UNDERSTANDING POLARIZATION AS A FOREGROUND FOR HI EPOCH OF REIONIZATION MEASUREMENTS

Saul Aryeh Kohn

A DISSERTATION

in

Physics and Astronomy

Presented to the Faculties of the University of Pennsylvania

in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

2018

Supervisor of Dissertation	Graduate Group Chairperson
James E. Aguirre	Whoever the graduate chair is
Associate Professor of Physics and Astronomy	Professor of Physics and Astronomy

Dissertation Committee:

Adam Lidz, Associate Professor of Physics and Astronomy

Masao Sako, Associate Professor of Physics and Astronomy

another professor, Assistant Professor of Physics and Astronomy

another professor, Professor of Physics and Astronomy

for my grandparents, endless sources of inspiration

UNDERSTANDING POLARIZATION AS A FOREGROUND FOR HI EPOCH OF REIONIZATION MEASUREMENTS

Acknowledgments

Acknowledgements require a certain mindset to be written well.

ABSTRACT

UNDERSTANDING POLARIZATION AS A FOREGROUND FOR HI EPOCH OF REIONIZATION MEASUREMENTS

Saul A. Kohn

James E. Aguirre

Abstracts are written last.

Contents

List of Tables

List of Figures

Part I

Introduction & Mathematical Formalisms

Chapter 1

The Epoch of Reionization

Chapter 2

Astrophysical Polarization

Chapter 3

Interferometry

Chapter 4

Instrumental Polarization

Chapter 5

Instruments

5.1 Instruments used in this work

5.1.1 The Donald C. Backer Precision Array for Probing the Epoch of Reionization (PAPER)

5.1.1.1 PAPER-32 redundant array

5.1.1.2 PAPER-32 polarized imaging array

5.1.1.3 PAPER-64

5.1.1.4 PAPER-128

5.1.2 The Hydrogen Epoch of Reionization Array (HERA)

5.1.2.1 HERA-19 Commissioning Array

5.1.2.2 HERA-47

5.1.2.3 Future HERA Build-Outs

5.2 Other current and future low-frequency interferometers

5.2.1 The Low Frequency Array (LOFAR)

5.2.2 The Murchinson Widefield Array (MWA)

Part II

Outer space in Fourier space

Chapter 6

Peering through the EoR Window

Chapter 7

Data Preparation and Processing

The data volume of interferometric measurements inherently scale as the square of the number of antennas in the array (N_{ant}). Not only does the sheer volume of data from large-N_{ant} arrays pose a problem for data storage, but also it requires precise and efficient efforts to quality assure (QA) the data.

In this chapter, I will outline some of the efforts involved in data preparation, preprocessing and QA that are required for an EoR power spectrum estimate.

7.1 Data Compression

The PAPER-128 correlator produced 288 MIRIAD files per night. Each of these contained 8126 baselines, and each baseline contained visibilities over 1024 98 kHz frequency channels and 56 10 s time integrations. The four instrumental polarizations were in separate files. In sum, each file was 4.2 GB which meant that each night 1.2 TB of data were recorded.

In order to efficiently transport the data over Gigabit Ethernet from the Karoo Radio Quiet Zone (KRQZ) to Cape Town, and from Cape Town under transatlantic cables to Philadelphia, some compression was required. It was also required that such a compression, while lossy, did not effect the targeted cosmological signal.

7.1.1 Delay–Delay-Rate Filtering

The compression algorithm implemented for PAPER observations, Delay–Delay-Rate (DDR) filtering, was introduced in Parsons and Backer (2009) described in Parsons et al. (2014), and we briefly review it below.

The geometric delay of a celestial signal, originating form direction $\hat{s}$, incident on an interferometric baseline described by vector $\vec{b}$, is

$$\tau_g = |\vec{b} \cdot \hat{s}|/c \tag{7.1}$$

where c is the speed of light. This relationship implies that τ_g is bounded for a given baseline

$$-|\vec{b}|/c \leqslant \tau_g \leqslant |\vec{b}|/c \tag{7.2}$$

Equation 7.2 therefore gives the maximum value of $|\tau_g|$ physically meaningful for a given array – the maximum baseline length in that array, divided by c. For PAPER, the maximum baseline length is 300 m, corresponding to $\max(|\tau_g|) = 1\mu$s. As reviewed in Chapter 6, the delay axis may be accessed by Fourier transforming a visibility along the frequency axis. Once in delay space, power at delays larger in magnitude than 1μs could be removed. With a sufficiently large frequency bandwidth, this would not produce aliased signal, according to the critical Nyquist rate. By using the 1μs as a delay bound for all visibilities, the frequency axes of all compressed visibilities remained the same (reduced in number from 1024 to 203), which while sub-optimal from a compression point of view, allowed for ease of programming at later stages.

