BAYESIAN INFERENCE AND THE ANALYTIC CONTINUATION OF IMAGINARY-TIME QUANTUM MONTE CARLO DATA

Mark JARRELL\textsuperscript{a}, J.E. GUBERNATIS\textsuperscript{b}

\textsuperscript{a}Department of Physics, University of Cincinnati, Cincinnati, OH 45221, USA
\textsuperscript{b}Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
Bayesian inference and the analytic continuation of imaginary-time quantum Monte Carlo data

Mark Jarrell \textsuperscript{a}, J.E. Gubernatis \textsuperscript{b}

\textsuperscript{a} Department of Physics, University of Cincinnati, Cincinnati, OH 45221, USA
\textsuperscript{b} Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Received January 1995; editor: D.K. Campbell

Contents:

1. Introduction 136 4.4. Other items and issues 156
2. The analytic continuation problem 138 5. Case study: the $D = \infty$ periodic Anderson model 159
2.1. Dynamical correlation functions and their analytic continuation 138 5.1. Qualifying the data 160
2.2. Examples of dynamical correlation functions 140 5.2. Qualifying the solution 165
3. Methods of Bayesian inference 142 5.3. Special considerations: two-particle spectra 169
3.1. Bayesian inference 143 6. Applications 170
3.2. Likelihood function 144 6.1. Anderson models 171
3.3. Entropic prior 145 6.2. Hubbard model 176
3.4. Bayesian inference revisited 147 6.3. Holstein model 178
3.5. Likelihood function revisited 150 6.4. Quantum spin models 182
4. Maximum entropy analysis 151 6.5. Other applications 185
4.2. Distributions and their normalizations 153 Appendix A. The Bryan algorithm 189
4.3. Bayesian inference once again 155 References 192

Note added in proof 195
Abstract

We present a way to use Bayesian statistical inference and the principle of maximum entropy to analytically continue imaginary-time quantum Monte Carlo data. We supply the details that are lacking in the seminal literature but are important for the motivated reader to understand the assumptions and approximations embodied in these methods. First, we summarize the general relations between quantum correlation functions and spectral densities. We then review the basic principles, formalism, and philosophy of Bayesian inference and discuss the application of this approach in the context of the analytic continuation problem. Next, we present a detailed case study for the symmetric, infinite-dimension Anderson Hamiltonian. We chose this Hamiltonian because the qualitative features of its spectral density are well established and because a particularly convenient algorithm exists to produce the imaginary-time Green's function data. Shown are all the intermediate steps of data and solution qualification. The importance of careful data preparation and error propagation in the analytic continuation is discussed in the context of this example. Then, we review the different physical systems and physical quantities to which these, or related, procedures have been applied. Finally, we describe other features concerning the application of our methods, their possible improvement, and areas for additional study.
1. Introduction

The inability to extract dynamical information from imaginary-time quantum Monte Carlo data has long been a factor limiting the usefulness of quantum simulations. Quantum Monte Carlo simulations generate static correlation functions $G(\tau)$, from which we want to extract the dynamical spectra $A(\omega)$ by inverting

$$G(\tau) = \int d\omega \frac{A(\omega)e^{-\tau \omega}}{1 \pm e^{-\beta \omega}}. \quad (1.1)$$

Linear response theory relates many of these spectral functions directly to experimentally measurable quantities. By being able to compute these quantities, we can connect the simulations more closely to experiment. Static correlation functions are easy to obtain, but dynamical information has generally been hard to obtain because of the lack of a consistently effective procedure to invert this transform, i.e., to analytically continue the imaginary-time data into real-frequency information. This state of affairs is frustrating because of the sub-optimal “rate of return” from relatively expensive simulations and is unfortunate because of the limited relevance of the readily computable quantities to experimentally measurable quantities. Recently, along with our colleagues, Devinder Sivia and Richard Silver, we published procedures [1–3] that provided a practical solution to this problem. Since those publications, these procedures have been adopted by several research groups, and some additional insights into their successful implementation have developed. In this report, we will present an extended tutorial on these procedures, giving their current state as we understand them, and also present a brief review of the spectrum of their applications. Our focus will be on the technically difficult issues for their proper use. To do this, we feel the potential user also needs a full understanding of their logical, probabilistic, and statistical framework.

Prior to our work, attempts to solve the analytic continuation problem fell into two broad classes. In one class were methods that modified the Monte Carlo algorithm [4]; in the other class were ones that used existing algorithms and attempted extracting information from the resulting data [5–8]. We are concerned with the latter approach. Of previous work in this class, most germane to our work are the ones by Schüttler and Scalapino [5], White et al. [7], and Jarrell and Biham [8].

Schüttler and Scalapino were the first to propose a least-squares approach and to identify the inherent difficulty of the problem: the extraction is similar to performing an inverse Laplace transform numerically, which is a well-known, ill-posed problem. With the necessarily noisy and incomplete Monte Carlo data, the unique determination of a spectral density is impossible. In the least-squares approach, a “best” solution is sought by constraining the solution with information on moments and sum rules, assumptions about smoothness, or the requirements of positivity [5,8]. Associated with each constraint is a Lagrange multiplier that becomes an unknown parameter of the solution. In general, the spectral functions produced by these methods are just qualitatively interesting.

The Bayesian-based maximum entropy approach we use has different qualities as it explicitly approaches statistical data analysis within the concepts of conditional probabilities (Bayesian logic) [9]. In this approach, the spectral density is regarded as a probability function, and what is generally extracted from the data is the most probable spectral density [1,3]. What is unique about the maximum entropy (MaxEnt) approach is the specification of the prior probability function of the solution in terms of the information-theory definition of entropy. In the absence of data, the resulting spectral density is the one that maximizes the entropy. From a practical point of view, the problem
is reduced to finding the most probable solution by maximizing the entropy when it is constrained by the least-squares problem. In a recent form of the method, classic maximum entropy [10,11], the Lagrange multiplier associated with this constraint is determined from Bayesian logic; as a result, the method has no parameters to adjust arbitrarily.

The Bayesian approach to data analysis is a way of incorporating prior knowledge about probabilistic relationships among the data, the variables, and the solutions. MaxEnt specifies the prior probability function of the solution and prohibits correlations between different frequency values unless they are warranted by the data. The least-squares approaches assume that the prior probability is uniform, which is to say it becomes incorporated into a normalization factor.

Since entropy is a relative function, MaxEnt contains the choice of a “default model” to set the zero and the maximum of the entropy. More importantly, the default model is the solution produced in the absence of data or relevant information in the data. It also represents an additional way to incorporate certain types of prior knowledge about the spectral density into the solution process. The absence of additional knowledge is manifested by a “flat model” which is a constant, independent of frequency. Since quantum simulation data lack very high frequency information, a model that incorporates proper high frequency behavior is often useful. In our calculations for the Anderson impurity Hamiltonian [2,12,13], for example, we mainly used perturbation theory to provide a default model. At high frequencies, our results are biased towards the default model, which is becoming exact, but at low frequencies, they exhibited an important universality not present in the default model, but present in the physics.

We feel the default model can allow perturbation theory and quantum Monte Carlo to be combined in a novel way; however, this model can be chosen in a variety of other ways. For example, if several moments of the spectral density are known, then MaxEnt can be used to determine the most likely default model based on this information. In a practical sense, the default model is a convenient means to incorporate prior information about the solution into the problem. More recent use of the procedures, however, adopt the flat model, and shift priorities from model determination to the determination of good data. With good data, experience has shown the relative insensitivity of the results on the model. To solve an ill-posed problem, nothing beats good data.

The MaxEnt approach has successfully been used to extract dynamical information from various simulations of several different Hamiltonians calculated by several different quantum Monte Carlo methods. Here, in a self-contained manner, we will present and discuss general considerations necessary to make the procedures work. We feel that simply applying the procedures in a “black box” manner is insufficient to obtain proper results.

In Sections 2, 3, and 4, we summarize the general relations between quantum correlation functions and spectral densities and the general features of MaxEnt. These sections are intended to be self-contained, summarizing information spread across various sources. In Section 5, we present a detailed case study for the symmetric, infinite-dimension Anderson Hamiltonian. We chose this Hamiltonian because the qualitative features of its spectral density are well established and because a particularly convenient algorithm exists to produce the imaginary-time Green's function data [14]. Then, in Section 6, we review the different physical systems and physical quantities to which these, or related, procedures have been applied. Finally, in Section 7, we describe other features concerning the application of our methods, their possible improvement, and areas for additional study.
2. The analytic continuation problem

The analytic continuation problem seeks to extract real frequency, dynamical information from imaginary-time correlation functions computed in quantum Monte Carlo simulations. The imaginary-time correlations functions are one- and two- particle Green's functions, and associated with these Green's functions are spectral densities. We seek these spectral densities because linear-response theory relates these functions to experimentally measurable quantities. To define the analytic continuation clearly, we will first define the relations between the correlation functions, represented by the Green's functions, and the spectral densities. Then, detailing the analytic continuation problem is straightforward. Our brief summary of the properties of various time-dependent correlation functions and their relation to spectral densities is from various sources [15–18].

2.1. Dynamical correlation functions and their analytic continuation

We consider a system described by a Hamiltonian $H$ and perturbed by an external field which couples to an operator $B$. The linear response of the system to this field is described by the retarded Green's function (in units of $\hbar = 1$)

$$iG_R(t) = \langle [B(t), B^\dagger(0)] \rangle_\beta, \quad t > 0$$  \hfill (2.1)

where the angular brackets denote thermal averaging, the operators $B^\dagger$ and $B$ are in the Heisenberg representation, and the positive (negative) sign is used for fermion (boson) operators. Often, experimentally interesting properties of the system, such as the optical conductivity, NMR relaxation rate, dynamic structure factor, etc., are related to the spectral density of this Green's function. With the Fourier transform of $G_R(t)$ defined by

$$\hat{G}_R(\omega) = \int_0^\infty \frac{dt}{2\pi} e^{i\omega t} G_R(t)$$  \hfill (2.2)

the associated spectral density is

$$A(\omega) = -\frac{1}{\pi} \text{Im} \hat{G}_R(\omega).$$  \hfill (2.3)

Because of the analyticity of the retarded Green's function in the upper-complex plane,

$$\hat{G}_R(\omega) = -\int_{-\infty}^{\infty} \frac{d\omega'}{\pi} \text{Im} \hat{G}_R(\omega') \frac{\Delta}{\omega - \omega' + i\eta}, \quad 0 < \eta \ll 1.$$  \hfill (2.4)

So with $A(\omega)$ and (2.4), the Fourier transform

$$G_R(t) = \int_{-\infty}^{\infty} d\omega \ e^{i\omega t} \hat{G}_R(\omega)$$  \hfill (2.5)

determines $G_R(t)$.

Most quantum Monte Carlo simulations at finite temperatures are done in imaginary time. The Wick rotation, $-it \rightarrow \tau$, converting real time $t$ to imaginary time $\tau$, transforms oscillatory exponentials $e^{itH}$,
which appear in the Heisenberg representation of the operators, into diffusive ones $e^{-\tau h}$. At large $t$, the real-time operator becomes highly oscillatory and Monte Carlo sampling inefficient as sampling on smaller and smaller times scales is necessary just to achieve proper self-cancellations [19].

What the finite-temperature quantum Monte Carlo procedures estimate are Green’s functions of the form

$$G(\tau) = \langle T_\tau B(\tau) B^\dagger(0) \rangle,$$

(2.6)

where $T_\tau$ is the imaginary-time-ordering operator. Because this Green’s function satisfies $G(\tau) = \pm G(\tau + \beta)$, its Fourier transform is given by

$$\hat{G}(i\omega_n) = \int_0^\beta d\tau e^{-i\omega_n\tau} G(\tau)$$

(2.7)

where $\omega_n$ is a Matsubara frequencies equal to $(2n + 1)\pi/\beta$ [2n\pi/\beta] for fermion [boson] operators. This transform is related to the spectral density in a manner similar to the relation between $G_R(\tau)$ and $A(\omega)$

$$\hat{G}(i\omega_n) = \int_{-\infty}^{\infty} d\omega' \frac{A(\omega)}{i\omega_n - \omega'}.$$  

(2.8)

The inverse Fourier transform is

$$G(\tau) = \frac{1}{\beta} \sum_n e^{i\omega_n \tau} \hat{G}(i\omega_n),$$

(2.9)

and applying this transform to (2.8), one can show that

$$G(\tau) = \int_{-\infty}^{+\infty} d\omega \frac{e^{-\tau \omega}}{1 \pm e^{-\beta \omega}} A(\omega)$$

(2.10)

with the positive [negative] sign for fermion [boson] operators. The solution of this equation for $A(\omega)$ is the principal subject of this paper. With quantum Monte Carlo estimates of $G(\tau)$, we seek to solve this equation for $A(\omega)$. From (2.3), (2.4), and (2.8), we begin the see the basis for the “analytic” prescription to do the analytic continuation [18]: $i\omega_n \rightarrow \omega + i\eta$.

While imaginary-time quantum Monte Carlo eliminated the difficulties associated with sampling highly oscillatory functions and made $G(\tau)$ relatively easy to obtain, obtaining $A(\omega)$ from $G(\tau)$ has its own difficulties. These difficulties arise because at large positive and negative frequencies the kernel of (2.10)

$$K(\tau, \omega) = \frac{e^{-\tau \omega}}{1 \pm e^{-\beta \omega}}$$

(2.11)

is exponentially small, so the large $|\omega|$ features of $A(\omega)$ depend upon subtle features in $G(\tau)$. To obtain $A(\omega)$ from $G(\tau)$ is thus akin to doing numerically a two-sided Laplace transform. This type of problem is well known to be ill-posed, and obtaining $G(\tau)$ from quantum Monte Carlo data compounds the problem as the data are incomplete and noisy.
The spectral density has several general features that we will exploit,

\[
\text{sign}(\omega) A(\omega) \geq 0 \quad \text{bosons,}
\]
\[
A(\omega) \geq 0 \quad \text{fermions.}
\]  \hspace{1cm} (2.12)

Additionally,

\[
\int_{-\infty}^{\infty} d\omega \ A(\omega) < \infty
\]  \hspace{1cm} (2.13)

which is often a manifestation of some physical sum rule. The boundedness means that we can always re-define the spectral density so that the integral is unity. These properties thus enable us to interpret \(A(\omega)\) as a probability function.

2.2. Examples of dynamical correlation functions

To give several examples of interesting correlation functions, it seems best to discuss them in the context of specific Hamiltonians. First, we will consider the single impurity (spin degenerate) Anderson model

\[
H = \sum_{k\sigma} \varepsilon_k n_{k\sigma} + \sum_{k\sigma} V_{kd} (c_{k\sigma}^\dagger d_{\sigma} + d_{\sigma}^\dagger c_{k\sigma}) + \varepsilon_d \sum_{\sigma} d_{\sigma}^\dagger d_{\sigma} + U n_{d\uparrow} n_{d\downarrow}
\]  \hspace{1cm} (2.14)

where \(\varepsilon_k\) is the band energy of an electron in state \(k\), \(\varepsilon_d\) is the orbital energy of the impurity, \(V_{kd}\) is the strength of the hybridization between the orbital and the conduction band, and \(U\) is the strength of the electro-static repulsion between two electrons both occupying the orbital state. The total spectral density has contributions from the orbital states and the conduction band. We will only be concerned with the impurity (orbital) state contribution.

We define \(G_{\sigma}(\tau > 0) = \langle d_{\sigma}^\dagger(\tau) d_{\sigma}(0) \rangle\) where \(d_{\sigma}\) is the operator that removes an electron from state \(d\) with spin \(\sigma\), and also define \(A_{\sigma}(\omega)\) as the corresponding spectral density. This quantity represents the probability of finding an electron with spin \(\sigma\) and energy \(\omega\) in the state \(d\). Next, defining \(A(\omega) = \sum_{\sigma} A_{\sigma}(\omega)/2\) and \(G(\tau) = \sum_{\sigma} G_{\sigma}(\tau)/2\) and then using (2.10), we can write

\[
G(\tau) = \int_{-\infty}^{+\infty} d\omega \frac{e^{-\tau\omega}}{1 + e^{-\beta\omega}} A(\omega).
\]  \hspace{1cm} (2.15)

The spectral density associated with the impurity states satisfies the sum rule

\[
\int_{-\infty}^{\infty} d\omega \ A(\omega) = 1.
\]  \hspace{1cm} (2.16)

Another interesting case for the single-impurity model arises for the two-particle Green’s function

\[
\hat{\chi}(\tau) \equiv 2 \int_0^\beta d\tau (d_{\uparrow}^\dagger(\tau) d_{\uparrow}(\tau) d_{\downarrow}^\dagger(0) d_{\downarrow}(0)) = \int_{-\infty}^{+\infty} \frac{d\omega \ \chi''(\omega) e^{-\tau\omega}}{\pi \ 1 - e^{-\beta\omega}}
\]  \hspace{1cm} (2.17)
where $\chi''(\omega)$ is the associated spectral density. Such spin correlation functions are the imaginary-part of the transverse magnetic susceptibility. The following sum rule exists
\[
\int_{-\infty}^{\infty} \frac{d\omega}{\pi} \frac{\chi''(\omega)}{\omega} = \chi(T)
\]  
(2.18)

where $\chi(T)$ is the magnetic susceptibility at temperature $T$. One can also show that
\[
\frac{1}{T_1 T} = K \lim_{\omega \to 0} \frac{\chi''(\omega)}{\omega},
\]  
(2.19)

where $K$ is some constant depending on the g-factor of the impurity nucleus, the Bohr magneton, and the coupling between the nuclear and d-electron spin. $T_1$ is the nuclear magnetic relaxation time.

