extra notes

Bai-Qiang Qiang

July 30, 2021

1 Map Making Setup

The map making procedure could be summarized in equation

$$\mathbf{d} = P\mathbf{m} + \mathbf{n} \tag{1}$$

where $\mathbf{d}$, P, $\mathbf{m}$, $\mathbf{n}$ are time-ordered data (TOD), pointing matrix, CMB map, and noise. The time-ordered data we collected is given by map signal $P\mathbf{m}$ plus noise $\mathbf{n}$. The pointing matrix P acting on the map gives the signal of the map at some specific position of sky where telescope is pointing at. Here we could assume that the noise has zero mean $\langle \mathbf{n} \rangle = \mathbf{0}$, since if it is not zero, we can always subtract its mean value to make it zero. And noise covariance matrix could be written as $N = \left\langle \mathbf{n}\mathbf{n}^\dagger \right\rangle$.

As we can see the map making model Eq.(1) mathematically is a standard linear regression problem, with *design matrix* being pointing matrix P, and *regression coefficients* are $\mathbf{m}$. Naturally, we want to estimate linear regression coefficients $\mathbf{m}$, with *generalized least square* (GLS) technique. The noise $\mathbf{n}$ is *heteroscedastic*: the variances N are different for various frequencies. Usually detectors have a $1/f$ noise pattern [5]. The *generalized least square* (GLS) will provide better estimation than *ordinary least square* (OLS) method, because the data is heteroscedastic so we would like to focusing on fitting the data with lower noise.

The GLS estimated map $\hat{\mathbf{m}}$ is given by

$$\hat{\mathbf{m}} = \arg\min_{\mathbf{m}} (\mathbf{d} - P\mathbf{m})^\dagger N^{-1} (\mathbf{d} - P\mathbf{m}) \tag{2}$$

and we could define

$$\chi^2(\mathbf{m}) \equiv (\mathbf{d} - P\mathbf{m})^\dagger N^{-1} (\mathbf{d} - P\mathbf{m}). \tag{3}$$

Therefore the estimated map $\hat{\mathbf{m}}$ is the one that minimizes $\chi^2(\mathbf{m})$. To find out the expression for $\hat{\mathbf{m}}$, we first take derivative with respect to vector $\mathbf{m}$

$$\begin{aligned}\frac{\partial}{\partial \mathbf{m}}\chi^2(\mathbf{m}) &= \frac{\partial}{\partial \mathbf{m}}(\mathbf{d} - P\mathbf{m})^\dagger N^{-1}(\mathbf{d} - P\mathbf{m}) \\ &= \frac{\partial}{\partial \mathbf{m}}\Big(\mathbf{d}^\dagger N^{-1}\mathbf{d} - \mathbf{d}^\dagger N^{-1}P\mathbf{m} - \mathbf{m}^\dagger P^\dagger N^{-1}\mathbf{d} + \mathbf{m}^\dagger P^\dagger N^{-1}P\mathbf{m}\Big) \\ &= -2P^\dagger N^{-1}\mathbf{d} + 2P^\dagger N^{-1}P\mathbf{m}\end{aligned} \tag{4}$$

then set it equal to zero $\frac{\partial}{\partial \mathbf{m}}\chi^2(\hat{\mathbf{m}}) = 0$. We get the *map making equation*

$$\hat{\mathbf{m}} = \Big(P^\dagger N P\Big)^{-1} P^\dagger N^{-1}\mathbf{d} \tag{5}$$

This is also called COBE method for map making.

2 Convient Properties for Map-making Equation

2.1 Unbiased linear estimator

A linear estimator means $\hat{\mathbf{m}}$ could be written as $\hat{\mathbf{m}} = W\mathbf{d}$, that is, it is a linear combination of $\mathbf{d}$. We say that the estimator is unbiased if

$$\begin{aligned}&\langle\hat{\mathbf{m}}\rangle = \mathbf{m} \\ \Rightarrow\ &\langle W\mathbf{d}\rangle = \mathbf{m} \\ \Rightarrow\ &\langle W(P\mathbf{m} + \mathbf{n})\rangle = \mathbf{m} \\ \Rightarrow\ &WP = I.\end{aligned} \tag{6}$$

At last step we used the property $\langle\mathbf{n}\rangle = 0$. The generalized least square estimator matrix $W = \big(P^\dagger N P\big)^{-1} P^\dagger N^{-1}$ satisfies the condition $WP = I$. Therefore $\hat{\mathbf{m}}$ is an unbiased estimated map.

2.2 Minimum variance linear estimators under the unbiased constraint

The covariance of the estimator $\hat{\mathbf{m}} = W\mathbf{d}$ is

$$\begin{aligned}\mathrm{Cov}[\hat{\mathbf{m}}] &= \mathrm{Cov}[W\mathbf{d}] \\ &= \mathrm{Cov}[WP\mathbf{m} + W\mathbf{n}] \\ &= \mathrm{Cov}[W\mathbf{n}] \\ &= WNW^\dagger\end{aligned} \tag{7}$$

Here we use a trick [4] and consider the matrix $W = W_{GLS} + W'$ where $W_{GLS} = \big(P^\dagger N P\big)^{-1} P^\dagger N^{-1}$ is the matrix for GLS estimation, and in order to satisfy the condition

$WP = I$, we should have $W'P = 0$. Then the covariance matrix

$$
\begin{aligned}
\mathrm{Cov}[\hat{\mathbf{m}}] &= W_{GLS} N W_{GLS}^{\dagger} + W' N W'^{\dagger} + W_{GLS} N W'^{\dagger} + W' N W_{GLS}^{\dagger} \\
&= \left(P^{\dagger} N P\right)^{-1} + W' N W'^{\dagger} + \left(P^{\dagger} N P\right)^{-1} P^{\dagger} W'^{\dagger} + W' P \left(P^{\dagger} N P\right)^{-1} \\
&= \left(P^{\dagger} N P\right)^{-1} + W' N W'^{\dagger}
\end{aligned} \tag{8}
$$

where the last line used condition $W'P = 0$.

