amplifier, which can be characterized by some input impedance). This perfect match occurs only if the impedance of the antenna is equal to the impedance of the resistor or amplifier to which it is connected. The antenna impedance is defined as the ratio of the current phasor to the voltage phasor at the input terminals of the antenna. When there is a mismatch between the impedance of the antenna \( Z_A \) and the power source \( Z_s \), then some of the incident voltage wave will be reflected. The ratio of the reflected and transmitted voltage will be

\[
\Gamma = \frac{Z_A - Z_S}{Z_A + Z_S}. \tag{4.7}
\]

For an antenna with a single input port, the scattering parameter or S-parameter, which describes the ratio of power reflected, is more often reported and easier to measure: \( S_{1,1} = |\Gamma|^2 \). In simulations, the full beam-pattern and reflection coefficient of an antenna with a dipole-feed receiver can be calculated by simulating a current source with a fixed input impedance at the terminals of the dipole, keeping track of the transmitted and reflected current, and calculating the far-field radiation pattern of the full system.

### 4.1.4 Noise and the System Temperature

The most important factor in designing a survey using a radio telescope is the radio telescope equation, which quantifies the time required to reach a desired signal to noise ratio. The noise level one can expect to achieve at each voxel of a map made by a radio telescope is

\[
\sigma = \frac{T_{\text{sys}}}{\sqrt{\Delta \nu \Delta t}} \tag{4.8}
\]
where $\Delta \nu$ is the frequency resolution of a given pixel, $\Delta t$ is the observing time spent at that pixel, and $T_{\text{sys}}$ is the typical effective temperature seen by the receiver. This equation is essentially the statistical statement that the variance of many uncorrelated samples drops as $1/N$. In this case, the number of samples is $N = 2\Delta \nu \Delta t$, from Nyquist sampling of the data at twice the frequency of the bandwidth. The variance of each sample is $2T_{\text{sys}}^2$, which follows from assuming that the electric field amplitude has a Gaussian distribution; a reasonable assumption because of the Central Limit Theorem. This equation can also be derived from the high photon limit of the variance of the occupation number of a thermal Boson distribution.

If there are $n$ receivers, each looking at different parts of the sky at different times, then the number of samples becomes $N = 2n\Delta \nu \Delta t$, so that

$$\sigma = \frac{T_{\text{sys}}}{\sqrt{n\Delta \nu \Delta t}}.$$  \hfill (4.9)

Solving for $\Delta t$, we find that the time to reach a desired noise level $\sigma$ is

$$\Delta t = \frac{T_{\text{sys}}^2}{n\Delta \nu \sigma^2}.$$  \hfill (4.10)

Along with some common sense cosmology, equation 4.10 tells us the strategy for a successful radio telescope experiment. For a cosmological survey, we want to observe the relevant scales on the sky (at low redshifts, we are particularly interested in the BAO scale), and we want to view as large a volume of the sky as possible to minimize sample variance. To minimize the thermal noise in the survey, one wants as many receivers as possible, with low system temperatures, observing for long periods of time.

The system temperature $T_{\text{sys}}$ includes contributions from many sources:

$$T_{\text{sys}} = T_{\text{rec}} + T_{\text{spill}} + T_{\text{CMB}} + T_{\text{Galaxy}} + T_{\text{source}},$$  \hfill (4.11)
where $T_{\text{rec}}$ is the thermal noise of the amplifier system, $T_{\text{spill}}$ is thermal radiation picked up from the small portion of the telescope’s beam that sees the warm ground, $T_{\text{CMB}}$ is the 2.7 Kelvin signal from the CMB, $T_{\text{galaxy}}$ is the brightness temperature of Galactic synchrotron and free-free emission, and $T_{\text{source}}$ is the brightness temperature contributed by the 21 cm signal, which will be negligible compared to the other terms. If the impedance of the antenna is not well matched to the amplifier, then only a fraction of the power from the sky will reach the amplifier, and the effective value of $T_{\text{rec}}$ will be multiplied by the inverse of that fraction. The only two terms that we can control are the amplifier noise $T_{\text{rec}}$ and the spill temperature $T_{\text{spill}}$. It is critical to minimize these terms as best as possible, given the $T_{\text{sys}}^2$ dependence of the mapping speed.

Two final complications must be mentioned. First, the radio telescope equation assumes uncorrelated Gaussian noise. This is a good approximation on short time-scales but tends to fail on longer time-scales, where $1/f$ noise becomes significant. This long time-scale noise can prevent one from reaching the level of noise implied by the radio telescope equation. We account for $1/f$ noise when we make our maps by treating long time-scale modes as unconstrained. The strategy is explained further in section 5.5. Second, the gain of the amplifier can drift over time, which would effectively add additional noise. A good way to protect against amplifier drift is to inject a small thermal noise calibration signal, designed to be stable over long time periods, somewhere before the amplifier. The noise cal signal is switched on and off at a fixed frequency so that it can be distinguished form the sky signal. Gain fluctuations can then be removed by dividing the total power by the measured power of the calibrator. Of course the calibrator introduces extra noise because it contributes power to $T_{\text{sys}}$ when it is on. The calibrator must also be measured either over a wide bandwidth or over long times so that dividing by the calibrator signal does not vastly increase the noise. However, if the calibrator and $1/f$ noise are treated carefully, one can reach a noise level very close to that predicted by the radio telescope equation.
4.1.5 Map Calibration

The amplifiers of the radio telescope will multiply the power measured by the antenna by some gain. Fluctuations in this gain can be removed by injecting a stable switching calibration signal into the system before the amplifiers. The switching signal can be found in the data, and dividing the total data by the magnitude of the calibration signal removes the gain variations. Though the calibration signal is designed to be stable, it will generally drift on very long time-scales. Therefore, it provides a relative calibration, but absolute calibration is best provided by observing an astronomical point source of known flux.

The idea is to periodically look at a strong point source during the course of a survey. The specific intensity in the vicinity of a strong point source can be modeled as

\[ I_\nu = I_{bg}^\nu + S \delta_D(\theta_0, \phi_0), \tag{4.12} \]

where \( I_{bg}^\nu \) is the specific intensity of the background (perhaps in Jy/Sr), \( S \) is the flux of the source (in Jy, for instance), and \( \delta_D(\theta_0, \phi_0) \) is a Dirac delta function at the location of the point source. The increase in power caused by pointing the antenna at the point source is easily measured in practice. First point the telescope at the point source and measure the power received by the antenna. Then point the antenna at a nearby empty part of the sky that is more than a beam-size away from the source. Subtracting the two measurements yields the power in the antenna due to the point source, which will be (assuming the source is unpolarized)

\[ P = \frac{1}{2} A_e S \Delta \nu, \tag{4.13} \]

where \( A_e \) is the effective collecting area of the antenna at the center of the beam-pattern, which I will henceforth refer to as the forward collecting area. The increase
in brightness temperature due to the point source is

\[ T_b = S A_e \frac{A_e}{2k}. \]  

(4.14)

If both the flux \( S \) of the point source and the forward collecting area of the antenna \( A_e \) are known, then the expected power or brightness temperature due to the source is also known. The noise calibration signal will have been switching during the point source measurement, so we now effectively have a measurement of the noise cal signal in brightness temperature units. This measurement of the noise cal brightness can be used to calibrate the rest of the data. The conversion factor from point source flux to brightness temperature is sometimes referred to as forward gain, \( G_f = \frac{A_e}{2k} \).

This calibration procedure depends on knowing the forward collecting area \( A_e \) or forward gain \( G_f \) of the telescope. Fortunately this can be calculated simply by mapping the telescope’s beam-pattern. The procedure is to again use the point source, but now to point the beam at a grid of locations surrounding the point source. This will map out \( A_e(\theta, \phi) \) up to a normalization constant. The normalization constant comes from equation 4.4, which ensures that \( \int A_e(\theta, \phi)d\Omega = \lambda^2 \). Most single receiver beams are well fit to a Gaussian profile. Applying this normalization to a Gaussian beam model of angular size \( \sigma \),

\[ A_e(\theta, \phi) = \frac{\lambda^2}{2\pi \sigma^2} e^{-\frac{\Delta \theta^2}{2\sigma^2}}, \]  

(4.15)

where \( \Delta \theta \) is the angular distance from the center of the beam. Therefore, for a Gaussian beam of angular size \( \sigma \),

\[ A_e = \frac{\lambda^2}{2\pi \sigma^2} \]  

(4.16)

and

\[ G_f = \frac{\lambda^2}{4\pi k \sigma^2}. \]  

(4.17)
4.1.6 Stokes Parameters

Most radio antennas have two receivers designed to be exclusively sensitive to radiation polarized in one of two orthogonal directions. Let’s label the voltage streams from these two polarizations $V_X$ and $V_Y$. The voltage streams of these two receivers can be multiplied by themselves or by each other (and possibly also phase shifted) and averaged over some binning time. For full polarization information, four quantities must be calculated

$$
\begin{align*}
P_{XX} &= \langle V_X V_X^* \rangle \\
P_{YY} &= \langle V_Y V_Y^* \rangle \\
P_{XY} &= \text{Re}\{\langle V_X V_Y^* \rangle\} \\
P_{YX} &= \text{Im}\{\langle V_X V_Y^* \rangle\}, 
\end{align*}
$$

(4.18)

where the averages are over the binning time. Phasor notation is used. The third line represents the average of the product of the voltages of the two polarizations. The fourth line is the average of the product of the two voltages after one has been phase shifted by 90 degrees. These quantities determine the full polarization information of the incoming radiation. They can be combined to form the common Stokes parameters:

$$
\begin{align*}
I &= \frac{1}{2} P_{XX} + \frac{1}{2} P_{YY} \\
Q &= \frac{1}{2} P_{XX} - \frac{1}{2} P_{YY} \\
U &= P_{XY} \\
V &= P_{YX}.
\end{align*}
$$

(4.19)

The Stokes parameter $I$ is the total power of the incoming radiation. $Q$ measures power that is linearly polarized in the $X$ (positive $Q$) or $Y$ (negative $Q$) direction, and $U$ measures power that is linearly polarized in the directions rotated 45 degrees from
the $X$ and $Y$ directions. The $V$ parameter measures circularly polarized power. For perfectly monochromatic radiation, $I^2 = Q^2 + U^2 + V^2$. However, because real radio telescopes operate over a finite bandwidth $\Delta \nu$, the phase of the incoming radiation can change on timescales $\Delta t \sim \frac{1}{\Delta \nu}$ [11], which are much shorter than the integration time. For incoherent radiation, the phase change results in constantly fluctuating polarization vectors, which tend to average to zero. Thus, for a real radio telescope

$$I^2 \geq Q^2 + U^2 + V^2.$$ (4.20)

The 21 cm signal is expected to be unpolarized ($Q$, $U$, and $V$ equal to zero). Synchrotron foregrounds however, may be around 10% linearly polarized. Astronomical sources are very rarely circularly polarized, but man-made radio frequency interference (RFI) often is. Measurements of $V$ can sometimes be used to separate and subtract RFI.

Real antennas will also be subject to polarization leakage. This occurs because receivers designed to pick up only radiation from one polarization will also pick up low levels of radiation from the orthogonal polarization. This may complicate foreground cleaning and will be discussed later.

### 4.2 Observations with the Parkes L-band Receiver

The Parkes Telescope is a 64 meter steerable radio telescope located in Parkes, New South Wales, Australia. The receivers are located 27 meters directly above the center of the dish, at the prime focus of the paraboloid. A 13 receiver array, operating between 1.2 and 1.5 GHz, was installed in the late 1990s to conduct the HIPASS survey, which surveyed the whole southern sky and in 2004 produced a catalog of low-redshift 21 cm galaxies [52]. The catalog was formed by searching for and identifying narrow frequency spikes in emission caused by HI in galaxies. In 2009, the idea of measuring structure by cross-correlating 21 cm maps with galaxy maps was demonstrated for
Figure 4-2: The relative location and sizes of the 13 beams on the sky. Taken from [4]

The first time: the 2-point correlation function was detected out to a separation of 3 Mpc by correlating the raw HIPASS maps (after some smooth foreground filtering) with 6dF galaxy maps [59].

The Parkes L-band array has 13 receivers and an excellent average system temperature of just 21 Kelvin, which makes it an ideal instrument for producing high quality low redshift 21 cm intensity maps. A low-level switched calibration signal modulated at 128 Hz is injected into each of the 13 beams to protect against gain fluctuations. The receivers form 13 simultaneous beams on the sky, whose relative orientations are shown in figure 4-2. Our group was granted a week of observing time, and in late April and early May of 2014 Yi-Chao Li, Cheng-Yu Kuo, and I went to Australia to make 21 cm intensity maps with the Parkes Telescope. Our survey strategy was to make maps overlapping the two long fields of the 2dF galaxy survey. The length of the fields made it possible to take data nearly constantly. Two spectrometers were used to analyze the data: one was MBCORR[83], which has a 64 MHz bandwidth, 1024 frequency channels, 2 second integration times, and forms the $XX$
and $YY$ quantities. This 2 second integration time is too long to see the switching of the calibration signal, but MBCORR measures the calibration signal over the full 64 MHz bandwidth via 128 Hz synchronous demodulation. We also attempted to use the HIPSR backend[66], which was new at the time. HIPSR offers a larger bandwidth and full Stokes parameters. Unfortunately there were problems with the binning of the calibrator in this data, and we are currently unsure whether our HIPSR data is usable. Descriptions of the analysis and results of Parkes MBCORR data will be discussed in chapters 5 and 7.

4.3 Observations with the Green Bank Telescope (GBT)

800 MHz Receiver

The Green Bank Telescope (GBT) is the largest fully steerable radio telescope in the world. The reflecting surface is a 100 by 110 meter section of a paraboloid. The focus of the paraboloid is located, not above the middle of the surface, but to one side. This unusual design allows the receivers to be housed at the end of an arm on the side of the dish rather than above the middle of the dish. When the telescope is operating, neither the receiver nor the arm blocks the aperture of the dish. The unblocked aperture results in a cleaner beam due to a lack of diffraction and scattering from the receiver cabin and its supports. The unblocked aperture also increases the forward gain, which gives the GBT excellent sensitivity to point sources. Figure 4-3 is an image of the GBT showing this off-axis design. The GBT is located in Green Bank West Virginia, in the National Radio Quiet Zone.

