Draft version August 13, 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 13, 2018*

## ABSTRACT

Observations of the Cosmic Microwave Background (CMB) are of immense value to modern cosmology. However, future CMB experiments must confront challenges in mission planning, hardware and analysis that arise from the sheer size of the time-ordered-data being recorded. These challenges are particularly significant for Antarctic and satellite experiments which depend on satellite links to transmit their data. We investigate using extreme digitisation to address these issues. Unlike lossless compression, extreme digitisation introduces additional noise into the data. We present optimal 1, 2 and 3 bit digitisation schemes and determine the degradation in temperature and polarisation power spectra caused by applying this process to the time-ordered-data (TOD). We find that 3 bit digitisation has a percent-level contribution to the map noise level. This is impressive considering that it would reduce the data volume by an order of magnitude. We argue that extreme 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 1964. 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. The most prominent science goal is the discovery of the imprint of inflationary gravitational waves. Additionally studies of CMB lensing and the Sunyaev-Zeldovich (SZ) effects will open up access to unprecedented insight. Moreover, through probing the relativistic number of species, the helium fraction and the neutrino mass sum, CMB experiments are a valuable counter-part to ground-based particle physics experiments.

The outstanding contribution of CMB science to modern physics has demanded a high 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. These include the compression of time-ordered data (TOD) into maps, bandpower estimation and the pseudo $C_l$ method.

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 time-ordered-data. Antarctica-based experiments that transmit a portion of their data via a satellite link have downsampled their TOD 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 and compresses it into a few bits. We apply extreme digitisation to the TOD and detail its effect on temperature and polarisation observations. We find that an optimal 3-bit digitisation scheme adds as little as $\sim 2\%$ to the map noise level. While the digtisation schemes described here are primarily laid out for ground-based experiments, the results we present here should also be considered by future space-based missions, which inevitably must incorporate lossy compression.

This work is structured as follows. We detail the arising challenges in handling large TOD in §2.1. We subsequently formulate extreme digitisation formerly in §2.2 and lay out the framework used to test its performance in §2.3. We detail how we estimate the power spectra from simulated maps in §3.1. We continue by presenting the noise induced through the dicitisation process in §3.2. We summarize 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.

The best observations sites for CMB measurements are at remote locations: in space and in the Antarctica. Experiments at these locations depend on satellite transmission. The next generation of ground-based experiments will aim to collect 2 million detector years worth 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

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

in the 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. 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 high frequencies will go unnoticed for longer.

Future space-based missions aim to exceed the detector count of Planck by at least an order of magnitude. It is questionable whether their telemetry specifications will allow for transmission of the data with Planck-style compression. Methods of storing large amounts of data on upcoming satellites will likely be prohibited by financial decisions: the amount of storage space required becomes financially relevant at the scales targeted. Missions will likely be left to design their own compression algorithms which will incorporate a combination of lossless and lossy compression techniques.

Beyond transmission challenges, mission planning is becoming exceedingly difficult. As noted by (S4 science book) 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. 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 maximal science output.

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.

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 (such as FLAC, run-length coding, Huffman coding, etc.) this will directly tackle transmission hurdles. While it needs to be investigated to what extent existing algorithms can be carried over, extreme digitisation has the possibility of solving planning and analysis problems.

Other science areas have demonstrated that extreme digitisation is a valuable compression technique. Jenet and Anderson (cite) explored the application of such compression to radio pulsar timing measurements with success. Recently Clearwater et al. (in prep.) 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) performance of searches for continuous gravitational wave searches using 1 and 2 bit data.

### 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 input signal. A theoretical framework to obtain these levels was laid out by Max in 1978. We review the key aspects of his work relevant for us 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, $x_k$, and the output levels, $y_k$. 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 is 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 may safely assume that ground-based 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 $y_i$'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 $x_k$'s and $y_k$'s were chosen appropriately.

This was carried out by Max. We incorporate his results by formulating the multi-level functions we use for

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

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 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

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

to the input signal. Finally the optimal 3 bit digitisation is described by the eight-level function

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

Other digitisation schemes can be thought of, which place the digitisation thresholds and output levels in a different way. 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 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 based on the results of Planck 2015. The key cosmological parameters are summarised in 1.