A similar geometric bound can be obtained by Fourier transforming the time axis of visibilities, provided that they were obtained in drift-scan mode (see Chapter 3). Parsons and Backer (2009) showed that the rate at which the geometric delay on an interferometric baseline changes is governed only by the position of the array on Earth, and the Earth's rotation:

$$\dot{\tau}_g = -\frac{\omega_\oplus \cos\delta}{c}\left(b_x \sin\alpha + b_y \cos\alpha\right) \tag{7.3}$$

where $\omega_\oplus$ is the angular frequency of the Earth's rotation, α and δ are the hour-angle and declination of a point on the celestial sphere, respectively, and $\vec{b} = (b_x, b_y, b_z)$ is the baseline vector expressed in equatorial coordinates.

For arrays not close to the geographic poles, $|b_y| \gg |b_x|$, there is a maximum rate of change (corresponding to (α, δ) = (0, 0)), producing a bound on $\dot{\tau}_g$:

$$-\omega_\oplus |b_y|/c \leqslant \dot{\tau}_g \leqslant \omega_\oplus |b_y|/c \tag{7.4}$$

for a 300 m East-West baseline, the maximum delay-rate is approximately $\max(|\dot{\tau}_g|) = 0.07\,\mathrm{ns\,s^{-1}}$. This delay-rate was not Nyquist sampled by a single PAPER file: requiring the previous and next files generated for that polarization to be appended on either side of each visibility's time axis to prevent aliasing from the decimation. For the large scale processing of months of data, this required a software pipeline described in Section 7.1.2.

There are also other issues with DDR compression, largely associated with instrument systematics. Delay transforms rely on the fact that the bright foregrounds that dominate the measured signal are spectrally smooth, and that the frequency response of the instrument is also spectrally smooth: this of course is the basis for the EoR window paradigm reviewed in Chapter 6. Likewise, delay-rate filtering assumes temporal smoothness. Radio Frequency Interference (RFI) signals created by human communications violate both models of smoothness, since they are typically confined to narrow bandwidths (creating sharp spikes along the frequency axis) and may be transient (creating sharp spikes along the time axis). This requires steadfast identification and flagging algorithms for RFI (see Section 7.2), and some variety of interpolation, fitting, or CLEANing across the flagged regions prior to compression.

7.1.2 Software Implementation

The first season of PAPER-128 data, due to a variety of circumstances, required compression on the computing cluster at the University of Pennsylvania. The raw data were stored on a high-volume drive that was able to connect with the cluster via a low-speed switch. The hardware capable of performing any sort of high-performance processing (i.e. holding the data in RAM) were ten “compute nodes" connected to the cluster via a high-speed switch, and mounted in an NFS architecture. The compute nodes could only hold ~ 10 PAPER-128 files in storage.

The processing stages for compression of a night of PAPER data, described below, required knowledge of the location and compression state of not only individual files, but also the neighbors-in-time of the file in question, in order to implement the DDR filter described above. To supervise the compression we created a MySQL database, which we interacted with via Shell and Python scripts. The database contained a table for the data files under processing and their compression state, a table of neighbor-relations, a table of file details, and a table of the processing nodes available. The schema of this database is shown in Figure 7.1.

To implement the compression, per file, the following steps were required:

