Draft version August 30, 2018
Preprint typeset using LaTeX style emulateapj v. 01/23/15

COSMIC MICROWAVE BACKGROUND POWER SPECTRA FROM FEW BIT TIMESTREAMS

L. Balkenhol[1] and C. L. Reichardt[1]

Draft version August 30, 2018

ABSTRACT

Observations of the Cosmic Microwave Background (CMB) are of significant value to modern cosmology and particle physics. Future CMB experiments face issues in hardware design, mission planning and analysis that arise from the vast size of time-ordered-data (TOD) being recorded. These challenges are particularly significant for Antarctic and satellite experiments which depend on satellite links to transmit their data. We explore the viability of reducing the TOD to few-bit numbers to address these issues. Unlike lossless compression, this technique introduces additional noise into the data. We present 1, 2 and 3 bit digitisation schemes and determine the degradation this compression causes in temperature and polarisation power spectra. We find that 3 bit digitisation has a percent-level contribution to the map noise level. We argue that such digitisation is a promising strategy for upcoming experiments.

Subject headings: cosmic background radiation — polarization — data compression

1. INTRODUCTION

Observations of the Cosmic Microwave Background (CMB) have played a key role in physics since its discovery (Penzias & Wilson 1965). Current and future CMB experiments will continue to deliver new insights by studying the temperature and polarisation information contained in the CMB. These will put tight constraints on cosmological models. Upcoming science goals include the discovery of inflationary gravitational waves and thorough studies of CMB lensing and the Sunyaev-Zeldovich (SZ) effects. Moreover, CMB experiments will measure the relativistic number of species and the neutrino mass sum (Abazajian et al. 2016).

High fidelity CMB science demands an excellent standard of data analysis. The CMB community has developed a variety of compression and computational techniques to manage the increasing influx of data, while maximising the science output (Tristram & Ganga 2007). These include the compression of time-ordered data (TOD) into maps (Tegmark 1997), bandpower estimation (Tegmark 1998) and the pseudo C_l method (Brown et al. 2005).

A growing hurdle for experiments at remote locations are the transmission limitations of satellite links. Space-based experiments have employed a combination of lossless and lossy compression techniques, including reduced bits in the TOD (Gaztanaga et al. 1998; Maris et al. 2003). Antarctica-based experiments that transmit a portion of their data via a satellite link have downsampled their data in the past to meet their telemetry requirements. They have yet to exploit few bit digitisation of the TOD. As we approach the next generation of ground-based experiments, Stage-4, and the launch of a new generation of space-based missions (liteBIRD, PIXIE, COrE+), we must treat the transmission bottleneck carefully. Without a review of the current compression techniques employed we are sure to lose information.

In this work we present the method of extreme digitisation, which compresses a rich digital input signal into a few bits. We apply extreme digitisation to the TOD and detail its effect on temperature and polarisation power spectra. We find that an optimal 3-bit digitisation scheme adds as little as $< 2\%$ to the map noise level. The digtisation schemes described here are laid out for ground-based experiments. Space-based missions must study carefully the effect of digitisation in their specific compression algorithms.

This work is structured as follows. In §2 we detail the challenges originating from handling large TOD, formulate the process of extreme digitisation and lay out the framework used to test its performance. Subsequently, we describe the power spectrum estimation employed and interpret the results obtained through this process in section §3. We summarise our findings in §4.

2. DIGITISATION

2.1. *Problem*

The science goals of upcoming CMB experiments naturally lead to a large influx of data To achieve the targeted sensitivity longer observations with more detectors are needed. In fact the number of detectors of ground-based experiments has been following a Moore's law like trend, doubling approximately every 2 years. This directly translates into an exponential growth in data volume (Abazajian et al. 2016)..

Excellent observation conditions for CMB observations can not only be found in space, but also at various locations on earth (Li et al. 2017; Kovac & Barkats 2007) among which is the South Pole. Space- and Antarctica-based experiments depend on satellite transmission. The next generation of ground-based instruments will aim to collect 2 million detector years of data. An experiment contributing to Stage-4 at the South Pole will face a data influx of $\sim$ O(10)Tb/d. However, the current transmission allocation for SPT3G is at 150Gb/d, which will likely only see a moderate increase in coming years. The transmission bottleneck is currently overcome by recovering the full data on hard drives with some latency and by transmitting a downsampled version of the data.

christian.reichardt@unimelb.edu.au
[1] School of Physics, University of Melbourne, Parkville, VIC 3010, Australia

The downsampling process loses high frequency information. Going into Stage-4 we anticipate that compression rates for transmission must increase by an order of magnitude. Continued use of downsampling will narrow the information window decisively - prohibiting high multipole moment science to be carried out on the transmitted dataset. This also means that any potential faults or errors in the experiment that only become visible in said range will go unnoticed for longer.