The variance $\mathrm{Var}[\hat{\mathbf{m}}_i]$ is diagonal elements of covariance matrix $\mathrm{Cov}[\hat{\mathbf{m}}]$

$$
\begin{aligned}
\mathrm{Var}[\hat{\mathbf{m}}_i] &= \left\{\left(P^{\dagger} N P\right)^{-1}\right\}_{ii} + \left\{W' N W'^{\dagger}\right\}_{ii} \\
&= \left\{\left(P^{\dagger} N P\right)^{-1}\right\}_{ii} + W'_{i,:} N W'^{\dagger}_{i,:}
\end{aligned} \tag{9}
$$

where $W'_{i,:}$ is the i^{th} row vector of W'. Since the noise covariance matrix N is a positive semi-definite matrix, therefore $W'_{i,:} N W'^{\dagger}_{i,:} \geq 0$. If $W' = 0$ Then we have $W = W_{GLS}$, and the variance $\mathrm{Var}[\hat{\mathbf{m}}_i]$ would have its minimum variance.

2.3 Minimize mean square error under constrain of unbiased linear estimator

The error is defined as the difference between estimated map and real one

$$
\begin{aligned}
\boldsymbol{\varepsilon} &\equiv \hat{\mathbf{m}} - \mathbf{m} \\
&= W\mathbf{d} - \mathbf{m} \\
&= (WP - I)\mathbf{m} + W\mathbf{n} \\
&= W\mathbf{n}
\end{aligned} \tag{10}
$$

where the last line used relation $WP = I$ for unbiased estimator W.

Now we need to minimize mean square error

$$
\begin{aligned}
\left\langle \boldsymbol{\varepsilon}^{\dagger} \boldsymbol{\varepsilon} \right\rangle &= \left\langle \mathrm{Tr}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\dagger}\right) \right\rangle \\
&= \left\langle \mathrm{Tr}\left(W \mathbf{n}\mathbf{n}^{\dagger} W^{\dagger}\right) \right\rangle \\
&= \mathrm{Tr}\left(W N W^{\dagger}\right) \\
&= \mathrm{Tr}(\mathrm{Cov}[\hat{\mathbf{m}}]) \\
&= \sum_i \mathrm{Var}[\hat{\mathbf{m}}_i]
\end{aligned} \tag{11}
$$

In the second line we used property $\boldsymbol{\varepsilon}^\dagger\boldsymbol{\varepsilon}$ is a scalar, so $\boldsymbol{\varepsilon}^\dagger\boldsymbol{\varepsilon} = \mathrm{Tr}\big(\boldsymbol{\varepsilon}^\dagger\boldsymbol{\varepsilon}\big) = \mathrm{Tr}\big(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\dagger\big)$. In the fourth line the trace is a linear operation and $\big\langle \mathbf{n}\mathbf{n}^\dagger \big\rangle = N$. The fifth line comes from Eq.(7). In Section 2.2 we have shown that the generalized least square matrix W_{GLS} minimizes $\mathrm{Var}[\hat{\mathbf{m}}_i]$ for each i, therefore it also minimizes the mean square error $\big\langle \boldsymbol{\varepsilon}^\dagger\boldsymbol{\varepsilon} \big\rangle = \sum_i \mathrm{Var}[\hat{\mathbf{m}}_i]$.

2.4 Maximum likelihood estimator for Gaussian noise

The previous properties does not depends on the noise distribution, if we assume that the noise has a multivariate normal distribution, $\mathbf{n} \sim \mathcal{N}(0, N)$ with mean 0 covariance N, its likelihood function will be

$$L(\mathbf{d};\mathbf{m}) = \frac{1}{\sqrt{(2\pi)^n |N|}} \exp\left(-\frac{1}{2}(\mathbf{d} - P\mathbf{m})^\dagger N (\mathbf{d} - P\mathbf{m})\right) \tag{12}$$

and log-likelihood

$$\log\left(L(\mathbf{d};\mathbf{m})\right) = -\frac{1}{2}(\mathbf{d} - P\mathbf{m})^\dagger N (\mathbf{d} - P\mathbf{m}) + \mathrm{cont.} \tag{13}$$

Maximizing this log-likelihood function with respect to $\mathbf{m}$, is equivalent to minimize $\chi^2(\mathbf{m}) = (\mathbf{d} - P\mathbf{m})^\dagger N (\mathbf{d} - P\mathbf{m})$, which is $\hat{\mathbf{m}}$.

