# MAXIMUM ENTROPY ANALYTIC CONTINUATION OF QUANTUM MONTE CARLO DATA

M. Jarrell
Department of Physics
University of Cincinnati
Cincinnati, Ohio 45221

*This article is submitted for Publication in*
**Advances in Physics**
(Ed. Kevin Bedell)

e-mail: Mark.Jarrell@uc.edu

**Abstract**

We present a pedagogical discussion of the Maximum Entropy Method which is a precise and systematic way of analytically continuing Euclidean-time quantum Monte Carlo results to real frequencies. Here, Bayesian statistics are used to determine which of the infinite number of real-frequency spectra are consistent with the QMC data is most probable. Bayesian inference is also used to qualify the solution and optimize the inputs. We develop the Bayesian formalism, present a detailed description of the data qualification, sketch an efficient algorithm to solve for the optimal spectra, and present a detailed case study to demonstrate the method.

# Contents


1 Introduction . . . 3
2 Formalism . . . 4
2.1 Green's Functions . . . 4
2.2 Bayesian Statistics . . . 5
2.3 Prior Probability . . . 6
2.4 Likelihood function . . . 8
2.5 Details of the MEM Formalism . . . 13
2.6 Model Selection . . . 16
2.7 Error Propagation . . . 17
3 Bryan's Method: a MEM algorithm . . . 17
3.1 Typical Algorithms . . . 18
3.2 Singular-Space Algorithm . . . 18
3.3 Selection of $\alpha$ . . . 20
3.4 Error Propagation . . . 21
4 Case Study . . . 22
4.1 Convergence of the Spectra . . . 22
4.2 Default Model Selection . . . 22
4.3 Error Propagation . . . 23
4.4 Two-Particle Spectra . . . 24
4.5 Annealing Method . . . 27
5 Conclusion . . . 28


## 1 Introduction

Most quantum Monte Carlo (QMC) simulations produce Green's functions $G(\tau)$ of imaginary time $\tau = it$. However, real-frequency results are crucial since most experiments probe dynamical quantities, including transport, densities of states, nuclear magnetic resonance, inelastic scattering, etc. Thus, the inability to extract real-frequency or real-time results from Euclidean (imaginary) time QMC simulations presents a significant limitation to the usefulness of the method. The relation between $G(\tau)$ and $A(\omega) = -\frac{1}{\pi}\mathrm{Im}G(\omega)$ is linear and surprisingly simple, for a Fermionic

single-particle Green's function $G$,

$$G(\tau) = \int d\omega \frac{A(\omega) e^{-\tau\omega}}{1 + e^{-\beta\omega}} \,. \tag{1}$$

Nevertheless, inversion is complicated by the exponential nature of the kernel $K(\tau, \omega) = e^{-\tau\omega}/(1 + e^{-\beta\omega})$[1]. For finite $\tau$ and large $\omega$ the kernel is exponentially small, so that $G(\tau)$ is insensitive to the high frequency features of $A(\omega)$. Equivalently, if we approximate both $G$ and $A$ by equal-length vectors and $K$ by a square matrix, then we find that the determinant of $K$ is exponentially small, so that $K^{-1}$ is ill-defined. Apparently, there are an infinite number of $A$ that yield the same $G$.

Previous attempts to address this problem include least-squares fits, Pade approximants and regularization. In the least squares method, Schüttler and Scalapino[1] approximated the spectrum with a set of box functions. The location and weight of these functions was determined by minimizing the least-squares misfit between the spectrum and the QMC data. However, as the number of box functions is increased to better resolve features in the spectrum, the fit becomes unstable and noisy. In the Pade method[2], $G$ (or rather its Fourier transform) is fit to a functional form, usually the ratio of two polynomials, which is then analytically continued formally by replacing $i\omega_n \rightarrow \omega + i0^+$. This technique works well when the data $G$ is very precise, as when analytic continuing Eliashberg equations, or when the fitting function is known *a priori*. However, it is generally unreliable for the continuation of less- precise QMC data to real frequencies. A more useful approach is to introduce regularization to the kernel, so that $K^{-1}$ exists. This method was developed by G. Wahba[3], and employed by White et al.[4] and Jarrell and Biham[5]. They used similar methods to minimize $(G - KA)^2$ subject to constraint potentials which introduce correlations between adjacent points in $A$ and impose positivity. However, these techniques tend to produce spectra $A$ with features which are overly smeared out by the regularization.

In the Maximum Entropy Method (MEM) we employ a different philosophy. Using Bayesian statistics, we define the posterior probability of the spectra $A$ given the data $G$, $P(A|G)$. We find the spectra which maximizes $P(A|G) \propto P(A)P(G|A)$ with the prior probability $P(A)$ defined so that $A$ has only those correlations that are required to reproduce the data $G$. To define the likelihood function $P(G|A)$, we take advantage of the statistical sampling nature of the QMC process.

In this chapter, we will present a short pedagogical development of the MEM to analytically continue QMC data. A more detailed review has been presented previously[6], and to the extent possible, we will follow the notation used there. This chapter is organized as follows: In Sec. 2, we will present the MEM formalism. In Sec. 3, the Bryan MEM algorithm will be sketched, which has been optimized for this type of problem. In Sec. 4, we will illustrate these techniques with the spectra of the Periodic Anderson model, described below, and finally in Sec. 5, we will conclude.

Throughout this chapter, we will illustrate the formalism and methods introduced with a simulation of the infinite-dimensional periodic Anderson model (PAM),

described by the Hamiltonian

$$H = \frac{-t^*}{2\sqrt{D}} \sum_{<ij>\sigma} (d^\dagger_{i\sigma} d_{j\sigma} + d^\dagger_{j\sigma} d_{i\sigma}) + V \sum_{i\sigma} (d^\dagger_{i\sigma} f_{i\sigma} + f^\dagger_{i\sigma} d_{i\sigma}) + \frac{U}{2} \sum_{i\sigma} (n^f_{i,\sigma} - \frac{1}{2})(n^f_{i,-\sigma} - \frac{1}{2}) \quad (2)$$

where $d_{i\sigma}$ and $f_{i\sigma}$ ($d^\dagger_{i\sigma}$ and $f^\dagger_{i\sigma}$) destroy (create) a d- and f-electron on site $i$ with spin $\sigma$, $U$ is the screened Coulomb-matrix element for the localized f-states, and $V$ characterizes the mixing between the two subsystems. We will study (2) on a simple hypercubic lattice of dimension $D \to \infty$ with hybridization $t = t^*/(2\sqrt{D})$ restricted to nearest-neighbors. We choose $t^* = 1$ as a convenient energy scale for the remainder of this chapter. The algorithm to solve infinite-dimensional lattice problems will be discussed in more detail in Chap. IV; however, the core of this algorithm is the Hirsch-Fye impurity algorithm[7]. Here the problem is cast into a discrete path formalism in imaginary time, $\tau_l$, where $\tau_l = l\Delta\tau$, $\Delta\tau = \beta/L$, $\beta = 1/k_B T$, and $L$ is the number of times slices. Euclidean-time Green's functions are measured on this discrete time domain.

## 2 Formalism

### *2.1 Green's Functions*

If this system is perturbed by an external field which couples to an operator $B$, then the linear response to this field is described by the retarded Green's function

$$G(t) = -i\Theta(t) \left\langle \left[ B(t), B^\dagger(0) \right]_\pm \right\rangle \quad (3)$$

where the negative (positive) sign is used for Boson (Fermion) operators $B$ and $B^\dagger$, and makes reference to the Dirac (anti)commutator. The Fourier transform of $G(t)$, $G(z)$ is analytic in the upper half plane, and its real and imaginary parts are related by

$$G(z) = \int d\omega \frac{\frac{-1}{\pi} \mathrm{Im} G(\omega)}{z - \omega} \,. \quad (4)$$

The Matsubara-frequency Green's function $G(i\omega_n)$ is obtained by letting $z \to i\omega_n$ in Eq. 4. This may be Fourier transformed to yield a relation between the Euclidean-time Green's function produced by the QMC procedure, and $\frac{-1}{\pi}\mathrm{Im}G(\omega)$

$$G(\tau) = \int d\omega \frac{\frac{-1}{\pi} \mathrm{Im} G(\omega) e^{-\tau\omega}}{1 \pm e^{-\beta\omega}} \,. \quad (5)$$

### *2.2 Bayesian Statistics*

We use our QMC algorithm to generate a set $\bar{G}^i_l$ of $i = 1, N_d$ estimates for the Green's function at each time slice $\tau_l = (l-1)\Delta\tau$, $l = 1, L$. Since many $A$ correspond

to the same data $\bar{G}$, we must employ a formalism to determine which $A(\omega)$ is the most probable given the statistics of the data and an prior information that we have about $A$. To quantify the conditional probability of $A$ given the data, and our prior knowledge, we use Bayesian statistics.

If we have two events $a$ and $b$, then according to Bayes theorem, the joint probability of these two events is

$$P(a,b) = P(a|b)P(b) = P(b|a)P(a)\,, \tag{6}$$

where $P(a|b)$ is the conditional probability of $a$ given $b$. The probabilities are normalized so that

$$P(a) = \int dbP(a,b) \quad \text{and } 1 = \int daP(a)\,. \tag{7}$$

In our problem, we search for the spectrum $A$ which maximizes the conditional probability of $A$ given the data $\bar{G}$,

$$P(A|\bar{G}) = P(\bar{G}|A)P(A)/P(\bar{G})\,. \tag{8}$$

Typically, we call $P(\bar{G}|A)$ the likelihood function, and $P(A)$ the prior probability of $A$ (or the prior). Since we work with one set of QMC data at a time, $P(\bar{G})$ is a constant during this procedure, and may be ignored. The prior and the likelihood function require significantly more thought, and will be the subject of the next two subsections.

### *2.3 Prior Probability*

We can define a prior probability for positive-definite normalizable spectra. For Bosonic Green's functions, we may define positive definite spectra if we redefine the kernel

$$K(\tau,\omega) = \frac{\omega\exp(-\tau\omega)}{1-\exp(-\beta\omega)} \text{ with } A(\omega) = \frac{-1}{\pi\omega}\mathrm{Im}G(\omega) \geq 0 \text{ for Bosons}\,. \tag{9}$$

For Fermionic Green's functions the spectra are already positive definite

$$K(\tau,\omega) = \frac{\exp(-\tau\omega)}{1+\exp(-\beta\omega)} \text{ with } A(\omega) = \frac{-1}{\pi}\mathrm{Im}G(\omega) \geq 0 \text{ for Fermions}\,. \tag{10}$$

In either case,

$$\int_{-\infty}^{\infty} d\omega A(\omega) < \infty\,, \tag{11}$$

which is usually a manifestation of a sum rule. These positive-definite normalized spectra $A$ may be reinterpreted as probability densities.