The Planck mission has demonstrated the merits of carrying out CMB observations from space (Planck Collaboration et al. 2018). Upcoming missions aim to exceed the detector count of Planck by at least an order of magnitude (Matsumura et al. 2014; Kogut et al. 2011; Delabrouille et al. 2018). However, strategies to meet the telemetry specifications of each mission appear to be non-settled. It is not clear whether the data compression knowledge developed during the Planck mission will guarantee optimal performance for future satellites. Each mission will have to carefully construct a compression algorithm through a combination of lossless and lossy techniques.

Beyond transmission challenges, mission planning is becoming exceedingly difficult. A full simulation of TOD over the entire parameter space of detection scenarios for numerous set-ups is the desired way to decide on Stage-4 configurations. Space-based missions must aim to carry out a similar analysis to optimise their science output. Given the shear size of TOD expected, this is not possible (Abazajian et al. 2016). We must rely on different planning strategies or aim to reduce the size of the TOD in order to maximise the productivity of planning and development stages and guarantee scientific excellence.

Operations on the TOD, such as noise-removal or map-making are a vital part of CMB data analysis. While we have not experienced the limitations of the accessible computational assets, the exponential growth of CMB data makes its analysis increasingly expensive. The community depends on the excellent facilities provided by the National Energy Research Scientific Computing Center (NERSC).

Extreme Digitisation would tackle the challenges mentioned above by reducing the size of the TOD by an order of magnitude. Together with already established lossless compression techniques (e.g. FLAC, run-length coding, Huffman coding, etc.) this will directly tackle transmission hurdles. Beyond that extreme digitisation has the potential of solving planning and analysis problems.

The Planck mission has demonstrated the advantages of using few bit compression of the TOD (Maris et al. 2003). Other science areas have also shown the power of extreme digtisation. (Jenet & Anderson 1998) explored the application of such compression to radio pulsar timing measurements with success. Recently (Clearwater et al. 2018) have investigated the advantages of using 1 and 2 bit data when searching for continuous gravitational waves using the Laser Interferometer Gravitational-Wave Observatory (LIGO).

2.2. *Extreme Digitisation*

Digitisation is a lossy compression technique. However, the induced noise depends on the number of bits used, the digitisation thresholds, and the output levels chosen. To minimise the noise induced through this process one must know the nature of the input signal. There is usually little concern around finding an optimal set of digitisation parameters for a given input signal. It is common to have access to large numbers of bits to store information, where changes in the digitisation scheme become insignificant to the distortion induced. It is in our interest however to consider extreme, i.e. few-bit, digitisation. To do so we review the key ideas laid out by Max (1960) below.

Digitisation discretises an input signal by sorting it into N appropriate ranges, such that an input between x_i and x_{i+1} produces an output at y_i. A digitisation scheme is described by the number of ranges, N, the endpoints of these ranges, x_i, and the output levels, y_i. Conventionally one chooses $x_1 = -\infty$ and $x_{N+1} = \infty$. In order to quantify the performance of a given digitisation scheme we define the distortion as

$$D = \left\langle (s - \hat{s})^2 \right\rangle$$

where s is the input and $\hat{s}$ the output signal. For an input signal that has at least some stochastic element to it we introduce the input amplitude probability density $p(x)$. This allows us to rewrite the above as

$$D = \sum_{i=1}^{N} \int_{x_i}^{x_{i+1}} (x - y_i)^2 \, p(x) dx$$

Seeing as we wish to minimise the distortion we differentiate the above with respect to x_i and y_i and set the derivatives to zero. We obtain the two equations

$$\frac{\partial D}{\partial x_i} = (x_i - y_{i-1})^2 \, p(x_i) - (x_i - y_i)^2 \, p(x_i) = 0 \qquad (1)$$

$$\frac{\partial D}{\partial x_j} = -2 \int_{x_i}^{x_{i+1}} (x - y_i) \, p(x) dx = 0 \qquad (2)$$

Rearranging equation 1 we deduce

$$x_i = \frac{y_i + y_{i+1}}{2} \qquad (3)$$

which informs us that an output level y_i must lie halfway between its delimiting thresholds x_i and x_{i+1}. We gain an additional condition from equation 2

$$\int_{x_i}^{x_{i+1}} (x - y_i) \, p(x) dx = 0 \qquad (4)$$

This implies that we should choose y_i, such that it halves the area underneath $p(x)$ in the interval from x_i to x_{i+1}.