For several other examples, we will consider the extended Hubbard model in one-dimension. The Hamiltonian is
\[
H = T + V
\]  
(2.20)

where
\[
T = -t \sum_{l, \sigma} (c_{l, \sigma}^\dagger c_{l+1, \sigma} + c_{l+1, \sigma}^\dagger c_{l, \sigma}),
\]  
(2.21)

\[
V = U \sum_l n_{l \uparrow} n_{l \downarrow} + V \sum_l n_l n_{l+1}.
\]  
(2.22)

The operators $c_{l, \sigma}^\dagger$ and $c_{l, \sigma}$ create and destroy an electron at lattice site $l$ with spin $\sigma$. The number operator for an electron at site $l$ with spin $\sigma$ is $n_{l, \sigma} = c_{l, \sigma}^\dagger c_{l, \sigma}$, and $n_l = n_{l \uparrow} + n_{l \downarrow}$.

To study the response of (2.20) to a small electric field, we will assume periodic boundary conditions to make our chain on $N$ sites into a ring and then thread the ring by a flux which we represent by a vector potential
\[
A_x(l, t) = A_0 \theta^{t-l-\omega t}.
\]  
(2.23)

From linear-response theory, the real part of the conductivity $\sigma(q, \omega)$ is related to the imaginary-time, current-current correlation function by [18,20]
\[
J(q, \tau) \equiv \langle j(-q, \tau) j(q, 0) \rangle = \int_{-\infty}^{\infty} d\omega K(\tau, \omega) \sigma(q, \omega)
\]  
(2.24)

where
\[
j(q, 0) = it \sum_{l, \sigma} (c_{l, \sigma}^\dagger c_{l+1, \sigma} - c_{l+1, \sigma}^\dagger c_{l, \sigma}) e^{iql},
\]  
(2.25)

\[
j(q, \tau) = e^{-\tau H} j(q, 0) e^{-\tau H},
\]  
(2.26)

and
\[
K(\tau, \omega) = \frac{\omega e^{-\tau \omega}}{\pi 1 - e^{-\beta \omega}}.
\]  
(2.27)

Typically, one would be interested in $\sigma(0, \omega) = \lim_{q \to 0} \sigma(q, \omega)$. 
The real part of the conductivity is an even function of frequency \( \sigma(q, \omega) = \sigma(q, -\omega) \) and satisfies an analog of the f-sum rule [18]. For the extended Hubbard model, this sum rule is [21]

\[
\lim_{q \to 0} \int_{-\infty}^{\infty} d\omega \sigma(q, \omega) = -\frac{\pi}{2} \left( \frac{T}{N} \right)
\]

(2.28)

For finite-sized systems this relation is not obeyed for the \( q = 0 \) response if periodic boundary conditions are used. In this case, a non-zero Drude weight exists at zero-frequency even though the exact solution for an infinite system predicts an insulating state. The source of this puzzling result has been traced to the diamagnetic contribution to the total current [22]. Quantum Monte Carlo simulations, for the most part, are done for chains long enough that the anomalous \( q = 0 \) response is not observed, i.e., the sum rule at \( q = 0 \) is satisfied. For numerical reasons, one usually computes the conductivity only for \( q = 2\pi/N \), the smallest non-zero value allowed for our chain.

We will now rewrite (2.27) by using two symmetries. The first is the general relation that \( K(\tau, \omega) = K(\beta - \tau, \omega) \). This relation also illustrates the important point that only data for \( 0 \leq \tau \leq \beta/2 \) independently enters the problem. The second symmetry is the physical relation that \( \sigma(\omega) = \sigma(-\omega) \). With these symmetries (2.27)

\[
K(\tau, \omega) = \frac{1}{4\pi} \frac{\omega}{1 - e^{-\beta \omega}} [e^{-\tau \omega} + e^{-(\beta - \tau)\omega}].
\]

(2.29)

In this form, the kernel of the resulting equation is an even function of \( \omega \) and a useful symmetry is built into the equation.

There is still more that one can do. For the Hubbard model, and other lattice models, one can compute the \( \rho \) dependence of the imaginary-time, single-particle Green’s function or the spin-spin correlation function. Then, by doing the analytic continuation, the location of peaks in the corresponding spectral densities can be determined as a function of \( \omega \) and \( \rho \). In this manner, the dispersion \( \omega(q) \) of quasi-particle excitations, if they exist, can be determined. This example and others do not exhaust the possibilities but clearly illustrate that with the analytic continuation of the quantum Monte Carlo data access to a broader spectrum of useful physical quantities is possible. We will begin by defining the framework for our procedures for doing the continuation.

3. Methods of Bayesian inference

Solving the integral equation

\[
G(\tau) = \int_{-\infty}^{\infty} d\omega K(\tau, \omega) A(\omega)
\]

(3.1)

is difficult. For the forward problem, i.e., “Given \( A \), what is \( G \)?,” \( G(\tau) \) is insensitive to the large \( |\omega| \) details of \( A(\omega) \). For the inverse problem, i.e. “Given \( G \), what is \( A \)?,” this insensitivity leads to an ill-posed problem, so with incomplete and noisy information about \( G(\tau) \), which is the best we can expect from a Monte Carlo simulation, an infinite number of solutions exist. The task is then to select from these solutions the one that is “best” by some criterion. One criterion is selecting the
spectral density that is the most probable. Another is an “average” spectrum. The guiding principles we will use towards this end originate with the methods of Bayesian statistical inference. The non-negativity and the normalizibility of the spectral density allow us to interpret it as a probability function. This fact permits us to use the principle of maximum entropy (MaxEnt) in conjunction with the Bayesian methods. We now discuss the various strategies for finding $A(\omega)$ given incomplete and noisy information about $G(\tau)$.

3.1. Bayesian inference

Given two events $a$ and $b$, Bayes’s theorem says [23]

$$\Pr[a, b] = \Pr[a|b] \Pr[b] = \Pr[b|a] \Pr[a]$$

(3.2)

where $\Pr[a]$ is the probability of $a$, $\Pr[a|b]$ is the conditional probability of $a$ given $b$, and $\Pr[a, b]$ is the joint probability function for $a$ and $b$. Also from probability theory [23], we have the condition for the marginalization of a variable

$$\Pr[a] = \int db \Pr[a, b]$$

(3.3)

and the conditions for normalization

$$\int da \Pr[a] = 1,$$

(3.4)

$$\int da \Pr[a|b] = 1.$$  

(3.5)

From (3.2) and (3.3), it also follows that

$$\Pr[a] = \int db \Pr[a|b] \Pr[b].$$

(3.6)

In the analytic continuation problem, our events are the functions $\tilde{G}(\tau)$ and $A(\omega)$, where $\tilde{G}(\tau)$ is our estimated (measured) value of $G(\tau)$. Our initial criterion for a best solution will be the function $A(\omega)$ that maximizes $\Pr[A|\tilde{G}]$. We take

$$\Pr[A|\tilde{G}] = \Pr[\tilde{G}|A] \Pr[A] / \Pr[\tilde{G}].$$

(3.7)

$\Pr[A|\tilde{G}]$ is called the posterior probability, $\Pr[\tilde{G}|A]$, the likelihood function, $\Pr[A]$, the prior probability, and $\Pr[\tilde{G}]$, the evidence. If we were to integrate (3.7) over $A$ and use (3.5), we would find that

$$\Pr[\tilde{G}] = \int \mathcal{D}A \Pr[\tilde{G}|A] \Pr[A].$$

(3.8)

From this point of view, the evidence is a normalization constant depending only on the likelihood function and the prior probability.

Eq. (3.7) transfers the problem of specifying the posterior probability to the problem of specifying the likelihood function and the prior probability. These latter functions are generally ones about which we can either make reasonable assumptions or have specific knowledge. Bayes's theorem
provided a means to compute the posterior distribution function from these other functions. What Bayes's theorem does not specify is how to select a solution from possible choices consistent with this posterior distribution. Several choices are possible, e.g. means and modes, that may lead to quite different results. Our initial criterion of the most probable solution, i.e., the mode, is many respects is arbitrary. On the other hand, this choice often leads to a well-defined optimization problem, which is comforting because of its familiarity.

3.2. Likelihood function

The likelihood function is a probability distribution for the data given in terms of its dependence on a set of parameters. This function is usually used in conjunction with the method of maximum likelihood in which this distribution is maximized with respect to the parameters. In general, the method of maximum likelihood is useful in parameter fitting problems involving only a few parameters because if the number of parameters is very large, then very close fits to the data can be obtained that are usually dominated by noise. The method of maximum likelihood does not specify how the likelihood function is to be determined. Most often, what is used is a known analytic form. In many cases, the physical system under consideration provides a specific theoretical model for this form. In the analytic continuation problem, a different strategy is needed as the distribution of the data is determined by the specific quantum Monte Carlo algorithm used, the manner in which the data is analyzed, the specific quantity measured, etc. The data only implicitly depends on the parameters of interest. The central limit theorem is used to provide an approximation for this function. In this subsection, we will discuss this approximate form for the likelihood function.

Using the data to find \( A(\omega) \) is type of parameter fitting problem in as much as we will determine a set \( A = (A_1, A_2, \ldots, A_N) \) of values of \( A \) at a number of discrete values \( \omega_i \) of \( \omega \), but the number \( N \) of parameters we determine is on the order of 200. From (2.10), we find that a given \( A \) makes a specific prediction for a set of \( G(\tau) \) values \( G = (G_1, G_2, \ldots, G_L) \) at various discrete values \( \tau_i \) of \( \tau \). In the method of maximum likelihood, the objective is to vary \( A \) so that the resulting \( G \) maximizes some likelihood function. With discrete values of \( \tau \) and \( \omega \), (3.1) becomes

\[
G_i = \sum_{ij} K_{ij} A_j
\]

(3.9)

where \( K_{ij} = K(\tau_i, \omega_j) \) and \( A_i = A(\omega_i) \Delta \omega_i \).

In maximum likelihood method, the \( \hat{G}(\tau) \) are regarded as random variables drawn from a likelihood function \( f(\hat{G}; G) \) specified by \( G \). For a single random sampling of \( \hat{G}(\tau) \), a set of random variables \( \hat{G}^{(j)} = (\hat{G}_1^{(j)}, \hat{G}_2^{(j)}, \ldots, \hat{G}_L^{(j)}) \) describing the values of \( \hat{G} \) at discrete values \( \tau_i \) of \( \tau \) is realized. For a single \( \hat{G}^{(j)} \), we can assign a number \( dP^{(j)} = f(\hat{G}^{(j)}; G) dG \). For \( M \) independent measurements, \( dP = \prod_{j=1}^{M} f(\hat{G}^{(j)}; G) dG \). The logarithm of the likelihood function is \( L = \sum_{j=1}^{M} f(\hat{G}^{(j)}; G) \). The method of maximum likelihood is based on the assumption that a single peak in the space of the parameters \( G \) dominates this function so that the mode \( \hat{G} \) is a reliable estimate of the parameters to be fitted and the width of the peak provides some measure of confidence in the fit.

From a Bayesian point of view, the method of maximum likelihood takes \( \Pr(A|\hat{G}) \propto \Pr(\hat{G}|A) \) so that the prior probability \( \Pr(A) \) is ignored. Hence, finding the \( A \) that maximizes the posterior probability is equivalent to finding the \( A \) that maximizes \( \Pr(A|\hat{G}) \propto e^{-L} \). The method does not specify the function \( f(\hat{G}^{(j)}; G) \). When the number of measurements becomes large, by central
limit theorem arguments, modes becomes means, etc., and the asymptotic behavior of the likelihood function becomes [24]

$$e^{-L} = e^{-\chi^2/2}$$

(3.10)

where

$$\chi^2 = \sum_{i,j} (\tilde{G}_i - G_i) [C^{-1}]_{ij} (\tilde{G}_j - G_j)$$

(3.11)

and

$$\tilde{G}_i = \frac{1}{M} \sum_{j=1}^{M} \tilde{G}_i^{(j)}.$$  

(3.12)

$C$ is the covariance matrix

$$C_{ik} = \frac{1}{M(M-1)} \sum_{j=1}^{M} (\tilde{G}_i - \tilde{G}_i^{(j)})(\tilde{G}_k - \tilde{G}_k^{(j)}).$$

(3.13)

Maximizing the likelihood is equivalent to minimizing $\chi^2$. A least-squares fitting procedure is thus a special case of the method of maximum likelihood with $L = \frac{1}{2} \chi^2$.

The least-squares method is seen to have several conditions for its proper application: independent measurements and a multi-variant Gaussian distribution of the data. It also makes an important assumption that the prior probability is unimportant or unknown. However, if many values of $A$ are used, then this method can lead to overfitting ($\chi^2 \approx 0$) and very noisy and non-unique results. Clearly, some sort of regularization is required. In the MaxEnt approach, this is supplied by the entropic prior.

3.3. Entropic prior

Since the spectral density $A(\omega)$ is non-negative and normalizable to unity, it can be interpreted as still another probability function. The principle of maximum entropy says that the values of a probability function $A(\omega)$ are to be assigned by maximizing the entropy expression [10,11]

$$S = - \int d\omega A(\omega) \ln[A(\omega)/m(\omega)].$$

(3.14)

The function $m(\omega)$ is called the default model. In the absence of data (or with the lack of information about the likelihood function),

$$\Pr[A|G] \propto \Pr[A]$$

(3.15)

Hence, the $A$ that maximizes the posterior probability is the same as the one that maximizes $S$ when

$$\Pr[A] \propto e^{aS}.$$  

(3.16)

The form for $S$ that is usually used is

$$S \equiv \int d\omega \left( A(\omega) - m(\omega) - A(\omega) \ln[A(\omega)/m(\omega)] \right)$$

(3.17)
which reduces to (3.14) if both \( A(\omega) \) and \( m(\omega) \) have the same normalization.

The entropy has a maximum value of zero when \( A(\omega) = m(\omega) \). When \( A(\omega) \neq m(\omega) \), \( S \) is negative. Just how negative is a measure of how much \( A(\omega) \) differs from \( m(\omega) \). In applications, we will use values of \( A \) and \( m \) (weighted by \( \Delta \omega \)) at discrete values of the frequency \( \omega \). In this case, we use

\[
S = \sum_i \left[ A_i - m_i - A_i \ln\left(\frac{A_i}{m_i}\right) \right],
\]

(3.18)

where \( A_i = A(\omega_i) \Delta \omega_i \) and \( m_i = m(\omega_i) \Delta \omega_i \).

What is being done is simply utilizing the prior information that the spectral density can be interpreted as a probability function. The important point is the specific functional form for this probability function is a theorem that follows from four axioms associated with such concepts as subset independence, coordinate invariance, system independence, and scaling. These concepts translate into requirements that the functional form of \( S \) should not depend on the type of data being analyzed and the units of \( A \), that extra knowledge about \( A \) in one part of frequency space should not affect \( A \) elsewhere, and that independent data should combine multiplicatively. This leads to an entropy defined relative to a default model \( m \). Relative entropy is required since it is invariant to coordinate transformations and the number of \( A_i \) used. If one does not want to accept these axioms, one does not have to accept this form. What is a bit mysterious is the implications of this form, particularly with respect to the requirement that independent data should combine multiplicatively. The meaning of the requirement is easily understood by using Gull and Skilling's kangaroo argument [25]. Although this argument can sometimes lead to non-intuitive results, its generality illustrates the effect and usefulness of the entropic prior.

In the kangaroo problem, we are given the information that 1/3 of kangaroos are blue-eyed and 1/3 are left-handed and are asked to estimate the proportion that are are both blue-eyed and left-handed. We start by setting up a table that assigns yet unknown probabilities to the various combinations of handedness and eye color

<table>
<thead>
<tr>
<th></th>
<th>Left-handed</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>T</td>
<td>p₁</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>p₂</td>
</tr>
<tr>
<td>eyes</td>
<td>p₃</td>
<td></td>
</tr>
<tr>
<td></td>
<td>p₄</td>
<td></td>
</tr>
</tbody>
</table>

Formally, we are seeking the values of the \( p_i \), given \( p_1 + p_2 = p_1 + p_3 = 1/3 \) where we get these relations from the use of (3.3)

\[
\Pr[\text{Blue,Left}] + \Pr[\text{NotBlue,Left}] = \Pr[\text{Left}]
\]

(3.19)

\[
p_1 + p_3 = 1/3,
\]

(3.20)

\[
\Pr[\text{Left,Blue}] + \Pr[\text{NotLeft,Blue}] = \Pr[\text{Blue}]
\]

(3.21)

We see that we have more unknowns than equations, meaning we may have an infinite number of possible solutions.

