Lecture Notes: Data Science in Radio Astronomy I

Objectives

This module introduces key data science techniques in radio astronomy, focusing on the analysis of spectral data from the 21cm hydrogen line. Students will explore the astrophysical significance of the 21cm line, its role in SETI, and the unique capabilities of the Allen Telescope Array (ATA). The module begins with an introduction to sampling theory, the Nyquist theorem, the binary file format, and handling large datasets. It then transitions to frequency-domain analysis, covering Fourier transforms (DFT and FFT) and power spectral density (PSD) to identify spectral features and mitigate noise. The final lecture focuses on advanced signal interpretation, including isolating regions of interest, understanding Doppler shifts, and constructing velocity profiles to study galactic rotation and dark matter. Each lecture incorporates a Jupyter Notebook coding exercise, where students simulate and analyze signals using Python.

Lecture 1: The 21cm Line and Digitization in Radio Astronomy

The Physics and Applications of the 21cm Hydrogen Line

- **Emitted by a hyperfine transition in neutral hydrogen:** The 21cm line is emitted due to a hyperfine transition in neutral hydrogen atoms, where the electron's spin orientation relative to the proton flips from aligned to anti-aligned.
- **Significant due to hydrogen's abundance and dust penetration:** This line is particularly significant because hydrogen is the most abundant element in the universe, and the 21cm wavelength penetrates dust clouds, allowing astronomers to peer through the gas that would be opaque to optical wavelengths.
- **Mapping hydrogen in galaxies:** By measuring the 21cm emission at different frequencies, astronomers can map the distribution of hydrogen in the Milky Way and other galaxies, revealing their spiral structures and overall shape.
- **Revealing galactic rotation and dark matter:** The Doppler shift of the 21cm line provides insights into the velocity of hydrogen gas, helping astronomers study galactic rotation curves and infer the presence of dark matter.
- **Key frequency for SETI:** The 21cm hydrogen line is considered a "universal frequency" for communication, as hydrogen is the most abundant element in the universe. It's a prime candidate for detecting signals from extraterrestrial intelligence.

Historical Context and Discoveries

- **Predicted in 1944:** The 21cm line was first predicted by Dutch astronomer Hendrik van de Hulst, who theorized that neutral hydrogen atoms could emit radio waves due to hyperfine transitions.

- **First Detected in 1951:** The 21cm line was first observed independently by Harold Ewen and Edward Purcell at Harvard, confirming van de Hulst's prediction.
- **Post War War II Development:** Advances in radar technology during World War II played a significant role in enabling the development of radio astronomy, which was crucial for detecting the 21cm line.
- **Birth of Radio Astronomy:** The detection of the 21cm line marked a milestone in the development of radio astronomy as a major field of study.
- **Mapping the Milky Way:** Early observations using the 21cm line provided the first detailed maps of the structure of the Milky Way, including its spiral arms.

The Role of the ATA in Modern Astronomy

- **Designed for SETI and Radio Astronomy:** The ATA is an array of 42 radio telescopes uniquely designed to perform both SETI observations and traditional radio astronomy studies.
- **Interferometric Capabilities:** The ATA combines signals from its individual antennas using interferometry, creating a large virtual telescope with enhanced resolution and sensitivity.
- **Phased Array for Simultaneous Observations:** Its phased array capabilities allow simultaneous observation of multiple targets, significantly improving efficiency and sensitivity for detecting faint signals.
- **Wide Frequency Coverage:** The ATA's unique log-periodic feeds enable observations across a wide range of frequencies (1–10 GHz), covering a significant portion of the "radio window" that passes through Earth's atmosphere. This range includes the 21cm hydrogen line at approximately 1.4204 GHz.
- **SETI Observations:** The ATA conducts dedicated SETI observations, searching for various signal types that could indicate the presence of extraterrestrial technology.
- **Contributions to Traditional Radio Astronomy:** The ATA also studies a wide array of astrophysical phenomena, including pulsars, fast radio bursts, quasars, and supernova remnants, enhancing our understanding of the universe.