To progress further we have to make an assumption about the distribution of input signals, $p(x)$. For our purposes we assume that CMB observations operate at low signal to noise. Furthermore we assume that the noise profile is Gaussian white noise[2], i.e. $p(x) = 1/\sqrt{2\pi} e^{-x^2/2}$. Given this assumption we can solve the problem using a numerical iterative procedure. One begins by picking y_1 and calculating the remaining x_i's and

[2] Please see the conclusion for a discussion of the effect of more realistic noise profiles.

y_i's using equation 3. Afterwards one observes whether this choice of values satisfy the conditions given by equation 4. If that is the case, the x_i's and y_i's were chosen appropriately.

We use the findings of Max to formulate the multi-level functions we use for our 1, 2 and 3 bit digitisation process. Given an input signal $s(t)$ a digitisation scheme using N bits returns the output $\hat{s}_N(t)$. For 1 bit digitisation we apply the sign function

$$\hat{s}_1(t) = \begin{cases} 1, & \text{for } s(t) > 0 \\ -1, & \text{for } s(t) \leq 0 \end{cases}$$

to the TOD. For 2 bit digitisation we apply the four-level function

$$\hat{s}_2(t) = \begin{cases} 1.51\sigma, & \text{for } s(t) \geq 0.9816\sigma \\ 0.4528\sigma, & \text{for } 0 \leq s(t) < 0.9816\sigma \\ -0.4528\sigma, & \text{for } 0.9816\sigma \leq s(t) < 0 \\ -1.51\sigma, & \text{for } 0.9816\sigma < s(t) \end{cases}$$

where σ is the standard deviation of the input signal. Finally the optimal 3 bit digitisation is described by the eight-level function

$$\hat{s}_3(t) = \begin{cases} 2.152\sigma, & \text{for } s(t) \geq 1.748\sigma \\ 1.344\sigma, & \text{for } 1.05\sigma \leq s(t) < 1.748\sigma \\ 0.756\sigma, & \text{for } 0.501\sigma \leq s(t) < 1.05\sigma \\ 0.245\sigma, & \text{for } 0 \leq s(t) < 0.501\sigma \\ -0.245\sigma, & \text{for } 0.501\sigma \leq s(t) < 0 \\ -0.756\sigma, & \text{for } 1.05\sigma \leq s(t) < 0.501\sigma \\ -1.344\sigma, & \text{for } 1.748\sigma \leq s(t) < 1.05\sigma \\ -2.152\sigma, & \text{for } 1.748\sigma < s(t) \end{cases}$$

The fact that the single bit digitisation scheme makes no references to the standard deviation of the signal is irrelevant. This affects the normalisation of the signal, which we can account for later. What is important is the spacing between digitisation thresholds and relative distance between the output levels chosen.

Completely different digitisation schemes can be thought of, which would for example reference the most drastic outlier in the dataset, or seek to place equal numbers of points into each digitisation level. However, given the assumptions made the schemes derived above are optimal. Additionally, they are simple enough to be easily implemented computationally.

2.3. *Methods*

To investigate the performance of the derived digitisation schemes we simulate many scans over CMB template maps at the timestream level. Each scan is performed by a single detector. We obtain control maps that use 64bit TOD and maps that have undergone 1, 2 and 3 bit digitisation at the timestream level. We calculate the temperature and polarisation power spectra of each map and determine the additional noise induced through the extreme digitisation process.

To create the template maps we use the healpix framework and the wealth of support available for it. We generate a realisation of I, Q and U maps with NSIDE = 4096 based on the results of Planck 2015. The key cosmological parameters assumed are $\Omega_b \approx 0.049$, $\Omega_c \approx 0.265$, $\Omega_m \approx 0.316$ and $h \approx 0.67$.