Certain limiting-case solutions can easily be found. First, we consider the case where blue eyes and left-handedness have maximum positive correlations. We find
Left-handed

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>1/3</td>
<td>0</td>
</tr>
<tr>
<td>eyes</td>
<td>0</td>
<td>2/3</td>
</tr>
</tbody>
</table>

If these qualities have maximum negative correlation, then we have

Left-handed

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>0</td>
<td>1/3</td>
</tr>
<tr>
<td>eyes</td>
<td>1/3</td>
<td>1/3</td>
</tr>
</tbody>
</table>

On the other hand, if the qualities are uncorrelated, then

Left-handed

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>1/9</td>
<td>2/9</td>
</tr>
<tr>
<td>eyes</td>
<td>2/9</td>
<td>4/9</td>
</tr>
</tbody>
</table>

This case corresponds to the natural assumption that the fraction of blue-eyed kangaroos that are left-handed is the same as that for the total population of kangaroos.

To obtain a solution to the problem, Gull and Skilling considered several commonly used regularizing functions, each yielding a different answer. They found

<table>
<thead>
<tr>
<th>Regularizing function</th>
<th>$p_1$</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-\sum p_i \log p_i$</td>
<td>0.11111</td>
<td>None</td>
</tr>
<tr>
<td>$-\sum p_i^2$</td>
<td>0.08333</td>
<td>Negative</td>
</tr>
<tr>
<td>$\sum \log p_i$</td>
<td>0.13013</td>
<td>Positive</td>
</tr>
<tr>
<td>$\sum \sqrt{p_i}$</td>
<td>0.12176</td>
<td>Positive</td>
</tr>
</tbody>
</table>

The entropic regularizer gives the same result as the case where we assumed that eye color and handedness were uncorrelated. The other regularizers produced solutions that implied some degree of correlation. These features are quite general: The entropic prior will not produce correlations in the results unless there is data to support them. Other choices will generally produce unjustified correlations.

3.4. Bayesian inference revisited

In this section we will discuss three different types of MaxEnt analysis: historic, classic, and Bryan’s approach. In Sections 3.2 and 3.3, we discussed specific choices for the likelihood function and the prior probability of $A$. With these choices represented by (3.10) and (3.16), we can now use Bayes’s theorem (3.7) as $\Pr[A|G] \propto e^Q$ where

$$Q = \alpha S - \frac{1}{2} \chi^2.$$  

(3.22)
For a given $A$, finding the $A(\omega)$ that maximizes $\Pr[A|\tilde{G}]$ is thus equivalent to maximizing $Q$ with respect to $A(\omega)$

$$\frac{\delta Q}{\delta A}_{A=A_0} = 0.$$  \hspace{1cm} (3.23)

The result, however, will depend on $\alpha$, and in what is often called historic maximum entropy, $\alpha$ is adjusted to make $\chi^2 = N$. This choice is the expected value of $\chi^2$ when the errors are Gaussian noise but is otherwise ad hoc. Therefore, the historic approach may be considered as a form of regularization.

Why choose the most probable spectral density? Why not chose the average spectral density? The primary reason is that the most probable spectrum can be found with standard methods of optimization. Furthermore, maximizing the posterior probability to find the most probable spectral density is consistent with the points of view that we are maximizing the likelihood function, subject to the constraints of the entropy, or maximizing the entropy, subject to the constraints of the data. Implicit in these viewpoints is the assumption that a single maximum dominates the posterior probability. Along with this expectation is also the expectation that the curvature about the maximum is a useful measure of the sensitivity of the solution to small changes in the solution. This curvature provides a means for error estimation. If the posterior probability function is dominated by a single symmetric sharp peak, then little difference will most likely exist between the mean and the mode. If the peak is not symmetric, then significant differences between the mean and the mode can exist, and the average result will become less likely. This situation will particularly be the case if the peaked is strongly skewed. The act of averaging can also “wash out” small features of the spectrum and overly broaden larger features. Additionally, the easily interpretable measure of the sensitivity of the solution to small changes in the result is lost. While a variance about the mean is easily calculated, its meaning for a strongly skewed distribution is less cogent than for a relatively symmetric one. In any case choosing the most probable result is what defines historic MaxEnt and is the basis for classic MaxEnt. As we will see later (Section 4.4) the maximum has several useful features that provide insightful information about the solution.

We will now describe what is called classic maximum entropy [10,11] which incorporates empirical Bayesian methods to fix hyperparameters such as $\alpha$. We start by rewriting (3.7) as

$$\Pr[A, \alpha|\tilde{G}] = \Pr[\tilde{G}|A, \alpha] \Pr[A, \alpha] / \Pr[\tilde{G}]$$

$$= \Pr[\tilde{G}|A, \alpha] \Pr[A|\alpha] \Pr[\alpha] / \Pr[\tilde{G}]$$

$$= \Pr[\tilde{G}|A] \Pr[A|\alpha] \Pr[\alpha] / \Pr[\tilde{G}]$$

(3.24) \hspace{1cm} (3.25) \hspace{1cm} (3.26)

to make explicit the dependence of our results on $\alpha$. The normalization factor $\Pr[\tilde{G}]$ equals $Z_L Z_S(\alpha)$ where the factors of $Z_L$ and $Z_S$ normalize the likelihood function $\Pr[\tilde{G}|A]$ and the entropic prior $\Pr[A|\alpha]$, respectively. Introduced is a new function $\Pr[\alpha]$, which we will usually take to be what is called the Jeffreys prior, $P[\alpha] \propto 1/\alpha$ [3,11,26] or to be a constant. (See Section 4.3.) With use of (3.10) and (3.16), (3.26) becomes

$$\Pr[A, \alpha|\tilde{G}] = \Pr[\alpha] \frac{e^{\nu}}{Z_L Z_S(\alpha)}.$$  \hspace{1cm} (3.27)

Using (3.3) and then (3.2), we can also write the posterior probability $\Pr[A|\tilde{G}]$ as
Pr\[A|\tilde{G}\] = \int d\alpha \ Pr\[A, \alpha|\tilde{G}\] \tag{3.28}
= \int d\alpha \ Pr\[A|\tilde{G}, \alpha\] Pr\[\alpha|\tilde{G}\]. \tag{3.29}

The distribution function Pr\[\alpha|\tilde{G}\] follows from functionally integrating (3.26) with respect to \(A(\omega)\)

\(Pr[\alpha|\tilde{G}] = \int DA \ Pr[\tilde{G}|A, \alpha] Pr[A|\alpha] Pr[\alpha] / Pr[\tilde{G}]\)

\(= Pr[\alpha] \int DA \ e^{\frac{Q}{Z_L Z_s}}. \tag{3.31}\)

If the number of data is large, then Pr\[\alpha|\tilde{G}\] is often sharply peaked at some value of \(\alpha = \hat{\alpha}\) \[11\], and (3.29) becomes

\(Pr[\alpha|\tilde{G}] \propto Pr[A|\tilde{G}, \hat{\alpha}]\). \tag{3.32}\)

With this \(\hat{\alpha}\), the \(A\) which maximizes \(Q\) becomes the classic maximum entropy solution. We remark that this solution is not equivalent to solving

\(\frac{\partial Pr[A, \alpha|\tilde{G}]}{\partial A} = 0, \tag{3.33}\)

\(\frac{\partial Pr[A, \alpha|\tilde{G}]}{\partial \alpha} = 0. \tag{3.34}\)

The second equation fixes \(\alpha\) with the value that maximizes Pr\[A, \alpha|\tilde{G}\] simultaneously with \(A\). In classic maximum entropy, \(\alpha\) is chosen by maximizing Pr\[\alpha|\tilde{G}\]. Basically, the parameter \(\alpha\) is a nuisance. To remove it from the problem, we attempted to integrate it out of the problem and to work with Pr\[A|\tilde{G}\] instead of Pr\[A, \alpha|\tilde{G}\]. This integration (3.31) simplifies because the Pr\[\alpha|\tilde{G}\] part of the integrand is expected to be sharply peaked. The parameter \(\alpha\) is chosen to be \(\hat{\alpha}\) the location of this peak, and with it we are able to obtain specific functional forms for Pr\[A|\tilde{G}\] and then with these forms find the most probable \(A\). \(\hat{\alpha}\) depends on \(A\) and vice versa, so the solution is found iteratively.

In our work, we started with historic maximum entropy. This method worked for many cases, but for some other cases we would consistently get unsatisfying results. Fortunately, about that time classic maximum entropy was developed. The ill-posed nature of the our problem, however, restricts the space in which we can search for a solution so that our effective number of data is usually small and this made our legitimate use of classic maximum entropy questionable, i.e., the assumption that Pr\[\alpha|\tilde{G}\] is sharply peaked was generally not the case.

To proceed, we adopted a Bayesian approach developed by Bryan[26] where \(\alpha\) is addressed by marginalization. In Bryan's method we find the \(\hat{A}_\alpha\) that satisfy (3.23) for each value of \(\alpha\). Then, we choose as our solution \(\hat{A}(\omega)\) defined by

\(\hat{A}(\omega) = \int d\alpha \hat{A}_\alpha(\omega) Pr[\alpha|\tilde{G}]\) \tag{3.35}\)

where Pr\[\alpha|\tilde{G}\] is given by (3.31). In the absence of data, this procedure produces the default model as the answer. We remark that this solution is not necessarily the one which maximizes Pr\[A|\tilde{G}\].
The Bayesian analysis gives rise to a set of maximum entropy maps, parameterized by \( \alpha \), each of which are the modes of the probability distribution \( \Pr[A|\alpha, \tilde{G}] \).

These are the major details of the maximum entropy method and its application to analytically continuing quantum Monte Carlo data. The Bayesian line of inference eventually placed \( \alpha \) on a equal footing with \( A(\omega) \). This allows \( A(\omega) \) to be determined without needing to adjust \( \alpha \) or using properties of \( \alpha \) not supported by the data. Along the way the assumptions and the approximations inherent to the approach are clearly specified.

3.5. Likelihood function revisited

Our discussion of the likelihood function in Section 3.3, while brief, was standard. By assuming we have a large number of data, the functional form for \( \Pr[\tilde{G}|A] \) is specified. The parameters in this form are familiar, and ones we normally compute in our statistical analysis of our simulation. More significantly the method of maximum likelihood assumes a dominant maximum. To be meaningful, the curvature of the parameter space around this maximum must be positive and to be useful it should also be large. In light of the principle of maximum entropy, we can examine the likelihood function from a different perspective. To do so, we ask, “What is the most probable likelihood function consistent with \( \tilde{G} \) and \( C \).

Before answering this question, we will first state a general result. If we have \( n \) functions \( g_k(x) \) and a probability density \( p(x) \) that are to satisfy

\[
\theta_k = \int dx \, g_k(x) \, p(x)
\]

then the normalized \( p(x) \) that maximizes the entropy \( \int dx \, p(x) \ln[p(x)] \) and at the same time satisfies these constraints is

\[
p(x) = \lambda_0 \exp \left[ - \sum_{k=1}^{n} \lambda_k g_k(x) \right]
\]

where the \( \lambda_i \)’s (Lagrange multipliers) are determined from the normalization condition

\[
\lambda_0 \int dx \, \exp[- \sum_{k=1}^{n} \lambda_k g_k(x)] = 1
\]

and the \( n \) equations

\[
\lambda_0 \int dx \, g_k(x) \exp \left[ - \sum_{k=1}^{n} \lambda_k g_k(x) \right] = \theta_k.
\]

Because the likelihood function is a probability function, we can find the most probable one by maximizing the entropy, with \( m(\omega) = 1 \), subject to the constraints

\[
\int d\omega \, A(\omega) = 1,
\]

\[
C_{ik} = \frac{1}{M(M-1)} \sum_{j=1}^{M} (\tilde{G}_i - \tilde{G}_i^{(j)}) (\tilde{G}_k - \tilde{G}_k^{(j)})
\]
where
\[ G(\tau) = \int d\omega K(\tau, \omega) A(\omega). \]  
\hspace{0.5cm} (3.42)

The first constraint requires normalization or equivalently specifies the zeroth moment of the distribution. The second constraint specifies the average and the variance, i.e., the first and second moments. Using (3.39), we find that
\[ \Pr[\tilde{G}|A] = e^{-x^2/2}/Z_L \]  
\hspace{0.5cm} (3.43)

which is precisely what we had before.

What's the point? The maximum likelihood argument is more familiar and as such seems more acceptable. The maximum entropy argument is less familiar. Should it thus be less acceptable? Specifying the likelihood function in a purely Bayesian way emphasizes the limited amount of information we have available from the simulation to specify the likelihood function. The Monte Carlo algorithms provide no direct information about the form of this distribution. With just zeroth, first, and second moment information, the principle of maximum entropy leads us to a normalized Gaussian. Under normal circumstances, this functional form can poorly approximate the actual distribution. Having accepted the Gaussian form, however, we are obligated to produce data as consistently as possible with it. Achieving the consistency is the hardest part of the problem. The QMC data must be monitored for the achievement of Gaussianly distributed data. We will discuss this point more fully later. Our experience has been that people tend to be most concerned about the entropic prior. Our belief is that the likelihood function should be a greater cause of concern. We are lead to the entropic prior by the desire to seem logically consistent; we are lead to the Gaussian form because we seem to have no better choice.

We also remark that in finding the likelihood function by maximizing the entropy we obtained the most likely function consistent with our prior knowledge of the zeroth, first, and second moments, i.e., we sought a normalized function with a specific mean and variance. The answer was a Gaussian which is a function for which the third and higher moments depend only on the zeroth, first, and second moments. As a choice of likelihood function, a Gaussian is consistent with our prior information and contains no additional information. This consistency does not justify the choice of a Gaussian for the likelihood function, it merely reinforces the kangaroo argument's illustration of the meaning and consequences of the entropic prior. The use of the Gaussian likelihood function is only justified if the data is consistent with it.

4. Maximum entropy analysis

4.1. Global maximum

The three flavors of Bayesian solutions for \( A \) presented in Section 3.4 all involve finding the maximum of the function \( Q = \alpha S - L = \alpha S - \frac{1}{2} x^2 \) as a function of \( A_i \). Easily shown is that
\[ \frac{\partial^2 S}{\partial A_i \partial A_j} = -\frac{\delta_{ij}}{A_i} = -\frac{\delta_{ij}}{\sqrt{A_i A_j}}. \]  
\hspace{0.5cm} (4.1)
From this expression we learn two things. One is that the curvature tensor of the entropy in the space defined by the $A_i$ equals $\sqrt{A_i A_j} \delta_{ij}$. We also learn that the entropy is a concave function of the $A_i$. Because $-\chi^2$ is also a concave function of the $A_i$, so is $Q = \alpha S - \frac{1}{2} \chi^2$, and therefore $Q$ has a unique maximum for a fixed value of $\alpha$. Understanding the nature and the consequences of this maximum is important to understanding the strength and difficulties of the procedures we use.

To study this maximum, we expand $Q$ about it for a fixed value of $\alpha$, and following the convention of Gull and Skilling [10,11], we do this expansion in a coordinate system, which we describe by the variable $X_i$, where the curvature of $S$ is flat. We start with

$$Q(A, \alpha) \approx Q(\hat{A}_a) + \sum_i \delta X_i \left[ \frac{\partial Q}{\partial X_i} \right]_{\hat{A}_a} + \frac{1}{2} \sum_{ij} \delta X_i \delta X_j \left[ \frac{\partial^2 Q}{\partial X_i \partial X_j} \right]_{\hat{A}_a}$$

$$Q(\hat{A}_a) + \sum_{ij} \delta X_i \left[ \frac{\partial A_i}{\partial X_i} \frac{\partial Q}{\partial A_j} \right]_{\hat{A}_a} + \frac{1}{2} \sum_{ijkl} \delta X_i \delta X_j \delta X_k \left[ \frac{\partial A_i}{\partial X_i} \frac{\partial A_j}{\partial X_j} \frac{\partial^2 Q}{\partial X_i \partial X_j \partial A_k \partial A_l} \right]_{\hat{A}_a}$$

where $\delta X_i = X_i - \hat{X}_i$, and

$$\frac{\partial^2 Q}{\partial A_i \partial A_j} = -\frac{\alpha \delta_{ij}}{\sqrt{A_i A_j}} - \frac{\partial^2 L}{\partial A_i \partial A_j}.$$  

Choosing the new coordinate system to be such that

$$\frac{\partial A_i}{\partial X_i} = \sqrt{A_i} \delta_{ij}$$  

and because

$$\left. \frac{\partial Q}{\partial A_i} \right|_{A_i = \hat{A}_a} = 0$$

by assumption, we can write

$$Q(A, \alpha) \approx Q(\hat{A}_a) - \frac{1}{2} \sum_{ij} \delta X_i \Gamma_{ij} \delta X_j$$

where $\Gamma$ is the positive-definite matrix $\Gamma_{ij} = \alpha \delta_{ij} + A_{ij}$ with

$$A_{ij} = \left[ \sqrt{A_i} \frac{\partial^2 L}{\partial A_i \partial A_j} \sqrt{A_j} \right]_{\hat{A}_a}$$

and

$$\frac{\partial^2 L}{\partial A_i \partial A_j} = [K^T \cdot C^{-1} \cdot K]_{ij} = \sum_k K_{ik} [C^{-1}]_{kl} K_{lj}.$$  

In this new coordinate system, the curvature of $Q$ in the immediate vicinity of a maximum is controlled by matrix $\Gamma$. If the eigenvalues of this matrix are small, the curvature around the maximum is flat. This flatness complicates finding the maximum and also introduces considerable uncertainty in the results.
Eq. (4.5) illustrates that if $\sqrt{A}_i$ were the metric for the problem, then

$$DA = \prod_i \frac{dA_i}{\sqrt{A}_i}$$

and our change of variables would be represented by

$$DA \rightarrow DX = \prod_i dX_i.$$  

(4.10)