3 Solving the Map Making Equation

The map making equation Eq.(5) derived from Generalized Least Square estimation,

$$\Big(P^\dagger N^{-1} P\Big)\hat{\mathbf{m}} = P^\dagger N^{-1}\mathbf{d} \tag{14}$$

If we define $A = P^\dagger N^{-1} P$ and $\mathbf{b} = P^\dagger N^{-1}\mathbf{d}$, then it could be written as $A\hat{\mathbf{m}} = \mathbf{b}$.

Based on current computation power, it is impossible to solve $\hat{\mathbf{m}}$ by calculating $\hat{\mathbf{m}} = \big(P^\dagger N^{-1} P\big)^{-1} P^\dagger N^{-1}\mathbf{d}$ directly, since the noise covariance matrix N is sparse in frequency domain, and pointing matrix P is sparse in (time by pixel) domain. In experiments currently under design, there may be $\sim 10^{16}$ time samples and $\sim 10^9$ pixels, so these matrix inversions are intractable. It is impossible to do these matrix multiplication directly and then take inverse. However, for a vector with size of map $\hat{\mathbf{m}}$, we could calculate $P^\dagger N^{-1} P\hat{\mathbf{m}} = A\hat{\mathbf{m}}$ by first taking Fourier transform $P\hat{\mathbf{m}}$ then inverse Fourier transform $N^{-1}P\hat{\mathbf{m}}$. This means it can be solved by the conjugate gradient method.

3.1 Preconditioner

To improve the performance of the conjugate gradient method, we could apply a preconditioner M to original problem $A\hat{\mathbf{m}} = \mathbf{b}$, which then becomes $M^{-1}A\hat{\mathbf{m}} = M^{-1}\mathbf{b}$. The preconditioner should reduce the condition number of original problem, so that the conjugate gradient method will converge faster. We want the preconditioner to capture as much information as possible from matrix A, but still keep it relative easy to calculate M^{-1}. For example, if $M = A$, $M^{-1}A\hat{\mathbf{m}} = M^{-1}\mathbf{b}$ would be solved immediately, but M^{-1} will be extremely difficult to calculate. We could simply choose $M = P^\dagger P$, and the operation $M^{-1}\mathbf{m} = \left(P^\dagger P\right)^{-1}\mathbf{m}$ is the average over each pixel of map $\mathbf{m}$.

For the conjugate gradient method, we need an initial guess map $\hat{\mathbf{m}}_0$. We can use zero vector $\hat{\mathbf{m}}_0 = \mathbf{0}$ as initial guess, but the simple binned map $\hat{\mathbf{m}}_0 = \left(P^\dagger P\right)^{-1}P^\dagger\mathbf{d}$ is a better choice (it is a the solution for white noise case $N \propto I$). (Papež et al. 2018[3]) showed that using $\hat{\mathbf{m}}_0$ as initial guess could improve performance significantly compare to zero vector $\mathbf{0}$ in come cases. As stated before, we can calculate $\left(P^\dagger P\right)^{-1}$ acting on any map-sized object, and $P^\dagger\mathbf{d}$ is indeed a map size object, so we could obtain simple binned map by calculating $\hat{\mathbf{m}}_0 = \left(P^\dagger P\right)^{-1}P^\dagger\mathbf{d}$ directly.

For the conjugate gradient method with simple preconditioner $M = P^\dagger P$, we have all we need. Next we only need to use conjugate gradient algorithm solve the problem.

3.2 Parameterized Conjugate Gradient Method

The above results have appeared in the literature. Here we now show that we can improve performance in some cases by using a parameterized version of the map making equation Eq.(14). The idea is that map making equation Eq.(14) is hard to solve due to noise covariance matrix is sandwiched between $P^\dagger P$. But if noise covariance matrix N is proportional to identity matrix I, then its solution is given by simple binned map $\mathbf{m}_0 = \left(P^\dagger P\right)^{-1}P^\dagger\mathbf{d}$, which could be solved directly. We can parameterize the noise covariance matrix N with a parameter η, such that initially $\eta = \eta_i$ we have $N(\eta_i) \propto I$ and in the end $\eta = \eta_f$ and $N(\eta_f) \propto N$, such that the final solution is what we want. We expect that the parameterized noise covariance matrix $N(\eta)$ would connect our initial guess $\hat{\mathbf{m}}_0$ and final solution $\hat{\mathbf{m}}$ as we change η from η_i to η_f.

Now instead of Eq.(14), we are solving

$$\left(P^\dagger N(\eta)^{-1}P\right)\hat{\mathbf{m}}(\eta) = P^\dagger N(\eta)^{-1}\mathbf{d} \tag{15}$$

Now question is how to find $N(\eta)$ such that $N(\eta_i) \propto I$ and $N(\eta_f) \propto N$? Since the non-white noise part of N is the difficult portion, we could think of it as a perturbation term, which adds upon the white noise. Initially the covariance is homoscedastic and solution is given by $\hat{\mathbf{m}}_0$, then we gradually add extra noise into this equation by changing η. At the end when $\eta = \eta_f$ we are solving equation Eq.(14).

Therefore we could separate noise covariance matrix into two parts $N = \tau I + \bar{N}$ where τ is the minimum eigenvalue of N. Then we define $N(\eta) = \tau I + \eta\bar{N}$, with parameter η represents the degree of *heteroscedasticity* which satisfies $\eta_i = 0$ and $\eta_f = 1$.