We simulate observing a $\sim 600\text{deg}^2$ patch of the sky. To do so we perform a number of constant elevation scans (CES), equally spaced in declination (DEC). We repeat the observation strategy 100 times with a slight offset in right ascension (RA) and DEC each time, such that all pixels within the patch are hit approximately uniformly. The speed at which we sweep across the survey area is adjusted to to produce a desired number of hits per pixel (hits per pixel) in the output maps.

While performing each CES the pixels being targeted are determined. The corresponding values from the template maps are then accessed and added to realisations of the detector noise of appropriate length. We assume the detector noise to be Gaussian white noise.

At this point we apply the digitisation schemes to the TOD. We compress the timestream into maps by averaging all hits falling into the same pixel. We produce 12 CMB maps in total: three control maps (I, Q, U) and nine maps with three each obtained from each digitisation scheme.

This process is carried out 6 times. We produce maps with approximately 800, 8,000, 80,000, 1,024,000, 10,240,000 and 102,400,000 hits per pixel. We assume that the detector is read out 200 times per second and has a noise level of $500\mu K\sqrt{s}$ for temperature and $\sqrt{2}\times 500\mu K\sqrt{s}$ for polarisation observations. The calculation outlined above has considerable computational requirements if we want to reach up to $\sim 10^8$ hits per pixel. To carry out this simulation we make use of parallelisation and the computing facilities provided by NERSC.

3. RESULTS

3.1. *Power Spectrum Estimation*

We use PolSpice to compute the TT, EE and BB power spectra of the reconstructed maps I, Q, U maps. When doing so we apodise the observed skypatch using a cosine mask with $\sigma_{\text{APOD}} = \sqrt{600}/2\text{deg}$ to minimise cut sky effects on the spectra.

Normalisation of the obtained power spectra is necessary. The digitisation schemes introduced in section §2 are designed to minimise the distortion of the TOD, not to preserve power at the map level. Calibration to the CMB is common for many experiments (e.g. SPT).

For the normalisation procedure we focus on the $\sim 10^8$hitsperpixel maps. We normalise the TT, EE, and BB power spectra from few bit TOD against their control counterparts. We consider the normalisation window cut out by the conditions

$$l > 35\,;\ \frac{C_l^{\text{S}}}{C_l^{\text{N}}} \geq 10$$

where C_l^{S} is the power spectrum of the input template maps and C_l^{N} the detector noise level. The lower limit is placed in accordance to the patch-size surveyed, the upper limit ensures appropriate signal to noise in the normalisation window.

The normalisation constants for each digitisation scheme and channel are obtained through this way are then applied to all lower hits per pixel simulations. A sample of the obtained power spectra are shown in figure 1.

