PHYSICAL REVIEW B VOLUME 44, NUMBER 12 15 SEPTEMBER 1991-II

Quantum Monte Carlo simulations and maximum entropy: Dynamics from imaginary-time data

J. E. Gubernatis
Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

Mark Jarrell
Department of Physics, University of Cincinnati, Cincinnati, Ohio 45221

R. N. Silver and D. S. Sivia
Theoretical Division and Manuel Lujan Jr. Neutron Scattering Center, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

(Received 25 February 1991)

We report the details of an application of the method of maximum entropy to the extraction of spectral and transport properties from the imaginary-time correlation functions generated from quantum Monte Carlo simulations of the nondegenerate, symmetric, single-impurity Anderson model. We find that these physical properties are approximately universal functions of temperature and frequency when these parameters are scaled by the Kondo temperature. We also found that important details for successful extractions included the generation of statistically independent, Gaussian-distributed data, and a good choice of a default model to represent the state of our prior knowledge about the result in the absence of data. We suggest that our techniques are not restricted to the Hamiltonian and quantum Monte Carlo algorithm used here, but that maximum entropy and these techniques lay the general groundwork for the extraction of dynamical information from imaginary-time data generated by other quantum Monte Carlo simulations.

I. INTRODUCTION

Using the method of maximum entropy,[1] we have extracted, with no adjustable parameters, spectral and transport properties from the imaginary-time correlation function data generated by quantum Monte Carlo (QMC) simulations of the nondegenerate, symmetric, single-impurity Anderson model.[2–4] The physical quantities obtained are found to be universal functions when the frequency and temperature are scaled by the Kondo temperature. This universality is a striking feature, not found in perturbation theory, but found in experiment, and provides the benchmark for our claim of successful extractions. In this paper, we present the details of our use of maximum entropy with QMC simulations. In a subsequent paper, we will discuss the physics obtained from our results.[5]

The inability to extract dynamical information from imaginary-time quantum Monte Carlo data has long been a factor limiting the usefulness of quantum simulations. Several attempts are published, and these attempts fall into two broad classes. In one class are methods that modify the Monte Carlo procedure;[6] in the other class are ones that use existing procedures and attempt to extract the information from the resulting data.[7–10] 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,[8] White *et al.*,[9] and Jarrell and Biham.[10]

Schüttler and Scalapino first proposed the least-squares approach and identified 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 requirement of positivity.[8–10] 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 maximum-entropy approach has different qualities as it explicitly approaches statistical data analysis within the concepts of conditional probabilities (Bayesian logic).[11] 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] What is unique about the maximum-entropy approach is the specification of four axioms that are used with Bayes's theorem to specify uniquely 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, hence the name of the method. 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 the recent form of the method, *classic maximum entropy,*[12] as distinct from an older form, *historic maximum entropy,* 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 data, variables, and the solution. Maximum entropy specifies the prior probability function of the solution and does so in such a way as to prohibit 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. Additionally, their solutions can have unpredictable, uncontrolled correlations that can distort the result from the true solution.

Since entropy is a relative function, maximum entropy contains the choice of a "model" to set the zero and the maximum of the entropy. More importantly, the model is the solution produced in the absence of data or of 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 Matsubara frequency. Since quantum simulation data lack very-high-frequency information, a model that incorporates proper high-frequency behavior is useful. In our calculations for the Anderson impurity model, we mainly used perturbation theory to provide a model. At high frequencies, our results are biased towards the model, which is becoming exact, but at low frequencies they exhibited important universality not present in the model, but present in the physics.

We feel the model allows perturbation theory and quantum Monte Carlo to be combined in a novel way; however, the model can be chosen in a variety of other ways. For example, if several moments of the spectral density are known, then maximum entropy can be used to determine the most likely model based on this information. In a practical sense, the model is a convenient means to incorporate prior information about the solution into the problem. The use of prior information can be essential to solving ill-posed problems satisfactorily.