Since 2008, our group has applied for and been granted observing time on the Green Bank telescope to make 21 cm intensity maps. The observations have utilized the 800 MHz receiver, covering the 700 to 900 MHz frequency range, which corresponds to the redshift range $1 \leq z \leq 0.58$. This redshift range is relevant for constraining dark energy models using the BAO signal, as discussed in [16]. The
Figure 4-3: The Green Bank telescope. When the 800 MHz receiver is operating, a boom extends out from the arm at the top left of the image and places the receiver at the prime focus of the paraboloid. This image is from the NRAO website: http://www.nrao.edu/pr/2013/GBTWVU/. The same off-axis design is used on roof-mounted satellite television dishes.
fields were also chosen to overlap the DEEP2 [23] and WiggleZ [26] galaxy redshift surveys. Matter clustering was detected by cross-correlating 21 cm intensity maps with these galaxy surveys [15, 49]. By treating the galaxy-HI cross-power spectrum as a lower bound and the HI-HI auto-power spectrum as an upper bound (biased by foreground residuals), constraints were placed on $b_{HI} \Omega_{HI}$ [86]. These results will be discussed in more detail in chapter 7. The back-end instrument used for most of the GBT analysis has been the Green Bank Ultimate Pulsar Processing Instrument (GUPPI) [69], which was developed for high precision pulsar timing. GUPPI supports integration times as short as 1 $\mu$s, and it supports full Stokes parameters. We used 1 ms integration times. As at Parkes, a switching calibration signal is injected into the receiver before the amplifier. Unlike with the Parkes MBCORR back-end, the fine time resolution of the data is sufficient to resolve the calibration signal in the digital data. The choice of a relatively short integration time was fortunate, because a de-dispersion search through the data found a transient signal from a Fast Radio Burst (FRB), likely from outside our Galaxy [50]. In addition to significant dispersion measure, GUPPI’s polarization data allowed us to detect that the FRB was polarized and exhibited significant rotation measure, indicating that the pulse had passed through a magnetic field.

The important point is that all these 21 cm maps and the discovered FRB were made using a single receiver. The receiver used a corrugated horn antenna. Corrugated horns are commonly used to illuminate dish reflectors like the GBT [33] because they can achieve very symmetric beam patterns with low cross-polarization. The 800 MHz GBT horn achieves good dish illumination and an excellent $T_{\text{spill}}$ of only about 5 Kelvin. The receiver has a low total system temperature of about 25 Kelvin. However, both the HI mapping speed and the expected rate of FRB discoveries could be improved if there were an array of multiple receivers at 800 MHz. The current corrugated horn, however, is not suitable for use in an array. The diameter at the wide end of the horn is quite large; a little over 1.8 meters or 4.8 wavelengths at 800
MHz. This is not a problem for a single feed, but if one wishes to construct an array, then it is desirable to minimize the size of the antennas so that more feeds can be fit in the focal plane.

4.4 Designing an 800 MHz Array for the GBT

Encouraged by the success of the 800 MHz GBT intensity mapping, we have sought to acquire larger maps with lower thermal noise. These maps would help us achieve our future science goals, which include:

- Achieving the sensitivity necessary to detect the linear (Kaiser) effect in redshift space distortions, which would allow us to break a degeneracy between HI bias and $\Omega_{HI}$.

- Detecting the 21 cm auto-power spectrum.

Both of these goals can be achieved more quickly if we can produce an array of multiple receivers to be mounted in the large GBT focal plane. The increase in mapping speed would obviously allow us to make higher quality maps more quickly, which will be useful for both goals. Increased mapping speed will also allow us to quickly broaden the angular extent of our maps. Increased angular extent is important for making a possible auto-power detection because 21 cm signal loss due to SVD foreground cleaning decreases when more angular modes are observed. Signal loss is smaller with increased angular area because the SVD technique determines the foreground modes from the map itself (we will see why this is necessary in chapter 6) rather than by using a pre-determined set of, for instance, smooth polynomials.

The goal of the multi-beam array, in the context of 21 cm mapping, is simply to minimize mapping time, which according to equation 4.10 can be achieved by increasing the number of receivers while maintaining as low a system temperature as possible. Simulations from GBT engineer S. Srikanth have shown that the focal
plane of the GBT has room to form additional high-quality beams around the location of the primary beam [81]. However, there are constraints on the size of any array that could be installed on the GBT. Since the GBT is a multi-wavelength telescope that is frequently switching receivers, the array must be small enough to be stowed away when not in use. Figure 4-4 shows the space available for the stowed array. The length of the array must be 3.02 meters or less, which makes it impossible to construct an array of more than 2 elements using the corrugated horn design. There are also weight and moment of inertia restrictions. The corrugated horn is a heavy and relatively long design. All of these factors make the corrugated horn unsuitable for an array.

For a multi-receiver array to be viable, we had to find a feed design with a similar or better spill temperature to the corrugated horn but with a significantly smaller
diameter. When a receiver is at or near the focus pointed at the middle of the GBT reflecting dish, the GBT dish occupies a cone with a half-angle of 39 degrees about the bore-sight of the receiver. Beyond those angles, the receiver will see the warm ground or the sky. Therefore, minimizing spill temperature is a matter of concentrating as much of the receiver’s beam-pattern within 39 degrees as possible. I tested various antenna designs using the Microwave Studio electromagnetic simulation software from CST. Microwave Studio is especially useful for testing multiple designs, since it has user-friendly tools for quickly constructing structures with complicated geometries. For each design, the far field beam-pattern was computed. An estimate of the spill temperature was computed for each design as:

\[
T_{\text{spill}} \approx 293 \, K \frac{\int G_{\text{rad}}(\theta, \phi)}{\int_{4\pi} G(\theta, \phi)},
\]

where the integral in the numerator is computed only over the range of angles between the edge of the GBT reflector and the horizon, as seen by the receiver. The horizon angles are a function of the elevation angle of the GBT, so the spill temperature was computed for three separate GBT elevation angles: 30, 60, and 90 degrees.

It is also critical that the final receiver be well-matched to the amplifier system, which is usually designed to have a 50 Ohm input impedance. In principle, it is possible to match the amplifier to any impedance using a transmission line impedance transformer, but this will only work perfectly at a single frequency. Therefore, it is important to find an antenna whose impedance is fairly constant over the bandwidth so that it can be well matched over the full 700 to 900 MHz bandwidth. Both the beam pattern and the matching to the amplifier system are calculated in a single Microwave Studio simulation. For the case of a dipole-feed antenna, the dipole is excited by a lumped element current source with a 50 Ohm internal impedance. For a horn-like antenna, which is designed to interface with a cylindrical waveguide that couples to dipole antennas via an orthomode transducer (OMT), the antenna is only simulated up to the interface with the cylindrical waveguide. The antenna is
excited at this interface by a forward traveling $TE_{11}$ wave. In both dipole-feed and horn antenna cases, the simulation calculates the reflection coefficient and the far-field beam pattern. The calculation is carried out using CST’s Time Domain Solver, which is based on the Finite Integration Technique [91]. A Gaussian pulse is applied at the port, and the fields are calculated until the energy remaining in the structure has decayed to a pre-determined value. The reflection coefficient at the port location and the far-field radiation pattern are then calculated as a function of frequency by Fourier transforming the time data.

Several possible feed antennas were considered. A natural candidate is the short backfire antenna (SBA) [27], which produces a relatively narrow beam with a diameter of just 2 wavelengths, considerable smaller than the existing horn antenna. The small cross-section and short length of the antenna make it suitable for the size and moment restrictions of the array. The prototypical SBA with a flat back-plane has a narrow bandwidth, but replacing the flat back-plane with a conical one can increase the bandwidth to our desired range [55]. In fact, the GBT already uses a modified SBA of this type for its lower frequency receivers [82]; see figures 4-4 and 4-5. However, the spill temperature of this SBA is about 10 Kelvin, twice that of the horn antenna. Given the $T_{sys}^2$ dependence of the mapping time, we decided to look for designs with lower $T_{spill}$. It should be noted that the higher spill temperature of the SBA is not necessarily a failing. Forward gain and thus point source sensitivity is increased when a feed antenna utilizes the entire reflector by strongly illuminating the whole reflecting dish. But any receiver that sees the whole dish will generally also see the warm ground. Designing a receiver is usually a trade-off between spill temperature and forward gain. However, 21 cm intensity mapping does not attempt to measure point sources, so we can focus solely on minimizing the spill temperature.

Another antenna that was simulated early on was the Winegard feed [72], which is a smaller diameter narrow-beam horn antenna developed by the Winegard Company. It has a 3 wavelength diameter. A traditional corrugated horn antenna has quarter
Figure 4-5: This picture shows the SBA used for the 600 MHz GBT feed. All pieces painted white are conductors. The structural towers holding up the dipoles and smaller reflectors are dielectric materials. The back reflecter is actually an obtuse angled cone. The conical back reflector increases the bandwidth compared to a flat reflector. The antenna operates like a leaky resonant cavity.
wavelength deep corrugations along the length of the inner cone that are perpendicular to the axis of the antenna (let's call these horizontal corrugations). The Winegard feed instead has 3 grooves on the outer lip of the antenna parallel to the antenna axis (vertical corrugations). A cross-section of the design can be seen in figure 4-6. However, simulations showed that the beam pattern of the Winegard feed is not narrow enough for our purposes and results in a similar spill temperature to the SBA.

Our final design is an SBA much like the 600 MHz GBT feed, but it also incorporates some features of the Winegard feed and a traditional corrugated horn. Like
a traditional horn antenna, we incorporate quarter wavelength corrugations, inspired by research showing that a traditional flat back-plane SBA can narrow its beam-pattern by adding corrugations to the rim [42]. We added corrugations to the GBT SBA design (scaled to 800 MHz). Guessing and checking various corrugated designs, I found that the lowest spill temperature was achieved when there was 1 horizontal corrugation at the mouth of the SBA and 1 vertical corrugation directly above it. A cross-section of the final design can be found in figure 4-7. This design achieved a simulated spill temperature just above 5 Kelvin over much of the band, creeping up to 10 Kelvin at 700 MHz and rising dramatically at the high frequency end of the band near 900 MHz. The rise at high frequencies was deemed acceptable, since a recent increase in RFI in the 850 to 900 MHz part of our band has forced us to throw out all data in that region. The calculated spill temperature as a function of frequency is shown in figure 4-8. Over much of our band, the corrugated SBA has a similar spill temperature to the corrugated horn, but its diameter is only about 2.77 wavelengths or 1.04 meters. This allows just enough room to fit a hex + 1 array of 7 receivers. A diagram of the array design can be found in figure 4-9.
Figure 4-8: Simulated spill temperature for our corrugated SBA receiver across the 800 MHz frequency band. The spill temperature is calculated at a GBT elevation angle of 30, 60, and 90 degrees. The points for the 60 and 30 degree elevation angle spill temperatures are practically identical.
Figure 4-9: Structural drawing of the final hex + 1 7-element array. Small holes in the receivers help save weight and have minimal effect on the beam-pattern.
Table 4.1: Estimate of $T_{sys}$ for both polarizations of a receiver in the proposed 7-element array.

<table>
<thead>
<tr>
<th>$f$ (MHz)</th>
<th>$T_{b}$</th>
<th>$T_{sky}$</th>
<th>$T_{spill}$</th>
<th>$T_{r}$ pol-0</th>
<th>$T_{r}$ pol-1</th>
<th>$T_{sys}$ pol-0</th>
<th>$T_{sys}$ pol-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>700</td>
<td>9.9</td>
<td>3</td>
<td>12.4</td>
<td>12.44</td>
<td>10.79</td>
<td>37.74</td>
<td>36.09</td>
</tr>
<tr>
<td>725</td>
<td>9.3</td>
<td>3</td>
<td>9.2</td>
<td>13.29</td>
<td>11.04</td>
<td>34.79</td>
<td>32.54</td>
</tr>
<tr>
<td>750</td>
<td>8.8</td>
<td>3</td>
<td>8.6</td>
<td>16.48</td>
<td>15.04</td>
<td>36.88</td>
<td>35.44</td>
</tr>
<tr>
<td>775</td>
<td>8.3</td>
<td>3</td>
<td>8.9</td>
<td>15.92</td>
<td>14.08</td>
<td>36.12</td>
<td>34.28</td>
</tr>
<tr>
<td>800</td>
<td>7.9</td>
<td>3</td>
<td>8.5</td>
<td>14.70</td>
<td>14.01</td>
<td>34.10</td>
<td>33.41</td>
</tr>
<tr>
<td>825</td>
<td>7.5</td>
<td>3</td>
<td>6.5</td>
<td>13.79</td>
<td>12.82</td>
<td>30.79</td>
<td>29.82</td>
</tr>
<tr>
<td>850</td>
<td>7.1</td>
<td>3</td>
<td>5.7</td>
<td>13.02</td>
<td>11.74</td>
<td>28.83</td>
<td>27.54</td>
</tr>
<tr>
<td>860</td>
<td>7.0</td>
<td>3</td>
<td>6.8</td>
<td>15.47</td>
<td>13.30</td>
<td>32.27</td>
<td>30.10</td>
</tr>
<tr>
<td>870</td>
<td>6.9</td>
<td>3</td>
<td>9.4</td>
<td>18.13</td>
<td>15.94</td>
<td>37.43</td>
<td>35.24</td>
</tr>
<tr>
<td>875</td>
<td>6.8</td>
<td>3</td>
<td>11.4</td>
<td>21.14</td>
<td>18.22</td>
<td>42.34</td>
<td>39.42</td>
</tr>
<tr>
<td>880</td>
<td>6.8</td>
<td>3</td>
<td>14.2</td>
<td>23.96</td>
<td>21.36</td>
<td>47.96</td>
<td>45.36</td>
</tr>
<tr>
<td>890</td>
<td>6.7</td>
<td>3</td>
<td>21.8</td>
<td>28.45</td>
<td>24.68</td>
<td>59.95</td>
<td>56.18</td>
</tr>
<tr>
<td>900</td>
<td>6.5</td>
<td>3</td>
<td>33.0</td>
<td>33.62</td>
<td>29.52</td>
<td>76.12</td>
<td>72.02</td>
</tr>
</tbody>
</table>

The design of the array was led by scientists at ASIAA in Taiwan: on the science side by Tzu-Ching Chang, and on the engineering side by Yuh-Jing Hwang. The array has been titled simply GBT-HIM (for Green Bank Telescope HI Intensity Mapping). A version of the receiver was built in Taiwan and shipped to Green Bank for testing. Tests at the Green Bank low RFI antenna range and indoor/outdoor test building verified the simulations of the amplifier temperature and the antenna beam-pattern. The 7-element array has been approved for construction as of 2016. An estimate of the full system temperature for a receiver in the proposed 7-element GBT-HIM array is presented in Table 4.1. The receiver averages a system temperature of 33 Kelvin from 700 to 850 MHz. Table 4.1 should be viewed as a conservative estimate of the system temperature. A further 3-5 Kelvin reduction in system temperature can be achieved by gold-plating the coaxial lines extending from the dipole antennas to the amplifier to reduce resistive loss.