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 (hpp) 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 13 maps in total: three maps (I, Q, U) that have not undergone few-bit digitisation at the timestream level for comparison purposes, nine maps with three each corresponding to each digitisation scheme and the hitmap.

This process is carried out 7 times, producing maps with different numbers of hpp. The simulation parameters are summarised in table 2. The calculation outlined above has considerable computational requirements if we want to reach up to $\sim 10^8$ hpp. To carry out this simulation we make use of parallelisation and the computing power provided by NERSC.

TABLE 1
Input Cosmological Parameters

| Parameter | Planck 2015 |
|---|---|
| $100\theta_{MC}$ | $1.04086 \pm 0.00048$ |
| $\Omega_b h^2$ | $0.02222 \pm 0.00023$ |
| $\Omega_c h^2$ | $0.1199 \pm 0.0022$ |
| $H_0$ | $67.26 \pm 0.98$ |
| $n_s$ | $0.9652 \pm 0.0062$ |
| $\Omega_m$ | $0.316 \pm 0.014$ |
| $\sigma_8$ | $0.830 \pm 0.015$ |
| $\tau$ | $0.078 \pm 0.019$ |
| $10^9 A_s e^{-2\tau}$ | $1.881 \pm 0.014$ |

Note. — Cosmological parameters used to create the template maps. Taken from Planck 2015: Cosmological Parameters.

TABLE 2
Assumed Survey Parameters

| Parameter | Value |
|---|---|
| NSIDE | 4096 |
| $f_{\text{readout}}[\text{Hz}]$ | 200 |
| $f_{\text{sky}}$ | $\sim 0.014$ |
| $\sigma^{\text{T}}_{\text{det}}[\mu\text{K}\sqrt{\text{s}}]$ | 500 |
| $\sigma^{\text{Pol}}_{\text{det}}[\mu\text{K}\sqrt{\text{s}}]$ | $\sqrt{2} \times 500$ |
| hpp | $(1, 10, 100) \times 800$ |
| | $(1, 10, 100) \times 1024000$ |

Note. — Parameters used in the simulated observation. The RA speed is adjusted to match the desired hpp.

## 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 to minimise cut sky effects on the spectra. Please see 3 for an overview of the parameters used in this process.

To normalise the the obtained power spectra we focus on the $\sim 10^8$hpp maps. We normalise each power spectrum originating from few bit TOD against its corresponding 64bit TOD counterpart. We place a lower limit on the normalisation window in multipole space by considering the size of the observed patch with

$$l = \frac{2}{\pi}\left(\frac{32400}{600}\right) \approx 35$$

TABLE 3
PolSpice Parameters

| weights | $\sigma_{\rm APOD}$[deg$^2$] | apodistation type | polarisation |
|---|---|---|---|
| cosine mask | $\sqrt{600}/2$ | cosine | Yes |

Note. — Parameters used when calling PolSpice to calculate the power spectra. Remaining parameters have been left at their default value.

where we have rounded up to the next largest integer. We find the upper bound on the normalisation window by demanding that

$$\frac{C_l^{\rm S}}{C_l^{\rm N}} \leq 10$$

where $C_l^{\rm S}$ is the power spectrum of the input template maps and $C_l^{\rm N}$ the detector noise level.

we close the normalisation window the first instance the above condition is met. The normalisation constants for each digitisation scheme and channel obtained through this way are then applied to all lower hpp simulations. A sample of the obtained power spectra are shown in figures 1, 2, and 3.

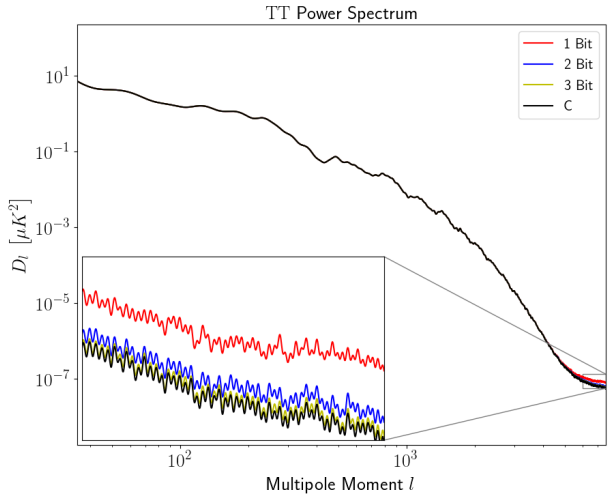


Fig. 1.— TT power spectrum reconstructed from a $\sim 10^8$hpp map.