Indeed, various arguments suggest the appropriateness of this metric.\(^1\) We will use it in what follows. In particular, the covariance of the $X_i$ about the maximum is

$$\langle \delta X_i \delta X_j \rangle = \int DX \delta X_i \delta X_j Pr[A|\bar{G}] \approx (\Gamma^{-1})_{ij}.$$  

(4.12)

In the space defined by the eigenvectors of $\Gamma$, the covariance $\langle \delta X_i \delta X_j \rangle \rightarrow (\delta x_i)^2 = 1/\gamma_i$ where $\gamma_i$ is the $i$-th eigenvalue of $\Gamma$. Thus, small eigenvalues of $\Gamma$ can lead to large variances in the result. We note we also have

$$\langle \delta A_i \delta A_j \rangle \approx (\sqrt{\bar{A}} \cdot \Gamma^{-1} \cdot \sqrt{\bar{A}})_{ij} = \sqrt{\bar{A}_i} [\Gamma^{-1}]_{ij} \sqrt{\bar{A}_j}.$$  

(4.13)

4.2. Distributions and their normalizations

We will now make more specific the forms of the various probability functions needed in our analysis. The easiest function to specify is the likelihood function. We have that

$$Pr[\bar{G}|A] = e^{-x^2/2}/Z_L$$

(4.14)

where

$$Z_L = \int DG e^{-x^2/2} = \int \prod_i dG_i e^{-x^2/2}.$$  

(4.15)

Because

$$\chi^2 = \sum_{i,j} (\bar{G}_i - G_i) [C^{-1}]_{ij} (\bar{G}_j - G_j)$$

(4.16)

the integration for the normalization constant $Z_L$ is a standard multi-variant Gaussian one and readily performed to yield

$$Z_L = (2\pi)^{N/2} \sqrt{\det C}.$$  

(4.17)

For the prior probability of $A$ given $\alpha$, we have

$$Pr[A|\alpha] = e^{\alpha S}/Z_S(\alpha)$$

(4.18)

\(^1\) We see no basis for the validity of these arguments. We use this metric to connect with the seminal literature. Additionally, its use simplifies some of the algebra. Its main effect is to change the curvature about the maximum, not its location. Recently, Skilling [27] has repudiated this metric.
where

\[ Z_S(\alpha) = \int \mathcal{D}A e^{\alpha S} = \prod_i \int \frac{dA_i}{\sqrt{A_i}} e^{\alpha S_i}. \]  

(4.19)

with

\[ S_i = A_i - m_i - A_i \ln(A_i/m_i). \]  

(4.20)

We will estimate \( Z_S(\alpha) \) by approximating the exponential by a Gaussian centered at \( A_i = m_i \). The value of \( S \) at the maximum was defined to be zero so

\[ \int \frac{dA_i}{\sqrt{A_i}} e^{\alpha S_i} \approx \int dX e^{-\alpha X^2/2} = \sqrt{2\pi}/\alpha. \]  

(4.21)

Consequently, we have that

\[ Z_S(\alpha) \approx \left(\frac{2\pi}{\alpha}\right)^{N/2} = \frac{(2\pi)^{N/2}}{\sqrt{\det \alpha I}}. \]  

(4.22)

This approximate value of \( Z_S \) is expected to be valid for large \( \alpha \).

The remaining probability function we need is posterior distribution of \( \alpha \) given the data \( \hat{G} \)

\[ \Pr[\alpha|\hat{G}] = \Pr[\alpha] \int \mathcal{D}A \frac{e^{\theta}}{Z_L Z_S(\alpha)}. \]  

(4.23)

Previously, we said that one usually takes \( \Pr[\alpha] \) to be the Jeffreys prior or to be some constant. Jeffreys argues that \( \alpha \) is a scale factor, and without prior knowledge, he asserts the choice of \( \Pr[\alpha] \) should be the least informative one and requires \( \Pr[\alpha] \) to be invariant with respect to a change in scale, i.e., \( \Pr[\alpha] \propto \Pr[\alpha] d(\alpha) = \Pr[\alpha] d(\alpha) \). Thus, \( \Pr[\alpha] \sim 1/\alpha \). For this \( \Pr[\alpha] \) to be a properly normalizable function, we must require \( \alpha \) to be restricted between some minimum and maximum positive values, \( 0 < \alpha_{\min} < \alpha < \alpha_{\max} < \infty \). This restriction is not a problem, nor is the choice of this function an issue. In general, the solution seems to have little dependence on whether we chose \( \Pr[\alpha] \) to be the Jeffreys prior or to be a constant.

We evaluate the integral for \( Z_L \) by again approximating the exponential in the integrand by a Gaussian centered around the maximum of its argument. In this case, the maximum is the maximum of \( Q \). We have that

\[ \Pr[\alpha|\hat{G}] \approx \Pr[\alpha] \frac{e^{Q(\hat{\alpha})}}{Z_L Z_S(\alpha)} \int \mathcal{D}X e^{-\frac{1}{2} \delta X (\alpha I + \Lambda) \delta X} \]  

\[ = \Pr[\alpha] \frac{e^{Q(\hat{\alpha})}}{Z_L Z_S(\alpha)} \frac{(2\pi)^{N/2}}{\sqrt{\det [\alpha I + \Lambda(\hat{\alpha})]}}, \]  

(4.24)

The above Gaussian approximation is expected to be valid if the curvature about the peak is sufficiently sharp.
4.3. Bayesian inference once again

The historic MaxEnt solution finds the solution to (3.23) for which $\chi^2 = N$, and there is not much more to say other than such a solution may not exist and the method tends to underfit the data [11,28]. The underfitting tends to produce smooth results (those which are close to the default model).

For classic MaxEnt, the situation is richer. Classic MaxEnt maximizes $\Pr[A, \alpha|\tilde{G}]$ with respect to $A$ and simultaneously maximizes $\Pr[\alpha|\tilde{G}]$ with respect to $\alpha$. Maximizing $\Pr[A, \alpha|\tilde{G}]$ with respect to $A$ is equivalent to solving the historic MaxEnt equation (3.23) for a fixed value of $\alpha$. Maximizing $\Pr[\alpha|\tilde{G}]$ with respect to $\alpha$ can conveniently be done by considering

$$\frac{\partial \log \Pr[\alpha|\tilde{G}]}{\partial \alpha} = 0.$$  \hspace{1cm} (4.25)

Using (4.24), we have that

$$\log \Pr[\alpha|\tilde{G}] = \text{constant} + \log \Pr[\alpha] + \frac{1}{2} \log \left[ \det \alpha I (\alpha I + A)^{-1} \right] + Q(\hat{A}_\alpha)$$

$$= \text{constant} + \log \Pr[\alpha] + \frac{1}{2} \text{Tr} \log \left[ \alpha I (\alpha I + A)^{-1} \right] + \alpha S(\hat{A}_\alpha) - L(\hat{A}_\alpha).$$ \hspace{1cm} (4.26)

The maximization condition leads to

$$-2\alpha S(\hat{A}_\alpha) = \text{Tr}[\alpha (\alpha I + A)^{-1}] + \frac{d \log \Pr[\alpha]}{d \log \alpha} - \text{Tr}\left[ \frac{d \log \alpha}{d \log \alpha} \alpha (\alpha I + A)^{-1} \right].$$  \hspace{1cm} (4.27)

In general, the logarithmic derivative of the prior probability with respect to $\alpha$ is small, being of order unity for the Jeffreys prior or smaller, and will be dropped from further consideration. The last term in the above expression depends on the rate that the logarithm of the eigenvalues of $A$ change with logarithmic changes in $\alpha$. It too is expected to small and is dropped from consideration. The defining equation for the classic MaxEnt solution is

$$-2\alpha S(\hat{A}_\alpha) = \text{Tr}[\alpha (\alpha I + A)^{-1}] + \frac{d \log \Pr[\alpha]}{d \log \alpha}.$$  \hspace{1cm} (4.28)

We note that in contrast to historic MaxEnt the value of $\chi^2$ is irrelevant to this choice of $\alpha$.

The summation on the right-hand side of (4.28) is defined to be $N_g$, the number of good measurements. More generally,

$$N_g = \text{Tr} \alpha (\alpha I + A)^{-1}.$$ \hspace{1cm} (4.29)

If $\lambda_i$ are the eigenvalues of $A$, then

$$N_g = \sum_i \frac{\lambda_i}{\alpha + \lambda_i}$$ \hspace{1cm} (4.30)

so when a $\lambda_i$ is larger than $\alpha$, this value of $\lambda_i$ contributes approximately 1 to the summation. Otherwise, it contributes a value of approximately zero. Accordingly, of the $N$ data values, $N_g$ is a measure of the number of effective independent measurements. When $N_g$ is large, the above condition is expected to constrain severely the possible values of $\alpha$ for which a solution is possible. Because of this, one expects $\Pr[\alpha|\tilde{G}]$ to be sharply peaked at some $\alpha = \bar{\alpha}$, and the approximation in (3.32) is justified.
We recall that our definition of the entropy fixed its maximum at a value of zero when \( A_i = m_i \). Hence, as the data pulls the solution away from the default model, \( S \) becomes a negative number. The quantity \(-2\alpha S\) is thus a positive number and provides a dimensionless measure of movement away from the default model. It also is a measure of the structure in the solution.

The difficulty that develops for the analytic continuation problem is that \( N_s \ll N \). The small value for \( N_s \) is a consequence of the ill-posed nature of the problem. Accordingly, while for many cases the classic MaxEnt strategy works well, for other cases we found Bryan’s method [26] to be more effective. This strategy is a Bayesian-based approach. Instead of a mode for the “best” solution, we select a mean defined by

\[
\langle A \rangle = \int \mathcal{D}A \, d\alpha A(\alpha) \, \text{Pr}[A, \alpha | \tilde{G}] \tag{4.31}
\]

\[
= \int \mathcal{D}A \, d\alpha A(\alpha) \, \text{Pr}[A | \tilde{G}, \alpha] \, \text{Pr}[\alpha | \tilde{G}] \tag{4.32}
\]

\[
\approx \int d\alpha \, \hat{A}_\alpha \, \text{Pr}[\alpha | \tilde{G}] \tag{4.33}
\]

The meaning of the last equation is that for a fixed value of \( \alpha \) we chose the \( \hat{A}_\alpha \) that maximizes \( \text{Pr}[A | \tilde{G}, \alpha] \), i.e., \( \mathcal{Q} \), and take as our solution the sum over all values of \( \alpha \) of these solutions weighted by the posterior probability of \( \alpha \). \( \text{Pr}[A | \tilde{G}, \alpha] \) is assumed to be sharply peaked. The integration over \( \alpha \) is done numerically with the approximation (4.24) used in the integrand. With this approach we obtain the classic MaxEnt result if \( \text{Pr}[\alpha | \tilde{G}] \) becomes sharply peaked and that we recover the model in the absence of data.

4.4. Other items and issues

One normalization constant about which we have said little is the evidence \( \text{Pr}[\tilde{G}] \). It, however, can convey much useful information about the maximum value of \( \mathcal{Q} \) [29]. To find a useful approximate expression for the evidence, we will first write

\[
\text{Pr}[\tilde{G}, A, \alpha] = \text{Pr}[\alpha] \, \text{Pr}[A | \alpha] \, \text{Pr}[\tilde{G}, \alpha | A] = \text{Pr}[\alpha] \, \text{Pr}[A | \alpha] \, \text{Pr}[\tilde{G} | A]. \tag{4.34}
\]

Integrating over \( A \) produces

\[
\text{Pr}[\tilde{G}, \alpha] = \text{Pr}[\alpha] \int \mathcal{D}A \, \text{Pr}[A | \alpha] \, \text{Pr}[\tilde{G} | A] \tag{4.35}
\]

and then using Bayes’s theorem we find

\[
\text{Pr}[\tilde{G} | \alpha] = \int \mathcal{D}A \, \text{Pr}[A | \alpha] \, \text{Pr}[\tilde{G} | A] \tag{4.36}
\]

Now approximating the integrand as a Gaussian centered at the maximum,

\[
\text{Pr}[A | \alpha] \, \text{Pr}[\tilde{G} | \alpha] = \frac{e^{\mathcal{Q}(\hat{A})}}{Z_L Z_3(\alpha)} e^{-\left(\frac{1}{2}\delta X (\alpha + A) \cdot \delta X\right)} \tag{4.37}
\]

for a given \( \alpha \).
\[
\begin{align*}
\Pr[\hat{G}|\alpha] &= \frac{e^{-Q(\hat{\lambda}_0)}}{Z_L} \frac{(2\pi)^{N/2}}{Z_S(\alpha) \sqrt{\det (\alpha I + \Lambda)}} \\
&= e^{\alpha S(\hat{\lambda})} e^{-L(\hat{\lambda}_0)} Z_L \sqrt{\frac{\det \alpha I}{\det (\alpha I + \Lambda)}} \\
&\approx e^{\alpha S(\hat{\alpha})} e^{-L(\hat{\lambda}_0)} Z_L \sqrt{\frac{\det \alpha I}{\det (\alpha I + \Lambda)}}.
\end{align*}
\]

Thus, for a fixed value of \(\alpha\), the evidence is approximately equal to the product of two factors. One factor is mainly the contribution from the likelihood function; the other is, often called the Ockham factor, modifies this result. The maximum, and ultimately the solution to the problem, results from a competition between the likelihood function fitting the values of the spectral density to good data points, which is the tendency for small alpha, and defaulting to the model, which is the tendency for large \(\alpha\) because of the Ockham factor. When the data is closely fitted, the maximum represents a fit to the data by a "physical model" with many parameters. The Ockham factor favors large values of \(\alpha\) and hence favors results fitting the data with a "physical model" with fewer parameters. From this point of view, the result is a simpler physical model, than the least-squares solution, in accord with Ockham’s centuries old dictum.

What happens if we were to maximize \(\Pr[\hat{G}|\alpha]\) with respect to \(\alpha\)? From (4.40), we have

\[
\log \Pr[\hat{G}|\alpha] = \text{constant} + \log \Pr[\alpha] + \frac{1}{2} \text{Tr} \log \left( \alpha I (\alpha I + \Lambda)^{-1} \right) \]

\[
+ \alpha S(\hat{\lambda}_0) - L(\hat{\lambda}_0),
\]

which apart from a different constant is the same as the expression in (4.26) for \(\log \Pr[\alpha|\hat{G}]\). In light of Bayes’s theorem, this similarity is not surprising. The condition for the maximum is

\[
-2\hat{\alpha} S(\hat{\alpha}) = N \sigma^2
\]

which is the Gull and Skilling criterion for choosing \(\alpha\) in classic MaxEnt. (For classic MaxEnt, we do not use Eq. 4.42, but instead we find numerically the \(\alpha\) that maximizes the probability. This result can significantly differ from the Gull and Skilling approximation.)

Often estimating the level of error is difficult. Bayesian analysis provides a way to check the consistency of the error estimate [3,11,29]. To do this, we rescale our error estimate by a factor of \(\sigma^2\), which operationally means replacing \(C\) by \(\sigma^2 C\). Because of the dependency of other quantities on \(C\), we also have \(L \rightarrow L/\sigma^2\), \(\Lambda \rightarrow \Lambda/\sigma^2\), and \(Z_L \rightarrow \sigma^N Z_L\). From another point of view, we are working with \(Q = \alpha S - L/\sigma^2\) and looking for the optimal value of \(\sigma\), hoping that it is approximately unity which means our original estimates of errors were well done. Classic maximum entropy told us how to choose \(\alpha\). Now, for a fixed value of \(\alpha\), we are use a similar procedure to fix \(\sigma\).

Making these substitutions into (4.40), we can write

\[
\log \Pr[\hat{G}|\alpha, \sigma] = \text{constant} + \frac{1}{2} N \log \frac{1}{\sigma^2} - \frac{1}{2} \text{Tr} \log (\alpha I + \Lambda/\sigma^2) - L/\sigma^2.
\]

Then maximizing this expression with respect to \(\sigma\) (or to \(1/\sigma^2\)) yields

\[
2L/\sigma^2 = N - \text{Tr}[\Lambda(\sigma^2\alpha I + \Lambda)^{-1}].
\]
On the right-hand side, we recognize the term involving the trace operation as being \( N_s \), and for the left-hand side we have
\[
2L/\sigma^2 = \chi^2.
\]
(4.45)

Thus, one arrives at the interesting result
\[
\chi^2 + N_s = N
\]
(4.46)

when the error rescaling parameter maximizes the evidence. Hence, \( \sigma \) is a simple measure of how much we must rescale \( \chi^2 \) so that it represents the number of data not contributing to the pulling of the maximum away from the default model.

Eqs. (4.42) and (4.46) represent an interesting duality. Because each has only one independent parameter, \( \alpha \) or \( \sigma \), one expects the solution obtained by constraining \( \chi^2 \) with the entropy to be consistent with the solution obtained by constraining the entropy with \( \chi^2 \). Eq. (4.42) quantifies how the good data contributes to the result; Eq. (4.46), how the bad data contributes to the \( \chi^2 \).