Assuming the conservative system temperature of 33 Kelvin over 150 MHz of
bandwidth, using 4.10, the array achieves 3 times the mapping speed of the corrugated horn receiver, with its 25 K $T_{sys}$ and 200 MHz bandwidth. If the region from 850 to 900 MHz continues to be unusable, then the array will achieve 4 times the mapping speed of the horn, since it will also be limited to a 150 MHz bandwidth. If a full 5 Kelvin can be reduced by gold-plating the coaxial lines, the increase in mapping speed is 4 times the full bandwidth horn and 5.5 times the 150 MHz bandwidth horn.

Unfortunately, though GBT-HIM passed its design review, it is currently a project in search of funding.

### 4.5 Survey of Upcoming Low Redshift 21 cm Experiments

As mentioned before, most upcoming instruments are interferometers. The reasons are cost and sensitivity. A single large dish is much more expensive to build than an array of small dishes. And, as shown in the previous section for the GBT, the optics for existing single dishes limit a focal plane array to just a few pixels. An interferometer can have a large number ($n \approx 1000$) of dishes, corresponding to $n$ effective beams on the sky. Except for HIRAX, which is not included there, this list is just an incomplete version of the exhaustive list from [12].

- **BAOBAB** a proposed array of 128 element array to operate between $0.58 \leq z \leq 1.37$ at the GBT or SKA site [64].

- **CHIME**: an interferometric array of five 20-by-100 meter cylinders that is now under construction. The pathfinder with two 20-by-37 meter arrays is already running [5].

- **FAST**: a 21-cm survey has been proposed [79] for the multi-beam 500 meter telescope in southwest China, similar to Arecibo.
• GBT — HIM: the proposed 7-element receiver for the 100 meter Green Bank Telescope described in 4.4 and [17].

• HIRAX: a proposed interferometric array of 1024 6-meter dishes, targeting the 400 to 800 MHz frequency range [54]

• MeerKAT: a Square Kilometer Array (SKA) precursor consisting of an array of 64 13.5 meter dishes in South Africa [38].

• SKA1 — MID: a planned Square Kilometer Array (SKA) Phase I configuration with 190 15 meter dishes in Northern Cape, South Africa [25].

• SKA1 — SUR: a planned Square Kilometer Array (SKA) Phase I configuration with 60 15 meter dishes, each equipped with 36-element phased array feeds, in Western Australia [25].

• Tianlai: a proposed array of eight 15-by-120 meter cylinders to be constructed in a remote part of north-west China [18]. Measured beam-patterns for the cylinder feed antennas and simulated beam-patterns for the cylinders can be found in [19]. Forecasts of constraints on dark energy and primordial non-Gaussianity from the array can be found in [94]. Dish and cylinder pathfinder arrays have already seen first light.

For forecasts of constraints on the power spectrum and cosmological parameters from these surveys, I refer the reader to [12].
Chapter 5

Making Maps From Parkes Radio Data

In this chapter, I describe the steps leading to the creation of 21 cm intensity maps from observations with the Parkes radio telescope. The analysis pipeline closely resembles the one laid out in [48] for a similar survey conducted with the Green Bank Telescope (GBT). The same code base is used, with a few modifications. The code is publicly available at https://github.com/kiyo-masui/analysis_IM. The mkpower branch is the most up to date with modifications for the Parkes analysis. Computations were done on the GPC supercomputer at the SciNet HPC Consortium. The most significant challenges presented by the Parkes analysis were as follows.

First, the Parkes data covers much larger fields of the sky than the GBT data. In order to exploit the large scales available in the angular direction, very large maps were made, spanning about 40 degrees in Right Ascension. However, the mapmaker needs to solve a linear equation using an inverse noise covariance matrix in map space. The memory requirements for performing operations on such a large matrix necessitated the development of parallel processing tools in the mapmaker code. This is described in section 5.4. The parallel code developed for this purpose is now also used for quickly making new GBT maps with the latest data.
Second, the Parkes telescope has 13 receivers with varying bandpass shapes. For GBT data, the time resolution was sufficient to see the switching of a stable broadband calibration noise diode. This was used for bandpass calibration, by assuming a flat spectral signal form the noise diode. Such a technique is impossible for Parkes data because the time resolution is too low to see the switching calibration signal. Only the total power from the calibration noise diode across the entire bandwidth was measured (by analog means) for each time bin. A new technique needed to be devised for bandpass calibration. In the end, bandpass and flux calibration was done separately for maps made from each beam. This is described in section 5.6.

5.1 Observation Strategy and Data Format

The Parkes 21-cm Multibeam Receiver [83] was used to map the two large contiguous fields of the 2dF Galaxy Redshift Survey [21] during a single week in late April and early May of 2014. The relative orientations of the 13 beams formed on the sky are shown in figure 4-2. During that week, 152 hours (1976 beam hours) of data on the two 2dF fields were collected. The Multibeam Correlator (MBCORR) backend was used with a 64 MHz bandwidth centered at 1315.5 MHz, a 62.5 kHz frequency bin size (so 1024 channels), and 2 second integration times. Due to high variance at the edges of the band, the lowest 10 MHz and highest 4 MHz were eventually removed from the final cross-power spectrum analysis described in chapter 7. The results therefore cover a redshift range of $0.057 < z < 0.098$. High variance and odd bandpass shapes were immediately evident in data from two of the YY beams and one of the XX beams (XX and YY refer to the two orthogonal linear polarizations); data from these polarizations and beams were therefore not used in any of the analysis. To minimize spurious signals from ground pickup, the telescope was positioned at a constant elevation angle during each field transit and scanned back and forth in azimuth as the field drifted through. Scans were made as the fields were rising and
Figure 5-1: The beam 1 hitmap for half of the 2dF1 field. Pixels where data was recorded are red, and blue pixels indicate positions on the sky that were never pointed at by the telescope. Lines with opposite slopes come from scans taken during rising and setting.

The radio maps corresponding to the 2dF field near the North Galactic pole (NGP) cover roughly 4$^h$30$'$ in right ascension and 11$^\circ$ in declination, centered at 12$^h$ and 0$^\circ$. The radio maps corresponding to the 2dF field near the South Galactic pole (SGP) cover roughly 6$^h$30$'$ in right ascension and 7$^\circ$ in declination, centered at 0$^h$40$'$ and -30$^\circ$.

Fluctuations in the gain of the amplifiers can potentially introduced a large noise term that is correlated with the sky. The MBCORR backend reduces gain fluctuations by injecting a flat broadband calibration signal, modulated by a 128 Hz square wave, into both polarizations of each beam. The signal is then amplified, downconverted using a local oscillator, filtered, digitized, and binned into 62.5 kHz frequency bins with a 2 second integration time. At the same time, the full bandwidth is measured by total power detectors. Synchronous demodulation of this signal at 128 Hz monitors the noise-cal power. The digitally sampled data is then divided by this measured noise-cal power to provide gain stability. Schematically, the data written by MBCORR is

$$D(\nu) = \frac{GB(\nu)P(\nu)P_{\text{cal}0}}{\langle GB(\nu)P_{\text{cal}} \rangle_{64 \text{ MHZ}}} = \frac{B(\nu)P(\nu)P_{\text{cal}0}}{P_{\text{cal}}} \quad (5.1)$$

where $G$ is the amplifier gain, $B(\nu)$ is the normalized bandpass shape, $P_{\text{cal}}$ is the
power of the noise-cal, $P_{\text{cal}0}$ is the pre-assumed power of the noise-cal, and $P(\nu)$ is the power measured by the digital sampler in the frequency band centered at $\nu$. The extra power from the noise-cal adds about 1 K to the approximately 21 K system temperature. Dividing by the small diode power introduces noise into the map which is highly correlated across all frequencies. Although this contribution to the noise is larger than the intrinsic thermal noise for the bandwidth of our maps (see Appendix A for details), it is removed in the first few modes of the foreground cleaning, described in Chapter 6.

5.2 RFI removal

Even in fairly radio quiet zones, where radio telescopes tend to be located, terrestrial radio frequency interference (RFI) is a reality for any modern radio experiment. In this section, we describe the first step of our analysis pipeline, which is a rough RFI cut.

The raw data from MBCORR is saved in FITS files that cover fairly short blocks of time, typically 3 minutes in length. These data blocks provide a conveniently manageable chunk of data to analyze. Our RFI flagging algorithm evaluates each data block separately. We calculate both the mean and the standard deviation of the data in each of the 1024 frequency channels. Each frequency channel can then be characterized by the ratio of standard deviation to mean, $\sigma/T_{\text{sys}}$. We can then look at the distribution of this noise to signal ratio over all the frequency channels. Channels that are extreme outliers are assumed to be RFI dominated, so we flag all of the data in that entire frequency channel. The high intrinsic spectral resolution of the data, 1024 channels across 64 MHz of bandwidth (so 62.5 kHz for each channel), allows for efficient identification and flagging of narrow-band RFI. After this RFI removal, the data are rebinned to 1 MHz channels (corresponding to a voxel depth of roughly 2.5 $h^{-1}$ Mpc at band center). After rebinning, most bands will not be flagged, since this
Figure 5-2: The left panel shows all of the data blocks for the XX polarization of beam 1 after RFI flagging and rebinning to 1 MHz frequency bands, and the right panel shows the same thing for the XX polarization of beam 10. The color scale shows the standard deviation divided by the mean ($\sigma/T_{\text{sys}}$ from equation 4.8) for each of the 64 frequency bins over the length of the block (typically 3 minutes). White spaces indicate masked frequency channels: this occurs at places where all 16 of the original 62.5 kHz frequency bins that contribute to a 1 MHz bin were masked by the RFI flagging algorithm. The vertical bands indicate variance present across the frequency band. Some of these are probably due to the beam passing through a bright point source, but the brightest ones are likely due to broadband RFI. There is also persistent narrow band RFI at the edges of the band, especially near 1290 MHz. The plot for beam 10 on the left illustrates the consistent high level of noise that led us to discard data from 3 of the 26 beams/polarizations.
would only be the case if all 16 of the 62.5 kHz channels contributing to that 1 MHz channel were flagged. Figure 5-2 shows the data after RFI flagging.

As explained in [48], the RFI flagger is fairly conservative in that it removes whole frequency channels rather than searching for the exact time at which the RFI occurred. Early efforts attempted to search for RFI in time and frequency and remove the specific frequencies and times which appeared troublesome and to keep the rest. But this method tended to often flag out areas of the map with bright point sources, mistaking them for RFI. The current RFI flagger, on the other hand, will not flag out bright point sources since we expect those to be smooth in frequency and thus to increase the variance of all the frequency channels. This also means that the RFI flagger will not be able to remove continuum RFI. However, any RFI in a block that is not flagged will contribute to a larger thermal noise estimation in the mapmaking stage, causing that block to be down-weighted when the map is made.

5.3 The Mapmaker Equation

In this section I use Einstein notation, with repeated indeces implying summation. Following [87], we assume a linear model in which the time ordered data consists of a contribution from the sky plus a contribution from noise.

\[ y_{\nu} = A_{t \nu} x_{\nu} + n_{\nu} \]  

(5.2)

The vector \( y_{\nu} \) represents the measured voltage data. \( A_{t \nu} \) is the pointing matrix for the telescope, which specifies where the telescope is looking when. The rows represent time and the columns represent pixels on the sky. In its simplest form, the rows of \( A_{t \nu} \) have a zero at all points except the pixel that includes the telescope’s pointing during that particular two second integration, which has a 1. One can also use a more complicated interpolation scheme, in which neighboring pixels get some weight, depending on the precise position and velocity of the telescope pointing during the
two second integration. The vector $\mathbf{x}_{\nu}$ is the map of the sky, which we would like to find. We do not attempt to deconvolve the map and instead choose to let $\mathbf{x}_{\nu}$ represent the sky convolved with the telescope beam. The vector $\mathbf{n}_{\nu}$ represents the noise term.

We assume that the noise $\mathbf{n}_{\nu}$ is Gaussian and therefore completely described by its covariance, $\mathbf{N}_{\nu \nu'}$.

\[ \mathbf{N}_{\nu \nu'} = \langle \mathbf{n}_{\nu} \mathbf{n}_{\nu'} \rangle \]  

Under this assumption, one can attain the maximum likelihood estimate for the map, $\tilde{\mathbf{x}}_{\nu}$, by minimizing $\chi^2$.

\[ \chi^2 = (\mathbf{y}_{\nu} - \mathbf{A}_{\nu} \tilde{\mathbf{x}}_{\nu}) \mathbf{N}_{\nu \nu'}^{-1} (\mathbf{y}_{\nu'} - \mathbf{A}_{\nu'} \tilde{\mathbf{x}}_{\nu'}) \]  

\[ \frac{d\chi^2}{d\tilde{\mathbf{x}}_{\nu}} = 0 = -2 \mathbf{A}_{\nu} \mathbf{N}_{\nu \nu'}^{-1} (\mathbf{y}_{\nu'} - \mathbf{A}_{\nu'} \tilde{\mathbf{x}}_{\nu'}) \]  

\[ \mathbf{A}_{\nu} \mathbf{N}_{\nu \nu'}^{-1} \mathbf{A}_{\nu'} \tilde{\mathbf{x}}_{\nu'} = \mathbf{A}_{\nu} \mathbf{N}_{\nu \nu'}^{-1} \mathbf{y}_{\nu'} \]  

\[ \tilde{\mathbf{x}}_{\nu} = (\mathbf{A}_{\nu} \mathbf{N}_{\nu \nu'}^{-1} \mathbf{A}_{\nu'})^{-1} \mathbf{A}_{\nu'} \mathbf{N}_{\nu \nu'}^{-1} \mathbf{y}_{\nu'} \]  

We shall refer to $\mathbf{A}_{\nu'} \mathbf{N}_{\nu \nu'}^{-1} \mathbf{y}_{\nu'}$ as the dirty map. It is analogous to an inverse variance weighted sum, except that it allows for possible off-diagonal correlations in time and frequency. The matrix $\mathbf{A}_{\nu} \mathbf{N}_{\nu \nu'}^{-1} \mathbf{A}_{\nu'}$ is analogous to the normalization factor used in computing an inverse variance weighted sum, $\sum 1/\sigma^2$. Just as $\sum 1/\sigma^2$ is the inverse of the total variance in the case without correlations, $(\mathbf{A}_{\nu} \mathbf{N}_{\nu \nu'}^{-1} \mathbf{A}_{\nu'})^{-1}$ is the total covariance of $\tilde{\mathbf{x}}_{\nu}$ in map space. The final map estimate, $\tilde{\mathbf{x}}_{\nu}$, shall be referred to as the clean map.