Eq.(15) then becomes

$$\left(P^\dagger(\tau I+\eta\bar{N})^{-1}P\right)\hat{\mathbf{m}}(\eta)=P^\dagger\left(\tau I+\eta\bar{N}\right)^{-1}\mathbf{d} \tag{16}$$

We require η being monotonically increase series $0=\eta_0<\eta_1<\cdots<\eta_n=1$. For some specific η_m, we use conjugate gradient method to solve equation $\left(P^\dagger N(\eta_m)^{-1}P\right)\hat{\mathbf{m}}(\eta_m)=P^\dagger N(\eta_m)^{-1}\mathbf{d}$ with simple preconditioner $P^\dagger P$, and using $\hat{\mathbf{m}}(\eta_{m-1})$ as the initial value. The initial guess is $\hat{\mathbf{m}}(\eta_0)=\mathbf{m}_0=\left(P^\dagger P\right)^{-1}P^\dagger\mathbf{d}$.

3.2.1 Choosing parameters η

The next question is how we choose these monotonically increasing parameters η. If we choose these parameters inappropriately, it would only makes it converge slower. Also we want to determine $\eta_1,\cdots,\eta_{n-1}$ before starting conjugate gradient iteration. That's because time ordered data $\mathbf{d}$ is very large, and we don't want to keep it in the system RAM during calculation. If $\eta_1,\cdots,\eta_{n-1}$ could be determined before the iterations, then we can first calculate $P^\dagger N(\eta)^{-1}\mathbf{d}$ for each η_m and store these map-sized objects in RAM, instead of the entire time-ordered data $\mathbf{d}$.

First let us try to find out our starting point η_1. What would be good value for η_1?

Here to simplify notation, I will use N_η to denote $N(\eta)$. The estimated map $\hat{\mathbf{m}}(\eta)=\left(P^\dagger N_\eta^{-1}P\right)^{-1}P^\dagger N_\eta^{-1}\mathbf{d}$ which minimizes

$$\chi^2(\mathbf{m},\eta)=(\mathbf{d}-P\mathbf{m})^\dagger N_\eta^{-1}(\mathbf{d}-P\mathbf{m}). \tag{17}$$

For some specific η value, the minimum χ^2 value is given by

$$\begin{aligned}\chi^2(\hat{\mathbf{m}}(\eta),\eta)&=\left(\mathbf{d}-P\hat{\mathbf{m}}(\eta)\right)^\dagger N_\eta^{-1}\left(\mathbf{d}-P\hat{\mathbf{m}}(\eta)\right)\\&=\mathbf{d}^\dagger\left[N_\eta^{-1}-N_\eta^{-1}P\left[P^\dagger N_\eta^{-1}P\right]^{-1}P^\dagger N_\eta^{-1}\right]\mathbf{d}\end{aligned} \tag{18}$$

Now let us see how $\chi^2(\hat{\mathbf{m}}(\eta),\eta)$ changes as we change η.

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\eta}\chi^2(\hat{\mathbf{m}}(\eta),\eta)&=\frac{\mathrm{d}}{\mathrm{d}\eta}\left(\mathbf{d}^\dagger N_\eta^{-1}\mathbf{d}\right)-\frac{\mathrm{d}}{\mathrm{d}\eta}\left(\mathbf{d}^\dagger N_\eta^{-1}P\left(P^\dagger N_\eta P\right)^{-1}P^\dagger N_\eta^{-1}\mathbf{d}\right)\\&=\mathbf{d}^\dagger N_\eta^{-1}\big[-\bar{N}+\bar{N}N_\eta^{-1}P\left(P^\dagger N_\eta P\right)^{-1}P^\dagger\\&\qquad-P\left(P^\dagger N_\eta^{-1}P\right)^{-1}P^\dagger N_\eta^{-1}\bar{N}N_\eta^{-1}P\left(P^\dagger N_\eta^{-1}P\right)^{-1}P^\dagger\\&\qquad+P\left(P^\dagger N_\eta^{-1}P\right)^{-1}P^\dagger N_\eta^{-1}\bar{N}\big]N_\eta^{-1}\mathbf{d}\end{aligned} \tag{19}$$

Simplify this expression with identity $\hat{\mathbf{m}}=\left(P^\dagger N_\eta^{-1}P\right)^{-1}P^\dagger N_\eta^{-1}\mathbf{d}$, and it yields

$$\frac{\mathrm{d}}{\mathrm{d}\eta}\chi^2(\hat{\mathbf{m}}(\eta),\eta)=-(\mathbf{d}-P\hat{\mathbf{m}}(\eta))^\dagger N_\eta^{-1}\bar{N}N_\eta^{-1}(\mathbf{d}-P\hat{\mathbf{m}}(\eta)). \tag{20}$$

Also notice that $\frac{\mathrm{d}}{\mathrm{d}\eta}\chi^2(\hat{\mathbf{m}}(\eta),\eta) = \frac{\partial}{\partial\eta}\chi^2(\hat{\mathbf{m}}(\eta),\eta)$ (the total derivative is the partial derivative), because by the definition of $\hat{\mathbf{m}}(\eta)$ it minimize $\chi^2(\mathbf{m},\eta)$ for some fixed η value, implies $\frac{\partial}{\partial\mathbf{m}}\chi^2(\hat{\mathbf{m}}(\eta),\eta) = 0$.