Sampling and the Nyquist Theorem

- **Sampling:** Sampling is the process of converting a continuous signal into discrete data points, which is essential for digital signal processing (DSP).
- **How Sampling Works:** The amplitude of the analog signal is sampled (measured) by an analog-to-digital converter (ADC) at constant time intervals, and the values are stored in a digital format. This allows sophisticated software algorithms to process the signal using DSP techniques.
- **Nyquist Theorem:** According to the Nyquist theorem, the sampling rate must be at least twice the highest frequency present in the signal to avoid aliasing. The relationship is described by the equation:

$$f_s \geq 2 * f_N \tag{1}$$

where f_s is the sampling frequency, and f_N is the Nyquist frequency, the highest frequency that can be sampled without distortion.

- **Aliasing and its Effects:** Aliasing occurs when the sampling rate is below the Nyquist frequency, causing false lower frequencies to appear in the data.

- **Demonstrating Aliasing:** The effect of aliasing can be observed in the GNU Radio Companion file `Nyquist.grc`. By increasing the signal frequency past twice the sample rate, a lower frequency appears in the Fourier transformed plot, demonstrating aliasing in real time.

Discussion Questions

1. **How does the ability of the 21cm line to penetrate dust contribute to our understanding of the galaxy?**

   - The 21cm line's ability to penetrate interstellar dust allows astronomers to observe regions of the galaxy that are invisible in optical wavelengths. This makes it possible to study the structure and dynamics of the Milky Way, including its spiral arms and dense molecular clouds, which are often hidden behind dust. By analyzing the 21cm line, we can map the distribution of neutral hydrogen and gain insights into galactic formation and evolution.

2. **What makes the ATA uniquely suited for both SETI and traditional radio astronomy?**

   - The ATA combines advanced interferometric capabilities with phased array technology, enabling simultaneous observation of multiple targets and enhanced sensitivity. Its wide frequency range (1–10 GHz) covers key regions of the radio spectrum, including the 21cm hydrogen line, allowing studies of phenomena such as pulsars, fast radio bursts, and potential extraterrestrial signals. Purpose-built for SETI, the ATA's unique design bridges traditional radio astronomy and the search for extraterrestrial intelligence.

3. **You are observing a signal between 25-35 Hz, with an analog filter to block frequencies outside this range. According to the Nyquist theorem, what is the minimum sample rate needed to accurately digitize this signal?**

   - According to the Nyquist theorem, the sampling rate must be at least twice the highest frequency in the signal. In this case, the highest frequency is 35 Hz, so the minimum sampling rate required is $2 \times 35 = 70$ Hz. Sampling at or above this rate ensures accurate digitization of the signal without aliasing.

Lecture 1 Resources

1. The Hydrogen 21cm Line - Hyperphysics
2. Milkyway in 21cm
3. Measurement of the Milky Way Rotation
4. Mapping Galactic Hydrogen
5. Analog-to-digital converters basics
6. Aliasing - Digital Signals Theory

Lecture 2: Data Storage and Frequency Analysis in Radio Astronomy

Data Storage in Radio Astronomy

- **Binary Format: Simple and Versatile:** Binary files store data in its rawest form, making them efficient for writing and reading large datasets. They preserve the exact numeric values of the observed data without additional overhead.
- **Why Binary is Used in AGISETI:** Binary files integrate seamlessly with GNU Radio and Python, making them ideal for educational purposes. They are easy for students to write and read without requiring specialized libraries.
- **Professional File Formats:** Formats like filterbanks, HDF5, and FITS are widely used in professional radio astronomy. These formats include metadata, compression options, and compatibility with advanced analysis tools, making them essential for large-scale research.
- **Filterbank Files:** Commonly used for SETI and pulsar observations, filterbank files organize spectrally resolved data with metadata like timestamps and frequency resolution. This file format will be used for data in the second data science module.
- **HDF5 Files:** Designed for managing large, complex datasets, HDF5 files support hierarchical organization, compression, and efficient data access, ideal for dynamic spectra and imaging.
- **FITS Files:** The standard in astronomy for imaging and spectral data, FITS files ensure detailed metadata and interoperability between analysis tools.
- **Choosing the Right Format:** Binary files are simple and efficient for educational purposes, while professional formats enhance usability, metadata inclusion, and research collaboration.