### 3.2. *Additional Noise*

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 64bit TOD (control) power spectra. We formulate

$$\frac{\Delta\sigma}{\sigma} = \frac{\sigma_{\rm map}^{\rm D} - \sigma_{\rm map}^{\rm C}}{\sigma_{\rm map}^{\rm C}} \tag{5}$$

where $\sigma_{\rm map}^{\rm C}$ is the map noise level of the control power spectra and $\sigma_{\rm map}^{\rm D}$ is the map noise level obtained from the power spectra originating from extremely digitised TOD. We assume that the digitisation process results in

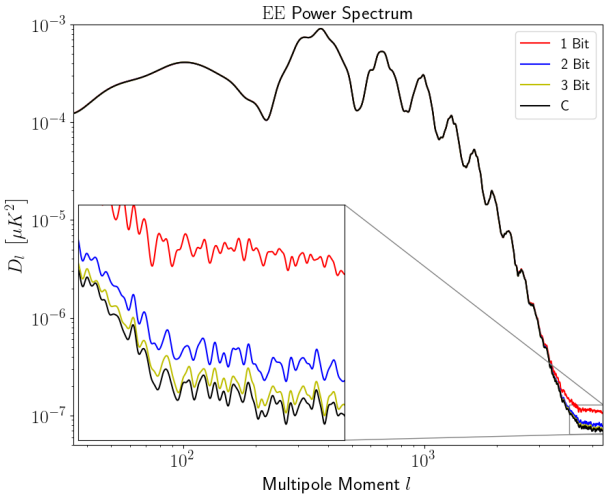


Fig. 2.— EE power spectrum reconstructed from a $\sim 10^8$hpp map.

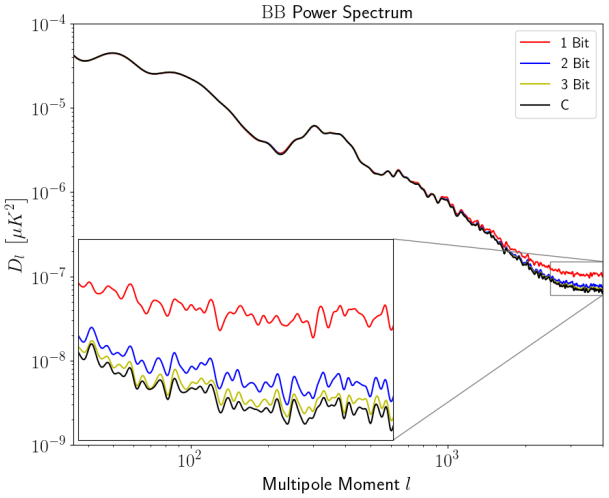


Fig. 3.— BB power spectrum reconstructed from a $\sim 10^8$hpp map.

adding a constant noise term to the 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 hpp and TT, EE and BB 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 figures 4, 5, and 6 for $TT$, $EE$, and $BB$ respectively.

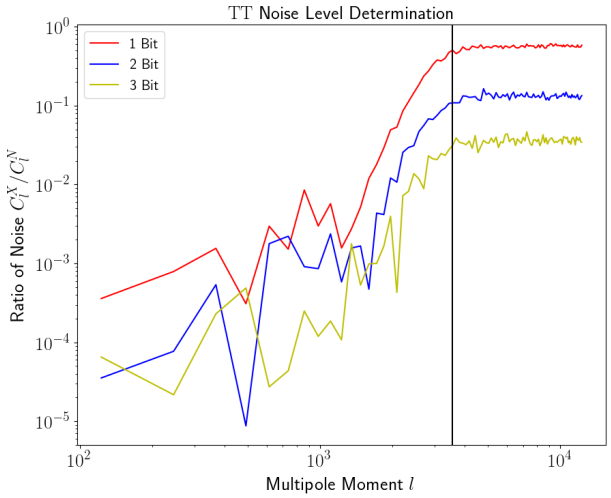


FIG. 4.— Calculated ratio of $C_l^X/C_l^N$ of the rebinned TT power spectrum. The vertical black line indicates from which point onwards data is used in calculating the equivalent noise level.