To further simplify the analysis, let's assume that the noise covariance matrix $N = \left\langle \mathbf{n}\mathbf{n}^\dagger \right\rangle$ is diagonal in the frequency domain. Therefore $\bar{N}$ and N_η are also diagonal in the frequency domain by definition, and all the diagonal elements are greater than or equal to zero, because covariance matrix is positive semi-definite. Also, we can conclude that matrix $N_\eta^{-1}\bar{N}N_\eta^{-1}$ is positive semi-definite matrix. Based on Eq. (20), we know that $\frac{\mathrm{d}}{\mathrm{d}\eta}\chi^2(\hat{\mathbf{m}}(\eta),\eta) \le 0$, so $\chi^2(\hat{\mathbf{m}}(\eta),\eta)$ is always decreasing as η changes from 0 to 1.

The fractional decrease of $\chi^2(\hat{\mathbf{m}}(\eta),\eta)$ at η is defined as

$$\begin{aligned}
-\frac{\delta\chi^2(\hat{\mathbf{m}}(\eta),\eta)}{\chi^2(\hat{\mathbf{m}}(\eta),\eta)} &= -\delta\eta\frac{1}{\chi^2(\hat{\mathbf{m}}(\eta),\eta)}\,\frac{\mathrm{d}}{\mathrm{d}\eta}\chi^2(\hat{\mathbf{m}}(\eta),\eta) \\
&= \delta\eta\frac{(\mathbf{d}-P\hat{\mathbf{m}}(\eta))^\dagger N_\eta^{-1}\bar{N}N_\eta^{-1}(\mathbf{d}-P\hat{\mathbf{m}}(\eta))}{\big(\mathbf{d}-P\hat{\mathbf{m}}(\eta)\big)^\dagger N_\eta^{-1}\big(\mathbf{d}-P\hat{\mathbf{m}}(\eta)\big)}
\end{aligned} \tag{21}$$

Here we put a minus sign in front of $\delta\chi^2(\hat{\mathbf{m}}(\eta),\eta)/\chi^2(\hat{\mathbf{m}}(\eta),\eta)$, such that it's non-negative. If we choose $\eta_1 = \eta_0+\delta\eta = \delta\eta$ such that $\eta_1 = \delta\eta$ is very small quantity. Then the fractional decrease from $\eta_0 = 0$ to $\eta_1 = \delta\eta$ is

$$-\frac{\delta\chi^2(\hat{\mathbf{m}}(0),0)}{\chi^2(\hat{\mathbf{m}}(0),0)} = \delta\eta\frac{1}{\tau}\frac{(\mathbf{d}-P\hat{\mathbf{m}}(0))^\dagger\bar{N}(\mathbf{d}-P\hat{\mathbf{m}}(0))}{\big(\mathbf{d}-P\hat{\mathbf{m}}(0)\big)^\dagger\big(\mathbf{d}-P\hat{\mathbf{m}}(0)\big)} \tag{22}$$

where we used the property $N_{\eta=0} = \tau I$.

We want $\left|\delta\chi^2(\hat{\mathbf{m}}(\eta_0),\eta_0)\right| = \chi^2(\hat{\mathbf{m}}(\eta_0),\eta_0)) - \chi^2(\hat{\mathbf{m}}(\eta_1),\eta_1)$ to be large such that it could converge fast. Which means $\chi^2(\hat{\mathbf{m}}(\eta_1),\eta_1)$ is much smaller than $\chi^2(\hat{\mathbf{m}}(\eta_0),\eta_0)$, or $\chi^2(\hat{\mathbf{m}}(\eta_1),\eta_1) \ll \chi^2(\hat{\mathbf{m}}(\eta_0),\eta_0)$. Then we would expect

$$-\frac{\delta\chi^2(\hat{\mathbf{m}}(0),0)}{\chi^2(\hat{\mathbf{m}}(0),0)} = 1 - \frac{\chi^2(\hat{\mathbf{m}}(\eta_1),\eta_1)}{\chi^2(\hat{\mathbf{m}}(0),0)} \approx 1^- \tag{23}$$

The upper bound is strictly smaller than 1. Now we could use Eq.(22) and let it equal to 1, then $\delta\eta = -\chi^2(\hat{\mathbf{m}}(\eta_0),\eta_0)/\frac{\mathrm{d}}{\mathrm{d}\eta}\chi^2(\hat{\mathbf{m}}(\eta_0),\eta_0)$. However if we apply this idea to $\eta_{m+1} = \eta_m + \delta\eta_m$ with $m \ge 1$, we would get

$$\delta\eta_m = -\chi^2(\hat{\mathbf{m}}(\eta_m),\eta_m)/\frac{\mathrm{d}}{\mathrm{d}\eta}\chi^2(\hat{\mathbf{m}}(\eta_m),\eta_m). \tag{24}$$

As mentioned before, we need to determine the entire series $\{\eta_i\}$ before conjugate gradient iterations, and we could not calculate $\hat{\mathbf{m}}(\eta_m)$ directly because of the difficulty of matrix inversions. Therefore we could not get $\delta\eta_m$ values in advance. That means we need to find another approach.