Fourier Transforms and the Frequency Domain

- **What is a Fourier Transform?** The Fourier transform is a mathematical tool that decomposes a signal into its constituent frequencies, providing insights into the frequency content of the signal.
- **Analogy with Optical Spectroscopy:** A Fourier transform can be thought of playing the role of a prism or diffraction grating in optical spectroscopy.
- **Importance in Radio Astronomy:** In radio astronomy, Fourier transforms are used to convert time-domain signals into the frequency domain, enabling analysis of spectral features like hydrogen line emissions and pulsar timing signals.
- **Discrete Fourier Transform (DFT):** While a traditional Fourier transform operates on mathematically defined signals, a discrete Fourier transform (DFT) converts a discrete, finite time-domain signal into its frequency-domain representation, showing how signal energy is distributed among discrete frequency components, allowing sampled data to be converted into the frequency domain.
- **Fast Fourier Transform (FFT):** The FFT is a computationally efficient algorithm for performing a discrete Fourier transform, essential for processing large datasets in real-time applications.
- **Demonstration of FFT:** The effects of a fast Fourier transform can shown using the included

GNU Radio Companion file `Fourier_Transform.grc`. This flowgraph adds 5 sinusoids with different frequencies and amplitudes together with a noise source, and visualizes the resulting signal in both the time-domain and frequency-domain. The individual frequency components that compose the signal can be clearly seen in the Fourier transformed data.

Introduction to Spectral Data

- **Definition of Spectral Data:** Spectral data represents how the power of a signal is distributed across different frequencies, obtained by applying a Fourier transform to a time-domain signal.
- **Applications in Astronomy:** Spectral data is essential for studying atomic emission and absorption lines, synchrotron radiation, and transient signals like fast radio bursts and pulsars.
- **Spectral Resolution:** The ability to distinguish closely spaced frequencies depends on observation duration, sampling rate, and the size of the FFT, with higher resolution enabling finer distinctions.
- **Relevance to Radio Astronomy:** Spectral data is used to map neutral hydrogen, detect narrowband signals in SETI, and differentiate between meaningful signals and noise.
- **Application in this Module:** In the lab accompanying these lectures, students will analyze spectral data of the 21cm Hydrogen emission line from the Milky Way. Analyzing spectral data reveals the Doppler shift of the hydrogen line, allowing students to distinguish emission from distinct spiral arms of the galaxy, and show that the galaxy is rotating.=

Power Spectral Density (PSD)

- **Definition of PSD:** PSD represents how the power of a signal is distributed across frequencies, providing a quantitative measure of signal strength in the frequency domain.
- **Amplitude vs. PSD:** Amplitude refers to the magnitude of a signal, representing its peak value at specific points in time or frequency. In contrast, PSD quantifies the average power content of a signal as a function of frequency, providing a measure of energy distribution over time.
- **Magnitude vs. Power:** Magnitude is the absolute value of the signal at a given point and is measured directly in units such as volts or intensity. Power is proportional to the square of the signal's amplitude, providing a measure of energy. In the frequency domain, power is often normalized by frequency to create a continuous distribution.
- **Relevance of PSD:** Expressing data as PSD is crucial for analyzing signals in noisy environments. By converting amplitude to power, PSD highlights weak features, such as the 21cm hydrogen line or pulsar signals, that may be buried in noise when viewed in the time domain.
- **Signal Width and Spectral Resolution:** The width of a signal in the frequency domain is inversely related to the duration of the observation in the time domain. A longer observation time results in a narrower signal peak. Spectral resolution is determined by the FFT size:

$$\Delta f = \frac{1}{T_{obs}} = \frac{F_s}{N_{FFT}}$$

A smaller Δf (higher spectral resolution) allows better distinction of closely spaced or narrow signals.

