Draft version August 8, 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 8, 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 the data. We investigate using extreme digitisation to address these challenges. 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 this process. In particular we show 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 (discovery paper). 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 give us access to previously unseen information. 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 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 (TOD). 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, future space-based missions must inevitably study lossy compression techniques.

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 the performance of this compression technique in §2.3. In §3 we present the results obtained. 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 the SPT3G is at $150 Gb/d$, which will likely only see a moderate increase in the coming years. The transmission bottleneck is currently overcome by recovering the full data with some latency and by transmitting a downsampled ver-

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

sion 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 expensive. 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. 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.

Operations on the TOD, such as noise-removal or map-making are a vital part of CMB data analysis. Through the exponentially growing size of TOD CMB data analysis is becoming 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 exploited lossless compression techniques (such as FLAC) 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 areas of science have already demonstrated that extreme digitisation can be a viable compression technique. For example Jenet and Anderson have shown the advantages of this compression technique to pulsar timing measurements. The computationally challenging searches for continuous gravitational wave searcher may also profit from few-bit digitisation as laid out by Clearwater et al. (in prep.).

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 the 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 that produces a signal out s_{out} form an input signal s_{in} we define the distortion as

$$D = \left\langle (s_{in} - s_{out})^2 \right\rangle$$

For a given input amplitude probability density $p(x)$ we may 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 assumptions we cannot solve proceed analytically, but can the problem using a numerical iterative procedure. One begins by picking y_1 and calculating the remaining x_i and y_i using equation 3. Afterwards one observes weather this choice of values satisfy the conditions given by equation 4. If that is the case, the x_k and y_k were chosen appropriately.

This was carried out by Max. We incorporate his results by formulating the multi-level functions we use for our 1, 2 and 3 bit digitisation process. Given an input signal $x(t)$ a digitisation scheme using N bits returns $\hat{x}_N(t)$. For 1 bit digitisation we apply

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

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

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

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

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

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

Other digitisation schemes can be thought of that place the digitisation thresholds and output levels in a different way. However, these are not of interest for now, given the ease at which the devised optimal schemes can be implemented and the fact that they will lead to worse performance.

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 ~ 600 deg^2 patch of the sky. To do so we perform a number of equally spaced constant elevation scans (CES) over the observation region. We repeat the observation strategy 100 times with a slight offset in RA and DEC each time, such that all pixels within the path are hit approximately uniformly. The speed at which we sweep across the survey field is set to a constant to produce a desired number of hits per pixel in the output maps.

While performing each CES the appropriate pixels that are 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 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.

The simulation parameters are summarised in table 2. It 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 power provided by NERSC.

TABLE 1
Input Cosmological Parameters

Parameter	Planck 2015
$100\theta_{MC}$	1.04086 ± 0.00048
$\Omega_b h^2$	0.02222 ± 0.00023
$\Omega_c h^2$	0.1199 ± 0.0022
H_0	67.26 ± 0.98
n_s	0.9652 ± 0.0062
Ω_m	0.316 ± 0.014
σ_8	0.830 ± 0.015
τ	0.078 ± 0.019
$10^9 A_s e^{-2\tau}$	1.881 ± 0.014

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

TABLE 2
Assumed Survey Parameters

NSIDE	f_{readout}[Hz]	f_{sky}	$\sigma^{\text{T}}_{\text{det}}[\mu\text{K}\sqrt{\text{s}}]$	$\sigma^{\text{Pol}}_{\text{det}}[\mu\text{K}\sqrt{\text{s}}]$
4096	200	~ 0.014	500	$\sqrt{2} \times 500$

Note. — Parameters used in the simulated sky strategy. The RA speed is tweaked to match the desired hits per pixel.

3. RESULTS

3.1. *Power Spectrum Estimation*

We use PolSpice to compute the TT, EE, BB power spectra of the reconstructed 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.

TABLE 3
PolSpice Parameters

weightfile	apodizesigma	apodizetype	polarization
cosine mask	$\sqrt{600}/2$	1	YES

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

To normalise the the obtained power spectra we focus on the $\sim 10^8$hpp run. 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 via

$$l = \frac{2}{\pi} \frac{32400}{600} \approx 35$$

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

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

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 hit per pixel simulations.

3.2. *Additional Noise*

To quantify how much the quality of the power spectra suffers due to the digitisation 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 power spectra. We formulate

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

where $\sigma_{\rm map}^{\rm F}$ 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 adding some constant independent of multipole moment to the power spectrum. If we are dominated by noise we can write

$$C_l^D = C_l^N + C_l^X$$

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 for different channels. The range considered involves any datapoints at multipole moments larger than the last point at which the detector noise is at least than 10 times the input power spectrum.

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.

The results for the 1, 2 and 3 bit digitisation schemes are summarised in figure 1. We would like to points 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 may 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 hits per pixel. The added noise therefore scales in the same fashion as the map noise level with the number of hits per pixel, 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.