Let us go back to Eq.(22). Since it is hard to analyze $\mathbf{d} - P\hat{\mathbf{m}}(\eta)$ under frequency domain, we treat it as an arbitrary vector, then the least upper bound of Eq.(22) is given by

$$-\frac{\delta\chi^2(\hat{\mathbf{m}}(\eta_0), \eta_0)}{\chi^2(\hat{\mathbf{m}}(\eta_0), \eta_0)} \leq \frac{\delta\eta}{\tau} \max(\bar{N}_f) \tag{25}$$

where $\max(\bar{N}_f)$ is the maximum eigenvalue of $\bar{N}$. We want $-\frac{\delta\chi^2(\hat{\mathbf{m}}(\eta_0),\eta_0)}{\chi^2(\hat{\mathbf{m}}(\eta_0),\eta_0)}$ being as large as possible, but it won't exceed 1. If we combine Eq. (23) and Eq. (25), and choose $\delta\eta$ such that the least upper bound is equal to 1, to make sure the process would not going too fast. Thus we have

$$\eta_1 = \frac{\tau}{\max(\bar{N}_f)} = \frac{\min(N_f)}{\max(N_f) - \min(N_f)}. \tag{26}$$

Here N_f and $\bar{N}_f$ are the eigenvalues of N and $\bar{N}$ in the frequency domain. If the condition number of noise covariance matrix $\kappa(N) = \max(N_f)/\min(N_f) \gg 1$, then $\eta_1 \approx \kappa^{-1}(N)$.

What about the other parameters η_m with $m > 1$? We use a similar analysis, letting $\eta_{m+1} = \eta_m + \delta\eta_m$ with a small $\delta\eta_m \ll 1$. First, let us find the least upper bound

$$-\frac{\delta\chi^2(\hat{\mathbf{m}}(\eta_m), \eta_m)}{\chi^2(\hat{\mathbf{m}}(\eta_m), \eta_m)} = \delta\eta_m \frac{(\mathbf{d} - P\hat{\mathbf{m}}(\eta_m))^\dagger N_{\eta_m}^{-1} \bar{N} N_{\eta_m}^{-1} (\mathbf{d} - P\hat{\mathbf{m}}(\eta_m))}{\big(\mathbf{d} - P\hat{\mathbf{m}}(\eta_m)\big)^\dagger N_{\eta_m}^{-1} \big(\mathbf{d} - P\hat{\mathbf{m}}(\eta_m)\big)} \tag{27}$$

$$\leq \delta\eta_m \max\left(\frac{\bar{N}_f}{\tau + \eta_m \bar{N}_f}\right) \tag{28}$$

The upper bound in the second line is a little bit tricky. Both matrix $\bar{N}$ and $N_{\eta_m}^{-1}$ can be simultaneously diagonalized in frequency space. For each eigenvector $\mathbf{e}_f$, the corresponding eigenvalue of the matrix on the numerator $N_{\eta_m}^{-1}\bar{N}N_{\eta_m}^{-1}$ is $\lambda_f = \bar{N}_f(\tau + \eta_m\bar{N}_f)^{-2}$, and the eigenvalue for matrix on the denominator $N_{\eta_m}^{-1}$ is $\gamma_f = (\tau + \eta_m\bar{N}_f)^{-1}$. Their eigenvalues are related by $\lambda_f = [\bar{N}_f/(\tau + \eta_m\bar{N}_f)]\gamma_f$. For any vector $\mathbf{v} = \sum_f \alpha_f \mathbf{e}_f$, we have

$$\frac{\mathbf{v}^\dagger N_{\eta_m}^{-1} \bar{N} N_{\eta_m}^{-1} \mathbf{v}}{\mathbf{v}^\dagger N_{\eta_m}^{-1} \mathbf{v}} = \frac{\sum_f \alpha_f^2 \lambda_f}{\sum_f \alpha_f^2 \gamma_f} = \frac{\sum_f \alpha_f^2 \gamma_f \bar{N}_f/(\tau + \eta_m \bar{N}_f)}{\sum_f \alpha_f^2 \gamma_f} \leq \max\left(\frac{\bar{N}_f}{\tau + \eta_m \bar{N}_f}\right). \tag{29}$$

Again assuming $\chi^2(\hat{\mathbf{m}}(\eta_{m+1}), \eta_{m+1}) \ll \chi^2(\hat{\mathbf{m}}(\eta_m), \eta_m)$, which we expect it to be satisfied for $\eta_m \ll 1$. That is because if $\eta \lesssim 1$, $\chi^2(\hat{\mathbf{m}}(\eta), \eta)$ would close to the minimum χ^2 which means $\chi^2(\hat{\mathbf{m}}(\eta_{m+1}), \eta_{m+1}) \lesssim \chi^2(\hat{\mathbf{m}}(\eta_m), \eta_m)$, which would violate our assumption. Luckily, the final result (**??**) is a geometric series, only the last few η_m values fail to satisfy this condition. Similarly, we could set the least upper bound equal to 1. Then we get

$$\delta\eta_m = \min\left(\frac{\tau + \eta_m \bar{N}_f}{\bar{N}_f}\right) = \eta_m + \frac{\tau}{\max(\bar{N}_f)}. \tag{30}$$

Therefore

$$\eta_{m+1} = \eta_m + \delta\eta_m = 2\eta_m + \frac{\tau}{\max(\bar{N}_f)} \tag{31}$$