3.2. *Additional Noise*

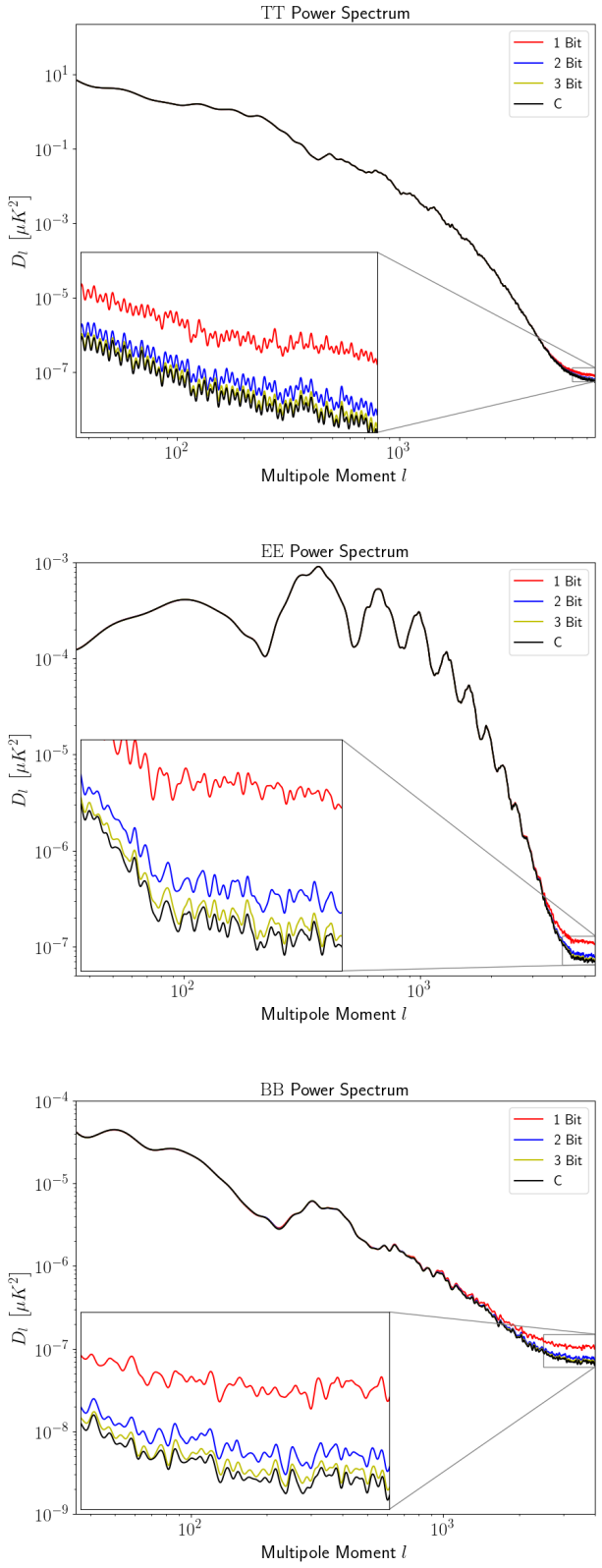

FIG. 1.— Reconstructed TT, EE, and BB power spectra. These originate from an observation with 102400000hitsperpixel. We observe that digitisation appears to add some constant to the noise level. As one would expect 3Bit digitisation performs the best, followed by 2Bit- and finally 1Bit digitisation.

To quantify how much the quality of the power spectra suffers from the digitisation process we compare the map noise levels inferred from the power spectra. These are put into the context of the map noise levels deduced from the control power spectra. We formulate

$$\frac{\Delta\sigma}{\sigma} = \frac{\sigma^{\mathrm{D}}_{\mathrm{map}} - \sigma^{\mathrm{C}}_{\mathrm{map}}}{\sigma^{\mathrm{C}}_{\mathrm{map}}} \tag{5}$$

where $\sigma^{\mathrm{C}}_{\mathrm{map}}$ is the map noise level of the control power spectra and $\sigma^{\mathrm{D}}_{\mathrm{map}}$ is the map noise level obtained from the power spectra originating from extremely digitised TOD. We assume that the digitisation process results in adding a constant noise term to each power spectrum. If we are dominated by noise we can write

$$C_l^D \approx C_l^N + C_l^X$$

Here C_l^D is the power spectrum originating from a digitised timestream, C_l^N is the detector noise level and C_l^X the additional noise induced through digitisation. We now progress equation 5 to

$$\frac{\Delta\sigma}{\sigma} = \sqrt{\frac{C_l^D}{C_l^N}} - 1 = \sqrt{\frac{C_l^N + C_l^X}{C_l^N}} - 1 = \sqrt{1 + \frac{C_l^X}{C_l^N}} - 1$$

The value of l at which we can safely assume to be noise dominated and apply the above framework varies between simulated observations of different hits per pixel and temperature and polarisation power spectra. The range considered involves any datapoints at multipole moments larger than the last point at which the detector noise is at least an order of magnitude larger than the template power spectrum, i.e.

$$\frac{C_l^{\mathrm{N}}}{C_l^{\mathrm{S}}} \geq 10$$

Before analysing C_l^X/C_l^N we rebin the power spectra to $\Delta l = 123$. This guarantees that the points in the noise tail are independent of one another, allowing us to extract an uncertainty for the above quantity. Plots for C_l^X/C_l^N are shown in figure 2.

The deduced additional noise for 1, 2 and 3 bit digitisation schemes are shown with respect to the hits per pixel in the maps in figure 3. We would like to point out three key results. Firstly, 3 Bit digitisation performs the best, followed by 2 Bit- and finally 1 Bit digitisation. This is what we expect, given that with each additional bit we retain more information. Secondly, for a fixed detector noise level the deterioration of the map quality scales in the same fashion as the number of hits per pixel. We see that $\Delta\sigma/\sigma$ is independent of the hits per pixel in the maps. Lastly, the added noise levels are astonishingly low. Keeping in mind that CMB detector sensitivity improves in steps of order of magnitude every few years, adding an extra percent-level noise term does not deteriorate the results appreciably. This is impressive, given that the use of an optimal 3 bit digitisation scheme will save approximately an order of magnitude in TOD volume.