Discussion Questions

1. **What advantages does analyzing a signal in the frequency domain provide compared to analyzing it in the time domain?**
   - The frequency domain provides a way to decompose a complex signal into its constituent frequencies, revealing patterns and features that are not easily visible in the time domain. For example, periodic signals appear as distinct peaks in the frequency domain, whereas in the time domain, they may be buried in noise. This is particularly useful in radio astronomy, where weak signals like the 21cm hydrogen line or pulsar signals can be distinguished from noise using frequency analysis.
2. **How is the Fourier transform similar to or different from the function of a prism in optical astronomy? What does this analogy tell us about the role of frequency analysis?**
   - A prism in optical astronomy disperses light into its constituent colors (frequencies), allowing astronomers to analyze spectral lines and deduce physical properties of stars and galaxies. Similarly, a Fourier transform breaks a signal into its frequency components, creating a "spectrum" of the signal. This analogy highlights that frequency analysis, like a prism, reveals hidden information about the signal's composition and behavior. In radio astronomy, this enables detection of specific signals like atomic hydrogen emission, synchrotron radiation, or artificial signals for SETI.
3. **Why is it necessary to carefully choose the observation duration, FFT size, and sampling rate when performing a radio astronomy observation?**
   - The observation duration, FFT size, and sampling rate all influence the frequency resolution and accuracy of the analysis. A longer observation duration improves frequency resolution, allowing narrower signals to be resolved. The FFT size determines the number of frequency bins: a larger FFT provides finer spectral resolution. The sampling rate must be high enough to satisfy the Nyquist theorem (at least twice the highest signal frequency) to avoid aliasing. Careful choice of these parameters ensures that signals of interest are resolved accurately without distortion or loss of information.

Lecture 2 Resources

1. Reading and Writing Binary Data in GNU Radio and Python
2. Introduction to Working with Filterbank Files
3. An Interactive Guide To The Fourier Transform
4. The Ultimate Guide to Frequency Analysis
5. Fourier Analysis and Radio Astronomy
6. What is Power Spectral Density?

Lecture 3: Advanced Signal Analysis in Radio Astronomy

Doppler Shifts and Velocity Profiles

- **Doppler Shifts in Astronomy:** The Doppler effect describes the change in frequency or wavelength of a wave in relation to an observer moving relative to the wave source. In astronomy, Doppler shifts are used to determine the radial velocity of astronomical objects.

- **Velocity Calculation:** The radial velocity (v) of the source can be determined using the formula:

$$v = c \cdot \frac{f_{\text{rest}} - f_{\text{observed}}}{f_{\text{rest}}}$$

  where:

  - v is the radial velocity of the source .
  - c is the speed of light (about 3×10^8 m/s or 300,000 km/s).
  - f_{observed} is the frequency observed by the telescope.
  - f_{rest} is the rest frequency of the source (1420.40575 MHz for hydrogen).

- **Redshift ($v > 0$):** Occurs when the source is moving away from the observer, causing the observed frequency (f_{observed}) to be lower than the rest frequency (f_{rest}).

- **Blueshift ($v < 0$):** Occurs when the source is moving toward the observer, causing the observed frequency (f_{observed}) to be higher than the rest frequency (f_{rest}).

- **Analyzing the Milky Way with Doppler Shifts:** By analyzing the Doppler-shifted 21cm hydrogen line at different galactic longitudes, students can map the motion of hydrogen gas and identify features of the Milky Way. The observed properties of the line and their physical causes include:

- **Line Strength (Intensity):** The strength of the 21cm line depends on the column density of neutral hydrogen gas along the line of sight. Regions with higher amounts of hydrogen produce stronger emission lines because there are more hydrogen atoms undergoing the hyperfine transition. Strong lines indicate denser gas clouds or longer paths through regions of neutral hydrogen.

- **Line Broadening:** The width of the line can be broadened due to multiple physical effects:

  - **Thermal Motion:** The random thermal motion of hydrogen atoms causes slight Doppler shifts in the emitted photons, leading to a broadened line. Higher gas temperatures result in greater thermal velocities and broader lines.
  - **Turbulent Motion:** Turbulence within the gas introduces additional velocity dispersion, further broadening the line. This is common in dynamic regions such as spiral arms.
  - **Bulk Motion of Gas Clouds:** If gas clouds along the line of sight are moving with slightly different velocities, their combined emission produces a broader line.

- **Spiral Arm Features:** Distinct peaks in the spectral data correspond to hydrogen gas in different spiral arms of the Milky Way. The Doppler-shifted frequencies reflect the radial velocities of gas in

these arms relative to the observer. Each spiral arm has its own characteristic velocity profile due to its motion within the rotating galaxy, and these features appear as separate peaks in the spectral data.

Curve Fitting in Spectral Data

- **What is Curve Fitting?** Curve fitting is the process of modeling observed data using mathematical functions to extract key parameters. In radio astronomy, curve fitting is especially used to analyze spectral lines and determine properties like center frequency, amplitude, and line width.