The final term $\tau/\max(\bar{N}_f) = \eta_1$ becomes subdominant after a few terms, and we see that the η_m increase like a geometric series. If written in the form $\eta_{m+1} + \tau/\max(\bar{N}_f) = 2\big(\eta_m + \tau/\max(\bar{N}_f)\big)$ it's easy to see that for $m \geq 1$, $\eta_m + \tau/\max(\bar{N}_f)$ forms a geometric series

$$\eta_m + \frac{\tau}{\max(\bar{N}_f)} = \left(\eta_1 + \frac{\tau}{\max(\bar{N}_f)}\right)2^{m-1} = \frac{\tau}{\max(\bar{N}_f)}2^m \tag{32}$$

where we used $\eta_1 = \tau/\max(\bar{N}_f)$. Note that $m = 0$ and $\eta_0 = 0$ also satisfy this expression and we've got final expression for all η_m

$$\eta_m = \min\left\{1,\ \frac{\tau}{\max(\bar{N}_f)}(2^m - 1)\right\} \tag{33}$$

Here we need to truncate the series when $\eta_m > 1$.

This is the main result. Eq. (33) tells us not only how to choose parameters η_i, but also when we should stop the perturbation, and set $\eta = 1$. For example, if noise covariance matrix N is almost white noise, then $\bar{N} = N - \tau I \approx 0$, and we would have $\frac{\tau}{\max(\bar{N}_f)} \gg 1$. This tell us that we don't need to use parameterized method at all, because $\eta_1 = 1$. Note that the vanilla conjugate gradient method with simple binned map as initial guess corresponds to choosing $\eta_0 = 0$ and $\eta_1 = \eta_2 = \cdots = 1$.

3.2.2 Intuitive Interpretation of η

In this section, let me introduce another way to understand the role of η. Our ultimate goal is to find $\hat{\mathbf{m}}(1)$ which minimizes $\chi^2(\mathbf{m})$ in Eq. (3). Since N is diagonal in frequency space, χ^2 could be written as a sum of all frequency mode $|(\mathbf{d} - P\mathbf{m})_f|^2$ with weight N_f^{-1}, such as $\chi^2(\mathbf{m}) = \sum_f |(\mathbf{d} - P\mathbf{m})_f|^2 N_f^{-1}$. The weight is large for low noise frequency mode (small N_f), and vice versa. Which means $\chi^2(\mathbf{m})$ would favor the low noise frequency mode over high noise ones. In other words the optimal map $\hat{\mathbf{m}}$ focusing on minimize the error $\boldsymbol{\varepsilon} \equiv \mathbf{d} - P\mathbf{m}$ in the low-noise part.

After introducing η, we minimize $\chi^2(\mathbf{m}, \eta)$ in Eq. (17) instead. For $\eta = 0$, $N^{-1}(0) \propto I$ the system is homoscedastic and the estimated map $\hat{\mathbf{m}}(0)$ does not prioritize any frequency mode. As we slowly increase η, we decrease the weight for the high noise modes, and focusing minimizing error for low noise part. If we start with $\eta_1 = 1$ directly, which corresponds to the vanilla conjugate gradient method, then the entire conjugate gradient solver will focus most on minimizing the low noise part, such that χ^2 would converge very fast at low noise region, but slowly on high noise part. It may be stuck at some local minimum point and hard to get to global minimum. However by introducing η parameter, we let the solver first treat every frequency equally, then as η slowly increases, it gradually give more focus to the lowest noise part.

4 Messenger Field Method

The messenger field method is a fixed point iterative solver introduced by Elsner and Wandelt (2013) [1] to solve Wiener filter. Later on Huffenberger and Næss (2018) [2] applied this method to map-making problem, and showed that in some cases messenger field is better than conjugate gradient with a simple preconditioner. Papež el al.(2018) [3] proved that messenger field is equivalent to apply a preconditioner to map making equation Eq.(5), and it can be solved using both fixed point iteration and preconditioned conjugate gradient methods. They showed that in some conjugate gradient with simple preconditioner outperforms messenger field method with both fixed point iteration and preconditioned conjugate gradient methods.

Messenger field method similarly separate noise covariance matrix $N = \bar{N} + T$, with $T = \tau I$ and τ being the minimum eigenvalue of N. Then there is a cooling parameter λ such that $N(\lambda) = \bar{N} + \lambda T$, with initial λ being a very large number and final λ being 1. As you might guess λ is related to η by $\lambda = 1/\eta$.

Before introducing messenger field method, let's first prove one identity

$$\begin{aligned}
&\left(P^\dagger T^{-1} P\right)^{-1} P^\dagger T^{-1} \big(T^{-1} + \bar{N}^{-1}\big)^{-1} T^{-1} P \\
=&\left(P^\dagger T^{-1} P\right)^{-1} P^\dagger T^{-1} \big(I + T\bar{N}^{-1}\big)^{-1} P \\
=&\left(P^\dagger T^{-1} P\right)^{-1} P^\dagger T^{-1} \big(I - T\bar{N}^{-1} + T\bar{N}^{-1}T\bar{N}^{-1} - \cdots\big) P \\
=&\, I - \left(P^\dagger T^{-1} P\right)^{-1} P^\dagger \cancel{T^{-1}T} \bar{N}^{-1} \big(I - T\bar{N}^{-1} + T\bar{N}^{-1}T\bar{N}^{-1} - \cdots\big) P \\
=&\, I - \left(P^\dagger T^{-1} P\right)^{-1} P^\dagger \bar{N}^{-1} \big(I + T\bar{N}^{-1}\big)^{-1} P \\
=&\, I - \left(P^\dagger T^{-1} P\right)^{-1} P^\dagger \big(\bar{N} + T\big)^{-1} P \\
=&\, I - \left(P^\dagger T^{-1} P\right)^{-1} P^\dagger N^{-1} P
\end{aligned} \tag{34}$$

where at third and fifth line we used expansion $(I + A)^{-1} = I - A + A^2 - \cdots$

