Penn Array Sensitivity & Data Analysis

13dec02 Brian S. Mason & Don Wells (NRAO)

1 Introduction

The purpose of this memo is to clarify Penn Array sensitivity estimates and present some thoughts about data analysis.

2 Sensivity Calculation

We outline the calculation of the RMS noise level in Penn Array map of a $1^\circ \times 1^\circ$ field with 6 hours of integration. We assume a Nyquist spacing of $4''$ and a beam FWHM of $8''$ (these are traditional numbers, approximately but not strictly compatible– in fact an $8''$ beam from the GBT is expected at $\sim 92.5\,\mathrm{GHz}$, for which $\lambda/D = 3.3''$). The beam volume for an $8''$ FWHM beam is $2\pi\sigma^2 = 72\,\mathrm{arcsec}^2$, so a $1\,\mathrm{deg}^2$ area contains 1.8×10^5 independent beams. If this area were mapped by a single detector, the detector would spend $6 \times 3600/(1.8 \times 10^5) = 0.12$ seconds per pixel for a noise level of $343/\sqrt{0.12} = 987\,\mu\mathrm{Jy}$ per beam on the sky. The noise in each detector is uncorrelated, so the final map noise is $987\,\mu\mathrm{Jy}/\sqrt{64} = 123\,\mu\mathrm{Jy}$.

Note that

- It is tempting but incorrect to calculate the noise per Nyquist pixel as these are highly correlated.
- You have to be careful with double counting (or "half counting") photons: because the $343\,\mu\mathrm{Jysec}^{1/2}$ figure (taken from the NSF filter proposal) already includes the fact that only 22% of the photons from one PSF fall on a given detector, the photon noise between detectors is uncorrelated. By subsequently adding the noise from four detectors in quadrature the sensitivity is as if you had a *single* detector which only accepted only 22% of the photons rather than four of these.

Sensitivities and data volumes for three other cases of Penn Array observing—all 6 hour integration times, and areas compatible with GBT slew rate limitations—are shown in Table 1. The data volume and N_{pix} estimates will be used in § 3.

3 Approach to Data Analysis

The Penn Array and instruments that follow in its footsteps will present astronomers with a tremendous increase in mapping speed over current capabilities. In order to best use this data our analysis algorithms should produce high-fidelity images of large regions of the sky. A key challenge is to effectively remove foreground (atmospheric) emission. In the past bolometers and bolometer arrays have used differencing and chopping techniques to this end. These

Field Size	Slew Rate [$'$/sec]	Samp. Rate [Hz]	N_{pix}	N_{data}	Sens [μJy]
$30'' \times 30''$	–	10	64	1.34×10^7	1
$5' \times 5'$	1	25	$5,620$	3.46×10^7	10
$1° \times 1°$	10	250	$810,000$	3.46×10^8	123
$8° \times 8°$	40	1000	5.18×10^7	1.39×10^9	$1,000$

Table 1: Cases of Penn Array sensitivities and data volumes. All assume: 6 hours of integration; Nyquist sampling in time; a per-detector sensitivity of $343\,\mu\mathrm{Jy\,sec}^{1/2}$; $\sim 50\%$ re-sampling ("overscanning") of the sky to improve atmosphere rejection; and reasonably quick slew rates for the same reason (except for the 1 FOV case). Note that the 80 to 100 GHz gain of the GBT (with phase III laser metrology) is between ~ 0.75 and ~ 1K/Jy so the μJy numbers roughly translate into μK.

approaches have the disadvantage that they limit the faithfully-reconstructable field of view to several times the maximum chop; in addition, optical or mechanical constraints on large-format systems often prevent a sufficiently rapid chop to be effective[1]. This leads us to consider a more optimal, but more computationally intensive, class of imaging algorithms.

The idea is that we seek to estimate N_{par} individual parameters (principally N_{pix} pixel values, N_{atm} atmosphere signal parameters, and N_{gain} bolometer gains) we represent as a vector $\mathbf{x}$. We do this using the $N_{data} = N_{samp} \times N_{detectors}$ data points. For Nyquist-sampled scans of the sky it is always the case that $N_{data} > N_{par}$ so the problem is overconstrained. The model parameters are further assumed to describe the data through a linear (or linearizable) system of N_{data} equations

$$\mathbf{d} = \mathbf{A}\mathbf{x} \tag{1}$$

Since any model can be linearized about the good solutions this is not a very restrictive assumption, however see below. The direct least-squares solution to the problem is

$$\mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{d} \tag{2}$$

Including optimal data weighting, this takes the form

$$\mathbf{x} = (\mathbf{A}^T\mathbf{C}_{TS}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{C}_{TS}^{-1}\mathbf{d} \tag{3}$$

where $\mathbf{C}_{TS}$ is the timestream (raw data) covariance matrix. The advantages of this approach include:

- The characteristic (scan-strategy dependent) differences in the time-stream signatures of the astronomy signal and the atmosphere are preserved and exploited, rather than destroyed as in most current bolometer array analyses which simply grid and average the data on the sky.

[1]For instance, BOLOCAM on the CSO cannot chop the tertiary for this reason. The GBT currently has no tertiary, and the secondary cannot be nodded faster than ~ 0.6 Hz— the primary more slowly than this.

- Direct solution of Eq 3 can be demonstrated to be the *optimal* image reconstructions in the χ^2 sense, and also returns the covariance matrix $[(\mathbf{A}^T\mathbf{C}_{TS}^{-1}\mathbf{A})^{-1}]$ of the sky pixels and other model parameters. This is useful for faint source detection projects, small-scale CMB measurements, or model fitting.

- Array "self-calibration" is a natural extension of the procedure (see below).