1. Copying the file from the storage volume to the cluster. For a single night of data, this required roughly 8 hours.

2. Copying the file from the cluster to the compute node. This required roughly 5 minutes.

3. Generate copy of the file, with metadata corrections. This required roughly 1 minute.

4. Delete the raw file.

5. RFI-flag the high frequency-resolution data. This required roughly 2 minutes.

6. Delete the metadata-corrected file.

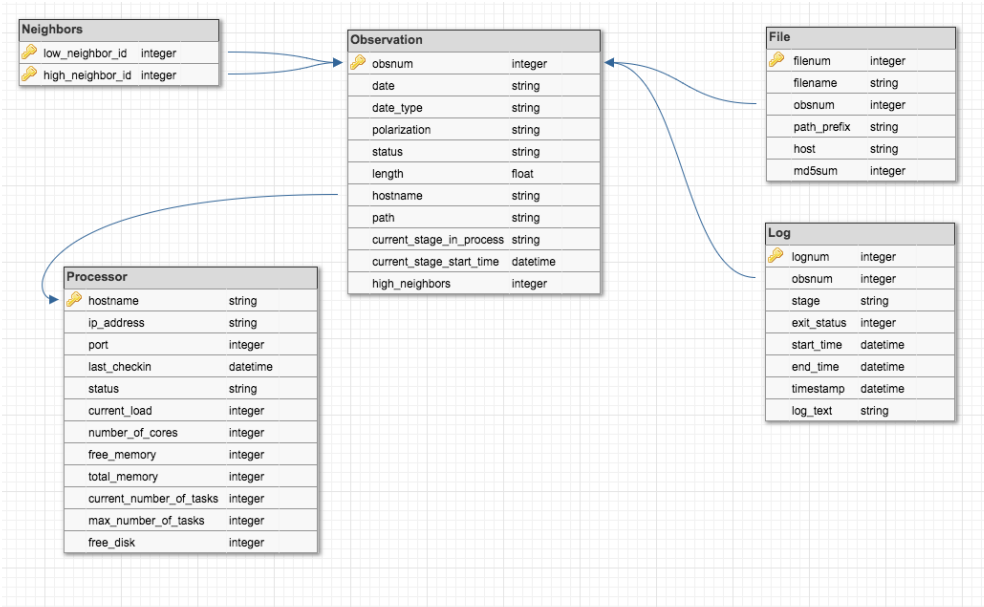

Figure 7.1: The schema of the database used to organize and implement PAPER data compression

7. Acquire time-neighbors to the file in question, and bring them to the RFI-flagged stage. The time required for this stage varied with cluster activity, but usually required roughly 20 minutes.

8. DDR filter the RFI-flagged data, using an high-tolerance iterative CLEAN. This required roughly 20 minutes.

9. RFI flag the compressed data (coarse flagging), saving the flags to a separate file. This required roughly 1 minute.

10. Apply the coarse RFI flags to the *uncompressed*, RFI-flagged data. This required roughly one minute.

11. DDR filter the now twice-RFI-flagged data, using a low-tolerance iterative CLEAN. This required roughly 120 minutes.

12. Copy compressed data to the cluster.

13. Delete the twice-RFI-flagged data.

14. If the once-RFI-flagged data are not required as neighbors, delete them.

15. If neighbors have already been compressed, delete them, otherwise begin their compression.

In total, this meant that across ten compute nodes, and efficient use of the fact that the neighbors could progress through the processing stages while the central file was being compressed, meant that it took roughly 20 to 24 hours to compress a night of observations.

7.2 Radio Frequency Interference

7.2.1 PAPER-128

7.2.2 HERA-19 and PAPER-19

7.3 Crosstalk in PAPER-64

7.4 Pre-Redundant Calibration QA

7.5 Post-Redundant Calibration QA

Chapter 8

Polarimetric Calibration

8.1 Redundant Calibration

8.2 Imaging Calibration

Chapter 9

The Ionosphere

Chapter 10

A view of the EoR window from the PAPER-32 imaging array

Chapter 11

A view of the EoR window from the HERA-19 commissioning array

Chapter 12

Deep integrations with PAPER-128

Part III

Expanding the potential of EoR measurements

Chapter 13

Higher-order correlation functions between the kSZ and 21cm fields during the EoR

Chapter 14

Deep Learning for 21cm Observations

Chapter 15

Conclusions

Appendices

Appendix A

Software

Software engineering and maintenance of existing codebases has been, generally speaking, historically undervalued and unappreciated (Muna et al. 2016). In this Appendix I would like to provide a brief description of the major software packages used in this work – without which, the work would not exist.

A.1 Astronomical Interferometry in Python (`aipy`)

The `aipy` software package (Parsons 2016) was developed by a team based largely at the University of California, Berkeley and led by Aaron Parsons. Developed under NSF funding for the PAPER experiment, it provides a Python API to interact with interferometric visibilities stored in the MIRIAD file format (Sault et al. 2011). It is able to efficiently query large MIRIAD files due the APIs closeness to the underlying C code. It also contains calibration, deconvolution, imaging and phasing code in Python, and interfaces with `HEALPix` (see Section A.5, below) as well as other astronomical Python packages.

`aipy` is maintained by the HERA software team, and can be found at: `https://github.com/HERA-Team/aipy`.