Our mapmaking code is broken up into two modules. First, the dirty mapmaker creates the dirty map $\mathbf{A}_{\nu'} \mathbf{N}_{\nu \nu'}^{-1} \mathbf{y}_{\nu'}$ and constructs the inverse noise covariance matrix, $\mathbf{A}_{\nu} \mathbf{N}_{\nu \nu'}^{-1} \mathbf{A}_{\nu'}$. The clean mapmaker then performs a linear solve to attain
our estimate for the clean map.

5.4 Parallelizing the Mapmaker Code

As previously mentioned, the Parkes fields are very long, and we would not like to truncate our maps very much so that we can access those long Fourier modes in our power spectrum analysis. The difficulty is in the construction of the inverse noise covariance matrix $A_{ti}N_{t\nu t'\nu'}^{-1}A_{t'i'}$ and the linear solution of equation 5.5, which involves matrix operations with this matrix. This matrix exists in map space, which is indexed by frequency, declination, and right ascension (RA), but the typical size of the Parkes maps, which were made with 0.08 degree pixels, is 64x100x500 in frequency, declination, and RA space. A double precision matrix in this space would require $10^{13}$ numbers and take up about 80 TB of memory. If we constrain our noise model not to include frequency correlations, then this is reduced by a factor of 64 to about 1 TB, still well beyond the RAM accessible to a single node of the GPC supercomputer. However, if there are no frequency correlations in the noise model, then each frequency is completely independent of the others, and equation 5.5 can be solved separately for each frequency. Each piece of the necessary inverse noise covariance matrix is only about 20 GB. Since the operations can be done separately, this problem lends itself perfectly to parallelization.

The dirty mapmaker module constructs the dirty map $A_{t'i'}N_{t'\nu t'\nu'}^{-1}y_{t'\nu'}$ and the inverse noise covariance matrix $A_{ti}N_{t'\nu t'\nu'}^{-1}A_{t'i'}$. The steps involved in this computation are:

1. The individual data blocks are pre-processed by subtracting the average and mean (this step is justified in section 5.5) and calculating the noise model $N_{t'\nu t'\nu'}$ for each data block (we assume no correlations between data blocks, so that the full noise matrix in time and frequency space is block diagonal).

2. The dirty map is constructed by looping through all the data blocks and apply-
ing the Pointing matrix $A_{t'}\nu'$ and the inverse Noise weighting $N_{t'\nu't'\nu'}^{-1}$ to the data $y_{t'\nu'}$.

3. The inverse noise covariance matrix $A_{t'i}N_{t'i't'i'}^{-1}A_{t'i'}$ is calculated in map space by looping through all the data blocks and applying their noise models to this matrix. For this process, the full inverse noise covariance matrix in map space must be stored in memory.

The most time consuming and memory intensive step by far is step 3. The parallelized dirty mapmaker implements the 3 steps as follows.

1. Each node completes step 1 and pre-processes all the data (this part could have been parallelized as well, but it was not a limiting factor in the run-time).

2. One of the nodes makes the dirty map, as in step 2 above.

3. Each node works separately on constructing its own frequency chunk of the noise covariance matrix. If the node is assigned more than one frequency and has more than one virtual core, threading is implemented so that different cores can work in parallel on different frequencies. When a node finishes its computation, its portion of the inverse noise covariance matrix is saved to the appropriate part of an HDF5 file that stores the full inverse noise covariance matrix.

The next module is the clean mapmaker, which solves the linear equation of equation 5.5 via Cholesky decomposition. Parallelizing this module was a simple matter of having each node load a separate frequency chunk of the inverse noise covariance matrix and solve for that frequency chunk of the clean map. The chunks of the clean map are then passed to the first node, and the first node saves the full clean map.

The code base for our analysis pipeline is written in Python. The intensive computational parts all use either NumPy or SciPy routines (which operate on matrices using BLAS or LAPACK), Fortran, or Cython. The parallelized code utilized the Message Passing Interface (MPI).
5.5 The Noise Covariance Model

There are several goals in constructing a noise model. First, we need the noise model to measure variance from the actual data, so that it can detect blocks with excess variance from RFI that has gotten past our initial flagging; it must report high noise for these blocks so that their contribution to the map estimate will be down-weighted. Second, it should suppress fluctuations that come from long time-scale $1/f$ noise. Third, it should model the noise as accurately as possible without exceeding computational time and memory limits. With approximately 200,000 time bins and 64 frequency channels, a full noise covariance matrix would mean constructing and inverting a $10^7$ by $10^7$ matrix, which would take up approximately 1 petabyte of memory. Instead, we choose to only model noise correlations between time bins within the same 3 minute scan. This makes the noise model block diagonal in time and allows the noise covariance matrix to be constructed and inverted independently for each scan. The RAM requirement becomes negligible because only the noise model of a single 3 minute scan needs to be stored in memory at any one time. Of course, this simplification does not come for free. In reality there is $1/f$ noise in the amplifier system temperature, which causes significant noise drift on timescales longer than a single scan. The correlations caused by the $1/f$ noise cannot be modeled with our block diagonal scheme. Instead, we explicitly subtract the time mean and slope from each scan to remove long time-scale drift from our maps. We account for this in our noise model by adding a large noise term to constant and linear time modes. This gives the mapmaking solution of equation 5.5 the freedom to change the slope and mean of each scan. These unconstrained modes, coupled with the overlapping web of scans, ensure that the map is stitched together in a self-consistent manner. The mapmaker equation forces optimal consistency where scans overlap. The resultant map has only the mean of the entire map missing, rather than having multiple unconstrained means on the angular scale of each single scan.

The noise model used for the Parkes maps is diagonal in frequency and block
diagonal in time. The first term is the thermal noise term from the radiometer equation. This term is diagonal in time and frequency.

\[ N_{t_0t_0'}^{\text{Thermal}} = \delta_{t,t'} \delta_{\nu,\nu'} \frac{T_{\text{sys}}(\nu)^2}{\Delta \nu \Delta t} \] (5.6)

For Parkes, the system temperature is about 21 Kelvin, the time bin is 2 seconds, and the frequency bin is 1 MHz. Therefore, the expected thermal noise term is about 220 mK². As explained in appendix A, the noise cal also adds coherent noise at about 20 times this variance. But rather than using these calculations for the thermal noise, we measure it from the time variance of each scan (the variance is calculated after the mean and slope are subtracted). The variance is measured independently for each frequency, so the thermal noise term can have a frequency dependence. Measuring the noise from the scan ensures that we down-weight any blocks or frequencies within blocks that are contaminated by RFI that survived the initial cut. However, it also artificially increases the noise on scans that cross bright point sources. This is not a big problem in terms of biasing the map, since every pointing at a bright source will have similarly high noise, but it does mean that the noise covariance matrix produced by the mapmaker will have artificially high noise near bright point sources. The problem can be mostly fixed by treating the first map as a rough estimate of the true sky and then running the analysis pipeline a second time. In this second run, the first map can be subtracted out of the time stream data to yield a much better estimate of the noise in each data block. A second more aggressive RFI flagging algorithm can then be run on each block. The mapmaker can then be run again, with the noise model estimate coming again from the data minus the first map.

The next term in the noise model accounts for the fact that we subtracted the mean and slope from each scan. In order to stitch the map together consistently, we must deweight these modes by adding a large noise to them. The deweighting allows the mapmaker to change the mean and the slope of each scan so that the map is maximally consistent where the scans overlap. The deweighting noise term is as
follows:

\[ N_{\nu'\nu}^{\text{Deweight}} = T_{\text{large}}^2 \delta_{\nu,\nu'} U_{t,i} \delta_{i,i'} U_{i',t'} \]  

(5.7)

The first column of \( U_{t,i} \) is a normalized constant vector, and the second column is an orthonormal linear slope vector. The rest of the columns of \( U_{t,i} \) are simply zero. \( T_{\text{large}} \) is chosen to be 1 Kelvin. This value was chosen to be much larger than the expected thermal noise but not so large that inverting the noise matrix is numerically unstable.

Our final time stream noise model is just the sum of the thermal and dweighting noise terms. As stated previously, it is block diagonal, with no correlations between different 3 minute scans. Within each 3 minute scan, the form of the noise model is:

\[ N_{\nu'\nu} = \delta_{t,t'} \delta_{\nu,\nu'} \frac{T_{\text{sys}}(\nu)^2}{\Delta \nu \Delta t} + T_{\text{large}}^2 \delta_{\nu,\nu'} U_{t,i} \delta_{i,i'} U_{i',t'} \]  

(5.8)

5.6 Calibrating the Maps

After mapmaking, we have 23 maps, all with slightly different uncalibrated bandpass shapes and flux scales. The following sections describe the bandpass and flux calibration of the maps.

5.6.1 Bandpass Calibration

The bandpass has three effects on the data. The most significant effect is a systematic multiplication of the true sky spectrum by the average bandpass shape. Secondly, fluctuations in the bandpass create noise in addition to the intrinsic thermal noise of the receiver. Finally, there is an effect where the presence of bright point sources adds ripples to the bandpass shape due to standing waves between the receivers and the dish (see section 3.5 of [13] for a discussion of this effect for Parkes). Our bandpass calibration scheme aims only to estimate the average bandpass shape over the entire
week long observation. Though it may be possible to account for variation of the
bandpass over time and near bright point sources by computing multiple bandpasses
for different times/locations, this risks biasing the data.

Each beam and polarization can have a different bandpass shape, in principle. Our
first step for estimating these bandpass spectra is to construct frequency covariance
matrices between maps made with different beams. The covariance matrices are
constructed as follows: let $M_i$ represent a map produced by beam $i$, arranged as
a matrix with frequency identified by the row and pixel by the column, and let $M_j$
represent a map of the same region of the sky, from beam $j$. Both maps are convolved
to a common beam resolution, and a frequency covariance matrix $C_{i,j}$ is constructed
as
\[
C_{i,j} = \frac{\bar{W}_i \circ M_i (\bar{W}_j \circ M_j)^T}{\bar{W}_i \bar{W}_j^T}
\]
where $\bar{W}_i$ represents a matrix of the inverse noise weights at each pixel for beam $i$,
factorized into a separable function of angle and frequency. The factorization of the
weights prevents frequency structure in the weights from influencing the covariance
matrix. $\bar{W}_i \circ M_i$ is the voxel-by-voxel product of the map with the factorized weights.
This covariance matrix is computed for all pairs of different beams, but only between
common polarizations (thus avoiding thermal noise bias and spurious correlations due
to polarized signal). Each covariance matrix, $C_{i,j}$, is decomposed into singular values
(SVD),
\[
C_{i,j} = U \Sigma V^T
\]
and the normalized left and right eigenvectors in $U$ and $V$ are ordered by their shared
singular values, which are the entries of the diagonal matrix $\Sigma$. The singular values
fall off rapidly, with the first eigenvalue almost two orders of magnitude larger than
the second.

The projections of the left and right first eigenvectors onto their maps reveal that
the first modes are due to diffuse emission over the full region of the maps. Therefore,
the first eigenvector of $V$ represents an estimate of the bandpass shape of beam $j$
multiplied by the frequency spectrum of the strong diffuse emission in the map, which
will be dominated by galactic synchrotron radiation. Similarly, the first eigenvector
of $U$ represents the diffuse emission spectrum multiplied by the bandpass of beam
$i$. This interpretation is also motivated by SVD analysis of GBT maps. For the
GBT maps, which are bandpass calibrated with a noise diode before mapmaking,
the first eigenvector has been found to capture the frequency spectrum of the diffuse
synchrotron emission across the map [85]. In figure 5-3, I plot the eigenvectors for
the XX polarization in the sub-map region centered at 33 degrees right ascension.

Let $\hat{u}_i(\nu)$ represent the average first eigenvector of beam $i$, computed by averaging
the first left eigenvector of $C_{i,j}$ over each other beam, $j \neq i$. We assume a Galactic
synchrotron spectral index of -2.7, which must be divided out to attain an estimate
of the bandpass from $\hat{u}_i(\nu)$. We compute a bandpass estimate for beam $i$ thusly:

$$\hat{B}_i(\nu) = \frac{\hat{u}(\nu)\nu^{2.7}}{\langle \hat{u}(\nu)\nu^{2.7} \rangle_{64\,MHz}}. \quad (5.11)$$

Equation 5.11 is the bandpass estimate that is used for each beam and polarization
and which is computed separately for each of the four sub-maps. The bandpass
estimates are similar but clearly distinct for each beam, as one would expect since
the amplifiers and filters are similar but not identical. The variance in each beam’s
bandpass estimate from equation 5.11 is negligible in any one region. The standard
deviation across the four sub-maps is slightly less than 3%. Figure 5-4 shows the final
XX polarization bandpass estimates for the sub-map region centered at 33 degrees
right ascension.