Skilling[8] argues that the prior probability for such an unnormalized probability density is proportional to $\exp(\alpha S)$ where $S$ is the entropy defined relative to some positive-definite function $m(\omega)$

$$\begin{aligned} S &= \int d\omega \left[A(\omega) - m(\omega) - A(\omega)\ln\left(A(\omega)/m(\omega)\right)\right] \\ &\approx \sum_{i=1}^{N} A_i - m_i - A_i \ln\left(A_i/m_i\right) , \end{aligned} \tag{12}$$

where $A_i = A(\omega_i)d\omega_i$, $i = 1, N$. Thus, the prior is conditional on two as yet unknown quantities $m(\omega)$ and $\alpha$

$$P(A|m,\alpha) = \exp\left(\alpha S\right) . \tag{13}$$

$m(\omega)$ is called the default model since in the absence of data $\bar{G}$, $P(A|\bar{G},m,\alpha) \propto P(A|m,\alpha)$, so the optimal $A = m$. The choice of $\alpha$ will be discussed in Sec. 2.5.

Rather than try to repeat Skilling's arguments here for the entropic form of $P(A|m,\alpha)$, we argue that this form yields the desired effects:

1. it enforces positivity of $A$,
2. it requires that $A$ only have correlations which are required to reproduce the data $\bar{G}$, and
3. it allows us to introduce prior knowledge about the the spectra (i.e. exact results at high frequencies) in the default model.

The first effect follows from the form of $P(A|m,\alpha)$, assuming that $m$ is positive definite. The third effect will be discussed in Sec. 4.5.

To illustrate the second effect, Gull and Skilling use their kangaroo argument[9]. Imagine we have a population of kangaroos. We know that one third of them are left handed and one third have blue eyes. The joint probabilities of left-handedness and eye color may be represented in a contingency table.

| | | Left Handed T | Left Handed F |
|---|---|---|---|
| Blue Eyes | T | $p_1$ | $p_2$ |
| | F | $p_3$ | $p_4$ |

We are given that $p_1 + p_2 = p_1 + p_3 = 1/3$, what is the fraction that are both blue eyed and left handed, $p_1$? Clearly, there is not enough information to answer this question. We must make some additional assumptions. If we assume that there is a maximum positive correlation between left handedness and blue eyes, then

| | | Left Handed T | Left Handed F |
|---|---|---|---|
| Blue Eyes | T | 1/3 | 0 |
| | F | 0 | 2/3 |

If these events have a maximum negative correlation, then

| | | Left | Handed |
|---|---|---|---|
| | | T | F |
| Blue | T | 0 | 1/3 |
| Eyes | F | 1/3 | 1/3 |

However, if we are forced to answer this question without the use of further information, a more natural assumption to make is that the events of handedness and eye color are uncorrelated, so that 1/9 of the kangaroos are both blue eyed and left handed.

| | | Left | Handed |
|---|---|---|---|
| | | T | F |
| Blue | T | 1/9 | 2/9 |
| Eyes | F | 2/9 | 4/9 |

This final answer is the one obtained by maximizing the entropy $S = -\sum_{i=1}^{4} p_i \ln p_i$ subject to the Lagrange constraints $\sum_{i=1}^{4} p_i = 1$, $p_1 + p_2 = 1/3$ and $p_1 + p_3 = 1/3$. All other regularization functions yield either positive or negative correlations between handedness and eyecolor.

To relate this to the analytic continuation problem, imagine that each $A_i$ is an independent event. If we maximize the entropy of $A$, subject to the constraint of reproducing the data $G = KA$, then the resulting spectrum is the one with the least correlations that is consistent with $\bar{G}$. If we identify a feature in the spectrum as a region of correlated $A_i$ (such as a peak) in deviation from the default model $m_i$, and such a feature emerges in the spectrum $A(\omega)$ and persists as the data $\bar{G}$ becomes more precise, then we have reason to believe that this feature is real. The choice of any other regularization function would produce artificial features in the data.

### *2.4 Likelihood function*

The form of the likelihood function is dictated by the central limit theorem, which for the purposes of this chapter may be illustrated with the following example. Suppose we use our QMC algorithm to generate $N_d$ measurements of the Green's function $\bar{G}_l^i$ (where $l$ is an integer between 1 and $L$, and $i$ an integer between 1 and $N_d$). According to the central limit theorem, if each of these measurements is completely independent of the others, then in the limit of large $N_d$, the distribution of $\bar{G}_l$ will approach a Gaussian, and the probability of a particular value $G_l$ is given by

$$P(G_l) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\chi^2/2}, \tag{14}$$

where $\chi^2 = \frac{1}{\sigma^2}\left(\frac{1}{N_d}\sum_{i=1}^{N_d}\bar{G}_l^i - G_l\right)^2 = \frac{1}{\sigma^2}\left(\left\langle\bar{G}_l\right\rangle - G_l\right)^2$, $\sigma^2 = \frac{1}{N_d(N_d-1)}\sum_i\left(\left\langle\bar{G}_l\right\rangle - \bar{G}_l^i\right)^2$ and the angular brackets indicate an average over the bins of data

Of course, in the QMC process each of the measurements is not independent of the others. Correlations exist between adjacent measurements ($\bar{G}_l^i$ and $\bar{G}_l^{i+1}$) in the QMC process, and between the errors of the Green's function at adjacent time slices ($\bar{G}_l^i$ and $\bar{G}_{l+1}^i$) at the same QMC step. The removal of these correlations is the most critical step in the MEM analytic continuation procedure.

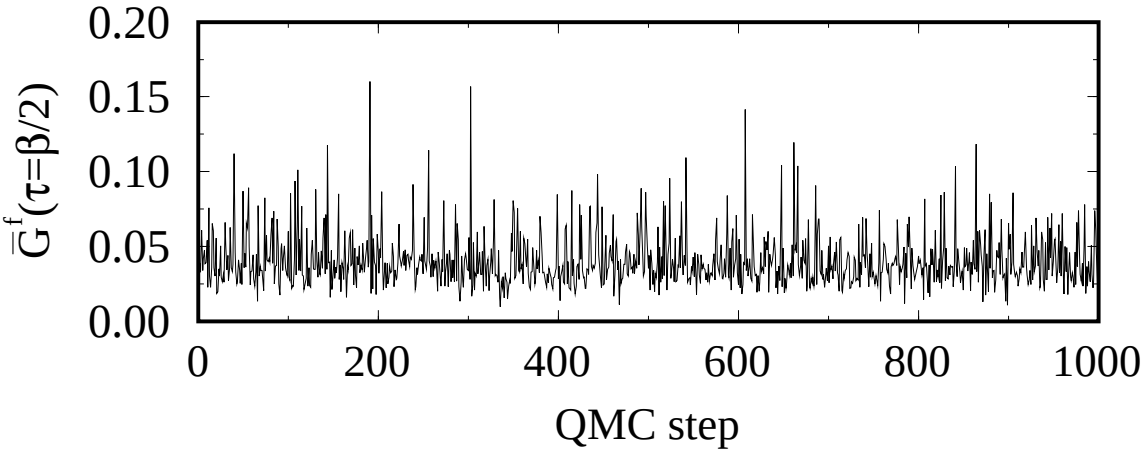


Figure 1: *Symmetric PAM f-electron local Green's function $\bar{G}^f(\tau = \beta/2)$ plotted as a function of the QMC step for $U = 2$, $V = 0.6$, and $\beta = 20$.*

Correlations between adjacent measurements are illustrated in Fig 1 where measurements of $\bar{G}^f(\tau = \beta/2)$ are plotted as a function of the QMC step. Clearly, the data from adjacent QMC steps is correlated and the data are skewed since the Green's function is bounded from below ($\bar{G}_l^i > 0$). As a result the data are not Gaussianly distributed, as shown in Fig. 2(a). Here, a histogram of the data is compared to a Gaussian fit. The deviations from a Gaussian are quantified by the moments of the distribution. The most relevant ones in the present case are the skewness (third moment) and kurtosis (fourth moment) which measure the degree of asymmetry around the mean and the pointedness (or flatness) of the distribution relative to the Gaussian [10]. The data are clearly not Gaussianly distributed, and display significant skew and kurtosis. To deal with this difficulty, we rebin the data. For example, we set $\bar{G}_l^1$ equal to the average of the first 30 measurements, $\bar{G}_l^2$ equal to the average of the next 30 measurements, etc. The distribution of this rebinned data is shown in Fig. 2b. It is well approximated by a Gaussian fit (the solid line).

The bin size (here, 30 measurements) must be chosen large enough so that the bin averages are uncorrelated, but small enough so that sufficient bins remain to calculate the likelihood function. To determine the smallest bin size that yields uncorrelated data we quantify the deviation of the distribution from a Gaussian by measuring moments of the distribution. Of course, because the data are a finite set, each of these measured moments has some standard deviation (proportional to $1/\sqrt{N_{bins}}$). Thus,

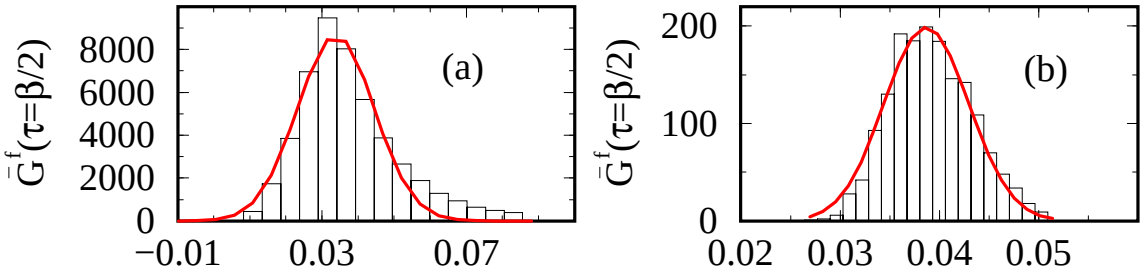


Figure 2: *Distribution of the data shown in Fig. 1 (a) and after rebinning (b). The solid line is a Gaussian fit. In (b) the data was processed by packing it sequentially into bins of 30 measurements each.*

one way to determine if the skewness and kurtosis of a distribution are acceptably small is to measure these values relative to what is expected from a Gaussian distribution. We will use such relative values. As the bin size increases, the relative kurtosis