The disadvantage is that this approach is more computationally intensive than simply gridding the data onto the sky. Consequently we will consider the computational feasibility of variants to this approach for the specific observing cases of Table 1. The key metrics in all cases are the number of data points N_{data} (the length of the data vector $\mathbf{d}$) and the number of model parameters N_{par} (the length of the model vector $\mathbf{x}$). The model parameters are principally $4'' \times 4''$ sky pixels; bolometer gain parameters; and atmosphere time-series coefficients. Assuming the bolometer gains vary significantly with a characteristic timescale of $\sim 0.1\,\mathrm{Hz}$, and that the atmosphere is a DC level across the array (a good approximation for a large telescope such as the GBT, at least with this size an FOV), then these two components contribute $\sim 138,000 + 22,000 = 160,000$ parameters in a 6 hour observation. The total number of model parameters is then $N_{par} \sim 160,000 + N_{pix}$. Note that the gain parameters are fairly sparse ($\sim 1\%$ or lower filling factor) but the atmosphere coefficents are not.

3.1 Self-Calibration

Some parts of the problem are not linearizable, for instance, the problem is intrinsically bilinear in data $\times$ gain. In this case Eqs 1, 2, or 3 can be alternately solved for the sky/atmosphere parameters and the gains. This iterative scheme is similar to the self-calibration cycle often applied to synthesis data. For bolometer systems, unlike aperture synthesis instruments, you *always* have a bright self-calibratable source: the atmosphere (also, the cal diode). The comparative ease of including instrument calibration is an advantage of this formalism, and is important since highly accurate calibration is needed to image faint structures against a bright foreground such as the atmosphere.

In the iterative scheme the majority of the "non-pixel" model parameters separate from the rest of the problem. That is, rather than a single

$$N_{par} = N_{pix} + N_{gain} + N_{atm} \sim N_{pix} + 138,000 + 22,000 \tag{4}$$

(for 6 hours) problem, you iterate on an $N = N_{pix} + 22,000$ problem and an $N = 138,000$ problem. (*what about data corrlations introduced by gain uncertainties?*). An alternate approach would be to iteratively solve the full (bilinear) system of equations by recasting Eq. 1 as

$$\mathbf{d} = \mathbf{A}(\mathbf{x}_i)\,\mathbf{x}_{i+1} \tag{5}$$

where $\mathbf{A}$ now depends on the previous model vector solution. If computationally feasible this is likely to be more optimal.

Note that the gain and sky parts of $\mathbf{A}$ are fairly sparse, but the atmosphere part is densely filled. If included in the full linear algebra solution these are the rate limiting factor for projects of this nature (hours integration; arcminutes square). In any case, the iterative solutions for projects of this size will be straightforward. Simple linear operations on the timestreams can give good initial values for all of the parameters which will greatly speed the convergence of the iterative schemes.

3.2 Feasibility

In its simplest form, the direct sky solution requires the inversion of $\mathbf{A}^T\mathbf{A}$, an $N_{par} \times N_{par}$ matrix. Experience with CBI data analysis on the CITA supercomputing cluster suggests that this will be feasible for problems with up to $N_{par} = 10^4$ in ~ 24 hours. Then even for the simplest formulation of the problem, the degree-scale ($1^\circ \times 1^\circ$ and $8^\circ \times 8^\circ$) maps are not amenable to direct solution. To this add that we have not considered the more optimal formulations (*e.g.*Eq. 3), which requires inversion of the $10^8 \times 10^8$ or larger timestream covariance matrix. This is unfeasible and even *storage* is a challenge. A more clever approach is needed.

We consider first the largest (degree field) problems. These have the advantage of being pixel dominated: $N_{pix} > N_{atm} + N_{gain}$. Since most of the entries in $\mathbf{A}$ only connect one or a few timestream samples with a sky pixel, the matrix is quite sparse. This offers the prospect of significant speedups and reduction of storage requirements. Note that $\mathbf{A}^T\mathbf{A}$ may not share this character and the approach is generally an interative solution of Eq. 1. There are many "industrial" packages, such as LSQR, to solve large, sparse linear least squares problems. Most of these allow constraints to be placed on the model parameters, which will help control degenerate modes in the fit (*e.g.*a DC offset in the sky pixel values versus such an offset in the atmosphere model) or calibrate the data. Many of these industrial codes are publicly available as `C`, `FORTRAN`, and even `MATLAB` codes, which will speed development. Our informal survey of the literature suggests that problems of the scale we are discussing are amenable to iterative solution.

The arcminute field problems may be amenable to more direct approaches such as some variant of Eq 2 or 3. Since these yield estimates of the pixel covariance matrix, this will be desirable, *e.g.*, for very deep galaxy searches. Note that these smaller problems are less pixel-dominated— the gain and atmosphere parameters play a bigger role when the rate of sky coverage is less.

The formulation of imaging above is quite general. We intend to develop it first by application to data from existing GBT dual-beam receivers. Some open issues we would like to address with this data:

- Find a good practical approximation for the timestream covariance $\mathbf{C}_{TS}$ — *e.g.*, can we break the problem up and neglect correlations between scans? In this case the noise can be taken to be constant over some timescale and FFT methods could be used to invert C_{TS} making that a

N_{data}^2 step. This will enable a more accurate pixel correlation matrix to be constructed, and more sensitive images to be made.

- How well do the gains separate from the other model parameters without unacceptable correlations being introduced? How well do the iterative schemes converage?

- Investigate alternate schemes to deal with the atmosphere, *e.g.*, the "timestream projection operator" approach used by Kuo *et al.* (astro-ph/0212289) to make images with scanned bolometer data from ACBAR.