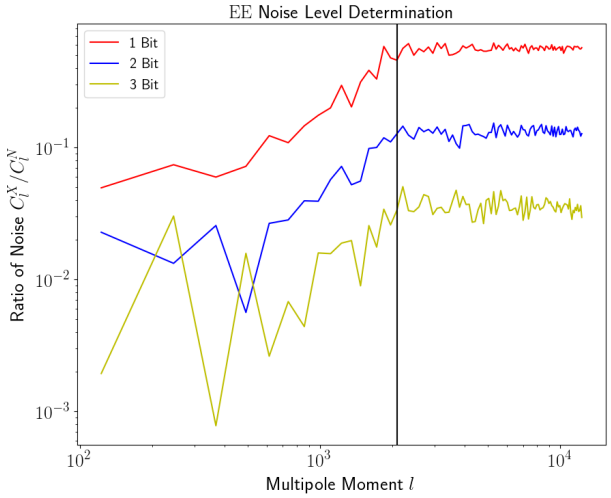


FIG. 5.— Calculated ratio of $C_l^X/C_l^N$ of the rebinned EE power spectrum.

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 7. 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.

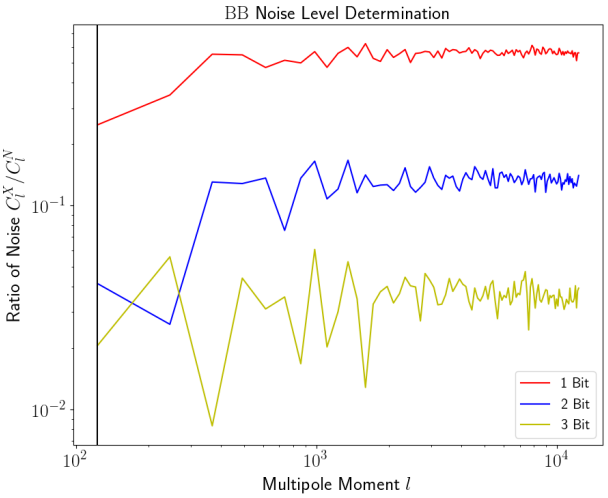


FIG. 6.— Calculated ratio of $C_l^X/C_l^N$ of the rebinned BB power spectrum.

This is what we expect, given that with each additional bit we retain more information. Secondly, for a fixed detector noise level, the additional percentage to the map noise level due to digitisation is independent of the number of hpp. The added noise therefore scales in the same fashion as the map noise level with the number of hpp, given a fixed detector noise level. 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.

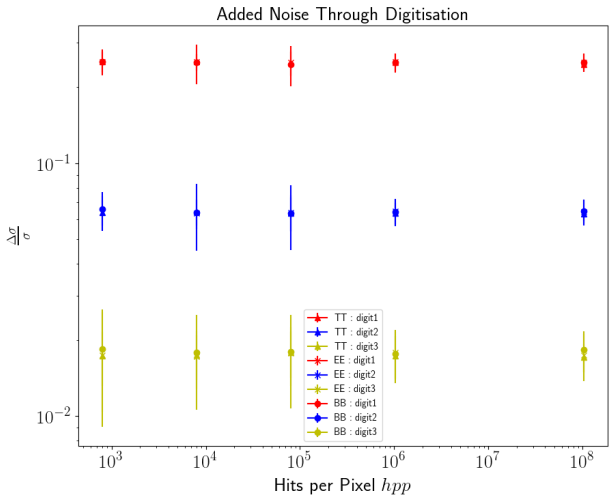


FIG. 7.— Addition to the map noise level due to digitisation.

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 $\infty$ to $x_N$ the 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 and data analysis 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.

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 this changes when moving to a more realistic noise profile. We do not expect this to alter the practicality of our results - even if a moderate change in the 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 rather entirely rely 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.