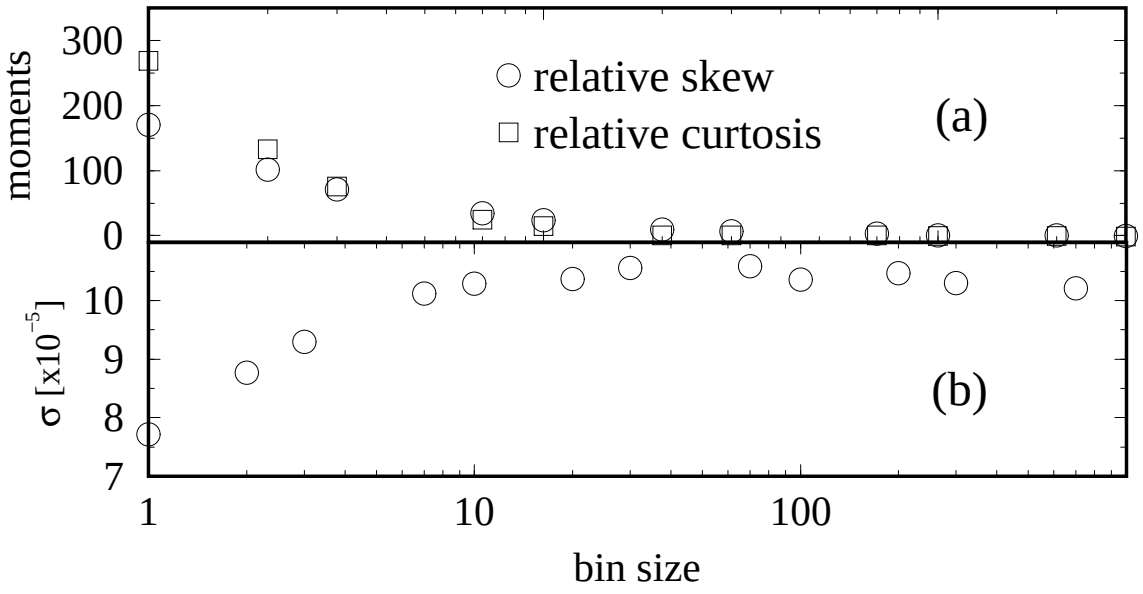


Figure 3: *Relative kurtosis and skew (a) and error bar (b) of the data shown in Fig. 1 as a function of bin size. Here the total amount of data is fixed, so increasing the bin size decreases $N_{bins}$ proportionately. As the bin size increases to about 30, the relative kurtosis and skew decrease to roughly zero and the error bar saturates, indicating that the bins are uncorrelated samples and that the data has become Gaussianly distributed.*

and skewness decrease monotonically, indicating the convergence of the distribution to a Gaussian. This behavior is shown in Fig. 3a for the $G(\tau = \beta/2)$ data.

In addition, Fig. 3b shows that the error estimate also converges as the bin size

increases. Here, the error estimate is given by

$$\sigma = \sqrt{\left(\left\langle \bar{G}^2 \right\rangle - \langle \bar{G} \rangle^2\right) / (N_{bins} - 1)} \tag{15}$$

where angular brackets indicate an average over the bins of data. Because correlations between successive Monte Carlo measurements always make this error estimate smaller than the actual value, this error estimate should initially increase monotonically with bin size, as shown. This behavior is easily understood by considering a perfectly correlated sample where the data in each bin is identical. Clearly, for this perfectly correlated sample, the error estimate would be zero. As the bins become uncorrelated, the error estimate increases. With independent data and a large number of equally sized bins, eventually $\sigma^2 \sim 1/N_{bins}$. However, with a fixed amount of data, as is typical with a QMC simulation, increasing the bin size decreases $N_{bins}$ proportionally, and the error estimate can saturate as illustrated in Fig. 3b. Thus, the saturation of the error estimate indicates that the correlations between Monte Carlo measurements, i.e., between bin averages, have been removed. The point at which saturation occurs in a plot like Fig. 3b provides a useful first estimate of the minimum bin size required to remove correlations between the bins. In general, one should perform this test for the Green's function at all times $\tau_l$; however, we have found it is often sufficient to perform this test at only a few times. For the remainder of this section, we will assume that the bin size is sufficiently large so that both the error estimate and the moments of the distribution have converged to values which indicate that the data are both statistically independent and Gaussian-distributed.

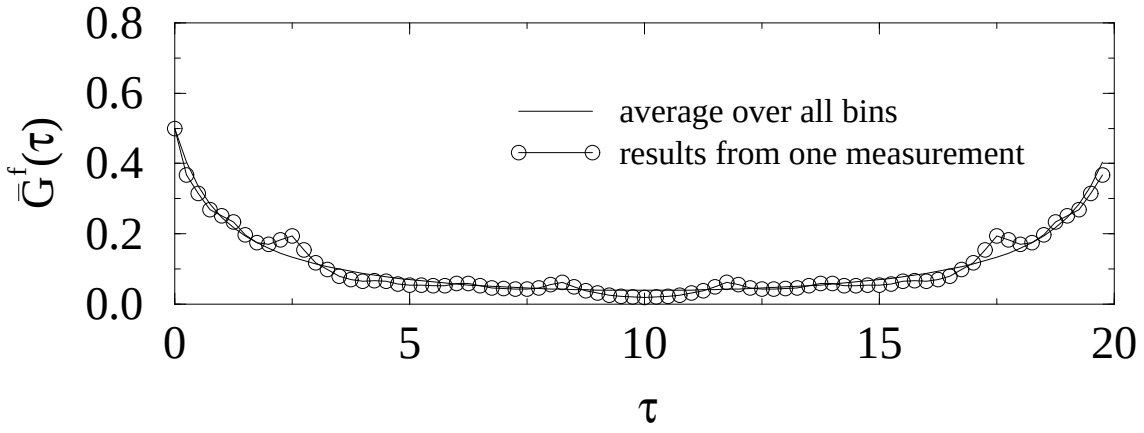


Figure 4: $\bar{G}^f(\tau)$ *from one measurement compared to* $\bar{G}^f(\tau)$ *obtained from the average over 800 bins of data, each containing 1520 measurements. If the result from a single measurement at a certain point differs from the essentially exact result obtained by averaging over many bins, then the results at adjacent points also deviate from the exact results.*

Now, only the errors in the Green's function $\bar{G}$ at adjacent time slices remain correlated. This correlation may be seen by comparing the results from a single measurement with those essentially exact values obtained from averaging over many

measurements. Such a comparison is shown in Fig. 4 where if the result from a single measurement differs from the essentially exact result at a certain value of $\tau$, then the results at adjacent values of $\tau$ also tend to deviate from the exact results in a similar way. These correlations of the error in Euclidean time are characterized by the covariance

$$C_{lk} = \frac{1}{N_{bins}(N_{bins}-1)} \sum_{j=1}^{N_{bins}} (\left\langle \bar{G}_l \right\rangle - \bar{G}_l^j)(\left\langle \bar{G}_k \right\rangle - \bar{G}_k^j) \,. \tag{16}$$

If $C$ is diagonal, then according to the central limit theorem, the likelihood function is $P(\bar{G}|A) = \exp[-\chi^2/2]$ where

$$\chi^2 = \sum_{l=1}^{L} \left( \frac{\bar{G}_l - \sum_j K_{l,j} A_j}{\sigma_l} \right)^2 . \tag{17}$$

and $\sigma_l^2$ are the diagonal elements of $C$. However, in general, the covariance matrix $C_{ij}$ is not diagonal because errors at different values of $\tau$ are correlated. To define a meaningful measure of how well $A_i$ reproduces the data, we must find the transformation $\mathbf{U}$ which diagonalizes the covariance matrix

$$\mathbf{U}^{-1}\mathbf{C}\mathbf{U} = \sigma_i'^2 \delta_{ij} \,. \tag{18}$$

Both the data and kernel are now rotated into this diagonal representation

$$\mathbf{K}' = \mathbf{U}^{-1}\mathbf{K} \qquad \bar{\mathbf{G}}' = \mathbf{U}^{-1}\bar{\mathbf{G}} \,. \tag{19}$$

and each measurement $\bar{G}_i'$ is statistically independent. Therefore, we can use

$$\chi^2 = \sum_{l} \left( \frac{\bar{G}_l' - \sum_j K_{l,j}' A_j}{\sigma_l'} \right)^2 . \tag{20}$$

to measure the misfit between the spectrum and the data and to define the likelihood function.

Since the set of data is finite, it is necessary to balance the need of removing the correlations in imaginary-time with the need of removing the correlations between Monte Carlo steps. To remove the correlations in Monte Carlo steps the bin size must be large; however, to calculate the covariance accurately, many bins of data are required. If there are not enough bins of data, then the covariance and (as shown in Fig. 5) its eigenvalue spectrum can become pathological. The reason for this pathology is that when we diagonalize the covariance matrix, we are asking for $L$ independent eigenvectors. We must have enough bins of data to determine these directions so that $N_{bins}$ must be greater than or equal to $L$. In fact, since the information contained in a given bin of data is not completely independent from the other bins, we must have $N_{bins} > L$. Otherwise, as shown in Fig. 5, where $L = 41$, the eigenvalue spectrum displays a sharp break when $N_{bins} < L$, indicating that only

a finite number of directions, less than $L$, are resolved. The small eigenvalues after the break are essentially numerical noise and yield artifacts in the spectra. Simply throwing away the small eigenvalues and their associated eigenvectors does not cure the difficulty since the small eigenvalues and eigenvectors contain the most precise information about the solution. Thus, the only reasonable thing to do is to increase the number of bins. Empirically, we find that we need

$$N_{bins} \geq 2L \tag{21}$$

in order to completely remove the pathology of the sharp break in the eigenvalues[11].

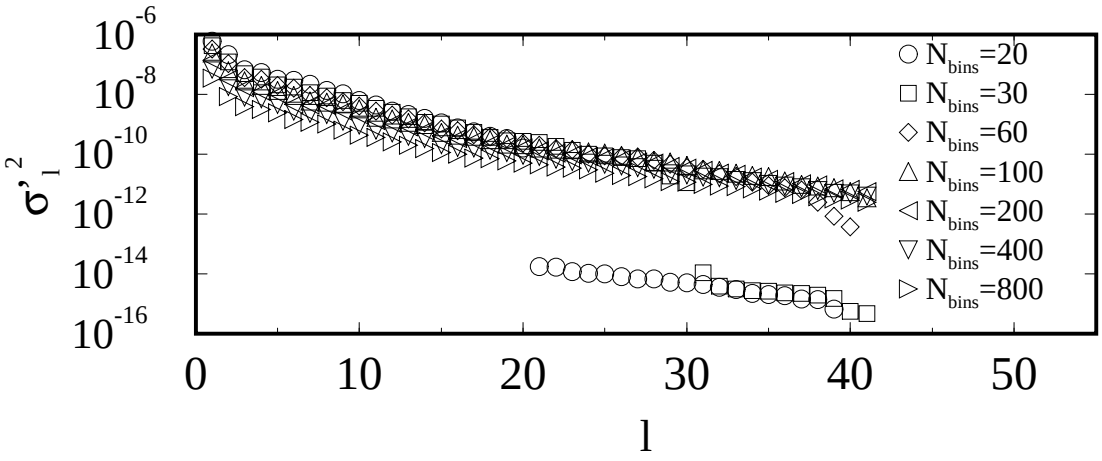


Figure 5: *Eigenvalue spectra of the covariance matrix of $G^f$ for different numbers of bins of data. Each bin contains 100 measurements and $L = 41$. When $N_{bins} \lesssim 2L$, $\sigma'^2_l$ develops a sharp break.*

We find that proper preparation of the data, removing correlations, is the most critical step in the MEM procedure. If the data are uncorrelated and $N_{bins} > 2L$, then the resulting spectra will be reliable (however, for weak data, it will show a significant bias towards the default model). If not, then the Gaussian form of the likelihood function is unjustified and the resulting spectra will generally have spurious features.