Extreme digitisation is a viable lossy compression technique when dealing with low signal to noise, large datasets with a signal that is slowly varying with respect

to the sampling rate. Under these conditions we have a locally flat, low signal, on top of which we add many noise realisations. In the limit of many samples these allow us to reconstruct the signal well. Furthermore a small signal will prevent saturation of the output, i.e. the inability of the digitised output to communicate any information where in the range ∞ to x_N an input signal lies.

4. CONCLUSIONS

In this work we have motivated the investigation of extreme digitisation as a technique in combating arising data challenges in CMB data analysis. The reduction of the TOD by an order of magnitude directly addresses the issues in data transmission faced by remote location observations. Benefits in mission planning, data analysis and hardware requirements are possible.

We have derived a set of optimal digitisation schemes and presented the level at which the induce noise into the temperate and polarisation power spectra. We find that an optimal 3 bit digitisation adds as little as $< 2\%$ to the map noise level for temperature and polarisation observations. As mentioned above this addition is insignificant given that the sensitivity of CMB experiments follows a Moore's law like trend. The benefit of this compression technique is the reduction of the TOD volume by an order of magnitude.

Future work investigating this compression technique must aim to understand the nature of the induced noise better. It is of great value to find the higher statistical moments, i.e. skewness and kurtosis of the added noise term. For this an analysis of the performance of cluster-finding algorithms on the digitised datasets is useful.

It should be laid out how the results changes when moving to a more realistic noise profile. We do not expect this to alter the practicality of our results - even if a different noise profile doubles the additional percentage to the map noise level extreme digitisation is still practical. Ideas on how to deal with 1/f noise, e.g. chunking of the data before applying extreme digitisation have already been investigated by Planck. These thoughts should be considered when designing the compression schemes of future space-based CMB missions, which will be unable to recover their full data with latency, but rely entirely on the transmitted data.

We thank the referee as well as Srinivasan Raghunathan and Federico Bianchini for valuable feedback on the manuscript. We acknowledge support from an Australian Research Council Future Fellowship (FT150100074), and also from the University of Melbourne. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. We acknowledge the use of the Legacy Archive for Microwave Background Data Analysis (LAMBDA). Support for LAMBDA is provided by the NASA Office of Space Science.

REFERENCES

Abazajian, K. N., et al. 2016, ArXiv e-prints, 1610.02743 1, 2.1
Brown, M. L., Castro, P. G., & Taylor, A. N. 2005, MNRAS, 360, 1262 1
Clearwater, P., Melatos, A., Bailes, M., Flynn, C., & Nepal, S. 2018, Submitted for publication. 2.1
Delabrouille, J., et al. 2018, J. Cosmology Astropart. Phys., 4, 014 2.1
Gaztanaga, E., Barriga, J., Romeo, A., Fosalba, P., & Elizalde, E. 1998, ArXiv Astrophysics e-prints 1
Jenet, F. A., & Anderson, S. B. 1998, PASP, 110, 1467 2.1
Kogut, A., et al. 2011, J. Cosmology Astropart. Phys., 7, 025 2.1
Kovac, J. M., & Barkats, D. 2007, ArXiv e-prints, 0707.1075 2.1
Li, Y.-P., Liu, Y., Li, S.-Y., Li, H., & Zhang, X. 2017, ArXiv e-prints, 1709.09053 2.1
Maris, M., Maino, D., Burigana, C., Mennella, A., Bersanelli, M., & Pasian, F. 2003, Mem. Soc. Astron. Italiana, 74, 488 1, 2.1
Matsumura, T., et al. 2014, Journal of Low Temperature Physics, 176, 733 2.1
Max, J. 1960, 6, 7 2.2
Penzias, A. A., & Wilson, R. W. 1965, ApJ, 142, 1149 1
Planck Collaboration, et al. 2018, ArXiv e-prints, 1807.06205 2.1
Tegmark, M. 1997, ApJ, 480, L87 1
Tegmark, M. 1998, in Eighteenth Texas Symposium on Relativistic Astrophysics, ed. A. V. Olinto, J. A. Frieman, & D. N. Schramm, 270 1
Tristram, M., & Ganga, K. 2007, Reports on Progress in Physics, 70, 899 1