### 5.6.2 Flux Calibration

Calibration of the total flux is achieved via periodic on/off scans of the astronomical point source J1939-6342 and beam mapping scans made with the point source
Figure 5-3: This image matrix shows the first eigenvectors of the XX polarization sub-maps centered at 33 degrees right ascension, made from the 12 good XX beams. The plot at row $r$ and column $c$ represents the first eigenvector of beam $c$, calculated from the SVD of $C_{r,c}$, plotted across the full 64 MHz bandwidth. The eigenvectors are expected to represent the product of the bandpass shape and the power law of the diffuse emission present in the map. The column-to-column variation reflects the slightly different bandpass shapes of the various beams.
Figure 5-4: The final bandpass estimates for the 12 good XX beams, plotted over the full 64 MHz bandwidth. The ripples in the bandpass shapes appear to have a period near 5.7 MHz, and are probably caused by standing waves between the receiver cabin and the reflecting dish [13].
PKS0043-42. In the on/off scans, the power of the switching noise-cal, $\dot{P}_{\text{cal}}$, is measured in Janskys by comparison to J1939-6342. The consistency of these measurements verifies the stability of the noise-cal to $\sim 1\%$ over the full week of observations. Measurements from [67] and [32] found the flux of J1939-6342 to be $16.4 \pm 1.0$ Jy, which translates to an absolute calibration uncertainty of $6.1\%$. To convert from Janskys to Kelvin, we require the forward gains of the beams. These are obtained by mapping the beams with PKS0043-42 and fitting the shapes to a Gaussian. Utilizing these fits and the fact that the total effective collecting area for any antenna is

$$\int_{4\pi} A_e(\theta, \phi) d\Omega = \lambda^2, \quad (5.12)$$

we find that each beam has a forward gain in our band of $G = 0.79 K Jy^{-1}$ with an uncertainty of $5\%$. Combining the point source flux, noise-cal stability, and forward gain errors in quadrature, we estimate a total flux uncertainty of $8\%$.

With all these calibration factors in hand, we apply a bandpass and flux calibration to the maps as follows:

$$M_i^C(\theta, \nu) = \frac{GM_i(\theta, \nu) \dot{P}_{\text{cal}}}{P_{\text{cal}0} \dot{B}_i(\nu)}. \quad (5.13)$$

The inverse noise weights produced by the mapmaker are also calibrated, following a similar procedure:

$$W_i^C(\theta, \nu) = \frac{W_i(\theta, \nu) P_{\text{cal}0}^2 \dot{B}_i(\nu)^2}{G^2 P_{\text{cal}}^2}. \quad (5.14)$$

In preparation for the next stage of the analysis, we average maps from the two linear polarizations to form unpolarized maps. The 13 beams are then averaged, using their inverse noise weights, into four groups. A: beams 1-3, B: beams 4-6, C: beams 7-9, and D: beams 10-13.

The calibrated maps at band-center before and after foreground cleaning can be seen in figure 5-5, along with their inverse noise covariance weights.
Figure 5-5: The two 21-cm sub-maps that overlap the 2dF SGP field are shown at band center. The three rows from top to bottom show: the maps before any foreground modes are removed; the maps after 10 modes are removed; and the inverse noise weights, which are roughly proportional to the time spent observing each pixel. The color scales, from left to right, refer to the maps from top to bottom. All beams have been combined, and the resolution has been degraded to 1.4 times the original beamsize. The point sources and diffuse galactic foregrounds have been strongly suppressed in the foreground cleaned maps, and the scale of the remaining fluctuations is consistent with thermal noise. It should be noted that the fluctuations on the right ascension edges of the cleaned maps saturate the scale, but their magnitude is consistent with thermal noise and sparse coverage. Some cross-hatched striping can be seen in the maps and weights due to the differing scan angles of the azimuthal scan strategy as the field rose and set. The noise implied by the weights is higher than thermal noise because the mapmaker’s noise estimation includes variance from the noise-cal measurement, residual RFI, and fluctuating foregrounds.
5.7 On the Optimality of the Parkes Maps

The large Parkes fields presented some challenges to our analysis pipeline. A second round of mapmaking has not been conducted, due partly to the long wait time for the high memory computing nodes. There is potentially much to be gained by a second round of mapmaking, which would minimize RFI contamination in the maps and would lead to more optimal noise weights.

Another issue is that the frequency binning was chosen to be coarse compared to the angular pixel size, which limits the maximum $k$ in the $k_\parallel$ direction. This choice was also made to save computing time and to minimize the hard drive space needed to store the inverse noise covariance matrix. Rerunning the maps with 8 times the frequency resolution, though computationally intensive, would resolve this issue.
Chapter 6

Cutting through the Foregrounds

The expected HI 21-cm signal is on the order of a mK or less. The primary foregrounds are due to synchrotron and free-free emission from the Milky Way, from unresolved point sources, and from bright point sources. At low redshifts, these foregrounds are a few Kelvin, three orders of magnitude larger than the signal. At the higher redshifts of the Epoch of Reionization, the relative brightness of the foreground increases to 5 or 6 orders of magnitude larger than the expected HI signal. Despite this, foreground removal is possible because the foregrounds are intrinsically smooth in frequency [53][44]. The HI signal, on the other hand, follows the clumpy distribution of the matter power spectrum in both the angular and frequency directions. This means that in the absence of instrumental effects, except for large scale line-of-sight clustering at low $k_\parallel$, the HI signal should be distinguishable in the frequency direction as small amplitude higher frequency ripples on top of the much larger smooth frequency spectrum of the foregrounds [53].
6.1 Identifying Foregrounds via Principal Component Analysis

Our method for identifying foregrounds is to perform a principal component analysis (PCA) on the maps. I shall often refer to this method of foreground removal as the SVD (standing for singular value decomposition) method. The reason for the SVD name will become clear in this section. The logic of the method can be explained with an idealized example, in which we’ll assume that our map of the sky contains only fluctuations from foregrounds and 21-cm signal, with no thermal noise.

Let $M_{\nu \theta}$ be a noiseless map of the sky fluctuations, organized as a matrix. Each angular pixel is associated with a different $\theta$ index (so $\theta$ actually rolls up all the right ascension and declination combinations into a single index), and each frequency channel corresponds to a different $\nu$ index. We form a frequency covariance matrix as follows:

$$C_{\nu \nu'} = \frac{M_{\nu \theta} M_{\nu' \theta}}{N_{\text{pix}}}$$  \hspace{1cm} (6.1)

I am again using Einstein notation, so there is an implied sum over all pixels in the map. In words, the covariance matrix is constructed by summing the outer products of the frequency structure at each individual pixel and then dividing by the number of pixels. The next step is to perform an eigenvalue decomposition of $C_{\nu \nu'}$. Because $C_{\nu \nu'}$ is a real symmetric matrix, this is guaranteed to produce a set of orthogonal eigenvectors and is exactly equivalent to the singular value decomposition (SVD) of $C_{\nu \nu'}$.

$$C_{\nu \nu'} = U \Sigma U^T$$  \hspace{1cm} (6.2)

We order these eigenvectors by the magnitude of their eigenvalues, which represent the quadrature average of the power in the map due to each vector. The columns of $U$ form an orthonormal basis that captures the frequency structure of the map that is coherent across pixels. Because the foregrounds dominate the power in the map,
Figure 6-1: Results of the SVD calculation of foreground modes on a simulated 30 degree by 30 degree map with the same frequency range as our GBT maps; 700 to 900 MHz over 256 frequency channels. The simulated data was provided by Le Zhang, and it follows the Santos foreground model and the 21-cm model of [97]. The plot shows the square root of the eigenvalues $\Sigma$, normalized to 1 for the largest eigenvalue. The foregrounds are completely described by the three largest eigenvectors, which are smooth functions of frequency. The 21-cm power spectrum can be reconstructed with minimal signal loss after de-projecting these 3 frequency modes from every pixel of the map.

they will dominate the modes with the highest eigenvalues. Since the foregrounds all have smooth frequency structure, they ought to be described by just a few smooth frequency eigenvectors. Simulations of this SVD foreground technique on maps with simulated 21-cm signal plus our best-guess foreground models have born this out; see figure 6-1.

The uniqueness of the SVD and the eigenvalue decomposition mean that the above analysis is equivalent to taking the SVD of the map $M_{\nu\theta}$ because

$$C_{\nu\nu'} = \frac{M_{\nu\theta} M_{\nu'\theta}}{N_{\text{pix}}} = \frac{U \Sigma M V^T V \Sigma M U^T}{N_{\text{pix}}} = \frac{U \Sigma_M^2 U^T}{N_{\text{pix}}} = U \Sigma U^T \quad (6.3)$$
where the columns of $U$ are the orthonormal frequency eigenvectors of the map’s SVD, $\Sigma_M$ is the matrix of singular values of the map, and the columns of $V$ are the orthonormal spatial eigenvectors of the map’s SVD. The eigenvalues $\Sigma$ of the frequency-frequency covariance are equal to the square of the singular values of the map divided by the number of pixels. From this analysis, one can see that the rank of the covariance matrix can be limited by the number of angular modes present in the map, not just the number of frequency channels. The foregrounds should be low rank, occupying just a few frequency and angular modes in the map. In the absence of noise, the rest of the map’s rank comes from the 21-cm signal and from the random correlations of the 21-cm signal with the foregrounds. The number of modes occupied by the 21-cm signal will scale with the angular size of the map. When one measures the foreground structure empirically from the map using this technique, rather than using some pre-fit functional form, it is critical to measure enough independent 21-cm signal modes on the sky to fill the rank of the covariance matrix $C_{\nu \nu'}$. This will minimize overlap of the few contaminated foreground modes with the 21-cm signal.

Our power spectrum analysis, described in the next section, consists of projecting out the largest foreground modes (the principal components) from each pixel in the map and then forming an estimate of the power spectrum from the remaining part of the map. Mathematically, we multiply the map $M_{\nu \theta}$ on the left by the deprojection operator:

$$\Pi_{f}^\nu = I - U_f S U_f^T$$  \hspace{1cm} (6.4)

Here, $I$ is the identity matrix and $S$ is a selection matrix indicating which modes to remove. $S$ is diagonal, with ones on the upper left corner for the modes that should be removed. Applying $\Pi_{f}^\nu$ to our map projects out the selected principal components along each line of sight.

Real maps will have thermal noise in addition to foregrounds and 21 cm signal, and the noise is in fact quite a bit brighter than the 21 cm signal in both the Parkes and GBT maps. If the SVD matrix is formed by correlating a map with itself, then
the noise will bias the determination of the foreground modes. In order to avoid this noise bias, we always correlate maps made with different data sets. For the GBT analysis, we correlate maps made with different data from different observation times. For the Parkes analysis, we correlate maps made by different beams.

The rosy picture painted by figure 6-1 is an extreme idealization, ignoring all of the instrumental effects associated with making real maps. Analysis of real maps has found that many more modes are contaminated by foregrounds and that most of them are not well approximated by smooth functions of frequency. Some of the likely reasons for this are explored in the next section.

6.2 Systematic Effects on Foregrounds

There are several instrumental effects which can convert the intrinsically smooth foregrounds into less smooth structures. Some of these effects, like systematic bandpass calibration errors, consistently alter each pixel in the same way. Any consistent effect across pixels is easily found using the PCA/SVD technique described in the previous section, and it does not increase the number of frequency modes that are contaminated by foregrounds. However, there are several effects which mix angular structure into frequency structure, causing the apparent foreground signal to differ depending on the pixel. These mode mixing effects cause more frequency modes to be contaminated by foregrounds, which means that more modes must be removed to get to the level of the HI signal. Below I list the important known effects that can introduce spurious structure into the intrinsically smooth foregrounds:

- **Mode mixing** due to the frequency-dependent telescope beam-pattern inevitably occurs. The term mode mixing refers to instrumental effects transferring intrinsically angular structure on the sky to frequency structure in the computed sky maps. For a single dish experiment with an idealized Gaussian beam, the effect is present but mild. The mode mixing occurs because the ob-
served map represents the convolution of the sky with the telescope beam, but
the beam is frequency-dependent: its full-width at half-power (FWHP) falls off
approximately as $\nu^{-1}$ due to diffraction. Now imagine a strong point source near
the edge of the beam. At low frequencies, the telescope will capture more of
the flux from that point source due to the larger beam-pattern. The presence of
angular structure (a nearby point source) thus creates artificial frequency struc-
ture. This mode mixing is more extreme if the beam has frequency-dependent
side-lobes. Far side-lobes can fluctuate very quickly as a function of frequency,
and this can cause angular structure in the far side-lobes to mix into rapidly
varying power as a function of frequency. The mode mixing effect can be mit-
igated by convolving the map to a common beam resolution, but the efficacy
of this procedure depends on how accurately the frequency dependence of the
beam-pattern can be measured. We emphasized the single dish case here, but
the situation is more extreme for interferometers because the effective beam
formed by correlating two receivers will resemble an interference pattern, which
has very high frequency-dependent side-lobes within the main envelope.

- A similar effect, which I will call polarized mode mixing occurs due to the
coupling of angular variation in polarized foregrounds and frequency dependent
polarization leakage of the telescope. This coupling mixes angular structure in
the polarized foregrounds to frequency structure in the maps. By polarization
leakage, I refer to any instrumental effect which causes intrinsically polarized
signal (Stokes Q, U, or V) to be measured as unpolarized Stokes I signal. These
effects can be significant because synchrotron radiation is often strongly po-
larized. An additional complication is that the polarized foregrounds will also
often exhibit Faraday rotation, whereby the linear polarization is rotated by an
angle equal to $RM\lambda^2$. Again, this effect increases the number of modes occupied
by the foregrounds.

- **Systematic errors in bandpass calibration.** The frequency response of the
amplifiers and filters in a radio receiver cause the intrinsic frequency structure to be multiplied by an instrumental response function, known as the bandpass. If the bandpass is stable, this simply changes the shape of the foreground signal but does not increase the number of modes occupied by the foregrounds.

- **Varying spill temperature.** For a single-dish instrument, mapping the sky involves moving the telescope. It is desirable to minimize variation in the spill temperature so that fluctuations in spill temperature do not appear as structure in the sky. In the GBT and Parkes surveys, we maintain a constant elevation angle. However, there will be some azimuthal dependence to the spill temperature due to local variation in elevation and albedo. This will show up at some level as false frequency structure in the sky.

Because the foregrounds are so large, these effects can potentially mask the HI signal even if they are only 0.1% effects. The precise frequency structure of the non-smooth components will depend upon difficult to measure telescope systematics and the angular structure of the foregrounds. Therefore, we suggest that the contaminated foreground modes should be found from the data, with no preset parametric model.

### 6.3 Parkes Foreground Removal

Our group has argued that rather than subtracting a set of predetermined smooth frequency functions from each pixel, it is better to determine the shape of the modified foregrounds through a principle component analysis of the entire map, which is dominated by the foregrounds modified by instrumental effects. Successful detection of HI depends on controlling instrumental effects so that the smallest number of frequency modes are contaminated by foregrounds [85].