### *2.5 Details of the MEM Formalism*

We will now construct the formalism to locate the most likely spectrum $\hat{A}$ and set the value of $\alpha$. The first step is to normalize the likelihood function $P(\bar{G}|A)$ and the prior $P(A|\alpha, m)$. Here it will be necessary to integrate over the space of all spectra $A_i$. This is done with Gaussian approximations to the integrals. Following Skilling and Bryan[12], we employ a measure $d^N A/\prod_i \sqrt{A_i}$ which amounts to a change of variables to a space where $S$ has no curvature[6].

For example, the normalized prior probability is

$$P(A|\alpha,m) = \frac{1}{Z_S}\exp\left\{\alpha\left(-\sum A_i \ln A_i/m_i - A_i + m_i\right)\right\} \tag{22}$$

where

$$Z_S = \int \frac{d^N A}{\prod_i \sqrt{A_i}}\exp\left\{\alpha\left(-\sum A_i \ln A_i/m_i - A_i + m_i\right)\right\}. \tag{23}$$

The integrand is maximized when $S = 0$, i.e. when $A = m$. We approximate the integral by expanding the argument of the exponent to second order around this maximum, $S \approx \frac{1}{2}\delta A^T \nabla\nabla S|_{A=m}\,\delta A = -\frac{1}{2}\delta A^T \{1/m\}\,\delta A$, where $\{1/m\}$ is the diagonal matrix with finite elements composed of $1/m_i$, and $\delta A$ is the vector $A - m$.

$$Z_S \approx \int \frac{d^N A}{\prod_i \sqrt{A_i}}\exp\left\{\alpha\left(-\frac{1}{2}\delta A^T \{1/m\}\,\delta A\right)\right\}. \tag{24}$$

We define a change of variables, so that $dy_i = dA_i/\sqrt{A_i}$ and find

$$Z_S \approx \int d^N y \exp\left\{\alpha\left(-\frac{1}{2}\delta y^T \{m\}^{1/2}\{1/m\}\{m\}^{1/2}\delta y\right)\right\} = (2\pi/\alpha)^{N/2} \tag{25}$$

The likelihood function must also be normalized

$$P(\bar{G}|A) = e^{-\chi^2/2}/Z_L \tag{26}$$

where

$$\chi^2 = \sum_l \frac{\left(\bar{G}'_l - \sum_i K'_{li}A_i\right)^2}{\sigma'^2_l} \tag{27}$$

where $K'$ and $\bar{G}'$ are the kernel and data rotated into the data space where the covariance is diagonal, and $\sigma'^2_l$ are the eigenvalues of the covariance. If we let $G_l = \sum_i K'_{li}A_i$, then

$$Z_L = \int d^L G \exp\left\{\frac{1}{2}\sum_{l=1}^{L}\frac{\left(\bar{G}'_l - G_l\right)^2}{\sigma'^2_l}\right\} = (2\pi)^{L/2}\prod_l \sigma'_l \tag{28}$$

Using Bayes theorem, we find

$$\begin{aligned} P(A,G|m,\alpha) &= P(G|A,m,\alpha)P(A|m,\alpha) \\ &= P(A|G,m,\alpha)P(G|m,\alpha) \end{aligned} \tag{29}$$

or

$$P(A|G,m,\alpha) \propto P(G|A,m,\alpha)P(A|m,\alpha) = \frac{\exp(\alpha S - \chi^2/2)}{Z_S Z_L} \tag{30}$$

Since the normalization factors $Z_S$ and $Z_L$ are independent of the spectrum, for fixed $\alpha$ and data, the most probable spectrum $\hat{A}(\alpha)$ is the one which maximizes $Q = \alpha S - \chi^2/2$. An algorithm to find this spectrum is discussed in Sec. 3. However, the question of how to select $\alpha$ and the default model remains.

*Selection of $\alpha$*

The selection of $\alpha$ strongly effects the choice of the optimal spectrum[13] since $\alpha$ controls the competition between $S$ and $\chi^2$. If $\alpha$ is large, then the entropy term is emphasized and the data cannot move the spectrum far from the model. If $\alpha$ is small, then the least square misfit between the spectrum and the data is minimized so that $\chi^2 \ll L$. The numerical error in the QMC data then begins to dominate the solution and the spectra displays random oscillations and noise. Thus, it is important to find a sensible way of selecting $\alpha$. Typically, $\alpha$ is selected in one of three ways.

*Historic MEM* [14, 12] In the historic method, $\alpha$ is adjusted so that $\chi^2 = L$. The justification for this is that if the spectrum is known and the data was repeatedly measured, then the misfit between the data and the spectrum $\chi^2 = L$ on average. However, the data are only measured once and the spectrum is not known *a priori*. Also, setting $\chi^2 = L$ tends to under fit the data since good data can cause structure in the spectrum which reduces $\chi^2$ from $L$. Thus, there is little reason to believe that $\alpha$ can be chosen without input from the data itself.

*Classic MEM* [13] A more appropriate method of setting $\alpha$ is to choose the most probable value, defined by maximizing

$$P(\alpha|\bar{G}, m) = \int \frac{d^N A}{\prod_i \sqrt{A_i}} P(A, \alpha|\bar{G}, m)\,. \tag{31}$$

The integrand

$$P(A, \alpha|\bar{G}, m) = P(A|\bar{G}, m, \alpha)P(\alpha) \propto \frac{\exp(\alpha S - \chi^2/2)}{Z_S Z_L} P(\alpha) \tag{32}$$

involves the prior probability of $\alpha$. Jeffreys[15] argues that since $\chi^2$ and $S$ have different units, $\alpha$ is a scale factor. He asserts that in lieu of prior knowledge, it should have the simplest scale invariant form $P(\alpha) = 1/\alpha$. Thus,

$$P(\alpha|\bar{G}, m) = \int \frac{d^N A}{\prod_i \sqrt{A_i}} \frac{\exp(\alpha S - \chi^2/2)}{Z_S Z_L \alpha} = \frac{Z_Q}{Z_S Z_L \alpha} \tag{33}$$

$Z_Q$ is calculated in a similar fashion to $Z_S$. We expand about the maximum of $Q$ at $A = \hat{A}$ so that $\exp\{\alpha S - \chi^2/2\} \approx \exp\{Q(\hat{A}) + \frac{1}{2}\delta A^T \left.\nabla\nabla Q\right|_{\hat{A}} \delta A\} = \exp\{Q(\hat{A}) + \frac{1}{2}\delta A^T \{\frac{1}{2} \left.\nabla\nabla\chi^2\right|_{\hat{A}} - \{\alpha/\hat{A}\}\}\delta A\}$. We again make a Gaussian approximation to the integral, and if $\lambda_i$ are the eigenvalues of $\frac{1}{2}\{A^{1/2}\} \left.\nabla\nabla\chi^2\right|_{\hat{A}} \{A^{1/2}\}$, then

$$P(\alpha|\bar{G}, m) = \frac{1}{W_\alpha} \prod_i \left(\frac{\alpha}{\alpha + \lambda_i}\right)^{1/2} \frac{e^{Q(\hat{A})}}{\alpha} \tag{34}$$

where

$$W_\alpha = \int \frac{d\alpha}{\alpha} \prod_i \left( \frac{\alpha}{\alpha + \lambda} \right)^{1/2} e^{Q(\hat{A})} \,. \tag{35}$$

The optimal $\alpha$, $\hat{\alpha}$ may be determined by the condition

$$\frac{\partial P(\alpha|\bar{G}, m)}{\partial \alpha} = 0 \,. \tag{36}$$

For strong data, $P(\alpha|\bar{G}, m)$ is dominated by the product and $\exp Q(\hat{A})$ so that

$$-2\hat{\alpha} S \approx \sum_i \frac{\lambda_i}{\hat{\alpha} + \lambda_i} \,. \tag{37}$$

Each $\lambda_i$ which is much greater than $\hat{\alpha}$ contributes one to the sum and hence one to the number of good observations in the data. If the number $N_{good} = -2\hat{\alpha} S$ is large, then $P(\alpha|\bar{G}, m)$ is very sharp the spectra corresponding to $\alpha = \hat{\alpha}$ is a good approximation of the spectra which has been properly averaged over $P(\alpha|\bar{G}, m)$.