A.2 Astronomy in Python (`astropy`)

`astropy` is an open-source and community-developed core Python package for Astronomy, containing a host of extremely useful utility functions and objects (Astropy Collaboration et al. 2013).

A.3 Common Astronomy Software Applications (`CASA`)

`CASA` is under active development, with the primary goal of supporting the data post-processing needs of the next generation of radio telescopes. It is developed by an international consortium of scientists based at the National Radio Astronomical Observatory (NRAO), the European Southern Observatory (ESO), the National Astronomical Observatory of Japan (NAOJ), the CSIRO Australia Telescope National Facility (CSIRO/ATNF), and the Netherlands Institute for Radio Astronomy (ASTRON), under the guidance of NRAO (McMullin et al. 2007).

A.4 Deep Learning packages

Experimentation with deep learning analyses of 21 cm simulated observations took place in Keras (Chollet et al. 2015), PyTorch (Paszke et al. 2017) and Tensorflow (Abadi et al. 2016).

A.5 Hierarchical Equal Area isoLatitude Pixelization of the sphere (`HEALPix`)

The `HEALPix` software, and its Python wrapper `healpy`, provide a pixelization which subdivides a spherical surface into pixels which each cover the same surface area as every other pixel. Pixel centers occur on a discrete number of rings of constant latitude. This

scheme makes natively spherical measurements, such as angular power spectra and wide-field images, simple and efficient to interact with (Górski et al. 2005).

A.6 pyuvdata

pyuvdata provides a Python interface to interferometric data. It can read and write MIRIAD and UVFITS file formats, as well as read CASA measurement sets and FHD (Sullivan et al. 2012) visibility save files (Hazelton et al. 2017).

pyuvdata is maintained by the HERA software team, and can be found at: https://github.com/HERA-Team/pyuvdata.

A.7 The Scientific Python Ecosystem (scipy)

Many of the above tools require at least one of the many packages under the scipy ecosystem. It is truly foundational to almost any scientific analysis that takes place in Python (Jones et al. 2001).

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X.: 2016, *ArXiv e-prints*

Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., Greenfield, P., Droettboom, M., Bray, E., Aldcroft, T., Davis, M., Ginsburg, A., Price-Whelan, A. M., Kerzendorf, W. E., Conley, A., Crighton, N., Barbary, K., Muna, D., Ferguson, H., Grollier, F., Parikh, M. M., Nair, P. H., Unther, H. M., Deil, C., Woillez, J., Conseil, S., Kramer, R., Turner, J. E. H., Singer, L., Fox, R., Weaver, B. A., Zabalza, V., Edwards, Z. I., Azalee Bostroem, K., Burke, D. J., Casey, A. R., Crawford, S. M., Dencheva, N., Ely, J., Jenness, T., Labrie, K., Lim, P. L., Pierfederici, F., Pontzen, A., Ptak, A., Refsdal, B., Servillat, M., and Streicher, O.: 2013, *A& A* **558**, A33

Chollet, F. et al.: 2015, *Keras*, `https://github.com/fchollet/keras`

Górski, K. M., Hivon, E., Banday, A. J., Wandelt, B. D., Hansen, F. K., Reinecke, M., and Bartelmann, M.: 2005, *ApJ* **622**, 759

Hazelton, B., Beardsley, A., Pober, J., Jacobs, D., Ali, Z., and Lanman, A.: 2017, *HERA-Team/pyuvdata: Version 1.1*

Jones, E., Oliphant, T., Peterson, P., et al.: 2001, *SciPy: Open source scientific tools for Python*, [Online; accessed <today>]

McMullin, J. P., Waters, B., Schiebel, D., Young, W., and Golap, K.: 2007, in R. A. Shaw, F. Hill, and D. J. Bell (eds.), *Astronomical Data Analysis Software and Systems XVI*, 376, p. 127, ASP, San Francisco, CA