In the practical use of (4.46), we find the maximum, calculate the corresponding \( \chi^2 \) and \( S \), and then choose \( \sigma \) so that
\[
\chi^2/\sigma^2 + N_s = N
\]
(4.47)
is satisfied. This value of \( \sigma \) is the error rescaling. It is a useful figure of merit (but not an modifier that “corrects” the data). From experience, we know that when \( \sigma \) deviates from unity by more than several tens of percent, we had better try to get better data. We also remark that (4.46) illustrates that in classic maximum entropy, \( \chi^2 < N \) and hence represents a solution that differs from historic maximum entropy by being fitted closer to the data. As a consequence, the features in a classic maximum entropy solution are usually sharper than those in historic maximum entropy solution.

We have said little about the default model. All our results explicitly depend on it, and to be precise, we should have written (3.26) as
\[
Pr[A|\tilde{G}, \alpha, m] = Pr[A|A, \alpha, m] Pr[A|\alpha, m] Pr[\alpha, m] Pr[m]/Pr[\tilde{G}|\alpha, m]
\]
(4.48)
\[
= Pr[\tilde{G}|A] Pr[A|\alpha, m] Pr[\alpha, m] Pr[m]/Pr[\tilde{G}|\alpha, m]
\]
(4.49)

We did not make the dependence explicit for notational convenience.

Natural questions are “What do we choose for the default model?” and “What effect does the default model have on the results?” The default model is both a nuisance and an opportunity. Unlike \( \alpha \), the default model is a parameter whose prior probability \( Pr[m] \) is not expected to be sharply peaked. It represents many parameters and finding a most a priori probable default model seems impractical. We have little reason to expect we can construct a good default model unless we have good a priori information about the solution. Also it seems unlikely we can easily obtain a spectral density by averaging it over the most probable spectral density times the probability of a model given the data. Thus, averaging the default model out of the problem appears impossible. We have to do something. In subsequent sections, we will give practical guidelines for choosing the default model and monitoring the sensitivity of the results to its choice. The choice of a good default model can positively impact the result, usually by choosing it to express prior information that we know about the solution (i.e. additional sum rules, exact high-frequency results, etc.). For now, we will simply
state that we routinely obtain results that are quite independent of the default model if we have good Monte Carlo data.

In the context of Bayesian inference, we would like to mention that one can generate formal procedures to prefer one default model over another. The simplest procedure compares the evidence of the solutions for two different models. The one with the greater evidence, particularly if it is significantly greater, points to the preferred model. More formal analysis will also show that the model with more parameters will be penalized by an Ockham factor. In any applications, we choose the least informative model, i.e., a constant over a finite range of frequencies and zero elsewhere. With good data, the results we obtain from this choice vary little from those obtained from another default model that more closely approximates the final result. If a good and an easily obtainable default model exist, one should of course use it.

5. Case study: the $D = \infty$ periodic Anderson model

We believe that it is possible to extract from the quantum Monte Carlo data spectral densities that are essentially free from artifacts caused by overfitting to the data and have only small and controllable amounts of statistical error. This extraction can be done by the systematic approach composed of careful qualification of both the QMC data and the image. We now will demonstrate this approach, using as an example the determination of f-electron spectral density of the single-particle Green's function for the symmetric, infinite-dimensional, periodic Anderson model (PAM).

We chose the symmetric PAM f-density of states because it is expected to have non-trivial features: a small hybridization gap centered at $\omega = 0$ at low temperatures and peaks associated with $fd$ hybridization at higher frequencies at all temperatures. Thus, the extent that we can resolve these features offers a substantial test of the method. This model is described by the following Hamiltonian:

$$H = \frac{-t^*}{2\sqrt{D}} \sum_{(ij)\sigma}(d_{i\sigma}^ad_{j\sigma} + d_{j\sigma}^ad_{i\sigma}) + V \sum_{i\sigma}(d_{i\sigma}^af_{i\sigma} + f_{i\sigma}^a d_{i\sigma}) + \frac{U}{2} \sum_{i\sigma}(n_{i\sigma}^f - \frac{1}{2})(n_{i\sigma}^f - \frac{1}{2})$$

(5.1)

where $d_{i\sigma}$ and $f_{i\sigma}$ ($d_{i\sigma}$ and $f_{i\sigma}$) 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 (5.1) on a simple hypercubic lattice of dimension $D$ 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 section.

The algorithm to solve infinite-dimensional lattice problems will be discussed in more detail in Section 6.2; however, it is based upon the Hirsch-Fye impurity algorithm [14]. Here the problem is cast into a discrete path formalism in imaginary time, $\tau$, where $\tau = l\Delta\tau$, $\Delta\tau = \beta/L$, and $L$ is the number of times slices. The values of $L$ used ranged from 40 to 160, with the largest values of $L$ reserved for the largest values of $\beta$ since the time required by the algorithm scales like $L^3$. No "sign problem" was observed at any filling.

In this section, our emphasis will be the qualification of the QMC data $G(\tau)$ and the qualification of corresponding spectral density determined from
Fig. 1. The PAM f-state Green's function \( G(\tau) \) for \( \tau = \beta/16 \) and \( \tau = \beta/2 \) plotted as a function of QMC step for \( U = 2 \), \( V = 0.6 \), and \( \beta = 20 \).

\[
G(\tau) = \int d\omega \frac{e^{-\tau\omega} A(\omega)}{1 + e^{-\beta\omega}}.
\] (5.2)

We used Bryan's MaxEnt method to determine \( A(\omega) \). The numerical algorithm to solve the MaxEnt equations is described in the Appendix. Our analytic continuation procedures consist of following steps:

1. Coarse grain the QMC data to remove the correlations that exist between measured values from successive Monte Carlo steps.
2. Diagonalize the covariance matrix of \( G(\tau) \) to remove correlations that exist between successive values of \( \tau \).
3. Analytically continue the data and systematically improve it until the spectral density stops changing.
4. Test the dependence of the result on the default model.
5. Estimate error.

The first two steps, which involve qualification of the data, are described in the next subsection, Section 5.1. The last three steps, which involve the qualification of the image, are described in Section 5.2. In general, these steps suffice. Sometimes numerically difficult to measure quantities like two-particle Green's functions require some additional consideration described in Section 5.3.

5.1. Qualifying the data

5.1.1. Correlations between Monte Carlo steps

As described in Section 3.2, the form of the likelihood function assumes that the data are Gaussian-distributed and composed of statistically independent samples. Typically, the data lacks both of these properties, and it necessary to remove correlations from the data at different Monte Carlo steps for a given value of \( \tau \) and at different \( \tau \) values for a given Monte Carlo step.

Here, we will first consider the correlations between Monte Carlo steps. In our example, these correlations can be seen in the single-particle Green's function data produced by the Monte Carlo process. These correlations are illustrated in Fig. 1 where measurements of \( G(\tau = \beta/2) \) and \( G(\tau = \beta/16) \) are plotted versus the Monte Carlo step number. These data are clearly correlated as adjacent measurements are similar in value. In addition, they are not Gaussianly distributed because the data are highly skewed since they are bounded from below by 0. The skewness in the data is more clearly
seen in the $G(\beta/2)$ values.

The histogram in Fig. 2 illustrates the non-Gaussian nature of the data more clearly. Especially for the $\tau = \beta/2$ data, the histogram shows significant deviation from a Gaussian fit (the solid lines). This deviation may be reduced by coarse-graining the data. By this we mean we average 2, 3, 4, etc. successive values of the data and use these averages instead of the original data. We call these groupings bins and refer to the number of such bins of data as $N_{\text{bins}}$. The consequence of this coarse-graining is shown in Fig. 3, where the sequential data points shown in Fig. 1 were placed into bins of 30 measurements each. Clearly, this procedure produces data which are much closer to being Gaussian-distributed.

It is also possible to quantify the deviation of the distribution from a Gaussian by measuring 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 [30]. Of course, because the data are a finite set, each of these measured moments has some standard deviation (proportional to $1/\sqrt{N_{\text{bins}}}$). Thus, 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 and skewness decrease monotonically, indicating the convergence of the distribution to a Gaussian. This behavior is shown in Fig. 4a for the $G(\tau = \beta/2)$ data. In addition, Fig. 4b shows that the error estimate also converges as the bin size increases. Here, the error estimate is given by

$$\sigma = \sqrt{(\langle G^2 \rangle - \langle G \rangle^2)/(N_{\text{bins}} - 1)}$$

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 increases. With independent data and a large number of bins, eventually $\sigma^2 \sim \cdots$
Fig. 5. $G(\tau)$ from one measurement compared to $G(\tau)$ 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.

$1/N_{\text{bin}}$. However, with fixed amount of data, as is typical with a QMC simulation, increasing the bin size decreases $N_{\text{bin}}$ proportionally, and the error estimate can saturate as illustrated in Fig. 4b. Thus, the saturation of the error estimate indicates that the correlations between Monte Carlo measurements, i.e., between bin averages, have been removed, and the point at which saturation occurs in a plot like Fig. 4b provides a useful first estimate of the bin size. In general, one should perform this test for all of the elements of the covariance; however, we have found it is often sufficient to test only a few of the diagonal elements. For the remainder of this section, we will assume that the bin size is sufficiently large 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.

5.1.2. Correlations in imaginary-time

In addition to correlations between measurements as a function of the Monte Carlo step, at a given Monte Carlo step the data are also correlated as a function of $\tau$. This correlation may be seen by comparing the results from a single measurement with those essentially exact values obtained from averaging over all measurements. Such a comparison is shown in Fig. 5 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 deviate from the exact results in a similar way.

Thus, in general, the covariance matrix $C_{ij}$ is not a diagonal matrix because measurements at different values of $\tau$ are correlated. If this is the case, then

$$\chi^2 = \sum_{l=1}^{L} \left( \frac{\tilde{G}_l - \sum_j K_{lj} A_j}{\sigma_l} \right)^2,$$

where $\sigma_l$ is obtained from the square-root of the diagonal elements of the covariance, is inappropriate to define the likelihood function because it assumes independent measurements.

To define a meaningful measure of how well $A_l$ reproduces the data, we must find the transformation $U$ which diagonalizes the covariance matrix.
\[(U^{-1}CU)_{ij} = \sigma_i^2 \delta_{ij}.\]  

(5.5)

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

\[K' = U^{-1}K, \quad \tilde{G}' = U^{-1}\tilde{G} \]  

(5.6)

and each measurement \(\tilde{G}'_i\) is statistically independent. Therefore, we can use

\[\chi^2 = \sum_i \left( \frac{\tilde{G}'_i - \sum_j K'_{ij}A_j}{\sigma_i^2} \right)^2 \]  

(5.7)

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 in Monte Carlo time. To calculate the covariance accurately, many bins of data are required; however, to remove the correlations in Monte Carlo time, the bin size must be large. If there are not enough bins of data, then the covariance and 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 \(N_{\text{bins}}\) 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 \(N_{\text{bins}} > L\). As shown in Fig. 6, where \(L = 41\), the eigenvalue spectrum displays a sharp break when \(N_{\text{bins}} < L\), indicating that only a finite number of directions, less than \(L\), are resolved. Empirically, we find that we need

\[N_{\text{bins}} \geq 2L \]  

(5.8)
in order to remove the pathology of the sharp break in the eigenvalues.\footnote{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 the $G(r)$. However, this procedure neglects the correlations in the error which are clearly visible in Fig. 5 and creates an incorrect likelihood function. We have found that this “elimination procedure” produces unpredictable results, especially when the data quality is marginal. In addition, discarding the off-diagonal elements causes the error propagation procedure appropriate for the problem, to be described in Section 5.2.3, to become invalid. The best “trick” appears to be making the bins large enough so that the data are uncorrelated and Gaussian-distributed, but having at least twice more bins of data than the number of imaginary-time slices.} The break results from some of the eigenvalues becoming much smaller than the rest. Simply throwing away the small eigenvalues, and their associated eigenvectors, does not cure the difficulty often caused by the break, i.e., artifacts in the solution.

5.2. Qualifying the solution

5.2.1. Convergence to the solution

Now that the correlations have been removed from the data, we can begin to process the data to obtain the spectral function. In this case study, we will use three default models: two non-informative models – the flat model $m(\omega) = \text{constant}$ and a simple Gaussian

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

(5.9)

and also a third one obtained from second-order perturbation theory in $U$ [31.39].

To minimize the effects of statistical error, the accuracy of the data needs to be increased until the spectral density has converged. This consideration is demonstrated in Fig. 7, where the accuracy of the data is increased 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 increased by a factor of two, 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 the previous subsection. As $N_{\text{bins}}$ 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 about a linewidth.

5.2.2. Default model selection

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 $\Pr[A, \alpha, m|\hat{G}] \propto \Pr[A, \alpha|\hat{G}, m]$. Then $\Pr[m|\hat{G}]$ is obtained by marginalizing $\Pr[A, \alpha, m|\hat{G}]$ over $A$ and $\alpha$. The effect of varying the default model parameters is 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 $\Pr[m|\hat{G}]$ of these default models, shown in the inset, is strongly peaked around $\Gamma \approx 1.6$. (We note that the normalization of $\Pr[m|\hat{G}]$ is unknown, since the prior probability of the default model and data are unknown). The resulting spectral densities are shown in Fig. 8b
Fig. 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 circles, was used.

Fig. 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), the posterior probability of the default model $\Pr(m|\tilde{G})$, produced by marginalizing the joint probability $\Pr(A, \alpha, m|\tilde{G})$ over $\alpha$ and $A$, is plotted as a function of $\Gamma$. The normalization of $\Pr(m|\tilde{G})$ is unknown because it depends upon the probability of the data and the prior probability of the default model which too are unknown. Thus, this posterior distribution may be used for comparison purposes only.

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,
Fig. 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 $Pr[m|\bar{G}]$ is plotted as a function of $V$ is shown in the inset to (a). Here, $Pr[m|\bar{G}]$ is sharply peaked around $V = 0.597$ which is a value several orders of magnitude larger than for the posterior probability shown in the inset of Fig. 8. The sharpness indicates that the perturbation theory default model yields a significantly more probable spectral density.

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.

Finally, one should compare the optimal spectral densities obtained with a variety of optimized default models shown in Fig. 10 where we plot the final spectral density 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, listed in the figure caption, 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 was used. This independence is one signature of good data!

5.2.3. Error propagation

As a final test of the quality of the spectral density, one can evaluate its error in different intervals of frequency. It is not possible to assign error bars to each point in the spectral density because the errors between different points are strongly correlated, and because $A_\lambda$ represents the probability within some region of finite width and hence lacks meaning at a specific value of $\omega$. Rather, it is only possible to assign error bars to integrated functions of the spectral density such as [32],

$$H = \int d\omega A(\omega) h(\omega).$$ (5.10)
Fig. 10. $A(\omega)$ generated using (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.

In the quadratic approximation, the probability of the the spectral density is

$$\Pr[A | \mathcal{G}, m, \alpha] \propto e^{-\frac{1}{2} \Delta A^T \cdot \nabla Q \cdot \Delta A},$$

and the covariance of the spectral density is

$$\langle \Delta A(\omega) \Delta A(\omega') \rangle = - (\nabla Q)^{-1}$$

so the error in the measurement of $H$ may be associated with the covariance of the spectral density

$$\langle \Delta A(\omega) \Delta A(\omega') \rangle$$

so the error in the measurement of $H$ may be associated with the covariance of the spectral density

$$\langle \Delta A(\omega) \Delta A(\omega') \rangle$$

The matrix $\nabla^2 Q$ is readily available because it is used as the Hessian of the Newton search algorithm typically used to find the spectral density. Thus, we may assign error bars to integrated functions of our spectral density.

In Fig. 10, we chose to assign error bars to the integrated spectral density ($h(\omega) = 1$) over different nonoverlapping regions. The 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. Note that in Figs. 10b and 10c the size of the error bars is roughly consistent with the deviation of the spectrum from the default model. One may show that the sensitivity to the default model is directly related to the size of the error bars.
5.3. Special considerations: two-particle spectra

Before leaving this section, we will next discuss the special difficulties associated with the calculation 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. 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.

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

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

(5.14)

Here, we multiplied the kernel and divided the dynamic susceptibility by \( \omega \) so that the kernel

\[
K(\tau, \omega) = \frac{\omega \left[ e^{-\tau\omega} + e^{-(\beta-\tau)\omega} \right]}{1 - e^{-\beta\omega}},
\]

(5.15)

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

The moments used to generate the default model are

\[
\frac{1}{2} \chi(\omega = 0) = \int_0^\infty d\omega \left( \chi''(\omega) / \omega \right),
\]

(5.16)

\[
\chi(\omega = 0) = \int_0^\infty d\omega \left( \chi''(\omega) / \omega \right) \omega \coth(\beta\omega / 2).
\]

(5.17)

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

\[
m(\omega) = \exp \left[ \lambda_0 + \lambda_1 \omega \coth(\beta\omega / 2) \right]
\]

(5.18)

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.

This procedure 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.

We will conclude this section by noting that while the systematic preparation of the data described in Section 5.1 and the qualification of the image described in Section 5.2 is time-consuming, we believe that they are as important to quality of the final result, as is an accurate MaxEnt code. In each case where we employed this systematic technique we produced images that are precise at low frequencies, and free from spurious (unjustified) features at all \( \omega \).
6. Applications

In this section, we review a select set of applications of Bayesian statistical inference methods to the determination of the spectral densities for a variety of physical models. We emphasize those applications that either illustrate a novel use of the methods or produce some interesting new physical results. We will first present results for the Anderson single-impurity model and then for the \( D = 2 \) and \( \infty \) periodic Anderson lattice models. For the single impurity model \((D = 0)\), the emphasis is on the universal behavior that follows from the Kondo effect. Producing this behavior was a severe test of our methods. For \( D = \infty \), the emphasis is on the mean-field physics, which is exact in this limit, and its possible connection with the behavior of lower dimensional \((D = 2 \text{ or } 3)\) systems.