*Bryan's Method* [17] However, typically we find that $N_{good} \ll L$. Then $P(\alpha|\bar{G}, m)$ is a broad and highly skewed distribution. For example, $P(\alpha|\bar{G}, m)$ for the data shown in Fig. 1 is plotted in Fig. 6. The distribution is wide, so many reasonable

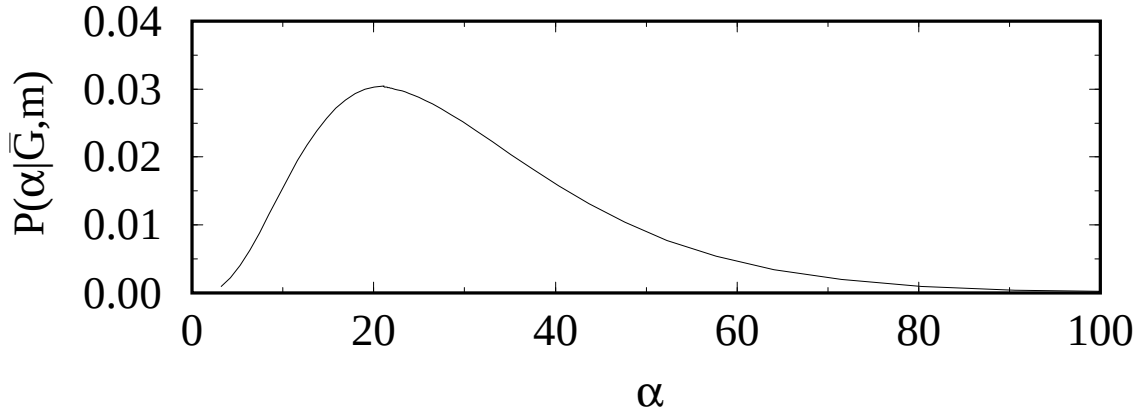


Figure 6: The posterior probability $P(\alpha|\bar{G}, m)$ as a function of $\alpha$ for the periodic Anderson model data presented in Fig. 1. Since $P(G|I)$ is unknown, the magnitude of the ordinate is also unknown. The distribution is wide, so many reasonable values of $\alpha$ exist. The distribution is also skewed, so the value of $\alpha$ at the peak is not representative of the mean.

values of $\alpha$ exist. The distribution is also skewed, so the value of $\alpha$ at the peak is not representative of the mean. To deal with this, Bryan[17] calculates the optimal spectrum $\hat{A}(\alpha)$ for each $\alpha$. The solution is taken to be

$$\bar{A} = \int d\alpha \hat{A}(\alpha) P(\alpha|\bar{G}, m) \,. \tag{38}$$

These three MEM methods will produce essentially identical results if the data are uncorrelated and precise. However, when the data are less precise but still uncorrelated, the method suggested by Bryan, averaging $\hat{A}(\alpha)$ weighted by $P(\alpha|G,m)$, generally produces more acceptable results and converges to a good result faster than the classic method and much faster than the historic method as the data is improved. A further advantage of the averaging is that it allows an accurate relative assessment of the posterior probability ($\int_0^\infty d\alpha P(m|G,\alpha)$) of the default model. This information is invaluable in determining which default model yields the most likely $A$.

### *2.6 Model Selection*

Bayesian statistics may also be employed to select the default model. I.e. if we must choose between different models, or set parameters used to define a default model function, then we choose these models or parameters based upon the posterior probability of the model

$$P(m|\bar{G}) = \int d\alpha P(\alpha|m,\bar{G})P(m)\,. \tag{39}$$

We see no *a priori* reason to favor one default model over an another, so we typically set the prior probability of the model $P(m)$ =constant. Then the integrand in Eq. 39 is given by Eq. 34 so that

$$P(m|\bar{G}) \propto W_\alpha\,. \tag{40}$$

Since the prior probability of the model is unknown, $P(m|\bar{G})$ determines only the relative probability of two models, and by inference the relative probability of their corresponding spectra.

### *2.7 Error Propagation*

To absolutely qualify the spectrum, we need to assign error bars to it. In the quadratic approximation, the probability of the spectral density is

$$P(A|\bar{G},m,\alpha) \propto e^{-\frac{1}{2}\delta A^T\cdot\nabla\nabla Q|_{\hat{A}}\cdot\delta A}\,, \tag{41}$$

thus the covariance of the spectral density is

$$\langle\delta A(\omega)\delta A(\omega')\rangle = -\left(\nabla\nabla Q|_{\hat{A}}\right)^{-1}\,. \tag{42}$$

It is not possible to assign error bars to each point in the spectral density since this matrix is generally not diagonal. Thus errors between different points are strongly correlated. Also, $A_i$ represents the spectral probability within some region of finite

width and hence lacks meaning at a specific value of $\omega$. However, it is possible to assign error bars to integrated functions of the spectral density such as [16],

$$H = \int d\omega A(\omega) h(\omega) \,. \tag{43}$$

where $h(\omega)$ is an arbitrary function of $\omega$. The error of $H$ may be associated with the covariance of the spectral density $\langle \delta A(\omega) \delta A(\omega') \rangle$

$$\left\langle (\delta H)^2 \right\rangle = \int \int d\omega d\omega' \, h(\omega) h(\omega') \left\langle \delta A(\omega) \delta A(\omega') \right\rangle \,. \tag{44}$$

The matrix $\nabla\nabla Q|_{\hat{A}}$ is readily available because it is used as the Hessian of the Newton search algorithm typically used to find the optimal spectral density.

## 3 Bryan's Method: a MEM algorithm

We will now sketch Bryan's numerical algorithm to find the optimal spectrum. For a more detailed description, we refer the reader to his paper[17]. We have found his algorithm to be very appropriate for the numerical analytic continuation problem for two reasons: First, due to the exponential nature of the kernel which relates $A$ to the data $\bar{G}$, we typically have $L \gg N_{good}$. Thus, the problem is usually "oversampled." Bryan tailored his numerical algorithm[17] to this type of problem by working in a reduced space whose dimension is determined by singular-value-decomposition of the kernel $K$ and is equal to the largest possible number of good singular values (i.e., numerically significant) which may parameterize the solution. The dimension of this space is usually much less than the number of $A_i$, and we found the computational advantage over methods that use the entire space determined by the number of $A_i$ to be significant. Second, for the analytic continuation problem, the approximation of setting $\alpha$ equal to its optimal value is questionable because of the wide range of reasonably acceptable values of $\alpha$. Bryan deals with this by calculating a result which is averaged over $P(\alpha|G, m)$.

### *3.1 Typical Algorithms*

What distinguishes Bryan's numerical algorithm from its predecessors is the way in which the space of possible solutions is searched. Typical algorithms search for an optimal $A$ by stepping through the entire space of $A$

$$A \rightarrow A + \delta A \tag{45}$$

with

$$\delta A = -(\nabla\nabla Q)^{-1} \nabla Q \,. \tag{46}$$

The Hessian $(\nabla\nabla Q)^{-1}$ is

$$(\nabla\nabla Q)^{-1} = (\alpha\nabla\nabla S - \nabla\nabla L)^{-1} = \left(\alpha\{A\}^{-1} - \nabla\nabla L\right)^{-1} . \tag{47}$$

where $\{A\}$ is a diagonal matrix with the elements of $A$ along its diagonal. It may conceptually be expanded using the binomial theorem so that $(\nabla\nabla Q)^{-1}$ may be written as a power series in $\{A\}\nabla\nabla L$. Thus, $\delta A$ may be written as a combination of $\{A\}\nabla Q = \{A\}\left(\alpha\nabla S - \nabla L\right)$, and powers of $\{A\}\nabla\nabla L$ acting on $\{A\}\nabla S$ and $\{A\}\nabla L$. Each of these vectors defines a direction in which the search can precede. Typically, between three and ten directions are used; however, these directions are often inappropriate for the problem at hand, because as mentioned earlier, the space of all possible solutions is too large for such oversampled data.

### *3.2 Singular-Space Algorithm*

To alleviate this problem, Bryan performs a singular-value decomposition (SVD) of the kernel $K$, i.e., $K = V\Sigma U^T$ where $U$ and $V$ are orthogonal matrices and $\Sigma$ is a diagonal matrix, and works in the resulting singular space. To see that this space still contains the solution, we consider

$$\nabla L = \frac{\partial F}{\partial A}\frac{\partial L}{\partial F} = K^T\frac{\partial L}{\partial F} \tag{48}$$

where $F = KA$. We see that $\nabla L$ lies in the vector space defined by the columns of $K^T$. We next perform a SVD on $K$ and assume the diagonal elements of $\Sigma$ are ordered from largest to smallest. The smallest elements are essentially zero (to the numerical precision of the computer) since the kernel is effectively singular. However, $s$ of the elements are assumed finite. Now the vector space spanned by the columns of $K^T$ is the same as the space spanned by the columns of $U$ associated with the non-singular values. Bryan calls this reduced space the *singular space*. Thus, to the precision that can be represented on the computer, $\{A\}\nabla L$ and all of the search directions formed by acting with $\{A\}\nabla\nabla L$ lie in the singular space spanned by the columns of $\{A\}U_s$, where $U_s$ is the singular space projection of $U$. The only direction not in this space is $\{A\}\nabla S$. Thus, Bryan's algorithm works in *at most* an $s+1$-dimensional subspace of the $N$-dimensional space of $A$.

In this singular space, the condition for an extremum of $Q$, $\nabla Q = 0$, is

$$\alpha\nabla S - \nabla L = 0 \rightarrow -\alpha\ln\left(A_i/m_i\right) = \sum_j K_{ji}\frac{\partial L}{\partial F_j} . \tag{49}$$

Thus, the solution may be represented in terms of a vector $u$

$$\ln\left(A/m\right) = K^T u . \tag{50}$$

Unless $K$ is of full rank, so that $s = N$, the components of $u$ will not be independent. However, since $K^T$ and $U$ share the same vector space and since most of the relevant search directions lie in the singular space, Bryan proposes that the solution be represented in terms of $U$ and $u$ as

$$A_i = m_i \exp \sum_n U_{in} u_n \,. \tag{51}$$

Thus, to the precision to which it may be represented on the computer and determined by SVD, the space $u$ must contain the solution defined by $\nabla Q = 0$, and the search can be limited to this $s$-dimensional space.

Bryan's algorithm precedes by first reducing all the relevant matrices to the singular space. With the definitions $K = V\Sigma U^T$ and $\log(A/m) = Uu$, the condition for an extremum becomes

$$-\alpha U u = U \Sigma V^T \frac{\partial L}{\partial F} \,, \tag{52}$$

or

$$-\alpha u = \Sigma V^T \frac{\partial L}{\partial F} \equiv g \,, \tag{53}$$

where each of these matrices and vectors has been reduced to the singular space. ($u$ is now a vector of order $s$, $\Sigma$ is an $s \times s$ diagonal matrix. etc.). Bryan then uses a standard Newton's search to find the solution in the singular space, starting from an arbitrary $u$. The increment at each iteration is given by

$$J \delta u = -\alpha u - g \,, \tag{54}$$

where $J = \alpha I + \partial g / \partial u$ is the Jacobian matrix, $I$ the identity matrix, and

$$\frac{\partial g}{\partial u} = \Sigma V^T \frac{\partial^2 L}{\partial F^2} \frac{\partial F}{\partial A} \frac{\partial A}{\partial u} \,. \tag{55}$$

With the definition $W = \partial^2 L / \partial F^2$ (which is just the diagonal matrix with elements $1/\sigma_l'^2$), $M = \Sigma V^T W V \Sigma$, and $T = U^T A U$. $M$ and $T$ are symmetric $s \times s$ matrices, the Jacobian $J = \alpha I + MT$, and

$$(\alpha I + MT)\, \delta u = -\alpha u - g \tag{56}$$

At each iteration $\delta u$ must be restricted in size so that the algorithm remains stable. Thus, another parameter $\mu$ (a Marquart-Levenberg parameter) is added

$$[(\alpha + \mu) I + MT]\, \delta u = -\alpha u - g \tag{57}$$

and adjusted to keep the step length $\delta u^T T \delta u$ below some the limit

$$\delta u^T T \delta u \leq \sum_i m_i \tag{58}$$

so the search is within the range of validity of a local quadratic expansion of $Q$.