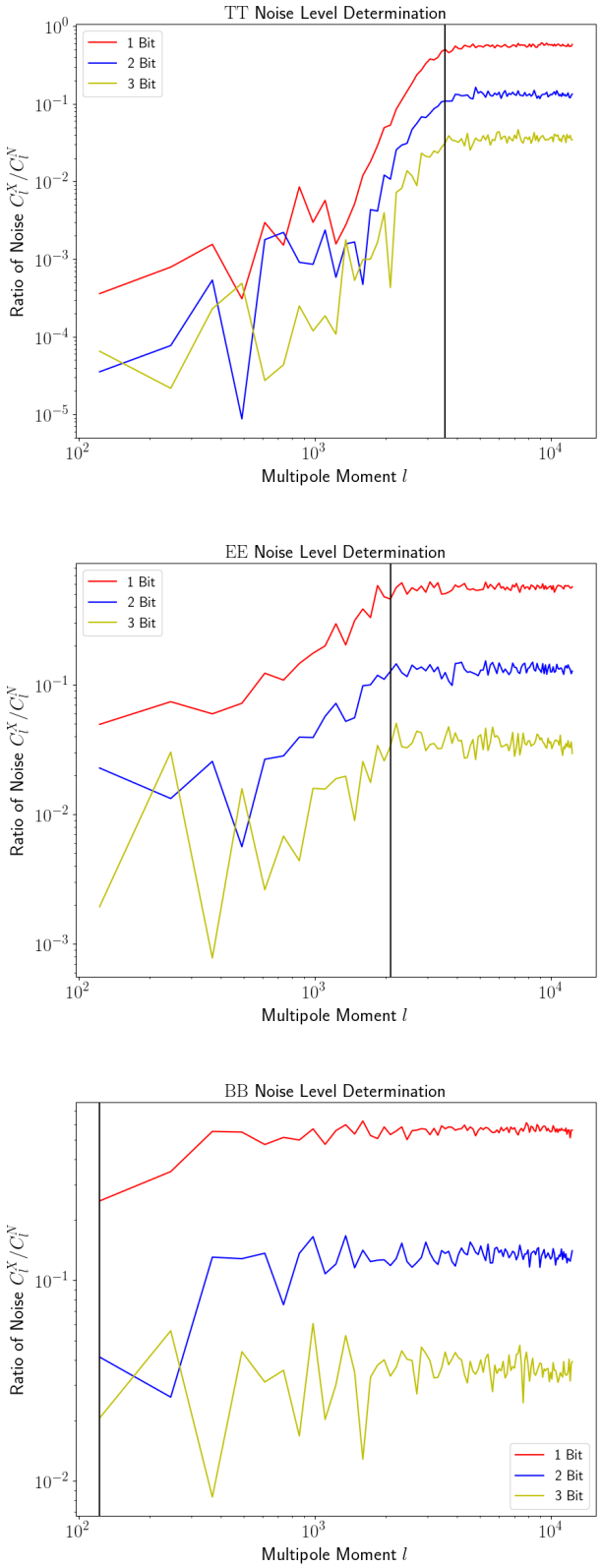

FIG. 2.— Calculated ratio of C_l^X/C_l^N of the rebinned TT, EE, and BB power spectra. The vertical black line indicates from which point onwards data is used to calculate the equivalent noise level. The above power spectra originate from a 80000hitsperpixel map. Increasing the hits per pixel in the map shifts the plateau to the right.

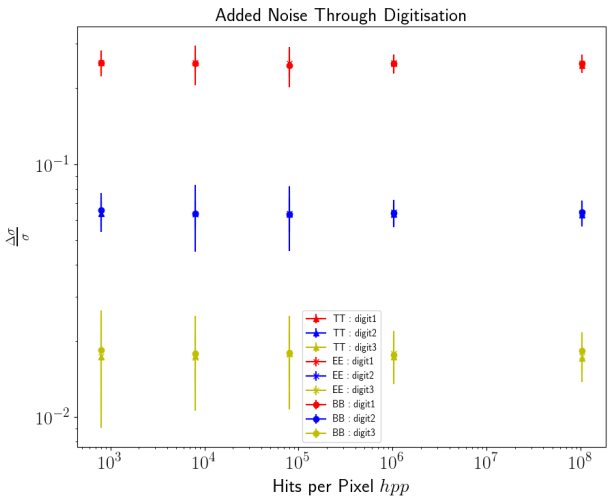

FIG. 3.— Addition to the map noise level due to digitisation.