Muna, D., Alexander, M., Allen, A., Ashley, R., Asmus, D., Azzollini, R., Bannister, M., Beaton, R., Benson, A., Berriman, G. B., Bilicki, M., Boyce, P., Bridge, J., Cami, J., Cangi, E., Chen, X., Christiny, N., Clark, C., Collins, M., Comparat, J., Cook, N., Croton, D., Delberth Davids, I., Depagne, É., Donor, J., dos Santos, L. A., Douglas, S., Du, A., Durbin, M., Erb, D., Faes, D., Fernández-Trincado, J. G., Foley, A., Fotopoulou, S., Frimann, S., Frinchaboy, P., Garcia-Dias, R., Gawryszczak, A., George, E., Gonzalez, S., Gordon, K., Gorgone, N., Gosmeyer, C., Grasha, K., Greenfield, P., Grellmann, R., Guillochon, J., Gurwell, M., Haas, M., Hagen, A., Haggard, D., Haines, T., Hall, P., Hellwing, W., Herenz, E. C., Hinton, S., Hlozek, R., Hoffman, J., Holman, D., Holwerda, B. W., Horton, A., Hummels, C., Jacobs, D., Juel Jensen, J., Jones, D., Karick, A., Kelley, L., Kenworthy, M., Kitchener, B., Klaes, D., Kohn, S., Konorski, P., Krawczyk, C., Kuehn, K., Kuutma, T., Lam, M. T., Lane, R., Liske, J., Lopez-Camara, D., Mack, K., Mangham, S., Mao, Q., Marsh, D. J. E., Mateu, C., Maurin, L., McCormac, J., Momcheva, I., Monteiro, H., Mueller, M., Munoz, R., Naidu, R., Nelson, N., Nitschelm, C., North, C., Nunez-Iglesias, J., Ogaz, S., Owen, R., Parejko, J., Patrício, V., Pepper, J., Perrin, M., Pickering, T., Piscionere, J., Pogge, R., Poleski, R., Pourtsidou, A., Price-Whelan, A. M., Rawls, M. L., Read, S., Rees, G., Rein, H., Rice, T., Riemer-Sørensen, S., Rusomarov, N., Sanchez, S. F., Santander-García, M., Sarid, G., Schoenell, W., Scholz, A., Schuhmann, R. L., Schuster, W., Scicluna, P., Seidel, M., Shao, L., Sharma, P., Shulevski, A., Shupe, D., Sifón, C., Simmons, B., Sinha, M., Skillen, I., Soergel, B., Spriggs, T., Srinivasan, S., Stevens, A., Streicher, O., Suchyta, E., Tan, J., Telford, O. G., Thomas, R., Tonini, C., Tremblay, G., Tuttle, S., Urrutia,

T., Vaughan, S., Verdugo, M., Wagner, A., Walawender, J., Wetzel, A., Willett, K., Williams, P. K. G., Yang, G., Zhu, G., and Zonca, A.: 2016, *ArXiv e-prints*

Parsons, A.: 2016, *AIPY: Astronomical Interferometry in PYthon*, Astrophysics Source Code Library

Parsons, A. R. and Backer, D. C.: 2009, *AJ* **138**, 219

Parsons, A. R., Liu, A., Aguirre, J. E., Ali, Z. S., Bradley, R. F., Carilli, C. L., DeBoer, D. R., Dexter, M. R., Gugliucci, N. E., Jacobs, D. C., Klima, P., MacMahon, D. H. E., Manley, J. R., Moore, D. F., Pober, J. C., Stefan, I. I., and Walbrugh, W. P.: 2014, *ApJ* **788**, 106

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A.: 2017

Sault, R. J., Teuben, P., and Wright, M. C. H.: 2011, *MIRIAD: Multi-channel Image Reconstruction, Image Analysis, and Display*, Astrophysics Source Code Library

Sullivan, I. S., Morales, M. F., Hazelton, B. J., Arcus, W., Barnes, D., Bernardi, G., Briggs, F. H., Bowman, J. D., Bunton, J. D., Cappallo, R. J., Corey, B. E., Deshpande, A., deSouza, L., Emrich, D., Gaensler, B. M., Goeke, R., Greenhill, L. J., Herne, D., Hewitt, J. N., Johnston-Hollitt, M., Kaplan, D. L., Kasper, J. C., Kincaid, B. B., Koenig, R., Kratzenberg, E., Lonsdale, C. J., Lynch, M. J., McWhirter, S. R., Mitchell, D. A., Morgan, E., Oberoi, D., Ord, S. M., Pathikulangara, J., Prabu, T., Remillard, R. A., Rogers, A. E. E., Roshi, A., Salah, J. E., Sault, R. J., Udaya Shankar, N., Srivani, K. S., Stevens, J., Subrahmanyan, R., Tingay, S. J., Wayth, R. B., Waterson, M., Webster, R. L., Whitney, A. R., Williams, A., Williams, C. L., and Wyithe, J. S. B.: 2012, *ApJ* **759**, 17