This search can be made more efficient if Eq. 57 is diagonalized, so that of order $s$ operations are required for each $\alpha$ $\mu$ pair. First, we diagonalize $T$

$$TP = P\Gamma \tag{59}$$

where $P$ is an orthogonal matrix and $\Gamma$ is diagonal with finite elements $\gamma_i$. Then we define

$$B = \{\gamma^{1/2}\}P^T M P\{\gamma^{1/2}\} \tag{60}$$

and solve the second eigenvalue equation

$$BR = R\Lambda \tag{61}$$

where $R$ is orthogonal and $\Lambda$ the diagonal matrix with finite elements $\lambda_i$. Finally, to diagonalize Eq. 57 we define

$$Y = P\{\gamma^{-1/2}\}R\,. \tag{62}$$

Then $Y^{-T}Y^{-1} = T$, and $Y^{-1}MY^{-T} = \Lambda$, so that

$$Y^{-1}\left[(\alpha+\mu)I + MT\right]\delta u = \left[(\alpha+\mu)I + \Lambda\right]Y^{-1}\delta u = Y^{-1}\left[-\alpha u - g\right] \tag{63}$$

which yields $s$ independent equations for $Y^{-1}\delta u$. Again, as these equations are iterated, $\mu$ must be adjusted to keep the step length

$$\delta u^T T \delta u = \left|Y^{-1}\delta u\right|^2 \leq \sum_i m_i\,. \tag{64}$$

### *3.3 Selection of $\alpha$*

The value $\alpha$ is adjusted so that the solution iterates to either a fixed value of $\chi^2$ (for historic MEM) or to a maximum value of $P(\alpha|G,m)$ given by Eq. 34 (for classic MEM). Then, $A$ is obtained from

$$A_i = m_i \exp\left(\sum_{n=1}^{s} U_{in}u_n\right)\,. \tag{65}$$

Alternatively, Bryan suggests that one may start the algorithm with a large $\alpha$ for which $P(\alpha|\bar{G},m)$ is negligibly small, and then iterate to $\alpha \approx 0$ so that the averaged spectrum may be approximated

$$\langle A\rangle = \int_0^\infty d\alpha P(\alpha|G,m)\hat{A}(\alpha) \tag{66}$$

where $\hat{A}(\alpha)$ is the optimal spectrum (that for which $\nabla Q = 0$) for the value of $\alpha$ specified in the argument. This latter step may be necessary when $P(\alpha|G,m)$ is not a sharply peaked distribution. In fact this is usually the case, as may be seen in Fig. 6.

### *3.4 Error Propagation*

As discussed in Sec. 2.7, it is possible to assign error bars to integrated functions of the spectrum $H = \int d\omega A(\omega)h(\omega)$

$$\left\langle (\delta H)^2 \right\rangle = \int\int d\omega d\omega' \, h(\omega)h(\omega') \left\langle \delta A(\omega)\delta A(\omega') \right\rangle \, , \tag{67}$$

where

$$\left\langle \delta A(\omega)\delta A(\omega') \right\rangle = - \left( \nabla\nabla Q|_{\hat{A}} \right)^{-1} \, . \tag{68}$$

This is the inverse of the Hessian of the algorithm discussed above. $\nabla\nabla Q|_{\hat{A}}$ and is easily calculated in terms of singular-space quantities

$$- \nabla\nabla Q|_{\hat{A}} = \{1/A\} U Y^{-T} \{\alpha I + \Lambda\} Y^{-1} U^T \{1/A\} \, . \tag{69}$$

Its inverse

$$- \left( \nabla\nabla Q|_{\hat{A}} \right)^{-1} = \{A\} U Y \left\{ \frac{1}{\alpha + \lambda} \right\} Y^T U^T \{A\} \tag{70}$$

may be used to calculate the error of $H$, $\sqrt{(\delta H)^2}$ for any $\alpha$. In principle, one should average the error over $P(\alpha|m, \bar{G})$; however, we find that it is generally adequate to calculate the error of the spectrum at the optimal $\hat{\alpha}$.

We close this section with several practical comments: On a workstation, finding the optimal spectrum by searching in the singular space requires only a few minutes of computer time. This efficiency is in sharp contrast with the amount of computer we needed[18] even on a "supercomputer" for standard Newton algorithms[12] or simulated annealing methods that use the full space of $A$. We found it essential to use 64 bit arithmetic to obtain stable results. Also, we use LINPACK's [19] singular-value decomposition routine to do the SVD and also to compute any eigenvalues and eigenvectors. The SVD routine in Numerical Recipes[20] and the EISPACK [21] eigenvalue-eigenvector routine RS are not as stable.

## 4 Case Study

In this section, we will demonstrate that it is possible to extract spectral densities from the quantum Monte Carlo data that are essentially free from artifacts caused by overfitting to the data and have only small and controllable amounts of statistical error. We will use as an example the electronic spectral densities of the infinite-dimensional periodic Anderson model (PAM). We have already qualified the local Greens function data to remove correlations using the procedure discussed in Sec. 2.4, so we can begin to process the data to obtain the single-particle density of states spectral function.

For the majority of this section, we will consider particle-hole symmetric data $G(\tau) = G(\beta - \tau)$, and spectra $A(\omega) = A(-\omega)$. This prior information may imposed

on the solution by constructing a symmetric kernel and default models. We will use three symmetric default models: two non-informative models — the flat model $m(\omega) = \text{constant}$ and a simple Gaussian

$$m(\omega) = \frac{1}{\Gamma\sqrt{\pi}} \exp[-(\omega/\Gamma)^2] \tag{71}$$

and also a third one obtained from second-order perturbation theory in $U$ [22, 23]. The kernel for symmetric Fermionic Green's functions may be modified to reflect the symmetry and the associated integral restricted to positive frequencies

$$G(\tau) = \int_0^\infty d\omega A(\omega) \frac{e^{-\tau\omega} + e^{-(\tau-\beta)\omega}}{1 + e^{-\beta\omega}} \,. \tag{72}$$

### *4.1 Convergence of the Spectra*

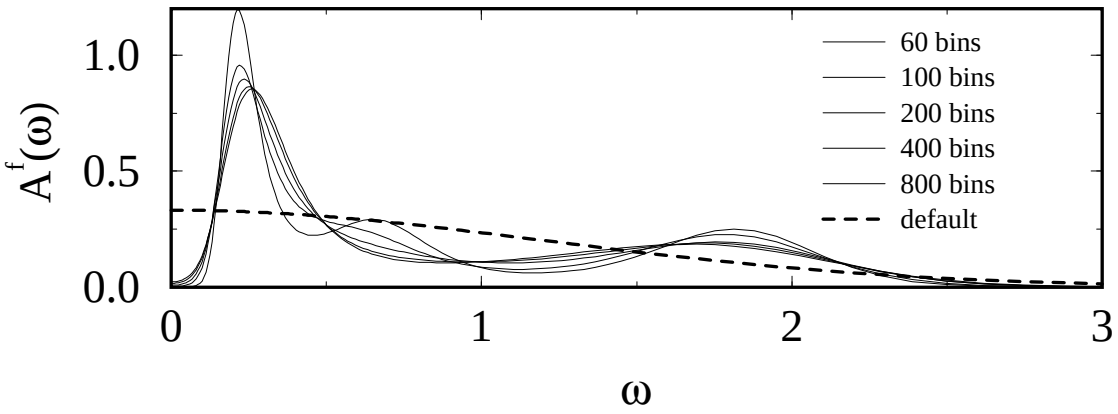


Figure 7: A sequence of spectral densities generated with increasingly accurate data. Every time the number of bins of data is doubled, the accuracy of the data increases by 41%. (The error is reduced by $1/\sqrt{2}$.) A Gaussian default model, the dashed line, was used.

To minimize the effects of statistical error, the accuracy of the data needs to be increased until the spectral density has converged. This is demonstrated in Fig. 7, where the accuracy of the data are improved by increasing the number of bins of data. Here, a Gaussian default model is used whose width $\Gamma = 1.6$ (chosen by an optimization procedure to be discussed below). Each time the number of bins of data is doubled, the accuracy of the data increases by 41%. The spectral densities corresponding to smallest number of bins of data have spurious features associated with overfitting. These features are associated with difficulties in calculating the covariance matrix, as discussed in Sec. 2.4. As $N_{bins}$ increases beyond $2L$, the spurious structure is quickly suppressed. By the time 800 bins of data have been used, the spectral density appears to be converged to several linewidths.

### 4.2 Default Model Selection

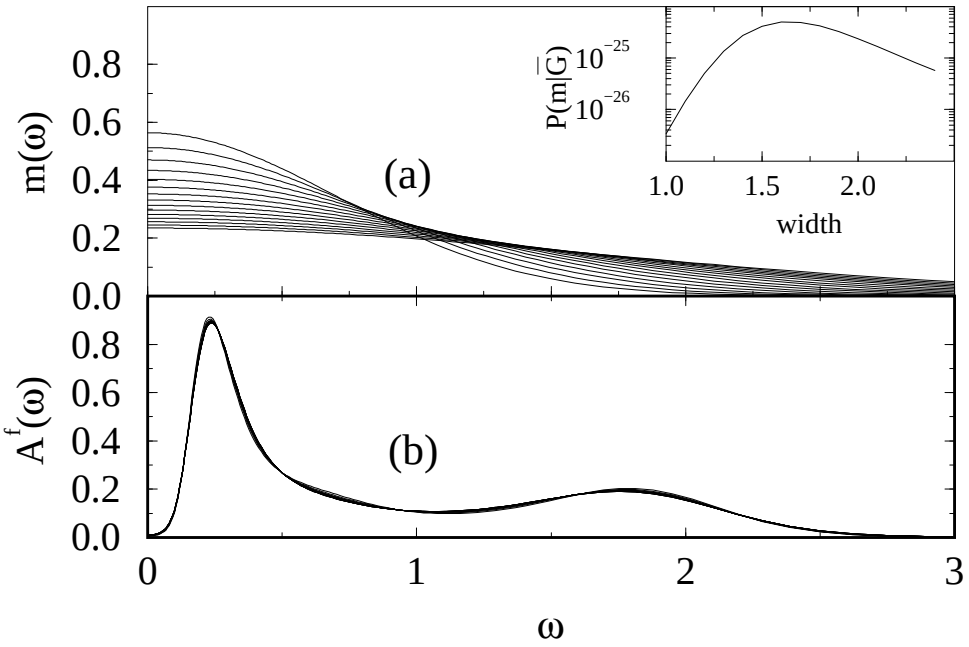


Figure 8: Test of the dependence of the spectral density upon the default model. The width $\Gamma$ of the Gaussian default model (a) is varied, producing a series of spectral densities (b). In the inset to (a) is the posterior probability of the default model $P(m|\bar{G})$, produced by integrating the joint probability $P(A, \alpha, m|\bar{G})$ over $\alpha$ and $A$, is plotted as a function of $\Gamma$. The normalization of $P(m|\bar{G})$ is unknown because it depends upon the probability of the data and the prior probability of the default model which are unknown. Thus, this posterior distribution may be used for comparison purposes only.