The foreground removal steps for the Parkes analysis are as follows. We construct
frequency-frequency covariance matrices as

\[ C_{A,B} = \frac{W_A \circ M_A (W_B \circ M_B)^T}{W_A W_B^T} \]  \hspace{1cm} (6.5)

using pairs of calibrated maps and factorized calibrated weights. Before a covariance is constructed, all frequency slices of the maps are convolved to a common beam resolution based on a frequency-dependent Gaussian beam model. To the extent that we can accurately model the beam, this step prevents the frequency dependent beam size from altering the shape of the SVD modes and from coupling angular variation in the maps into frequency variation. For each of our four sub-maps, we form six covariance matrices using all pairs of different beam groups: AB, AC, AD, BC, BD, CD. We then perform a singular value decomposition on each covariance matrix

\[ C_{A,B} = U \Sigma V^T \]  \hspace{1cm} (6.6)

where the columns of \( U \) and \( V \) are the orthonormal spectral modes of \( M_A \) and \( M_B \) respectively, ordered by their shared singular values. The singular values in \( \Sigma \) roughly correspond to the squared power present in the map from each frequency mode. Because foregrounds dominate the power in the map and are much more coherent between different frequencies than the 21-cm signal, the modes with the largest singular values (the principal components) are considered to be foreground modes. By only correlating maps made by different groups of beams, we have avoided any thermal noise bias in the determination of these foreground modes. Our foreground cleaning procedure projects out a fixed number of these spectral foreground modes from all pixels of the 21-cm maps. The number of foreground modes to remove is determined via simulations, as detailed in Section 6.4. Our SVD foreground removal process is formally described in [85] and is used in the analysis of [49]. Note that those examples use or assume a single receiver, which necessitates correlating separate season maps to avoid the noise bias that we avoid here by correlating maps made
from different beams.

Although the frequency modes with the highest singular values are dominated by foregrounds, there will inevitably be loss of 21 cm signal when these modes are removed from the maps. This loss of power must be accounted for in the cross-power spectrum by the application of a transfer function, as described in 6.4.

6.4 Cross-Power Spectrum Estimation

In this section, I describe the method for estimating the cross-power spectrum between our foreground cleaned Parkes 21-cm intensity maps and the 2dF galaxy overdensity maps. In order to compensate for signal loss from the foreground cleaning and to estimate error bars, we simulate the dark matter power spectrum and draw mock galaxy and HI maps from this simulation. We then use a Monte Carlo method to calculate the error bars, running 100 of these simulated galaxy and HI maps through our power spectrum pipeline. Subsection 6.4.1 describes these simulations, and Subsection 6.4.2 describes the estimation of the cross-power spectrum and its errors. The full formalism of this process is described in [85].

6.4.1 Simulations

The non-linear dark matter power spectrum at $z = 0.08$ is computed using the HALOFIT routine (based on the method described in [78]) of the Cosmic Linear Anisotropy Solving System (CLASS, [10]). We use Planck 2015 [1] values for the input cosmological parameters. In comoving Cartesian $k$-space, 100 random Gaussian dark matter realizations are drawn from this dark matter power spectrum for each of the four sub-map regions. In order to introduce redshift space distortions (RSDs), the $z$-axis is chosen to be the line-of-sight direction (flat-sky approximation). The Fourier amplitudes are then multiplied by the angle-dependent factor $b(1 + \mu^2)/\sqrt{1 + (k\mu\sigma_v/H_0)}$, where $b$ is the bias of the matter tracer, $\mu$ is the cosine
of the angle between the line-of-sight and the $k$-vector, $\sigma_v$ is the dispersion of the velocity field, $H_0 = 100 \, h \, \text{km s}^{-1} \, \text{Mpc}^{-1}$, and $\beta = f(z)/b$, with $f(z)$ being the dimensionless growth rate. Non-linear RSDs, known as fingers-of-god, are introduced on small scales through the denominator and linear RSDs, known as the Kaiser effect, are implemented through the numerator [58, 40]. We create an HI fluctuation map and an optical galaxy overdensity map from each of the 100 dark matter realizations. For the dispersion of the velocity field, we use $\sigma_v = 500 \, \text{km/s}$, from the fit of [34] to the 2dF correlation function. We choose $b_{\text{HI}} \sim 0.85$, supported by [46]. The optical galaxy bias $b_g$ is expected to be $\sim 1$ for 2dF galaxies at the redshifts of interest [20], so a bias of 1.0 is used for the galaxy over-density maps. Each HI map is also multiplied by an assumed average HI brightness temperature, $T_b = 0.064 \, \text{mK}$. This value comes from the ALFALFA survey [46] measurement of the comoving HI density, $\Omega_{\text{HI}}$, through the equation [15]

$$T_b(z) = 0.39 \frac{\Omega_{\text{HI}}}{10^{-3}} \left[ \frac{\Omega_m + \Omega_\Lambda (1 + z)^{-3}}{0.29} \right]^{-1/2} \left[ \frac{1 + z}{2.5} \right]^{1/2} \, \text{mK}. \quad (6.7)$$

All cosmological quantities are in units of today’s critical density. $\Omega_m$ and $\Omega_\Lambda$ are the the matter and dark energy densities at the present epoch. We use Planck 2015 [1] values again for these. The k-space maps are then converted to regular comoving coordinates. These steps imply a cross-power spectrum in redshift space given by Equation, with a cross-correlation coefficient of unity.

For the next stage of the analysis, the simulated maps must be converted to telescope coordinates of frequency, right ascension, and declination. This requires a fiducial cosmology (we again use Planck 2015) and a gridding scheme. In order to preserve the z-axis as the line-of-sight direction, the conversion of transverse lengths into angular distances in right ascension and declination assumes a constant radial distance, independent of the redshift – for this purpose, the radial distance that halves the volume of the survey is chosen. The maps are then interpolated onto
evenly spaced frequency intervals. An unclustered mock galaxy catalog, following the survey selection function, is added to each galaxy over-density map to approximate the effect of galaxy shot noise. A set of galaxy density maps without this shot noise contribution is also kept. One set of the HI fluctuation maps is kept unaltered, and a second set is convolved with a Gaussian beam of width 1.4 times the largest Parkes beam, equal to the resolution of the common-beam-convolved real radio maps.

6.4.2 Cross-power spectrum

The procedure for estimating the cross-power spectrum of a pair of HI fluctuation and galaxy overdensity maps is as follows. First, the maps are multiplied by their weights: the selection function for the galaxy map, and the inverse noise weights for the HI map. Then, they must be converted from right ascension, declination, and frequency coordinates to physical comoving coordinates. This conversion is the reverse of the procedure described in the second paragraph of Subsection 6.4.1. The z-direction is chosen to be the line of sight. To convert angular distances to transverse distances, the same radius is used for all redshifts – the radius that halves the volume of the survey. This approximation results in a slightly distorted map that occupies a cube in Cartesian comoving coordinates, with the z-direction corresponding to the line-of-sight. Lastly, the map is interpolated onto evenly spaced coordinates in the z-direction.

The 3D power spectrum can then be estimated as

\[ P = \frac{\mathcal{F}(M_{\text{HI}} \circ C^{-1}) \circ \mathcal{F}(M_g \circ S)^*}{N}, \]

where \( \mathcal{F}(X) \) represents the 3D Fourier transform of map \( X \) in comoving coordinates, \( \circ \) denotes element-wise multiplication, \( M_{\text{HI}} \) is the HI fluctuation map, \( M_g \) is the galaxy number overdensity map, \( C \) is the diagonal of the noise covariance of the observed 21-cm map, \( S \) is the galaxy survey selection function, and \( N = \)
\[ \sum_{i,j,k} (C^{-1} \circ S)_{ijk} \] is a normalizing factor, where \( i, j, k \) denote the map voxel coordinates. The 3D power spectrum is then averaged over the azimuthal angle with respect to the line-of-sight and binned to form a 2D power spectrum in \((k_{\perp}, k_{\parallel})\) space. The 2D cross-power spectrum of the real HI and galaxy maps is a robust estimate of the cross-power spectrum, because thermal noise and residual foregrounds are not expected to correlate with the optical survey and tend to average to zero due to the azimuthal average and binning. However, the cross-power will be systematically low due to signal loss from foreground cleaning and convolution with the telescope beam. This loss of power must be accounted for by the transfer function, which is calculated from the simulated maps.

To estimate the transfer function, we follow the procedure described in detail by [85]. The transfer function is calculated in \((k_{\perp}, k_{\parallel})\) space, which is a natural basis given that the foreground cleaning operates in the line-of-sight \( k_{\parallel} \) direction and the beam convolution operates in the \( k_{\perp} \) direction. Schematically, the transfer function is given by

\[
T_\alpha = \frac{\langle Q\left( \mathcal{F}\left\{ [\Pi_{M+s}(M_{\text{pks}} + X_{\text{HI}}^c) - \Pi_{M}M_{\text{pks}}] \circ C^{-1}\right\} \circ \mathcal{F}\{X_{g}^{\text{sn}} \circ S\}^{*}\right)\rangle_\alpha}{\langle Q(\mathcal{F}\{X_{\text{HI}} \circ C^{-1}\} \circ \mathcal{F}\{X_{g} \circ S\}^{*})\rangle_\alpha},
\]

(6.9)

where the operator \( Q()_\alpha \) represents the azimuthal averaging of the 3D k-vectors into \((k_{\perp}, k_{\parallel})\) bins indexed by \( \alpha \), \( M_{\text{pks}} \) is the real Parkes 21-cm map, \( X_{\text{HI}} \) is the simulated HI 21-cm fluctuation map, \( X_{\text{HI}}^c \) is that same simulated map convolved with a Gaussian beam, \( X_{g} \) is the simulated galaxy over-density map, \( X_{g}^{\text{sn}} \) is the simulated galaxy over-density map with galaxy shot noise, and the angled brackets indicate averaging over the 100 simulations. The \( \Pi \) operator represents the removal of the spectral foreground modes, which are calculated via the SVD procedure described in Section 6.3. The \( \Pi_{M} \) operator removes foreground modes that are determined from the common beam convolved real map, and the \( \Pi_{M+s} \) removes foreground modes that are determined...
from the sum of the convolved simulated HI map and the common-beam-convolved real map. The numerator of the transfer function can be considered our best estimate of the cross-power we would expect to measure from the Parkes instrument if our model for the simulated power spectrum were correct. It includes the effects of signal loss from foreground cleaning, residual foregrounds, thermal noise, galaxy shot noise, and beam convolution. When cleaning the simulated map, it is essential to calculate the foreground modes from the sum of the real map and the simulated signal, since the HI signal perturbs the measured foreground modes [85]. The subtraction of $\Pi M_{\text{pks}}$ decreases the variance in the estimation of the numerator. The denominator of the transfer function can be considered our best estimate of the simulated cross-power that would be measured in the absence of foregrounds, thermal noise, and shot noise. Therefore, the transfer function represents the fraction of signal retained in the cross-power after foreground removal and beam convolution.

Our estimate of the observed 2D cross-power spectrum, now compensated for signal loss by the transfer function, is

$$\widehat{P}_\alpha = \frac{Q\left(\mathcal{F}\{[\Pi M_{\text{pks}}] \circ C^{-1}\} \circ \mathcal{F}\{M_g \circ S\}^*\right)}{NT_\alpha} .$$

Similarly, we can calculate each simulated recovered 2D cross-power spectrum:

$$P_{\alpha}^{\text{sim}} = \frac{Q\left(\mathcal{F}\{[\Pi M_{s}(M_{\text{pks}} + X_{\text{HI}})] \circ C^{-1}\} \circ \mathcal{F}\{X_g \circ S\}^*\right)}{NT_\alpha} .$$

The mean of this quantity over 100 simulations represents the 2D cross-power spectrum we would expect to estimate, after compensating for signal loss with the transfer function, if our model for the cross-power were correct. The variance provides an estimation of the expected errors, incorporating thermal noise, foreground residuals, sample variance, signal loss from foreground cleaning, and galaxy shot noise. The number of foreground modes to remove is chosen to minimize this variance. If too
few modes are removed, residual foregrounds boost the variance. If too many modes are removed, most of the signal is also removed and the small transfer function in the denominator boosts the errors. We find that removing 10 SVD modes minimizes this variance.

To display our final results, we average the power spectrum to 1D bins. The average to 1D is weighted by the inverse variance of each $k$-bin across the 100 simulated recovered 2D power spectra. The observed cross-power spectrum, cleaned by removing 10 SVD modes, is shown in Fig. 7-3; we display only 1D Fourier modes for which we have full 2D angular coverage. The uncertainty assigned to each bin of the final observed 1D power is the corresponding standard deviation calculated from the 100 simulated 1D power spectra. The full covariance of the binned 1D power spectrum over the 100 simulations is also checked; we find no significant correlations, so the error bars at each point of figure 7-3 are independent. Each of the four sub-maps is analyzed independently of the others – the final result is an average of these.

As a null test for correlations between residual foregrounds and 2dF galaxies, we randomly shuffle the redshift slices of our 21-cm maps and compute the cross-power spectrum between these shuffled maps and the 2dF maps – we find the cross power is consistent with zero on all scales of interest.
Chapter 7

Results

In this chapter, I present the results of analysis of 21-cm intensity maps from Parkes and GBT.

7.1 GBT

The latest published results from intensity mapping with the GBT come from [49] and [86], which I will summarize here. In [49], GBT HI intensity maps from 190 hours of data on two fields are cross-correlated with the WiggleZ galaxy survey [26]. This cross-power spectrum is compared to the theoretical dark matter power spectrum to constrain the quantity $\Omega_{HI}b_{HI}r$ (using equation 3.85 and 3.92) at $z \approx 0.8$, where $r$ is the HI-galaxy cross-correlation coefficient, $b_{HI}$ is the HI bias, and $\Omega_{HI}$ is the comoving neutral hydrogen fraction at $z \approx 0.8$. In [86], the Schwarz inequality, which constrains $|r| \leq 1$, is utilized so that the cross-power measurement of $\Omega_{HI}b_{HI}r$ can be treated as a lower bound on $\Omega_{HI}b_{HI}$. The foreground cleaned auto-power spectrum is then considered an upper bound on $\Omega_{HI}b_{HI}$, since it may be biased by residual foregrounds. A Bayesian inference with conservative priors on the cross-correlation coefficient and the residual foreground levels constrains $\Omega_{HI}b_{HI} = [0.62^{+0.23}_{-0.15}] \times 10^{-3}$ at 68% statistical confidence, plus an additional 9% estimated systematic calibration.
Figure 7-1: From [86]. The auto-power spectrum of the two GBT fields are shown in blue and green. The solid line shows the power spectrum before foreground removal. Blue and green dotted lines show the power spectrum of the thermal noise in both maps. The noise bias is avoided in the auto-power spectra by cross-correlating maps made with different parts of the data set. Error bars come from thermal noise plus sample variance. The dashed red line is the lower bound on the power spectrum set by the cross-power measurement.