After apply preconditioner $P^\dagger T^{-1}P$ to the map making equation Eq.(5), we get:

$$
\begin{aligned}
&\Big(P^\dagger T^{-1}P\Big)^{-1}\Big(P^\dagger N^{-1}P\Big)\hat{\mathbf{m}} = \Big(P^\dagger T^{-1}P\Big)^{-1}P^\dagger N^{-1}\mathbf{d}\\
\Rightarrow&\hat{\mathbf{m}} - \Big(P^\dagger T^{-1}P\Big)^{-1}P^\dagger T^{-1}\big(T^{-1}+\bar{N}^{-1}\big)^{-1}T^{-1}P\hat{\mathbf{m}} = \Big(P^\dagger T^{-1}P\Big)^{-1}P^\dagger N^{-1}\mathbf{d}\\
\Rightarrow&\hat{\mathbf{m}} = \Big(P^\dagger T^{-1}P\Big)^{-1}P^\dagger T^{-1}\big(T^{-1}+\bar{N}^{-1}\big)^{-1}\big[T^{-1}P\hat{\mathbf{m}} + \big(T^{-1}+\bar{N}^{-1}\big)TN^{-1}\mathbf{d}\big]\\
\Rightarrow&\hat{\mathbf{m}} = \Big(P^\dagger T^{-1}P\Big)^{-1}P^\dagger T^{-1}\big(T^{-1}+\bar{N}^{-1}\big)^{-1}\big[T^{-1}P\hat{\mathbf{m}} + \big(I+\bar{N}^{-1}T\big)(\bar{N}+T)^{-1}\mathbf{d}\big]\\
\Rightarrow&\hat{\mathbf{m}} = \Big(P^\dagger T^{-1}P\Big)^{-1}P^\dagger T^{-1}\big(T^{-1}+\bar{N}^{-1}\big)^{-1}\big[T^{-1}P\hat{\mathbf{m}} + \bar{N}^{-1}\mathbf{d}\big] \qquad (35)
\end{aligned}
$$

where the second line we used identity Eq.(34).

To add cooling parameter λ, we only need to change T to λT and N to $N(\lambda)$. Then we could write it as fixed point iteration form

$$
\begin{cases}
\mathbf{t}_i = \big((\lambda T)^{-1}+\bar{N}^{-1}\big)^{-1}\big[(\lambda T)^{-1}P\hat{\mathbf{m}}_i + \bar{N}^{-1}\mathbf{d}\big]\\
\hat{\mathbf{m}}_{i+1} = \Big(P^\dagger(\lambda T)^{-1}P\Big)^{-1}P^\dagger(\lambda T)^{-1}\mathbf{t}_i
\end{cases} \qquad (36)
$$

This is fixed point iteration form of messenger field method. It's solving map making equation Eq.(5) with preconditioner $P^\dagger(\lambda T)^{-1}P$

$$
\Big(P^\dagger(\lambda T)^{-1}P\Big)^{-1}P^\dagger(\bar{N}+\lambda T)^{-1}P\hat{\mathbf{m}} = \Big(P^\dagger(\lambda T)^{-1}P\Big)^{-1}P^\dagger(\bar{N}+\lambda T)^{-1}\mathbf{d} \qquad (37)
$$

substitute $T=\tau I$

$$
\tau\Big(P^\dagger P\Big)^{-1}P^\dagger\Big(\tau I + \frac{1}{\lambda}\bar{N}\Big)^{-1}P\hat{\mathbf{m}} = \tau\Big(P^\dagger P\Big)^{-1}P^\dagger\Big(\tau I + \frac{1}{\lambda}\bar{N}\Big)^{-1}\mathbf{d} \qquad (38)
$$

since multiplying a constant won't change the condition number, it's equivalent to solve map making equation with perturbation parameter $\eta = 1/\lambda$ and simple preconditioner.

References

[1] F. Elsner and B. D. Wandelt. Efficient Wiener filtering without preconditioning. A&A, 549:A111, January 2013.

[2] Kevin M. Huffenberger and Sigurd K. Næss. Cosmic microwave background mapmaking with a messenger field. *The Astrophysical Journal*, 852(2):92, jan 2018.

[3] J. Papež, L. Grigori, and R. Stompor. Solving linear equations with messenger-field and conjugate gradient techniques: An application to CMB data analysis. A&A, 620:A59, November 2018.

[4] Cosma Shalizi. Weighted and generalized least squares.

[5] Max Tegmark. CMB mapping experiments: A designer's guide. Phys. Rev. D, 56(8):4514–4529, October 1997.