After results for the Anderson models, we discuss results for the Hubbard model in one, two, and infinite dimensions, with the range of physics spanning calculations of analytically continued, quantum Monte Carlo derived Luttinger liquid dispersion relations to calculations of transport behavior in infinite dimensions. Next is a discussion of the Holstein model in two dimensions. In this model, the electrons are coupled to the phonons and the opening of a charge-density wave gap in the spectral density is studied. The availability of the QMC generated spectral density provided an opportunity to compare the predictions of various levels of approximations to the Eliashberg equations for the problem. Interestingly, the evaluation of the approximate theory requires the delicate, standard, analytic continuation by Padé approximant methods as suggested by Vidberg and Serene [33].

Finally, we present results for quantum spin systems and then for several different classes of additional interesting applications. For quantum spins systems the determination of the Haldane gaps and spin-wave dispersion relations is illustrated. The other applications illustrate the wide class of problems to which the methods presented can be applied.
6.1. Anderson models

6.1.1. Impurity models

For single impurity problem, the following is a general form of the $N$-fold degenerate Anderson model:

$$ H = H_0 + H_1, $$

where

$$ H_0 = \sum_{km} \epsilon_k n_{km} + \sum_{km} V_{km} (c_{km}^\dagger f_m + f_m^\dagger c_{km}) + \sum_m \epsilon_m n_m, \quad H_1 = \frac{1}{2} \sum_{m,m'} U_{mm'} n_m n_{m'}. $$

Here, $c_{km}^\dagger$ creates a state in the conduction band with the energy $\epsilon_k$, and $f_m^\dagger$ creates a state $m$ at the impurity site with the unrenormalized energy $\epsilon_m$. The hybridization of the conducting band with the impurity site $V_{km}$ does not mix states with different $m$. The operators $n_{km}$ and $n_m$ are the number operators for the conduction band and orbitals at the impurity site. It is generally assumed that the conduction band is infinitely wide and structureless; therefore, $V_{km}$ is neither energy nor channel dependent. This assumption leads to the simple relation for the impurity level half-width $\Gamma = \pi D(0) V$, where $D(0)$ is the energy density of states per channel at the Fermi energy. In (6.2), the symmetric matrix $U_{mm'}$, with the additional condition $U_{mm} = 0$, represents the Coulomb repulsion between two electrons occupying different orbitals at the impurity site. We have that $2 \leq m \leq N$.

When $N = 2$, these channels represent an up or down electron spin state, and we have what is called the spin-degenerate model. This model is the standard one studied for years in connection with the Kondo effect in dilute magnetic alloys and more recently with the anomalous behavior of heavy fermion and Kondo insulating materials. Recent theoretical investigations of this model use the numerical renormalization group [34] and the exact Bethe ansatz solution [35]. These methods focus primarily on the computation of Kondo temperature $T_K$ and exploring the extent of universality and mixed valence behavior as exhibited by such physical quantities and the static magnetic susceptibility. More recently, the numerical renormalization group technique was extended to the computation of spectral functions [36]. In general, this method works best at low temperatures and is difficult to apply to models other than impurity models. For the present model, increasing $N$ restricts its utility to low temperatures [37,38]. This model was the initial testing ground for our MaxEnt procedures which are quite complementary to the numerical renormalization group approach.

For years, the general, but not directly substantiated, belief was that besides single-particle side peaks at $\omega \sim U/2$, corresponding to the adding or subtraction of a single electron to the system, that spectral density of the model also had a central many-body peak at the Fermi energy ($\omega = 0$), called the Kondo resonance, which would be a universal function of $\omega/T_K$. As $T$ approaches 0, the value of this central peak was expected to approach $1/\pi \Gamma$, the so-called unitarity limit, where the magnetic moment on the impurity is completely screened by the conduction electrons. Near $T = 0$ the model is a Fermi liquid. Figs. 12, 13, and 14 illustrate these features convincingly.

In Fig. 12, we show the results of Silver et al. [2] for the $\pi \Gamma A(\omega)$ as function of $\omega$ scaled by the Kondo temperature $T_K$. In this figure, the symmetric form of the model is assumed ($\epsilon_1 = \epsilon_1 = -U/2$) for which the model displays particle-hole symmetry. Here we are solving
Fig. 12. Study of Kondo scaling of the spectral density at fixed $T/T_K = 1.5$ and for the values of $u = U/\pi \Gamma$ indicated. QMC-ME labels the quantum Monte Carlo and maximum entropy results. They are universal functions of $\omega/T_K$ (independent of $u$) for $\omega/T_K < 20$ and are non-universal for larger $T/T_K$ in the vicinity of the $\omega = U/2$ peaks. H labels the prediction of the theory of Horvatic, Sokcevic, and Zlatic [39] for the same values of $u$, with higher values of $\pi\Gamma A(0)$ corresponding to the larger $U$. H is distinctly non-universal.

Fig. 13. QMC-ME labels the quantum Monte Carlo and the maximum entropy results for fixed $u = 2.0$ and varying $T/T_K$ as indicated. The Kondo peak is approximately independent of $u$ for $\omega/T_K$ less than 20. DS labels the Doniach-Sunjic [40] expression for the shape of the Kondo peak at zero temperature.

$$G(\tau) = \int_{-\infty}^{+\infty} d\omega \frac{e^{-\tau\omega}}{1 + e^{-\beta\omega}} A(\omega)$$

for $A(\omega)$ where $G(\tau) = \sum_\sigma G_\sigma(\tau)/2$ and $A(\omega) = \sum_\sigma A_\sigma(\omega)/2\pi$.

Fig. 12 shows the universality of the Kondo resonance. $\pi\Gamma A(\omega)$ is plotted as a function of $T/T_K$ for different sets of interaction parameters. At low frequency, as the temperature is lowered, the curves collapse one upon another. In Fig. 13, the movement of the value of the central peak to the unitarity limit is clearly indicated. Fig. 12 and other results provided a benchmark for the range of validity of the perturbation theory. The breakdown begins around $u \sim 1$, whereas previously, it was conjectured to occur around $u \sim 3$. Of course, the perturbation theory results do not display the universality, but this particular theory provided a good default model $m(\omega)$ and was used in predicting the QMC obtained $A(\omega)$.

For dilute alloys, knowing the single-particle spectral density, one can use standard expressions relating moments of the spectral density to static transport properties [42,43]. Universality of the central Kondo peak leads to universality in these properties. In Fig. 14, we show the predicted resistivity as a function of $T/T_K$ compared the universal curve obtained from experiment. The agreement
Fig. 14. The resistivity $\rho(T)$, scaled by its zero temperature value, as a function of $T/T_K$. The solid line is the experimental curve for the dilute magnetic alloy $La_{1-x}Ce_xB_8$ whose Kondo temperatures span several orders of magnitude [41] and the markers are the predicted values using the spectral density determined form the Monte Carlo simulations of the single-impurity Anderson model. The markers are for simulations with four different values of $U$.

Fig. 15. $f(\omega) = \frac{\pi T_K}{2\chi(T)}\frac{\chi''(\omega)}{\omega}$ plotted versus $\omega/T_K$ for various values of $u$ when $T/T_K = 1.5$ and $\Gamma = 0.5$.

between theory and experiment spans three decades of temperature, from the Fermi liquid regime, where the proper $(T/T_K)^2$ dependence occurs, to the high temperature regime, where the expected $T \ln T$ dependence was found. In addition, the coefficient of the parabolic fall-off at low temperatures was found to agree to within 10% of the value predicted by the exact Bethe ansatz solution of the model. The results in this figure represents the only successful calculation of the complete range of this universal behavior by a single calculational approach.

By computing two-particle Green’s functions as a function of $\tau$, dynamic transport behavior can also be obtained [12]. In Fig. 15, we show a scaled form of the imaginary part of the dynamic transverse magnetic susceptibility as a function of $T/T_K$ for different values of the model parameters. Again, the universality of the result is striking. In obtaining these results, the default model was a particular approximation due to Salomaa [44] that had only a single free parameter model. We chose the parameter so that the evidence $Pr[\hat{G}|m]$ was maximized. At this point, the universality of
our results was most clearly evidenced. Also implied by the figure, with the use of (2.19), is the
universality of the NMR relaxation rate $T_1$. A surprising, but a not yet fully explained, result is the
NMR relaxation rate being proportional to the resistivity.

Besides the symmetric, spin-degenerate ($N = 2$) Anderson single-impurity model [45–47], studies
have also occurred for asymmetric parameterizations, $N$ up to 6, and for the application of an external
magnetic field. These studies push the model closer to realizable experimental situations. For example,
the degeneracy of the lowest multiplets of orbitals of many rare-earth or transition metal impurities
in metallic hosts have spin degeneracies higher than two. By assigning different values of $\epsilon_m$ to
these states, one can mimic the splittings of these states by spin-orbit and crystal-field effects. The
results of QMC studies on these richer parameterizations of the model have illustrated, besides the
persistent universality, the sensitivity of the peak value of Kondo resonance to $T$ and $N$ but the
relative insensitivity of the peak position. The relative insensitivity of the peak location seems to
be a characteristic feature of the model. As the model becomes more asymmetric and the average
occupancy of the impurity level drops below the particle-hole symmetric value of 1, the location of
the peak does move slowly, and seemingly proportionately, to the decrease in impurity charge.

6.1.2. Periodic models

The Anderson impurity and lattice models (and the related Kondo models) serve as paradigms
for heavy fermion systems. The impurity model correctly describes the formation and subsequent
screening of a magnetic moment on the f-impurity orbital. This behavior is reflected in the magnetic
and transport properties of dilute magnetic alloys. However, most heavy fermion materials are not
disordered alloys, but rather a regular lattice of rare-earth ions embedded in a metallic matrix. In
these materials, the regularity of the rare-earth ions and their intersite correlations become important.
These effects are modeled by the periodic Anderson model (PAM) in $D$ dimensions with

$$H = -\frac{t^*}{2\sqrt{D}} \sum_{(i)\sigma} \left( d_{i\sigma}^\dagger d_{i\sigma} + d_{i\sigma}^\dagger d_{i\sigma} + \sum_{i\sigma} (\epsilon_d d_{i\sigma}^\dagger d_{i\sigma} + \epsilon f f_{i\sigma} f_{i\sigma}) \right)$$
$$+ U \sum_{i} (n_{f i1} - 1/2)(n_{f i1} - 1/2) + V \sum_{i\sigma} (d_{i\sigma}^\dagger f_{i\sigma} + h.c.)$$

(6.4)

where $d(f)_{i\sigma}^{(1)}$ destroys (creates) a d(f)-electron with spin $\sigma$ on site $i$.

Thus far, most QMC simulations of this model, and the related Kondo lattice model, have been
confining to one-dimension and to the calculation of static (thermodynamic) properties of the system
[48]. However, recently, a new approach [49–51] based on an expansion in inverse dimensionality
$1/D$ about the point $D = \infty$ was proposed to study such strongly correlated lattice models. In
this limit, the requirement of a finite total energy per site makes it necessary to rescale nonlocal
interactions by an appropriate power of $1/D$ [49,50]. While some quantities, e.g. spin-exchange,
will reduce to the corresponding mean field theory results [50], the interactions in the model, like the
screened Coulomb repulsion, remain nontrivial even in this limit. The basic idea of this approach is
to map infinite-dimensional correlated lattice models onto an appropriate self-consistently embedded
Anderson impurity. This mapping becomes exact as the number of nearest neighbors to a given
site goes to infinity. Then, the resulting self energy becomes purely local and may be written as a
functional $\Sigma[G, U]$ of a site-excluded local f-state Green's function $G_f$ and the interaction strength
$U$. The introduction of $G = [G_{ii}^{-1} + \Sigma]^{-1}$ is necessary to avoid overcounting diagrams on the site $i$. 
Fig. 16. (a) Local neutron structure factor $S(\omega)$, and (b) $f$-electron single particle density of states $A(\omega)$ for the infinite-dimensional periodic Anderson model when $T = 0.05$ and $\epsilon_f = \epsilon_d = 0$ for various values of $V$ and $U$. The approximate coincidence of the gap in each figure indicates that it is proportional to $T_0$. (Since the plots correspond to different values of $T/T_0$, one does not expect the plots to coincide exactly).

In contrast to standard skeleton perturbation theory, the QMC yields an essentially exact solution to all orders in $U$, and thus it is necessary to subtract the interaction on the site where $\Sigma$ is calculated [52,53].

Most QMC calculations involving this model have been confined to the symmetric limit. Here, the model is expected to develop gaps [54] in the charge and spin spectra and become an insulator (a so-called Anderson insulator) at temperatures below the Kondo scale, $T_0$. Jarrell et al. [31] employed a QMC simulation in the infinite-dimensional limit, together with the Bryan MaxEnt method to calculate the spectra of the PAM. As shown in Fig. 16, as the temperature is lowered both the single-particle density of states and the dynamic spin susceptibility, first develop features reminiscent of what is seen in the impurity problem. The single-particle density of states, develops a Kondo-like peak and the dynamic susceptibility develops a spin resonance both centered at zero frequency. However, as the temperature is lowered further ($T < T_0$) both of these dynamic quantities develop gaps. Thus, the density of states (DOS), specific heat, and susceptibility are heavy-fermion-like for $T \gg T_0$ and insulator-like for $T \ll T_0$. These properties are consistent with experimental observations on so-called Anderson insulators [55]. The gap for charge excitations may be inferred from the $f$-electron density of states. Since this gap results from particle-hole excitations across this gap, it is roughly twice the size of the gap in $A(\omega)$ and is roughly equal to the spin gap.

The relative size of the charge and spin gaps, however, appears to depend upon dimensionality. For example, in the one-dimensional Kondo lattice, the gap was calculated with NRG [56] and also inferred from thermodynamic measurements [48]. The charge gap was always much larger than the spin gap, especially when the $fd$ magnetic exchange is small. More recently, the charge and spin gaps were calculated using QMC and maximum entropy methods for the two-dimensional model [57]. Here too, the charge gap was larger than the spin gap. Apparently, the non-local charge and spin
excitations found in finite dimensions, but not in $D = \infty$, are responsible for the larger value of $\Delta_c/\Delta_t$.

6.2. Hubbard model

The Hubbard model [58] for a single tight-binding band in $D$ dimensions is described by the Hamiltonian

$$\begin{align*}
H &= -\frac{t^*}{2\sqrt{D}} \sum_{(i,j),\sigma} (c_{i,\sigma}^\dagger c_{j,\sigma} + c_{j,\sigma}^\dagger c_{i,\sigma}) \\
&\quad + \sum_i \left[ \epsilon (n_{i,\uparrow} + n_{i,\downarrow}) + U(n_{i,\uparrow} - 1/2)(n_{i,\downarrow} - 1/2) \right].
\end{align*}$$

(6.5)

Here, $c_{i,\sigma}$ ($c_{i,\sigma}^\dagger$) destroys (creates) an electron of spin $\sigma$ on site $i$ of a hypercubic lattice of dimension $d$, $n_{i,\sigma} = c_{i,\sigma}^\dagger c_{i,\sigma}$, and $t^* = 1$ sets the energy scale.

The Hubbard model has been an enduring problem for strongly correlated electrons in condensed matter physics. It is believed at least to qualitatively describe some properties of high temperature superconductors [59], for which the planar electronic degrees of freedom may be mapped onto a two-dimensional single-band Hubbard model [60]. Despite the simplicity of the model, no exact solutions exist except in one dimension, where the knowledge is rather complete [61]. The unusual properties in one dimension also give rise to discussions of whether the behavior in 2D could be related to the 1D case [62]. This stimulated a large amount of activity in simulations of the two-dimensional model.

QMC, combined with MaxEnt, has been used to study the dynamics of both one- and two-dimensional Hubbard models. In a series of calculations, several authors have established many of the basic dynamic properties of the two-dimensional model, including the opening of both spin and charge (Mott) gaps in the single-particle spectrum of the half-filled model [7,63]. More recently, Preuss et al. [64], have performed a rather complete study of the dynamics of the one-dimensional model. The calculated the single-particle spectral function $A(p, \omega)$ and were able to calculate the single-particle excitation spectrum. They found that it could be characterized by cosine-like bands. Then, by calculating the two-particle charge and spin spectra, they found the expected separation of the spin and charge excitations. The dispersion of their charge excitation spectrum agreed well with Bethe Ansatz results [65].

Recently, Jarrell et al. have used methods similar to those discussed in Section 6.1.2 above to simulate the infinite-dimensional Hubbard model. Dynamical properties of the model are obtained using a modified version of the MaxEnt technique. To obtain the single-particle density of states, Jarrell et al. [66,67] used the finite-$U$ non-crossing approximation (NCA) [68] result for the infinite-dimensional Hubbard model as a default model at high temperatures where it is essentially exact [69]. At low temperatures, the NCA single-particle density diverges at the Fermi surface and cannot be used as a default model. Thus, for these temperatures Jarrell et al. used the numerically continued result of the next higher temperature as a default model. The posterior probability of the final result is employed to determine which default model should be used. Generally, they found that the crossover temperature between using the NCA default model and the higher-temperature continuation lies at $T \approx 2T_0$, where $T_0$ is the Kondo-like scale for this model [67]. It should be noted here that the
Fig. 17. (a) The evolution of the single particle density of states of the infinite-dimensional Hubbard model with increasing $U$ in the paramagnetic state ($1/T = 7.2$) of the Hubbard model. (b) The antiferromagnetic susceptibility of the Hubbard model when $U = 1.5$. In the inset, the scaling behavior $\chi_{AF} \propto (T - T_F)^{-1}$ is illustrated.