- **Why Use Curve Fitting?** Curve fitting enables astronomers to model complex signals and extract meaningful information. By approximating spectral features with well-defined mathematical functions, curve fitting helps quantify key properties of the data, separate signals of interest from noise, and compare observed results to theoretical models.

- **The Gaussian Function for Spectral Lines:** Spectral lines often exhibit a Gaussian shape due to the combined effects of thermal motion, instrumental broadening, and natural line emission. The Gaussian function is described as:

$$I(f) = A \exp\left(-\frac{(f - f_c)^2}{2\sigma^2}\right)$$

  where:

  - $I(f)$: Intensity at frequency f,
  - A: Amplitude of the Gaussian,
  - f_c: Center frequency of the Gaussian,
  - σ: Standard deviation, which is related to the line width.

- **Physical Meaning of Fit Parameters:**

  - **Center Frequency (f_c)**: Corresponds to the Doppler-shifted position of the line. By comparing f_c to the rest frequency, we can calculate the radial velocity of the gas using the Doppler formula.
  - **Amplitude (A)**: Reflects the intensity of the signal, which is proportional to the column density of hydrogen gas along the line of sight. Larger amplitudes indicate a larger amount of gas in the radial direction.
  - **Line Width (σ)**: Provides insights into the physical conditions of the gas, related to the thermal properties and motion of the gas.

- **Curve Fitting in Practice:** In this module, curve fitting will be applied to spectral data of the 21cm hydrogen line. By fitting a Gaussian to the observed spectral line, the center frequency will reveal the radial velocity of the gas, the amplitude will provide an estimate of the hydrogen column density, and the line width will indicate the temperature, turbulence, or dispersion in the gas.

Noise Reduction and Welch's Method

- **Importance of Noise Reduction:**
  - Noise reduction techniques are essential for detecting faint signals in radio astronomy and SETI. It can improve signal-to-noise ratio (SNR) and help distinguish weak signals from the noise.
- **Overview of Welch's Method**
  - One noise reduction technique we'll explore here is Welch's method, which is especially useful for spectral data. Welch's method estimates the power spectral density by dividing the time-domain signal into smaller, overlapping segments, then applies a window function to each segment to reduce edge effects. The fast Fourier transform is then computed for each segment, which converts from the time-domain to the frequency-domain, then averages the power spectra of all segments, reducing random noise at the cost of frequency resolution.
  - **Advantages:** Reduces noise across the entire frequency range, unlike filters which focus on specific regions. Improves the visibility of narrowband signals in noisy data, which is very important in SETI searches, and produces a smoother and more accurate PSD by suppressing random variations.
  - **Disadvantages:** Segmenting the signal into smaller parts inherently reduced the spectral resolution on the FFT; shorter segments means fewer frequency bins, making it harder to distinguish closely spaced signals.
  - **Key Parameters:**
    - **Segment Length:** Shorter segments enhance noise reduction, but lower spectral resolution; longer segments improve spectral resolution but reduce the effectiveness of noise reduction.
    - **Overlap:** Ensures continuity between segments and reduces data loss.

Discussion Questions

1. **What physical properties of hydrogen gas can be determined from the shape and position of the 21cm line, and how do these relate to the motion of the Milky Way?**

   The center frequency of the 21cm line provides information about the radial velocity of the gas through the Doppler effect. The amplitude (line strength) reveals the column density of hydrogen gas along the line of sight, indicating denser regions like spiral arms. The line width can indicate thermal motion (broader lines for hotter gas), turbulent motion, or multiple gas clouds moving with slightly different velocities. By analyzing these properties, astronomers can map the motion of gas in the Milky Way and identify features like spiral arms.

2. **Suppose you observe two hydrogen lines with similar amplitudes but very different widths. What might this tell you about the physical conditions of the gas?**

   If two hydrogen lines have similar amplitudes but different widths, the broader line likely originates from gas with higher thermal motion, more turbulence, or multiple gas clouds at slightly different velocities. In contrast, the narrower line suggests colder gas with lower thermal velocities and less turbulence.

Lecture 3 Resources

1. Doppler Effect Simulator
2. Hydrogen Line Profile Database - University of Bonn
3. Spectral Line Broadening
4. Scipy Curve Fit Documentation
5. Video: Welch's method for smooth spectral decomposition