Extreme digitisation is a viable lossy compression technique when dealing with large volume, low signal to noise, slowly varying datasets. Making these assumptions the signal small and constant for a considerable number of datapoints. For such an interval the many samplings of the noise result in producing appropriate numbers in each output bin. The addition of more bits helps by classifying the size of the noise more and more appropriately and hence approaching a good estimate for the value of the signal faster. It is necessary to remain in the low signal to noise regime, as otherwise a dominant signal would saturate the digitisation scheme and for a given scheme only values in the range $\pm|N_{\rm hits} y_N|$ can be represented, given that we have $N_{\rm hits}$ datapoints. We must demand the data to be slowly varying such that many datapoints can be used to reconstruct a single significant step of the input signal. Finally a large data volume is clearly favourable for the reasons illustrated above: it allows for repeated sampling and therefore more accurate reconstruction of the weak signal.

4. CONCLUSIONS

In this work we have demonstrated the power of digitising CMB TOD from 64-bit down to few-bits. The additional noise induced through this step was at the percent-level for chosen optimal digitisation schemes for temperature and polarisation observations.

The motivation behind employing extreme digitisation is the reduction of TOD by an order of magnitude. This primarily addresses challenges in data transmission that will face upcoming CMB experiments at remote locations. Benefits in planning and analysis seem feasible.

Future work on investigating this compression technique must aim to understand the nature of the induced noise better. For this the performance of cluster-finding algorithms is interesting.

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 fully 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 Mi-

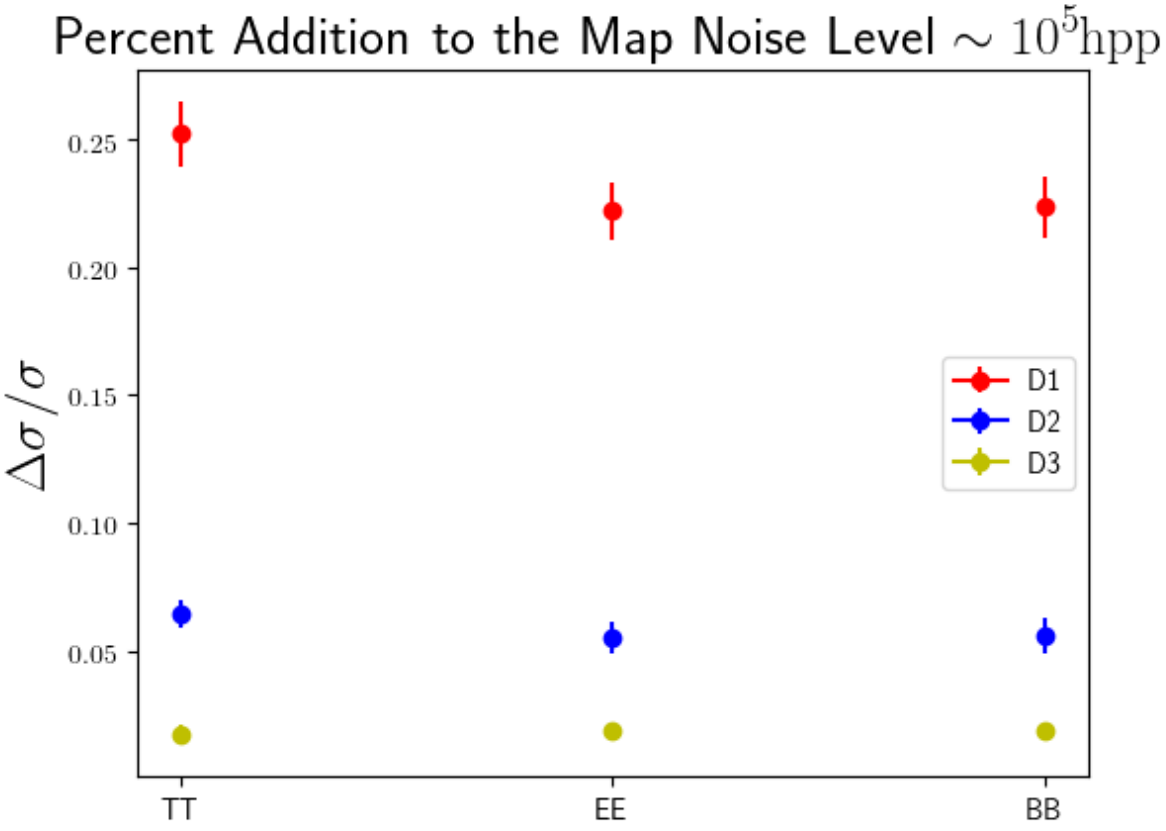

FIG. 1.— Addition to the map noise level due to digitisation.

TABLE 4
ASSUMED SURVEY PARAMETERS

Experiment	Sky coverage	Polarized Noise level (μK-arcmin)	$1/f$ knee	Beam FWHM (arcmin.)
CMB Stage III				
SPT-3G	6%	3.0	200	1.2
Simons Array	36%	9.5	200	3.5
CMB Stage IV	55%	1.3	100	4.0

NOTE. — Key numbers about the planned stage III and IV experiments. The sky coverage percentages are after galactic cuts. Unless otherwise noted, the Fisher matrix forecasts in this work use these numbers. All forecasts also allow for beam and calibration uncertainties as noted in the text.

crowave Background Data Analysis (LAMBDA). Support for LAMBDA is provided by the NASA Office of Space Science.