NCA default model is the exact result for sufficiently high temperatures, where the process begins. Once the density of states is determined, the self energy $\Sigma(\omega)$ may then be found by (numerically) inverting the Fadeev function $w(z)$ in the relation

$$A(\omega) = \text{Re} \left\{ w(\omega + \mu - \Sigma(\omega)) \right\} / \sqrt{\pi}. \quad (6.6)$$

The infinite-dimensional model displays much of the physics found in the three-dimensional model [53,66]. As shown in Fig. 17, at half-filling ($\delta = 1 - \langle n \rangle = 0$), correlation induced Mott-Hubbard gap in the single-particle density of states and a divergent antiferromagnetic susceptibility $\chi_{AF} \sim (T - T_F)^{-1}$ exist. Away from half-filling, the model develops a narrow peak of half-width $\approx T_0$ in the one-particle density of states near the Fermi energy (Fig. 18). As shown in the inset to Fig. 18, this peak is associated with reduction of the screened local moment $T\chi_{ii}$ and with the enhancement of the effective electron mass.

As shown in Fig. 19, Bulut et al. [71] find a very similar peak in the DOS of the two-dimensional Hubbard model, suggesting that this dynamically induced (Kondo-like) feature appears for all Hubbard models with $D \geq 2$. Moreover, Bulut et al. also analytically continued for the spectral function $A(k, \omega)$. Off half filling, they find a quite strong renormalization from the free dispersion $\sim \cos k$, especially close to the Fermi level. Here they observed a dynamically generated quasi-particle which is relatively dispersive in most parts of the Brillouin zone but rather flat in a region slightly above the X-point, in remarkable agreement with experimental data for the cuprates.

Additional agreement with cuprate experimental data is found in the infinite-dimensional Hubbard model results. As shown in Fig. 20 the resistivity found in the infinite-dimensional model is linear over a very wide temperature range. Also shown in this figure is a similar linear anomaly in the NMR rate. These anomalies are observed in the normal state of the cuprate superconductors [70],
Fig. 18. The evolution of the density of states with temperature when $U = 4$ and $\delta = 0.1878$. The development of a sharp peak at the Fermi surface is correlated with the reduction of the screened local moment $T\chi(T)$, as shown in the inset.

Fig. 19. The evolution of the density of states of the two-dimensional Hubbard model with temperature when $U = 8t$ and $\delta = 0.13$ (from [71]). Here $N(\omega)$ represents the density of states. suggesting that the cuprates and the Hubbard model share some common physics. It was suggested that the common physics involves Kondo-like screening of the local spins [67].

6.3. Holstein model

The Holstein Hamiltonian is a simple model of itinerant electrons interacting with local phonon degrees of freedom which provides a context for examining the competition of charge-density wave (CDW) and superconducting (SC) phases. The Hamiltonian is given by

$$H = -t \sum_{(ij)\sigma} (c_{i\sigma}^\dagger c_{j\sigma} + c_{j\sigma}^\dagger c_{i\sigma}) - \mu \sum_{i\sigma} n_{i\sigma}$$

$$+ g \sum_{i\sigma} n_{i\sigma} (a_i + a_i^\dagger) + \omega_0 \sum_i a_i^\dagger a_i. \quad \text{(6.7)}$$
Here $c_{i\sigma}^+$ is an operator which creates an electron of spin $\sigma$ on site $i$, $n_{i\sigma}$ is the fermion number operator, and $a_i^+$ creates a local phonon excitation at site $i$. $t$ is an electron transfer matrix element, $\omega_0$ is the phonon frequency, and $g$ is the electron-phonon coupling constant. Choosing $t = 1$ sets the scale of energy, and in the studies discussed below a square lattice this choice and periodic boundary conditions are also assumed.

The Holstein Hamiltonian has been examined by a variety of analytic and numerical approaches with focuses ranging from understanding normal state properties, like small and large polaron formation, to studying the possible highly correlated CDW and SC states. To study the formation of the CDW state, calculating the spectral density of electrons $A(p, \omega)$ is useful. For the noninteracting case ($g = 0$), an electron has energy $\omega = \epsilon(p)$, so the spectral density is a set of $\delta$-functions. For interacting systems, the spectral density typically has a broader, more incoherent distribution, and one is interested especially in identifying whether any sharp quasiparticle features remain at the Fermi surface. Indeed, if an insulating or superconducting gap is opened, this also will be reflected in the structure of $A(p, \omega)$.

Engelsberg and Schrieffer [72] calculated the spectral density of a model similar to the Holstein Hamiltonian using the Migdal approximation [73] which assumes a bare phonon propagator. Their results show a quasi-particle picture, namely a $\delta$-function peak at $\omega = 0$ and broadened distributions at $|\omega| \geq \omega_0$. Marsiglio [74] used a similar approach with higher order of approximations and found a rather different picture. The quasi-particle peak was washed away and the spectral density spread out for a large region in $\omega$. By dressing the phonon propagator, we obtained results similar to those of Marsiglio [74]: the sharp structures are smeared out; however, the spectral density does not have the expected gap at low temperatures. To recover these features, Niyaz et al. [75] needed to employ the gapped Migdal-Eliashberg equations. The important point is that numerical calculations show a transition to a CDW state; however, these numerical techniques only computed static properties. By
calculating the spectral function \( A(p, \omega) \), Niyaz et al. were able to study the formation of CDW gaps at half-filling.

The spectral density is related to the electron Green's function \( G(p, \tau) \) by

\[
G(p, \tau) = \int_{-\infty}^{\infty} d\omega \frac{e^{-\tau\omega} A(p, \omega)}{1 + e^{-\beta\omega}},
\]

where \( 0 < \tau < \beta \). The goal is to solve for \( A(p, \omega) \) from the QMC data \( \hat{G}(p, \tau) \) for \( G(p, \tau) \).

The formation of the CDW gap was found to depend on several parameters. Increasing lattice size required going to lower temperature to observe the gap opening. The observation confirmed earlier studies on the 2D Hubbard model that suggested the electron correlations lengths saturating on small lattices as temperature was lowered [761]. The formation of the CDW state also depends on the coupling constant \( g \). For larger interaction strength, we expect a higher transition temperature since the electrons tend to be more localized. In Fig. 21, we show the spectral density at the Fermi surface calculated by Niyaz et al. as a function of \( \omega \) for \( \omega_0 = 1 \) and \( g^2 = 1.5 \) for a \( 4 \times 4 \) lattice at \( \beta = 2, 5, 8 \) and 12. At \( \beta = 2 \), the central peak is already split, and at \( \beta = 5 \), a gap is about to open. At \( \beta = 8 \) and 12, we see fully opened gaps.

In Fig. 22, we also plot spectral densities at the Fermi surface calculated by Niyaz et al. versus frequency at various \( g^2 \) values while keeping the temperature and \( \omega_0 \) constants \( (\beta = 5 \) and \( \omega_0 = 1 \)). Their results show that at weak coupling \( (g^2 = 0.5) \), a quasiparticle central peak is present. But as
Fig. 22. The spectral density vs $\omega$ with $\beta = 5$, $\omega_0 = 1$, and $p = (0, \pi)$. (a) For a weak coupling $g^2 = 0.5$, a quasiparticle peak is present. (b) As the interaction strength is increased to $g^2 = 1.0$, the central peak present at weak coupling splits. (c) A gap is about to open at $g^2 = 1.5$. (d) At $g^2 = 2.0$, a large gap is present.

$g^2$ increases, the central peak splits and a gap opens for $g^2 > 1.5$. Again, we notice that the size of the CDW gap is approximately $2g^2/\omega_0$.

Finally, the effect of $\omega_0$ on the formation of the CDW state was also investigated. While keeping $2g^2/\omega_0$ constant and lowering temperature, the trend to a CDW state was the central peak splitting and separating. However, even at $\beta = 12$, a gap did not yet open. This closure indicated that the CDW state is suppressed as $\omega_0$ is increased while keeping $g^2/\omega_0$ constant.

By comparing spectral densities, Niyaz et al. were able to gauge the validity of the Eliashberg approximation. Their numerical results show that at high temperatures and weak couplings, the spectral density has a quasiparticle central peak and secondary structures at $\omega = \pm \omega_0$, in good agreement with the Migdal equations with a bare phonon propagator. On the other hand, at low temperature and strong couplings, QMC results clearly show the opening of a gap. The Migdal equation with dressed phonon propagator does show the diminishing of the quasiparticle central peak as the temperature is lowered, but it fails to show a gap. The gapped Migdal-Eliashberg equations are required to yield qualitatively similar result to the QMC simulations. Despite this qualitative agreement, these gapped equations give a higher transition temperature than that of the QMC simulation. The comparisons show the gapped Migdal-Eliashberg equations have a stronger CDW correlation than the QMC results.

One obvious source of the disagreement could be the use of the bare phonon propagator in the gapped equations. The origin of the difference might also be the neglect of the vertex corrections. In comparing the diagrammatic calculations with QMC results for two-dimensional repulsive $U$ Hubbard model, it was necessary to use an effective $\tilde{U} = 2$ in the diagrammatic calculation to match the
simulation results at \( U = 4 \). Thus, to get an agreement for the Holstein model, it was suggested that one might need a similar renormalization of the coupling constant. However, preliminary investigations suggested this is unlikely. The point is that the size of the gap at low \( T \) is in good agreement between the gapped Migdal-Eliashberg equations and the QMC when the same \( g \) is input to both. Thus, an attempt to fix up \( T_c \) as a simple change of \( g \) to \( \tilde{g} \) will produce trouble at \( T = 0 \).

An attractive feature of the Holstein model is simulations for cases other than the half-filled case can be performed without the infamous “fermion sign problem,” and besides determining the spectral density of the single-particle electron Green’s function associated, one can determine the spectral density associated with the phonon Green’s function as supporting information on what is happening. Indeed, such simulations have been done [77], and the development of gaps in the spectrum has been tracked in both of these cases. At half-filling, the phonon spectra shows considerable softening as the CDW gap opens. Away from half-filling, when a next-nearest neighbor hopping is added to the model to enhance pairing, it is the SC gap that apparently develops at a temperature close to the one at which the s-wave pairing susceptibility is rapidly growing. Both findings are in agreement with other QMC and analytic studies of the model.

6.4. Quantum spin models

The dynamics of quantum spin systems are particularly interesting. This interest has been generated in part by Haldane’s conjecture [78] for one-dimensional systems and by the discovery of the cuprate superconductors which, when undoped, are prototypical two-dimensional Heisenberg antiferromagnets.

6.4.1. One-dimensional Heisenberg model

The dynamics of an one-dimensional antiferromagnetic spin chain has received considerable attention since Haldane [78] made the novel prediction that the dynamic spin correlation function \( S(q, \omega) \) of half-integral spin chains was gapless, whereas for integral-spin chains it was gapped. In addition, several materials, such as \( \text{CuCl}_2\text{2N(C}_3\text{H}_5) \) (\( S = 1/2 \)) and \( \text{CsNiCl}_3 \) (\( S = 1 \)), are nearly one-dimensional antiferromagnets to which much experimental effort has been devoted. However, until recently, no reliable theoretical results for \( S(q, \omega) \) existed.

Deisz et al. [79] employed a combination of QMC and MaxEnt analytic continuation to make the first quantitatively correct calculations of \( S(q, \omega) \) for quantum Heisenberg spin chains. They found that for \( S = 1 \) and \( S = 2 \) a low frequency gap exists, whereas for \( S = 1/2 \) and \( S = 3/2 \) no gap is found. These results are summarized in Fig. 23 and are a direct confirmation of the Haldane conjecture [78]. In addition, these authors obtained an excellent fit within statistical and systematic error to experimental neutron scattering measurements of \( S(q, \omega) \) for \( \text{CuCl}_2\text{2N(C}_3\text{H}_5) \) [80].

6.4.2. Two-dimensional Heisenberg model

One of the most powerful applications of QMC is to act as a control for analytic approximations. For example, two relatively successful and stimulating analytic techniques have been developed to treat the 2D quantum Heisenberg model. First, Arovas and Auerbach [81] have used a Schwinger-Boson representation of the spin degrees of freedom. Second, Tyc et al. [82] found an approximate mapping between the quantum antiferromagnet and the rigid rotor model. The success of both of these models is somewhat surprising since these two analytic results are quite different.
M. Jarrell, J.E. Gubernatis/Physics Reports 269 (1996) 133–195

Fig. 23. Spin structure factor \( S(q = \pi, \omega) \) for 64-site antiferromagnetic spin chains of various spin magnitudes. The reduction in \( \omega = 0 \) intensity seen in \( S(q, \omega) \) for \( S = 1/2 \) and \( 3/2 \) is due to finite-size effects. The data have been scaled to fit on the same plot.

Makivic et al. [83] compared the predictions for the neutron spin structure factor \( S(k, \omega) \) of these two theories, with results from a massively parallel QMC simulation in what is essentially the thermodynamic limit (64 by 64 spins). Here, the classic MaxEnt procedure was used to analytically continue the spin structure factor. They found that the model proposed by Tyc et al. was qualitatively consistent with the QMC results. Specifically, it was verified that the dynamic structure factor near \( q = k - (\pi, \pi) = 0 \) can be cast into the scaling form:

\[
S(q, \omega) (1 - \exp(-\beta \omega)) / \omega = \omega_0^{-1} S(q = 0) f(q\xi) \Phi(q\xi, \omega / \omega_0),
\]

with a temperature dependent correlation length \( \xi(T) \) and frequency \( \omega_0(T) = \gamma_{q=0} \). The dynamic scaling hypothesis also implies scaling for long wavelength spin wave frequencies \( \omega_q = \omega_0(T) \Omega(q\xi) \) and lifetimes \( \gamma_q = \omega_0(T) \Gamma(q\xi) \). Fig. 24(a) shows the scaling plot of \( \omega_q \) and \( \gamma_q \) (b). The solid lines are plots of the scaling functions of Tyc et al., with scaling parameters fit to their data.

The results of Makivic et al. are also in excellent agreement with experiment. In experiment, the spin wave spectrum \( \omega_q \) may be measured using time-of-flight neutron spectroscopy. For example, the recent neutron scattering results for \( \text{La}_2\text{CuO}_4 \) of Hayden et al. [84] are compared to Makivic et al.'s results in Fig. 24(c). \( \text{La}_2\text{CuO}_4 \) is thought to be an excellent realization of a two-dimensional spin-1/2 quantum antiferromagnet. The only parameter which had to be adjusted to fit the two data sets was the near-neighbor exchange \( J \).

6.4.3. Two-dimensional XY model

The XY model is believed to describe the physics of granular superconductors and Josephson junction arrays in which the characteristic size of the superconducting grains is small, as well as the physics of two-dimensional spin systems with an easy-plane exchange anisotropy. Very recently, Kawashima and Gubernatis [86] developed a new class of algorithms which greatly improve the efficiency of the simulations of such models. The algorithm used for the XY model is related to the recently proposed loop algorithms [87]; however, it differs in that an in-plane quantization axis is used.
Fig. 24. (a) The scaling of spin wave frequencies $\omega_q = \omega_0 \Omega(q\xi)$, (b) the scaling of spin-wave linewidths $\gamma_q = \omega_0 \Omega'(q\xi)$ for the two-dimensional spin-1/2 quantum Heisenberg model. The solid lines correspond to the scaling functions of Ref. [82]. (c) Spin-wave frequencies and linewidths at $T = 0.35J$. The solid line is the $T = 0$ spin-wave dispersion, uniformly renormalized by the quantum factor of $Z_c = 1.18$ [83]. The integers on the abscissa enumerate points on the Brillouin zone of a $64 \times 64$ system, along the triangle bounded by $k_1 = (0, 0)$, $k_2 = (\pi, \pi)$, and $k_3 = (\pi, 0)$. The inelastic neutron scattering data [84] (Exp) are scaled by the bare effective exchange coupling $J \approx 0.13$ eV.

Fig. 25. $S(k = (\pi/4, \pi/4), \omega)$ when $T = 0.125$ and $J = 1$ for a $32 \times 32$ system size. The low-frequency peak is from single-magnon processes whereas the upper peak is probably due to multi-magnon processes.

Kawashima et al. [88] use the imaginary-time output of this code and analytically continue for $S(k, \omega)$ using the Bryan algorithm. Below the Kosterlitz-Thouless transition temperature, the quantum XY model displays only spin excitations in the structure factor $S(k, \omega)$ as shown in Fig. 25. The lower peak is due to the creation of single-magnon spin excitations; whereas, the small upper peak is probably due to multi-magnon processes. The fact that multi-magnon processes seem to be resolved here, but not in the two-dimensional Heisenberg calculation, is probably due to the improved accuracy of the new algorithm used to generate the data.