One may also test the dependence of the spectral density on the default model by changing its parameters or by using different models. The best model is the one with the largest posterior probability, calculated by assuming that the prior probability of the default model is flat, so that $P(A, \alpha, m|\bar{G}) \propto P(A, \alpha|\bar{G}, m)$. Then $P(m|\bar{G})$ is obtained by integrating $P(A, \alpha, m|\bar{G})$ over $A$ and $\alpha$. The effects of varying the default model parameters are shown in Fig. 8a where the same data set is analytically continued with Gaussian default models whose widths satisfy $1.0 < \Gamma < 2.4$. The posterior probability $P(m|\bar{G})$ of these default models, shown in the inset, is peaked around $\Gamma \approx 1.6$. (We note that the normalization of $P(m|\bar{G})$ is unknown, since the prior probability of the default model and data are unknown). The resulting spectral densities are shown in Fig. 8b and are found to depend only weakly upon the default model. It is also possible to optimize the perturbation theory default model and hence to optimize the corresponding spectral densities. The effect of this optimization is shown in Fig. 9 where the same data in Fig. 8 is analytically continued with perturbation theory default models shown in Fig. 9a. In the optimization of the default model, the *df*-hybridization $V$ is treated as a variational parameter. The

corresponding posterior probabilities are shown in the inset. This probability function is even more strongly peaked and of significantly larger magnitude than the one found for the Gaussian default model. These features indicate that the result found with the perturbation theory result is much more probable than the one found with the Gaussian default model.

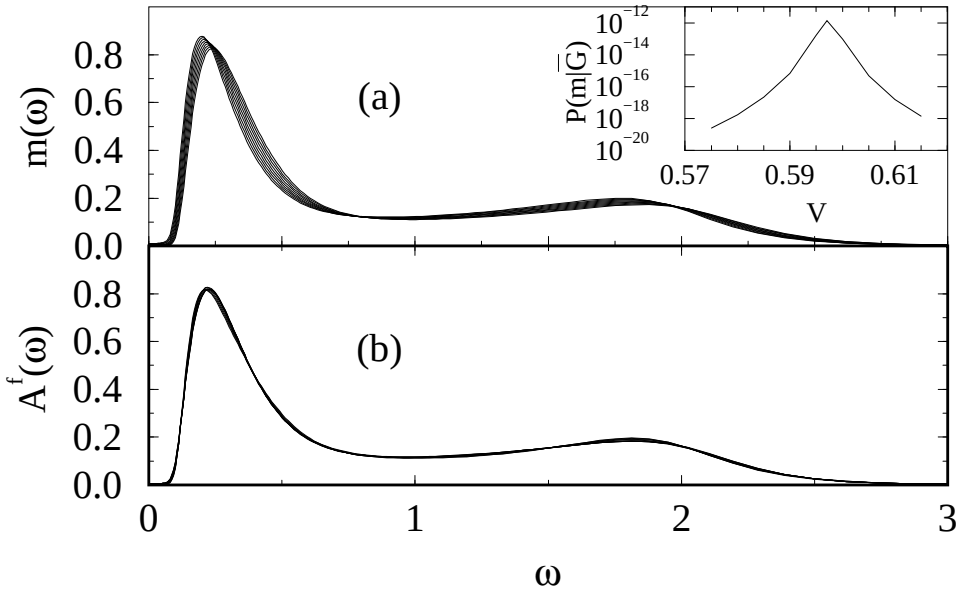


Figure 9: Test of the dependence of the spectral density upon the default model. (a) The default model found from perturbation theory as the *fd*-hybridization $V$ strength is varied; (b) the series of spectral densities produced by the default models in (a). The posterior probability of the default model $P(m|\bar{G})$ is plotted as a function of $V$ is shown in the inset to (a). Here, $P(m|\bar{G})$ is sharply peaked around $V = 0.597$ at a value which is several orders of magnitude larger than that for the posterior probability shown in the inset of Fig. 8. The perturbation theory default model yields a significantly more probable spectral density than the Gaussian default model.

### *4.3 Error Propagation*

In Fig. 10, we compare the optimal spectral densities obtained with the optimal perturbation theory, Gaussian, and flat default models. (The flat default model, with no adjustable parameters, is not optimized.) The posterior probabilities for each result indicate that the perturbation theory default model produces by far the most probable spectral density. However, we note that the qualitative features of the spectral density change little with the default model even though a large variety of default models were used. This independence is one signature of good data!

As a final test of the quality of the spectral density, one can evaluate its error in different intervals of frequency. In Fig. 10, we chose to assign error bars to the integrated spectral density ($h(\omega) = 1$) over different nonoverlapping regions. The

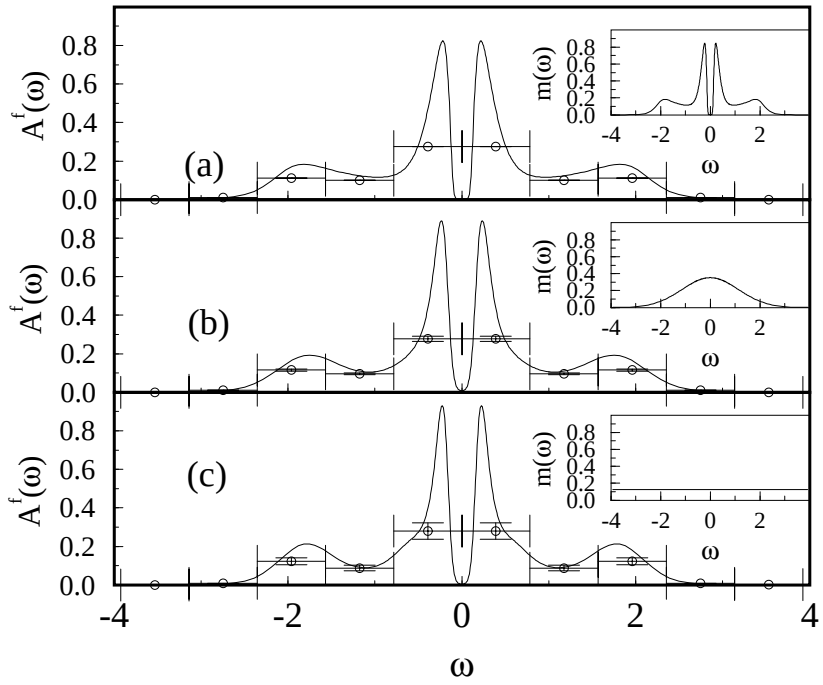


Figure 10: The f-electron density of states $A^f(\omega)$ generated using (a) a perturbation theory, (b) a Gaussian, and (c) a flat default model. These models are shown as insets to each graph. The data points indicate the integrated spectral weight within 10 non-overlapping regions of width indicated by the horizontal error bar. The vertical error bar indicates the uncertainty of the integrated weight within each region.

width of the region centered at each error bar is indicated by the horizontal spread of the error bar, the spectral weight within this region is indicated by the value of the data point, while the estimate of the uncertainty is indicated by the vertical spread. The perturbation theory default model yields the most precise spectra at all frequencies, consistent with the posterior probabilities of the models.

### *4.4 Two-Particle Spectra*

There are special difficulties associated with the calculation of spectral densities associated with two-particle Green's functions. These difficulties include noisier and more correlated data and the lack of a good default model. The latter problem stems from the traditional difficulties of performing perturbation theory for two-particle properties.

As an example, we will analytically continue the local f-electron dynamic spin susceptibility $\chi''(\omega)$ of the symmetric PAM. The Monte Carlo data $\chi(\tau) = 2\langle S^-(\tau)S^+(0)\rangle$

is related to $\chi''(\omega)$ by

$$\chi(\tau) = \int_0^\infty d\omega \frac{\omega[e^{-\tau\omega} + e^{-(\beta-\tau)\omega}](\chi''(\omega)/\omega)}{1 - e^{-\beta\omega}} \,. \tag{73}$$

We modified the kernel to account for the symmetry of the data $G(\tau) = G(\beta - \tau)$ and the spectrum $\chi''(\omega) = -\chi''(-\omega)$

$$K(\tau, \omega) = \frac{\omega[e^{-\tau\omega} + e^{-(\beta-\tau)\omega}]}{1 - e^{-\beta\omega}} \,. \tag{74}$$

Note that the kernel is non-singular at $\omega = 0$ and the spectral density $\chi''(\omega)/\omega$ is positive definite.

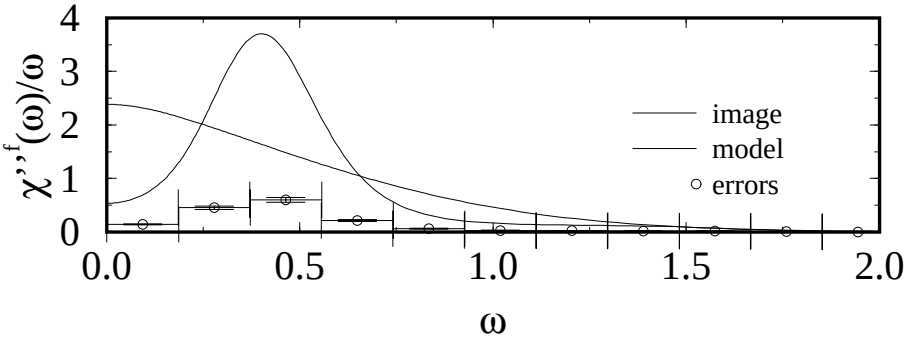


Figure 11: $\chi''(\omega)/\omega$ for $V = 0.6$, $U = 2$ and $\beta = 20$ for the PAM generated using a default model defined by two moments of the spectral density. The data points indicate the integrated spectral weight within 10 non-overlapping regions of width indicated by the horizontal error bar. The vertical error bar indicates the uncertainty of the integrated weight within each region.