Figure 7-1 shows the GBT auto-power spectrum. Figure 7-2 shows the 68% and 95% confidence intervals placed on the HI auto-power spectrum. These results represent the state-of-the-art in low-redshift 21-cm foreground removal. No other measurements have yet succeeded in removing foregrounds at this level.
Figure 7-2: From [86]. Confidence intervals on the constrained value of the power spectrum are shown. Dark gray shows the 68% confidence intervals, and light gray shows the 95% confidence intervals. The dotted red line is the lower bound set by the cross-power. The solid red line represents the 95% confidence interval that should be reachable based on thermal noise only, in the absence of any foreground residuals.
7.2 Parkes

The observed galaxy-HI cross-power spectrum is shown on the left-hand side of Fig. 7-3. We report an $11.9\sigma$ detection of cross-power. A simple model for the expected cross-power spectrum is the simulated CLASS HALOFIT dark matter power spectrum, with an assumed scale-independent galaxy bias $b_g = 1.0$ [20], HI bias $b_{HI} = 0.85$ [46], and mean 21-cm brightness temperature $T_b = 0.064$ mK [46]. The dashed black curve displays this model. In true comoving space, this model cross-power is

$$P_{HI,g}(k) = T_b b_{HI} b_g r P_{\delta\delta}(k),$$

(7.1)

where $P_{\delta\delta}(k)$ is the CLASS dark matter power spectrum, and $r$ is the galaxy-HI correlation coefficient, defined as

$$r = \frac{P_{HI,g}}{\sqrt{P_{HI,HI}P_{g,g}}}. \quad (7.2)$$

The true cross-power spectrum will also contain shot noise, but we do not attempt to model it, since it depends on the typical HI content of 2dF galaxies, which is unknown. It should be noted that the galaxy-HI correlation coefficient obeys the Schwarz inequality: $-1 \leq r \leq 1$. A value of $|r| < 1$ would mean that HI fluctuations and galaxy over-densities are not simple multiples of each other in Fourier space; one would find phase differences or fluctuating amplitude differences within each k-bin of their Fourier transformed maps. For the dashed black curve of the model, we assume $r = 1$ on all scales.

However, using the Halo Occupation Distribution (HOD) model [7, 98], one can deduce that the expected behavior of the cross-correlation coefficient is scale-dependent. In the HOD model, the power spectrum is the sum of three terms:

$$P(k) = P^{2h}(k) + P^{1h}(k) + P^{SN}(k),$$

(7.3)
Figure 7-3: Left: Observed 1D cross-power averaged over the four Parkes fields and cleaned by removing 10 SVD modes. A circle denotes positive power and a × denotes negative power. The grey line is the mean of the simulations, for which we assume $b_{\text{HI}} = 0.85$ and $T_b = 0.064$ mK, given by the ALFALFA measurement of $\Omega_{\text{HI}}$ [47]. The dashed black line is the corresponding dark matter power spectrum scaled as Eq 3.93. Plotted error bars are 1-σ, derived from the Monte Carlo simulations described in section 6.4.2. Right: The purple points are the average of the auto-power spectra of the 2dF galaxies in the regions that overlap our Parkes maps. Errors are the standard deviation of the mean over the four sub-map regions. The dashed black line is the simulated dark matter power spectrum. The solid grey line is the expected shot noise signal, simulated from 100 unclustered mock catalogs that follow the survey selection function. The green points are the 2dF auto-power data minus the simulated shot noise.
$P^{2h}(k)$ is the 2-halo term, which comes from matter tracers that occupy separate halos; $P^{1h}(k)$ is the 1-halo term, due to clustering of matter tracers within the same halo; and $P^{SN}(k)$ is the shot noise term. On large scales, the 2-halo term is dominant. On intermediate scales, the 2-halo term falls off, and the 1-halo term and shot noise begin to dominate the power spectrum. On the smallest scales, the 1-halo term will also fall off due to the finite extent of the halos, and shot noise will dominate.

Let us now consider the degree to which HI and galaxies will correlate for each of these terms. Since galaxies and HI should both be contained within halos, and the distribution of halos will trace the underlying dark matter density field, we expect $r^{2h} \approx 1$. On the other hand, it is likely that HI and optically selected galaxies have a tendency to occupy different halos. [37] studied the group membership of over 740 overlapping optical galaxies from SDSS and HI galaxies from ALFALFA. They found that only 25% of HI galaxies appear to be associated with an optically identified group, compared to half of optical galaxies. This tendency for HI to occupy different halos suggests that both shot noise and 1-halo clustering may not correlate between HI and optically selected galaxy populations: we expect $r^{1h} < 1$ and $r^{SN} < 1$. Therefore, $r$ is thought to be close to unity on large scales and to fall off on smaller scales, as shot noise and 1-halo clustering begin to dominate the power spectrum.

Now, let us analyze our measured cross-power spectrum. As previously indicated, the dashed black line in the left panel of Fig. 7-3 shows the power spectrum of Equation 7.1 (which includes no shot noise term) with $r = 1$. Redshift space distortions, which modify the power spectrum according to Equation 3.93, are also included; this curve is a binning of this distorted 2D cross-power to 1D with isotropic weights. The grey line shows the signal we’d realistically expect to measure if our model were accurate. It represents the average recovered cross-power spectrum from the 100 simulated galaxy and HI map pairs, including the effects of our window function, thermal noise, residual foregrounds, galaxy shot noise, compensation for signal loss, and anisotropic weighting. The deviations of the data from the grey line in the left panel of Fig. 7-3
indicate disagreement with the simple model of Equations 7.1 and 3.93. In summary, the cross-power is well below the expectation of our model at all scales except for the largest scale \( k \sim 0.07 \ h \text{Mpc}^{-1} \). The two negative points between 0.1 \( h \text{Mpc}^{-1} \) and 0.3 \( h \text{Mpc}^{-1} \) are slightly troubling, since there is no physical reason to expect an anti-correlation between HI and optical galaxies on large scales. However, the deviations from zero are not very significant, and those points are consistent with low positive clustering, or even the simulation line at \( \sim 2\sigma \). The three highest \( k \) points are all statistically significant, and the last two show a substantial decrement in power compared to the model.

For comparison, the right panel of Fig. 7-3 shows the auto-power spectrum of the 2dF galaxies that overlap our Parkes fields. We estimate the shot noise contribution to this power spectrum by averaging the power spectrum of 100 unclustered mock catalogs that follow the survey selection function; the shot noise estimate is the grey curve. The purple points show our calculated 2dF auto-power spectrum, and the green points show this same power spectrum after subtracting the estimated shot noise contribution. The galaxy power spectrum after shot noise removal shows a similar decrement in clustering at high \( k \) to the HI-galaxy cross-power spectrum, but the effect is not as drastic at the two highest \( k \)-bins. Our 2dF power spectrum roughly agrees with the graphed or tabulated 2dF power spectra of [20] and [88] at the points where they overlap, but the overlap is mostly at low \( k \). The smallest scales analyzed in those papers are \( k \sim 0.185 \ h \text{Mpc}^{-1} \) and \( k \sim 0.6 \ h \text{Mpc}^{-1} \) respectively. The analysis of [60] displays the ratio of the 2dF power spectrum to model fits, extending to \( k \sim 1 \ h \text{Mpc}^{-1} \). Their plots show a decrease in 2dF power relative to the models at small scales, which is similar to the effect that we find. However, they ascribe this effect to aliasing from coarse binning. Due to the rather coarse frequency binning of our maps, it is conceivable that aliasing is also responsible for some of the low power we observe at small scales. In order to test this, we bin the galaxy maps first to the same frequency resolution as our HI maps and then with a factor of eight finer
frequency resolution and calculate the power spectrum for both cases. We find a nearly identical galaxy power spectrum, indicating that aliasing is not an issue.

It is likely that some of the low HI-galaxy clustering that we see on small scales is due to the correlation coefficient dipping below 1, since shot noise and 1-halo clustering become more prominent in the power spectrum at high $k$. A reasonable way to test this is to split the galaxies by color, under the hypothesis that the HI content of red galaxies is lower than that of blue galaxies. If this hypothesis is true, we would expect the HI-red galaxy correlation coefficient to drop more rapidly at small scales than the HI-blue galaxy correlation coefficient. Following the k-corrected color splitting method used in [20], we split the 2dF galaxies into red and blue populations and analyze the HI-galaxy cross-power spectra. In our fields, red galaxies account for approximately a third of the total 2dF population, and blue galaxies account for two thirds. The cross-power spectra with the blue and red galaxies are shown in the left panel of Figure 7-4. The two cross-powers are quite similar, except that
there is significantly more power at $k \sim 1.5 \ h\text{Mpc}^{-1}$ in the HI cross-power with the blue galaxies compared to the HI cross-power with the red galaxies. The statistical significance of the difference is about $2.4 \sigma$. This result favors the picture that the cross-power spectrum of HI and optical galaxies at $k \sim 1.5 \ h\text{Mpc}^{-1}$ is dominated either by shot noise or the 1-halo term. Shot noise and 1-halo clustering is seen more strongly when correlating HI with blue galaxies because blue galaxies contain a much larger fraction of the HI and are more likely to occupy the same halos.

The results of our HI cross-power spectra with red and blue galaxies are qualitatively consistent with the findings of [57]. Their analysis of the projected cross-correlation function and auto-correlation function of red and blue SDSS galaxies and HI-selected ALFALFA galaxies reveals that the HI-blue cross-correlation coefficient is close to unity at all scales. On the other hand, the HI-red cross-correlation coefficient is unity at large separations, but it begins to drop at separations smaller than $\sim 5 \ h^{-1}\text{Mpc}$, indicating that the presence of a red galaxy decreases the probability of finding a nearby HI-galaxy, relative to statistically independent dark matter tracers. As noted by [57], this result may reflect the fact that red galaxies tend to preferentially inhabit high density halos [96], which usually have lower fractions of HI gas, as seen in studies of individual groups and clusters [36, 80, 37], hydrodynamic simulations [90], and empirical fits to the halo mass function [56]. A lack of HI mass for many red galaxies is also found in simulations by [92], using a semi-analytical galaxy formation model run on the Millennium simulation. A cross-power spectrum analysis of this simulation reveals a similar scale-dependence and color-dependence to the correlation coefficient.

The analysis of [57] suggests that optically selected blue galaxies and HI galaxies tend to occur in the same environments or perhaps are the same galaxies. To test this, we plot the cross-power spectrum of red and blue galaxies on the right panel of Fig. 7-4, suspecting that it may be similar to the cross-power spectrum of HI and red galaxies. In fact we find that the cross-power spectrum of red and blue galaxies
is more similar to the full galaxy auto-power spectrum with shot noise subtracted. This makes sense, since the blue and red galaxies are disjoint sets. The side-by-side comparison of red-blue galaxy cross-power spectrum and the HI-red cross-power spectrum in Fig. 7-4 suggests that the overlap of HI with red galaxies is even weaker than the overlap of red and blue galaxies.

Three points can be drawn from our results. First, the small-scale clustering amplitude is much lower than the HALOFIT prediction, as seen in both the galaxy-HI power spectrum and the shot-noise subtracted galaxy auto-power spectrum. Second, the galaxy-HI cross-correlation coefficient is scale-dependent and color-dependent, probably due to 1-halo clustering and shot noise. HI appears to be much more strongly associated with blue galaxies than red galaxies. Third, HI-galaxy clustering may be somewhat suppressed at 0.1 \( h \, \text{Mpc}^{-1} \) and 0.3 \( h \, \text{Mpc}^{-1} \) scales.
Chapter 8

Continuing Work

8.1 Parkes Auto-Power Constraints

A similar analysis to the Bayesian inference of the GBT auto-power spectrum is now being conducted on the Parkes data. Since the Parkes measurement extends to smaller scales, one of the goals in this analysis is to characterize the level of the 21-cm shot noise.

8.2 Simulating Instrumental Effects on Foreground Cleaning

There is a considerable difference in the SVD characteristics of real data from the Parkes and GBT telescopes compared to the SVD characteristics of simulated 21-cm signal plus foreground. The situation for real data is best illustrated in figure 8-1, from [85], showing the SVD singular values and principal eigenectors from GBT data. The number of foreground modes that are contaminating by signal is higher than the prediction of simulations, and the frequency structure of the SVD modes is not smooth except for the single largest mode.

Ben Hoscheit and I are attempting to reproduce the SVD characteristics of the
Figure 8-1: From [85]. These plots show the SVD analysis of the frequency-frequency covariance matrix formed by correlating pairs of 21-cm intensity maps of the GBT wide field. The map pairs cover the same location on the sky but are made with different sub-seasons of the data, to avoid thermal noise bias. Left: The solid line shows the square root of the SVD singular values, ordered from largest to smallest and normalized such that the largest singular value is set to 1. The dashed line is an estimate of the SVD spectrum of the noise and variable systematic effects. It represents the SVD of the difference of sub-season maps. The singular values of the dashed noise estimate is also normalized to the largest value of the solid line SVD spectrum. Right: The frequency eigenvectors associated with the SVD spectrum of the GBT maps. The eigenvectors are ordered by their associated singular values, with the largest one on the top. The first vector resembles a smooth power law, but note that there is some non-smooth frequency structure. The second mode comes from mode mixing of angular structure into frequency structure due to the frequency-dependent main-beam and can be removed by a common-beam convolution, assuming a Gaussian beam-model. The projection of the third mode onto the angular structure of the map correlates strongly with polarized emission in the map. Its frequency structure also resembles the frequency structure of measured GBT polarization leakage patterns. Therefore, this mode is likely due to leakage of polarized intensity into the unpolarized maps. The origin of the rest of the modes is unknown. Approximately 30 foreground modes are removed in the analysis of 7.1 to reach the level of signal and thermal noise. The foreground modes that are removed do not resemble smooth polynomials.
GBT data with simulated 21-cm signal and foregrounds provided by Le Zhang, following the Santos model described in section 3.5.2. Since the components of this model lack bright point sources, we have added unclustered bright point sources to our simulation, following the empirically fit flux distribution of [39]. Again following Santos, we assign random power law indexes to each point source from a Gaussian distribution with an average brightness temperature index given by

$$\alpha = -2.83 - 0.13 \log_{10}(S[\text{mJy}]/1000),$$

and a standard deviation of

$$\delta \alpha = 0.2.$$ (8.2)

We have attempted to include possible mode-mixing effects from the frequency-dependent beam. We initially convolved our simulated map with a Gaussian beam with a frequency-dependent beam-width: $\sigma \propto \nu^{-1}$. But since a real beam will have frequency-dependent side-lobes as well as a frequency-dependent main beam, we have attempted to model a more realistic beam-pattern by convolving the map with a 100-meter Airy beam-function; see figure 8-3 for a plot of this frequency-dependent beam-function. The SVD characteristics of maps convolved with the Gaussian beam are shown in figure 8-2, and the SVD of the maps convolved with the Airy beam is shown in figure 8-4.