Kawashima et al.'s calculations are the first numerically exact results for the dynamics of the quantum XY model and may be used to test analytic approximations. For example, the dispersion...
Fig. 26. Comparison of the spin-wave dispersion as calculated from QMC results (circles) and spin wave theory (solid line) \[89\] when \( T = 0.125 \).

(the location of the lower peak) is in rough agreement with spin-wave theory \[89\], provided that the spin stiffness (the initial slope of the the plot) is renormalized by about 5\%. Thus, an estimate of the spin-stiffness renomalization, due to quantum fluctuations, is 1.05.

6.5. Other applications

6.5.1. Ground state energies

With ground state (projector) methods, quantum Monte Carlo can estimate correlation functions of the form

\[
h(\tau) = \langle \Psi_T | e^{-\tau H} | \Psi_T \rangle
\]

(6.10)

where \( \Psi_T \) is a trial wavefunction. This correlation function can be re-expressed as the Laplace transform of pseudo-spectral-function \( c(E) \)

\[
h(\tau) = \int_{-\infty}^{\infty} dE \, c(E) e^{-\tau E}
\]

(6.11)

that is defined in terms of the overlap integral of \( |\Psi_T\rangle \) and the exact eigenfunctions \( |\Phi_i\rangle \) and eigenvalues \( E_i \) of \( H \)

\[
c(E) = \sum_i \delta(E - E_i) |\langle \Psi_T | \Phi_i \rangle|^2.
\]

(6.12)

If \( E(\tau) \) is defined by

\[
E(\tau) = -\frac{d \ln h(\tau)}{d\tau}
\]

(6.13)

then \( E_0 \) is

\[
E_0 = \lim_{\tau \to \infty} E(\tau).
\]

(6.14)

Typically, for fermion simulations, the signal-to-noise ratio for this estimate increases exponentially at large \( \tau \) because of the "sign problem." Caffarel and Ceperley \[90\] used the method of maximum
entropy to circumvent this problem and to accelerate the convergence of quantum Monte Carlo simulations to accurate estimates of $E_0$ and $|\Phi_0\rangle$. Their objective was to use a much as possible information in the correlation function is a way consistent with probability theory to exploit fully the information contained at small $\tau$ to avoid the large statistical errors at large $\tau$.

In their approach, besides quantum Monte Carlo estimates of just $h(\tau)$, they also used estimates of the first and second moments of $c(E)$. Because their test problem was the computation of the ground-state of a LiH molecule, which has a relatively large gap to the first excited state, they found it advantageous to treat the ground state energy $E_0$ as a variational parameter on equal footing with the $c(E)$. Using historic MaxEnt, they demonstrated an accelerated convergence to the ground-state, a significant advantage in using at least the first moment of $c(E)$, and a reduction in signal-to-noise ratio of the result as $\tau$ increases, i.e., they were able to obtain tighter error bounds than they by using the standard extrapolation methods.

6.5.2. Single-electron solvation

The physical problem is the venerable electron in a liquid helium bubble. The electron motion and thermal fluctuations cause the bubble to behave as an oscillating dipole, and the problem is to calculate the absorption spectra associated with the system. Using path-integral Monte Carlo methods, Gallicchio and Berne [91] computed the mean-squared electron displacement as a function of imaginary-time which is related in standard linear fashion to the absorption spectrum. They regularized the solution of this problem by means of a MaxEnt-based method. Their procedures, methods, and philosophy are different from those presented here. While their results also show small spurious features, overall they seem consistent with analytic results. Namely, their results exhibit the expected low frequency peak and long-tailed high frequency behavior.

7. Concluding remarks

We have attempted to summarize the way in which we use Bayesian statistical inference and the principle of Maximum Entropy to analytically continue imaginary-time quantum Monte Carlo data. We tried to supply the details that are lacking in the seminal literature but are important for the motivated reader to understand the assumptions and approximations embodied in these methods. There are several points worth re-emphasizing. First, the Bayesian and the MaxEnt methods we described did not originate with us. Second, little in the specifics of the methods are tied to the analytic continuation problem. This problem is terribly ill-posed, and the described methods are only one possible approach to solve such problems. The construction of such methods is an active area of research in applied mathematics and the physical sciences. In these fields, different philosophies exist, and often bitter disagreements exist between different schools of thought. In essence, we adopted one school of thought and were fortunate to discover what it took to make the approach work on the specific class of problems in which we were interested.

Data analysis techniques and procedures, similar in purpose to the ones we described, are essential to the proper application of the method, and this report documents as fully as we can what we actually do to perform the results of the analytic continuation we publish. Many other researchers are finding our basic approach useful; however, at times we develop uneasy feelings with the absence of published details about the procedures and techniques others use. MaxEnt methods are being
used, but it is unlikely the details are the same as those we described. From our own experience, the temptation to take short-cuts is great. Instances exist where the problem seems under sufficient control so "short-cuts" appear warranted. Repeatedly, however, when we do this, cases will arise whose results do not seem right. Then, by exercising proper care, we eventually find the data quality is poor and are faced with improving it to improve our confidence in our results. That is to say, we have to revert to "standard procedures." Short-cuts are dangerous.

An advantage of the described approach is the laying out of the points at which assumptions and approximations are made and points at which prior information can be inserted. The choice of the Gaussian likelihood function is an assumption, and because the data is often naturally inconsistent with this assumption, using it is one of the approximations. The choice of the entropic prior is both an assumption and an opportunity to incorporate of prior information through the default model. The main prior information captured is positivity and normalizibility. The assumption is the reasonableness of entropy function. The kangaroo argument seems to illustrate why one might want to use the entropic prior. In the presence of data, does the argument remain as compelling? We doubt if anyone knows for certain.

For classic MaxEnt, we discussed how the assumption that $\Pr[\alpha|\tilde{G}]$ was sharply peaked was inappropriate for the analytic continuation problem. This recognition lead us to a different strategy based on a method due to Bryan [26]. His strategy assumes that $\Pr[A|\tilde{G}, \alpha]$ is sharply peak. We have never tested this assumption. Most likely this condition is only approximately true.

The detailed analysis of the MaxEnt methods is directed at fixing the Lagrange multiplier (i.e., $\alpha$) that is present if one were simply constraining the likelihood function by the entropy function. All attempts to regularize an ill-posed problem by constraints are faced with a similar problem. The Bayesian approach to the choice of $\alpha$, i.e., fixing the Lagrange multiplier, often called a hyperparameter, is quite general and in many respects not specific to the choice of entropy as the regularizing function. Choosing the entropy has the distinct advantage of insuring positivity.

Recently, our colleague, Richard Silver [92], has proposed another Bayesian approach to inference problems that he has labeled quantum statistical inference (QSI). The MaxEnt approach makes no assumption on the continuity of derivatives in the result. Silver argues that in many cases one knows a priori the result is smooth so one should step beyond MaxEnt and build that information into the regularization procedure. Regularizing by a smoothness constraint is not new but using Bayesian methods to fix the Lagrange multiplier and a Legendre transformation to a different solution space are. His approach awaits tests on problems as difficult as the analytic continuation problem. The approach is interesting and might be a step forward.

Our techniques for data analysis are not absolute. We would gladly adopt better methods, but some set of methods to test for statistical independence and for Gaussian behavior are essential. Sometimes this quality in the data might be impossible to produce because of the quantum Monte Carlo algorithm, the physical model, and the region of parameter space being simulated. On the other hand, the data analysis is not the bottleneck.

The bottleneck issue is the proper computation of the covariance matrix. Computing it is required by the particular assumption made for the likelihood function. The data is made to approximate this assumption by reliance on the law of large numbers to approach the proper form. The approach scales roughly only as the square root of the number of independent Monte Carlo measurements, and often quite a large number of Monte Carlo steps are needed to produce an independent measurement. Once independence is under control a large number (500 to 1000) of measurements might be needed to
achieve a good approximation and sufficiently small error to obtain accurate results.

As mentioned earlier, setting the off-diagonal elements of the covariance matrix to zero, which is a polite way saying that one is ignoring them, it a tempting but potentially dangerous choice. Muramatsu [93] has argued that cases may exist in which the off-diagonal elements are small and comparable to the numerical error of their calculation. Hence, by not setting them to zero, noise is built into the calculation and can corrupt the results. His results seem to support this observation. In a different context, we recently had a similar experience with small off-diagonal elements approaching the likely numerical error, but setting them to zero produced disastrous results. Clearly, more is need to be learned on this issue, which is an important one. Another problem that can arise with the covariance matrix is a number of rows (or columns) becoming numerically identical and making the matrix highly singular [94]. Experience here is again mixed. In our case, the difficulty was traced to poor data. Both sets of difficulties point to the need to estimate the error on the covariance matrix. We have only explored this superficially.

Another problem that can develop is the parameter $\alpha$ becoming too small in either the historic or classic MaxEnt solutions. If this occurs, problems with Bryan's method will also exist. When $\alpha$ is small, e.g. less than 1, the various Gaussian approximations (which are one way of performing a low order steepest descent approximation) become invalid, and typically the results have unexplainable features. A small $\alpha$ promotes a solution fitted primarily to the data. One purpose of all these methods is to avoid doing this. The possibility of various Gaussian approximations being inaccurate is the apparent source of Skilling recent abandoning of MaxEnt [27]. The value of the normalization constant $Z_0$ for the entropic prior seems to be the biggest problem. While the entropy's definition depends on the default model $m$, within the steepest descent approximation, the normalization constant does not. It is unlikely that simply adding the next order correction, or numerically performing the integral (with or without the entropic metric), will placate the concerns. When we have encountered small $\alpha$ situations, they were accompanied by poor data quality. Again, we cannot say that this situation will always be the case. Clearly, the asymptotic nature of the theory with respect to the size of $\alpha$ is another factor to bear in mind. When $\alpha$ is not small, the solutions appear valid in all cases we can benchmark.

Repeatedly, we pointed to the need for good data. Without it, the methods will not work. In Monte Carlo simulations, the law of large numbers usually is adequate to insure that this goal will eventually be reached. Often quite a price for computer time is required. Some quantum Monte Carlo algorithms perform better than others. The standard worldline quantum Monte Carlo method should be avoided [95].

Acknowledgments

Most of this article was written while the authors were visiting the Institute for Theoretical Physics in Santa Barbara, California. They thank the Institute for its hospitality. They also thank R.N. Silver for numerous useful, helpful, and lively discussions. The work of J.E.G. was supported by the US Department of Energy. M.J. thanks H. Akhlaghpour, O. Biham, N. Bulut, W. Chung, D.L. Cox, J. Deisz, M. Makivic, D. Sivia, M. Vekic, and J.W. Wilkins for helpful discussions. His work was supported by the Office of Navel Research grant N00014-92-J-1686, NSF grants DMR-9406678 and DMR-9357199, and the Ohio Supercomputer Center.
Appendix A. The Bryan algorithm

In the interest of completeness, we will now sketch Bryan's numerical algorithm to find the maximum of $Q$ for the different values of $\alpha$ needed to evaluate the MaxEnt solutions. For a more detailed description, we refer the reader to his paper [26]. 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 many more "good" data values than we have "good" parameters with which specify the solution. Thus, the problem is usually "oversampled." Bryan tailored his numerical algorithm [26] 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 $A_i$ determined, 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 a wide range of reasonably acceptable values of $\alpha$. Bryan deals with this by calculating the result which averaged over $\Pr[\alpha|G,m]$, the posterior probability of $\alpha$.

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 $A$

$$A \rightarrow A + \delta A \quad (A.1)$$

with

$$\delta A = -(\nabla \nabla Q)^{-1} \nabla Q. \quad (A.2)$$

The Hessian $(\nabla \nabla Q)^{-1}$ of this search may conceptually be expanded using the binomial theorem

$$(\nabla \nabla Q)^{-1} = (\alpha \nabla S - \nabla L)^{-1} = (\alpha \{A\}^{-1} - \nabla L)^{-1} \quad (A.3)$$

where $\{A\}$ is a diagonal matrix with the elements of $A$ along its diagonal. $(\nabla \nabla Q)^{-1}$ may be written as a power series in $\{A\} \nabla L$. Thus, $\delta A$ may be written as a combination of $\{A\} \nabla = \{A\} (\alpha \nabla S - \nabla L)$, and powers of $\{A\} \nabla L$ acting on $\{A\} \nabla S$ and $\{A\} \nabla L$. Each of these vectors defines a direction in which the search can proceed. 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. 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

$$\nabla L = \frac{\partial F}{\partial A} \frac{\partial L}{\partial F} = K^T \frac{\partial L}{\partial F} \quad (A.4)$$

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 \( \{ A \} U_i \), where \( U_i \) 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 \( Q \),

\[
\nabla Q = \alpha \nabla S - \nabla L = 0 \rightarrow -\alpha \ln(A_i/m_i) = \sum_j K_{ji} \frac{\partial L}{\partial F_j}.
\]

(A.5)

Thus, the solution may be represented in terms of a vector \( u \)

\[
\ln(A/m) = K^T u.
\]

(A.6)

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

\[
A_i = m_i \exp \sum_n U_{in} u_n.
\]

(A.7)

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

\[
-\alpha Uu = U \Sigma V^T \frac{\partial L}{\partial F},
\]

(A.8)
or

\[
-\alpha u = \Sigma V^T \frac{\partial L}{\partial F} \equiv g,
\]

(A.9)

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

\[
J \delta u = -\alpha u - g,
\]

(A.10)

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

\[
\frac{\partial g}{\partial u} = \Sigma V^T \frac{\partial^2 L}{\partial F^2} \frac{\partial F}{\partial A} \frac{\partial A}{\partial u}.
\]

(A.11)
With the definition \( W = \partial^2 L / \partial F^2 \) (which is just the diagonal matrix with elements \( 1/\sigma_i^2 \)), \( M = \Sigma^{1/2} W \Sigma \), and \( T = U^T A U \). \( M \) and \( T \) are symmetric \( s \times s \) matrices, the Jacobian \( J = \alpha I + MT \), and

\[
(\alpha I + MT) \delta u = -\alpha u - g. \tag{A.12}
\]

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

\[
[(\alpha + \mu) I + MT] \delta u = -\alpha u - g \tag{A.13}
\]

and adjusted to keep \( \delta u \) below some designated limit so that the search is within the range of validity of a local quadratic expansion of \( Q \).

The values of \( \mu \) and \( \alpha \) are adjusted so that the solution iterates to either a fixed value of \( \chi^2 = 2L \) (for historic MaxEnt) or to a maximum value of \( \Pr[\alpha | G, m] \) (for classic MaxEnt). Then, \( A \) is obtained from

\[
A_i = m_i \exp \left( \sum_{n=1}^{s} U_{in} u_n \right). \tag{A.14}
\]

Alternatively, Bryan suggests that one may start the algorithm with a large \( \alpha \) for which \( \Pr[\alpha | G, m] \) is negligibly small, and then iterate to \( \alpha \approx 0 \) so that the averaged image may be approximated

\[
\langle A \rangle = \int_0^\infty \text{d} \alpha \Pr[\alpha | G, m] A(\alpha) / \int_0^\infty \text{d} \alpha \Pr[\alpha | G, m] \tag{A.15}
\]

where \( A(\alpha) \) is the optimal image (that for which \( \nabla Q = 0 \)) for the value of \( \alpha \) specified in the argument. This latter step may be necessary when \( \Pr[\alpha | G, m] \) is not a sharply peaked distribution. In fact this is usually the case, as may be seen in Fig. 27 where \( \Pr[\alpha | G, m] \) is plotted for the data shown in Fig. 9.

These three MaxEnt methods will produce essentially identical results if the data is good. (Historic MaxEnt, which iterates to a fixed value of \( \chi^2 \), can be made significantly different, of course, depending upon the value of \( \chi^2 \) chosen.) However, when the data is of marginal quality, the method suggested by Bryan, averaging \( A(\alpha) \) weighted by \( \Pr[\alpha | G, m] \) generally produces more acceptable results. A further and much more significant advantage of the averaging is that it allows an accurate relative assessment of the posterior probability (\( \int_0^\infty \text{d} \alpha \Pr[\alpha, G | m] \)) of the averaged model. As discussed in Section 5, this information is invaluable in determining which default model yields the most likely \( A \).

We close with several practical comments: On a workstation, finding the global maximum by searching in the singular space typically takes a few minutes or less of computer time. This efficiency is in sharp contrast with the amount of computer needed even on a “supercomputer” for standard Newton algorithms 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 [96] singular-value decomposition routine to do the SVD and also to compute any eigenvalues and eigenvectors we need. The SVD routine in Numerical Recipes [97] and the EISPACK [98] eigenvalue-eigenvector routine RS are not as stable. Additionally, the Numerical Recipes routines for solving the MaxEnt equations are less effective than the approach just described.
Fig. 27. The posterior probability $\Pr[\alpha|G, m]$ as a function of $\alpha$ for the periodic Anderson model density of states data presented in Fig. 9. Since $\Pr[G]$ 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.

References


[70] For reviews of relevant experiments involving the Cuprate superconductors, see C.H. Pennington and C.P. Slichter, Physical Properties of High Temperature Superconductors, edited by D.M. Ginsburg Vol.2; N.P. Ong, ibid.; Y. Iye, ibid. Vol. 3.


Note added in proof

Since we submitted our manuscript, several papers that use an analytic continuation method in conjunction with quantum Monte Carlo simulations have been published or have been accepted for publication. Hopefully, the list below is complete. We apologize for any omissions. We have included the titles of the papers for annotation purposes.