An alternative to perturbation theory is to construct a default model by using different moments of the spectral function that can be calculated from the data as constraints to the principle of maximum entropy. The moments used to generate the default model are

$$\frac{1}{2}\chi(\omega = 0) = \int_0^\infty d\omega (\chi''(\omega)/\omega) \,. \tag{75}$$

$$\chi(\tau = 0) = \int_0^\infty d\omega \, (\chi''(\omega)/\omega) \, \omega \coth(\beta\omega/2) \,. \tag{76}$$

The (unnormalized) model is then generated by maximizing the entropy subject to these constraints imposed with Lagrange multipliers $\lambda_0$ and $\lambda_1$ and is easily found to be

$$m(\omega) = \exp[\lambda_0 + \lambda_1 \omega \coth(\beta\omega/2)] \tag{77}$$

where $\lambda_0$ and $\lambda_1$ are determined by the constraint equations above.

Clearly this procedure may be generalized to utilize an arbitrary number of measured moments and often provides a better default model than perturbation theory. However, as shown in Fig. 11, the final spectral density can differ significantly from the default model when defined in this way. Nevertheless, the error bars indicate that the spectral density is trustworthy.

### *4.5 Annealing Method*

Occasionally we have reason to calculate a series of spectra for a variety of temperatures (i.e. for the calculation of transport coefficients). If this set is sufficiently dense, then starting from a perturbation theory default at high temperature, we may use the resulting spectra as a default model for the next lower temperature. As far as we know, this procedure has no Bayesian justification; however, it has significant physical motivation. At sufficiently high temperatures, perturbation theory often becomes exact. Thus, this annealing procedure may be initialized with an essentially exact result. Furthermore, as the temperature is lowered, we expect the high frequency features of many spectra to freeze out (this is an essential assumption behind the numerical renormalization group method). Thus, the QMC is only required to supply information about the low-frequency features. Since QMC is a discrete sampling procedure in Euclidean time, according to Nyquist's theorem QMC only provides information below the Nyquist frequency $\omega_N = \pi/\Delta\tau$. Thus, the perturbation theory provides the high-frequency information, the QMC the low-frequency information, and MEM provides a natural method for combining these information sources.

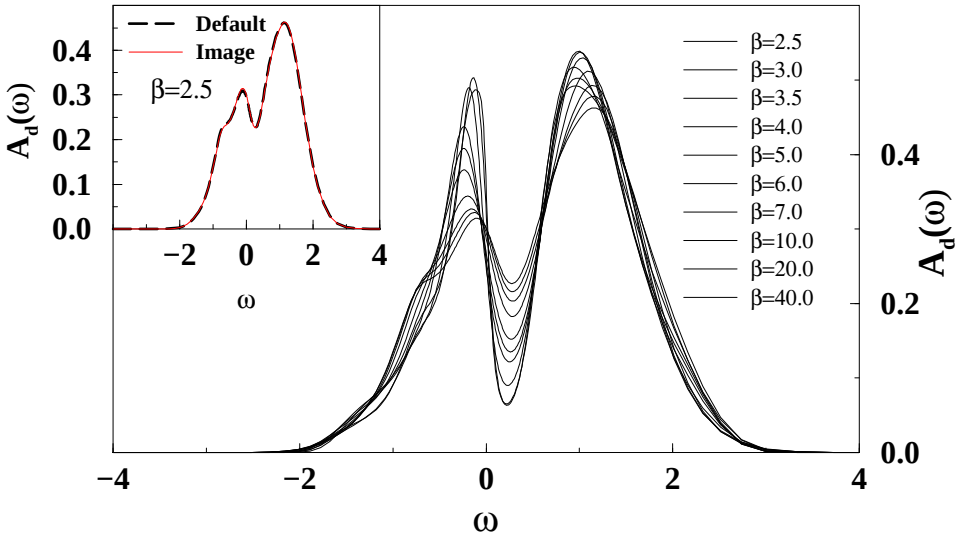


Figure 12: *The evolution of the d-electron density of states of the asymmetric PAM when* $U = 1.5$, $V = 0.6$ $n_d = 0.6$, *and* $n_f = 1.0$. *At high temperatures, as shown in the inset, the spectra is in essentially exact agreement with second-order perturbation theory. In addition, the d-electron states far from the Fermi surface are weakly renormalized by the strong electronic correlation on the f-orbitals. Thus, as the temperature is lowered, the low-frequency spectra change continuously, whereas the high frequency features change very little.*

For example, the evolution of the d-electron density of states of the asymmetric PAM is shown in Fig. 12. At high temperatures, as shown in the inset, the spectra is in essentially exact agreement with second-order perturbation theory. In addition, the d-electron states far from the Fermi surface are weakly renormalized by the strong electronic correlation on the f-orbitals. Thus, as the temperature is lowered, the low-

frequency spectra change continuously, whereas the high frequency features change very little.

We conclude this section by noting that while the systematic preparation of the data of the described in Sec. 2.4 and the qualification of the spectrum described in this section is time-consuming, we believe that it is as important to quality of the final result, as is an accurate MEM code.

## 5 Conclusion

The Maximum Entropy Method is a precise and systematic way of analytically continuing Euclidean-time quantum Monte Carlo results to real frequencies. Due to the exponential nature of the kernel which relates the spectra and the data, there are many $A$ which correspond to the same $\bar{G}$. With the MEM we employ Bayesian statistics to determine which of these is most probable. Bayesian inference is also used to assign error bars to integrals over the spectrum and optimize the default model.

The posterior probability of the spectrum is given by the product of the prior probability and the likelihood function. The entropic nature of the prior insures that the only correlated deviations from the default model which appear in the spectrum are those which are necessary to reproduce the data. The form of the likelihood function is determined by the central limit theorem, assuming that the data are statistically independent and Gaussianly distributed. Insuring these preconditions is the most critical step in the MEM procedure, and requires that the data be systematically rebinned and that the data and the kernel be rotated into the space in which the covariance of the data is diagonal.

Once the data has been properly characterized, we calculate the optimal spectrum using Bryan's algorithm which searches for a solution in the reduced singular space of the kernel. Bryan's method is more efficient than conventional techniques which search the entire spectral space. For any search algorithm three different techniques can be employed to set the Lagrange parameter $\alpha$ which determines the relative weight of the entropy and misfit: the historic, classic or Bryan's averaging technique. With precise uncorrelated data, each returns essentially the same spectrum, but with less-precise uncorrelated data, Bryan's technique yields the best results. Also, as the QMC data are systematically improved, images produced with Bryan's technique appear to converge more quickly than those produced by the other techniques.

Together, the techniques discussed in this chapter provide a powerful, accurate, and systematic approach to the analytic continuation problem. In each case where we have employed these techniques we have been able to produce spectra that are precise at low frequencies, and free from spurious (unjustified) features at all $\omega$.

We would like to acknowledge useful conversations and fruitful collaborations with H. Akhlaghpour, O. Biham D.L. Cox, C. Groetsch, J.E. Gubernatis, C. Jayaprakash, H.R. Krishnamurthy, L. Robinson, R.N. Silver, D. Sivia, and A.N. Tahvildarzadeh. This work was supported by the National Science Foundation grants DMR-9406678

and DMR-9357199, the Office of Naval Research grant N00014-95-1-0883 and by the Ohio Supercomputer Center.

## References

1. H.-B. Schüttler and D.J. Scalapino, Phys. Rev. Lett. **55**, 1204 (1985); Phys. Rev. B **34**, 4744 (1986).
2. H.J. Vidberg and J.W. Serene, J. Low Temp. Phys. **29**, 179 (1977).
3. G. Wahba, SIAM Journal on Numerical Analysis **14**, 651 (1977).
4. S.R. White, D.J. Scalapino, R.L. Sugar, and N.E. Bickers, Phys. Rev. Lett. **63**, 1523 (1989).
5. M. Jarrell and O. Biham, Phys. Rev. Lett. **63**, 2504 (1989).
6. M. Jarrell, and J.E. Gubernatis, Physics Reports Vol. **269** #3, p133-195, (May, 1996).
7. J.E. Hirsch and R.M. Fye, Phys. Rev. Lett. **56**, 2521 (1986).
8. J. Skilling, in *Maximum Entropy and Bayesian Methods* edited by J. Skilling (Kluwer Academic, Dordrecht, 1989), p. 45.
9. S.F. Gull and J. Skilling, IEE Proceedings **131**, 646 (1984).
10. For a discussion of kurtosis and skewness, as well as a discussion of the probability that a distribution is Gaussian, see *Numerical Recipes*, W.H. Press, S.A. Teukolsky, W.T. Vettering, and B.P. Flannery (Cambridge University Press, 1989), chap. 14.
11. It is tempting to disregard (i.e., set to 0) the off-diagonal elements of the covariance matrix as an alternative method of alleviating this pathology. Then, the eigenvalues will simply be the well-defined variance of $\bar{G}(\tau)$. However, this procedure neglects the correlations in the error which are clearly visible in Fig. 4 and yeilds an incorrect likelihood function. We have found that this procedure produces unpredictable results, especially when the data quality is marginal.
12. J. Skilling and R.K. Bryan, Mon. Not. R. astr. Soc. **211**, 111, (1984).
13. S.F. Gull, in *Maximum Entropy and Bayesian Methods* edited by J. Skilling (Kluwer, Dordrecht, 1989), p. 53.
14. S.F. Gull and G.J. Daniels, Nature, **272**, 686 (1978).
15. H. Jeffreys, **Theory of Probability**, (Oxford, Claredon Press, 1939); see also E. Jaynes, IEEE Trans. Sys. Sci. Cyb. Vol. SSC-4, (1993).
16. J. Skilling, in *Maximum Entropy and Bayesian Methods*, edited by J. Skilling (Kluwer Academic, Dordrecht, 1989), p. 455.
17. R.K. Bryan, Eur. Biophys. J. **18**, 165 (1990).
18. . M. Jarrell, J.E. Gubernatis, and R.N. Silver. Phys. Rev. B, **44**, 5347-50 (Sept. 1991).
19. J.J. Dongarra, C.B. Moler, J.R. Bunch, and G.W. Stewart, *LINPACK User's Guide* (SIAM, Philadelphia, 1979).
20. W.H. Press, S.A. Teukolsky, W.T. Vettering, and B.P. Flannery, *Numerical*

*Recipeies in Fortran, Second Edition* (Cambridge University Press, Cambridge, 1992).
21. B.T. Smith, J.M. Boyle, Y. Ikebe, V.C. Klema, and C.B. Moler, *Matrix Eigensystems Routines — EISPACK Guide* (Springer-Verlag, New York, 1976).
22. M. Jarrell, H. Akhlaghpour, and Th. Pruschke, Phys. Rev. Lett. **70**, 1670 (1993); Phys. Rev. B. **51**, 7429-40 (15, March 1995).
23. B. Horvatic, D. Sokcevic, and V. Zlatic, Phys. Rev. B **36**, 675 (1987).