The mode-mixing effects of side-lobes are demonstrated clearly by the Airy beam convolution. It is likely that effects of this kind are happening with GBT data, but they appear not to be the dominant effect. We are still far from reproducing the GBT SVD spectrum, which has many more high eigenvalue modes without smooth frequency structure. In future work, we will simulate realistic bandpass calibration errors, which may reproduce the lack of smoothness found in the GBT SVD modes. A longer term goal is to realistically model polarization leakage effects. Code developed by Richard Shaw for the analysis of polarization leakage in interferometers [76] attempts
Figure 8-2: The SVD spectrum of simulated maps with 21-cm signal, foregrounds, and point sources, convolved with a frequency-dependent Gaussian beam based on fits to the measured GBT beam with a Gaussian function. Two realizations of Gaussian thermal noise are added to the otherwise identical maps. Left: The square root of the SVD singular values, ordered from largest to smallest and normalized such that the largest singular value is set to 1. Right: The frequency eigenvectors associated with the SVD spectrum of the simulated maps. The first 5 smooth modes are due to foregrounds: the Gaussian beam convolution adds one smooth foreground mode (there are only 4 smooth foreground modes if no beam convolution is applied to the simulated maps). The eigenvectors are ordered by their associated singular values, with the largest one on the bottom (opposite convention of figure 8-1). The modes after the 5 largest are dominated by noise and 21-cm signal.
Figure 8-3: The hypothetical Airy beam of a 100 meter circular aperture. This represents the beam that would be formed by an ideal receiver that uniformly illuminated the Green Bank Telescope. The beam-pattern as a function of angular distance from the center of the beam is displayed for three frequencies in the GBT bandwidth: 700 MHz (blue), 800 MHz (green), and 900 MHz (red). Each beam-pattern is normalized to 1.0 at the center for this plot. Significant mode-mixing can occur when bright structure is in the side-lobes of the beam due to the strong frequency dependence of the side-lobes.
Figure 8-4: The SVD spectrum of simulated maps with 21-cm signal, foregrounds, and point sources, convolved with a 100-meter Airy beam. Two realizations of Gaussian thermal noise are added to the otherwise identical maps. Left: The square root of the SVD singular values, ordered from largest to smallest and normalized such that the largest singular value is set to 1. Right: The frequency eigenvectors associated with the SVD spectrum of the simulated maps. The first 10 modes are due to foregrounds, modified by the systematic effect of the Airy beam convolution. The eigenvectors are ordered by their associated singular values, with the largest one on the bottom (opposite convention of figure 8-1). The first 10 modes are fairly smooth and are easily associated with the smooth foregrounds. The Airy beam is responsible for 5 of these modes: modulation of the smooth power law by the frequency-dependent side-lobes of the Airy beam is visible in the frequency structure of these modes. If a Gaussian beam is used instead, only the first 5 modes smooth modes are present, and the next 5 modes are not at all smooth, apparently dominated by the noise and 21-cm signal. The singular values of these 5 non-smooth modes are also lower with the Gaussian beam-model.
to simulate realistic full-polarization foreground maps. This can be combined with measurements our team has made of the leakage patterns of the GBT [43]. If these effects are still unable to reproduce the GBT SVD spectrum, then we must re-consider our assumptions about the near-separability of the angular and frequency structure of the foregrounds.

8.3 Improving GBT Maps with the GBT-HIM Array

We hope to find funding for the 7-element GBT-HIM array described in section 4.4. In the successful GBT-HIM design review, Tzu-Ching Chang presented forecasts on constraining the BAO wiggles as well as the projected number of FRB detections for a hypothetical 1000 hour survey with GBT-HIM. These forecasts on power spectrum sensitivity to the BAO wiggles can be seen in figure 8.3. Based on the single FRB detection with the horn antenna we also estimate 6 FRB detections in 1000 hours (assuming a Euclidean distribution of FRB sources), following the formula of [22].
Figure 8-7: Forecasts of relative errors (a) and absolute errors (b) on the power spectrum at the location of the BAO wiggles. The purple curve is a hypothetical survey conducted with the current horn receiver, and the other colors show various times and angular coverages with the 7-element array. Only thermal noise is accounted for, so these errors should be viewed as an idealization. Foreground residuals may dominate over thermal noise.
Chapter 9

Conclusions

The 21-cm intensity mapping technique is a new method for efficiently mapping large scale structure (LSS) out to high redshifts by detecting the aggregate 21-cm emission from neutral hydrogen in many galaxies. Early efforts with the Parkes and GBT radio telescopes have succeeded in detecting cosmic structure by cross-correlating 21-cm intensity maps with optical galaxy maps. Both surveys do not have the signal-to-noise to make a detection using only the auto-power spectrum.

Parkes 21-cm maps crossed with the 2dF galaxy survey show a lack of small-scale clustering, which appears to also be present, to a lesser degree, in the 2dF galaxy auto-power spectrum. The Parkes results also show that the galaxy-HI correlation coefficient is scale-dependent and color-dependent. This may indicate that HI and optically selected galaxies do not principally occur in the same halos. If this is the case, then cross-correlating HI maps with red galaxies, in particular, may be a way of measuring 2-halo clustering without a large contribution from 1-halo clustering and shot noise. Placing constraints on the size of the shot noise term in 21-cm intensity mapping with the Parkes auto-power spectrum is the subject of continuing work.

Analysis of GBT intensity maps have constrained the parameter $b_{HI} \Omega_{HI}$ at $z \sim 0.8$. The auto-power spectrum is close to the thermal noise level expected for our survey time at high-k, but it shows possible signs of residual foreground bias at low-k.
The spectrum of the foregrounds detected in the GBT and Parkes data also indicate that foreground removal may be more difficult than simulations suggest. Understanding the origin of this effect through simulation is also the subject of continuing work.

21-cm intensity mapping has the potential to map most of the comoving volume of the observable Universe, even to redshifts in the Dark Ages, before the formation of the first stars and galaxies. Low redshift surveys, at $0 \leq z \leq 3$, can constrain dark energy models by measuring the expansion history. The BAO standard ruler should be fairly simple to measure with intensity mapping. Measuring the growth factor as a function of redshift will be more difficult, since it is degenerate with a potentially time-dependent HI fraction or HI bias. However, it should be possible to constrain the growth factor by detecting redshift space distortions.

The next few years will be an exciting time for low-redshift 21-cm intensity mapping, as many dedicated instruments come online (see section 4.5).
Appendix A

Calibrator Noise

A.1 MBCORR Backend

For the Parkes multibeam receiver, a -35 dB coupler is mounted on the circular waveguide of each beam at the point before the waveguide enters the dewar [83]. This coupler is positioned at 45° to both outputs (polarizations) of the orthomode transducer. The couplers are used to inject a low-level switched calibration signal equally to both polarizations. The cal signal comes from a stable noise diode modulated at 128 Hz. The sky plus cal signal for each beam then enters the dewar, is amplified, mixed to an intermediate frequency, fed by low-loss cables to the control room, filtered to a 64 MHz bandwidth (presumably not filtering away the low frequency 128 Hz diode modulation), and then fed to the digital samplers. The amplifier and filters multiply the power by a gain, $G$, and a frequency dependent bandpass, $B(\nu)$, defined such that $\langle B(\nu) \rangle_{64 \text{MHz}} = 1$. The digital samplers also include total-power detectors and synchronous demodulators which are used in conjunction with the injected switched noise source to measure the total power (which is $\langle G* B(\nu) *[T_{signal(\nu)}+1/2T_{cal(\nu)}] \rangle_{2\text{sec},64\text{MHz}} = \langle G* B(\nu) * T_{sys(\nu)} \rangle_{2\text{sec},64\text{MHz}}$) and the total modulated power (which is just the noise cal signal, $\langle (G* B(\nu) * T_{cal(\nu)}) \rangle_{2\text{sec},64\text{MHz}}$). This effectively measures $T_{sys}$ in units of $T_{cal}$ if one assumes that the noise diode is
stable. In the following calculations, I assume that the gain and bandpass are stable
on the 2 second integration timescale.

The data is now digitally sampled and binned by the correlator. The final data
format divides the sampled data by the coherently measured $T_{\text{cal}}$ and then multiplies
by an assumed calibration temperature, $T_{\text{cal0}}$.

$$T_{\text{data}}(\nu, t) = \frac{\langle B(\nu) * T_{\text{sys}}(\nu, t) \rangle_{2\text{sec},62.5kHz}}{\langle T_{\text{cal}} \rangle_{2\text{sec},64\text{MHz}}} T_{\text{cal0}} \quad (A.1)$$

The native frequency bin size is 62.5 kHz. The noise cal measurement is taken
over the full 64 MHz bandwidth. Note that this equation assumes that $T_{\text{cal}}$ is
a flat function of frequency, so that $\langle G * B(\nu) * T_{\text{cal}}(\nu) \rangle_{2\text{sec},64\text{MHz}} = G * \langle B(\nu) * T_{\text{cal}} \rangle_{2\text{sec},64\text{MHz}} = G * \langle T_{\text{cal}} \rangle_{2\text{sec},64\text{MHz}}$. Our pipeline bins the signal to larger 1 MHz
frequency bins before making maps.

$$T_{\text{rebinned}}(\nu, t) = \frac{\langle B(\nu) * T_{\text{sys}}(\nu, t) \rangle_{2\text{sec},1\text{MHz}}}{\langle T_{\text{cal}} \rangle_{2\text{sec},64\text{MHz}}} T_{\text{cal0}} \quad (A.2)$$

For the Parkes central beam, $T_{\text{sys}} \approx 21 K$ and $T_{\text{cal}} \approx 1.2 K$ [4].

### A.2 Extra Noise Induced

In the following calculations, I ignore the amplifier gain and assume that it is stable
enough over each 2 second integration to contribute negligible error to the measure-
ments.

The measurement of the noise diode can be thought of as the cal on power minus
the cal off power, each measured over a total time of 1 second:

$$\langle T_{\text{cal}} \rangle_{2\text{sec},64\text{MHz}} = \langle T_{\text{signal}} + T_{\text{cal}} \rangle_{1\text{sec},64\text{MHz}} - \langle T_{\text{signal}} \rangle_{1\text{sec},64\text{MHz}} \quad (A.3)$$

Using the radiometer equation, and assuming cal on and off measurements are
uncorrelated (I think this is fair, since the bandwidth is much higher than the 128 Hz
noise diode switching),

\[
\sigma_{\text{cal}}^2 = 2 \frac{T_{\text{signal}}^2 + [T_{\text{signal}} + T_{\text{cal}}]^2}{2 \text{sec} \times 64 \text{MHz}} \approx \frac{1}{16} \frac{T_{\text{sys}}^2 + 1/4T_{\text{cal}}^2}{2 \text{sec} \times 1 \text{MHz}} \approx \frac{1}{16} \frac{T_{\text{sys}}^2}{2 \text{sec} \times 1 \text{MHz}} \quad (A.4)
\]

Now we can calculate errors for equation A.2. I will assume that the errors in the measurement of \( T_{\text{cal}} \) are uncorrelated with the errors in the \( \langle T(\nu, t) \rangle_{2 \text{sec}, 1 \text{MHz}} \) measurement. I think this is justified by the bandwidth difference and the fact that \( T_{\text{cal}} \) subtracts two measurements of \( T_{\text{signal}} \).

\[
\sigma_{\text{rebbined}}^2 = \frac{T_{\text{cal}0}^2}{(T_{\text{cal}}/2 \text{sec}, 64 \text{MHz})^2} \frac{\langle B(\nu)T_{\text{sys}}(\nu) \rangle_{2 \text{sec}, 1 \text{MHz}}^2}{2 \text{sec} \times 1 \text{MHz}} + \frac{\langle B(\nu)T_{\text{sys}}(\nu) \rangle_{2 \text{sec}, 1 \text{MHz}}^2T_{\text{cal}0}^2}{(T_{\text{cal}}/2 \text{sec}, 64 \text{MHz})^4} \sigma_{\text{cal}}^2
\]

If we assume \( T_{\text{sys}} \) is an approximately flat function of frequency and use the fact that \( \langle T_{\text{sys}} \rangle \approx 21K \) and \( \langle T_{\text{cal}} \rangle \approx T_{\text{cal}0} \approx 1.2K \),

\[
\sigma_{\text{rebbined}}^2 \approx \langle B(\nu) \rangle_{2 \text{sec}, 1 \text{MHz}}^2 \left[ \frac{T_{\text{sys}}^2}{2 \text{sec} \times 1 \text{MHz}} + \frac{T_{\text{sys}}^2}{T_{\text{cal}0}^2} \frac{1}{16} \frac{T_{\text{sys}}^2}{2 \text{sec} \times 1 \text{MHz}} \right] \approx 20 \frac{\langle B(\nu) \rangle_{2 \text{sec}, 1 \text{MHz}}^2T_{\text{sys}}^2}{2 \text{sec} \times 1 \text{MHz}} \quad (A.5)
\]

So, the calibration scheme increases the noise on each pixel by a factor of about 4.5 compared to a perfectly stable radiometer with no noise cal.

It should be noted that the dominant noise term in equation A.6, from the error in the noise diode measurement, will be perfectly correlated across all frequencies and will likely follow the frequency structure of Tsys. This should be strongly suppressed by the foreground cleaning, and the first mode is likely to remove nearly all of this term. So it should still be possible to clean to the naive radiometer thermal noise level. Estimates of the noise from the mapmaker, though, would be dominated by this term, even if the foreground bias is removed by doing a second round of map making.
Bibliography