The maximum-entropy approach has successfully been used to extract dynamical information from various simulations of several different Hamiltonians calculated by several different quantum Monte Carlo methods.[2–4,13–16] Here, in a self-contained manner, we will present and discuss general considerations necessary to make the method work. Additional information may be found in our preliminary study of the feasibility of the approach.[1] Simply applying the method in a "black-box" manner is insufficient to obtain proper results. For concreteness, we present detailed considerations in the context of our previous work on extracting spectral and transport properties from simulations of the single impurity Anderson model.[2–4]

In Sec. II and III, we summarize general relations between quantum correlation functions and spectral densities, and general features of maximum entropy. These sections are intended to be relatively self-contained, summarizing information spread across various sources. In Sec. IV, we describe our test problem, the nondegenerate, symmetric, single-impurity Anderson model, and the quantum Monte Carlo algorithm used to compute its properties. We chose this model 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. Then, in Sec. V, we highlight important technical details about the application of maximum entropy to this Hamiltonian, describing measures taken to generate statistically independent, Gaussian-distributed data. An important feature is the computation of the covariance among the thermodynamic Green's-function values at different imaginary times. These correlations have been ignored in all previous work, but we find them to be essential for obtaining the proper solution. We also discuss issues helpful for judging the validity of the solution. In Sec. VI, we present a detailed study of the application of maximum entropy, illustrating several of the points in Sec. V and focusing on the differences between our spectral densities and those of the model. The principal difference is universality. Complete descriptions of the results and their physics will be given elsewhere.[5] In addition, we discuss a compelling feature of maximum entropy, its ability to estimate whether the error in the data is properly estimated. In cases where we achieved our best results, our data was the most consistent with this inferred estimate. Finally, in Sec. VII, we describe other features concerning the application of our methods, their possible improvement, and areas for additional study.

II. DYNAMICAL CORRELATION FUNCTIONS

We briefly summarize properties of various time-dependent correlation functions and their relation to spectral densities, drawing together bits and pieces from various sources.[8,17–19] We consider a system described by a Hamiltonian H and examine the time-dependent correlation between two operators A and B. Quantum Monte Carlo allows us to calculate the imaginary-time correlation function

$$G_{AB}(\tau)=\langle A(-i\tau)B\rangle = \frac{1}{Z}\mathrm{Tr}[e^{-\beta H}A(-i\tau)B] , \quad (1)$$

where

$$A(t)=e^{itH}Ae^{-itH} ,$$

and

$$Z=\mathrm{Tr}e^{-\beta H}$$

is the partition function at temperature $k_BT=1/\beta$ and $0\le\tau\le\beta$.

Experimental methods allow us to measure the response of the system in terms of real-time correlation functions of the form

$$S_{AB}(t)=\langle A(t)B\rangle \quad (2)$$

for the correlation function and

$$\chi_{AB}(t)=i\langle[A(t),B]_-\rangle \tag{3}$$

for the linear response function. The scattering function is the Fourier transform of $S_{AB}(t)$

$$S_{AB}(\omega)=\int_{-\infty}^{+\infty}dt\, e^{i\omega t}S_{AB}(t)\ , \tag{4}$$

with $S_{AB}(-\omega)=\exp^{-\beta\omega}S_{AB}(\omega)$. Other quantities, like the frequency-dependent conductivity and magnetic susceptibility, are given by the *retarded* part ($z=\omega+i0^+$) of the two-sided Laplace transform of $\chi_{AB}(t)$

$$\chi_{AB}(z)=\sigma\int_{-\infty}^{+\infty}dt\,\theta(\sigma t)e^{izt}\chi_{AB}(t)\ , \tag{5}$$

where z is a complex frequency with $\mathrm{Im}z\neq 0$, $\sigma=\mathrm{sgn}(\mathrm{Im}z)$ and

$$\theta(s)=\begin{cases}1, & s>0\\ 0, & s<0\ .\end{cases}$$

Another correlation function (Green's function)

$$\phi_{AB}(t)=i\langle[A(t),B]_\pm\rangle \tag{6}$$

with its corresponding two-sided Laplace transform

$$\phi_{AB}(z)=\sigma\int_{-\infty}^{+\infty}dt\,\theta(\sigma t)e^{izt}\phi_{AB}(t) \tag{7}$$

is useful (for fermionic operators). In this definition, the choice of commutator depends of whether A and B anticommute (upper sign) or commute (lower sign). ϕ_{AB} is not directly related to G_{AB}, but both S_{AB} and χ_{AB} can be expressed in terms of ϕ_{AB}.

These connections are most easily expressed in terms of various spectral functions

$$\chi''_{AB}(\omega)=\frac{1}{2i}[\chi_{AB}(z=\omega+i0^+)-\chi_{AB}(z=\omega-i0^+)]$$
$$=\frac{1}{2i}\int_{-\infty}^{+\infty}dt\,e^{i\omega t}\chi_{AB}(t) \tag{8}$$

and

$$\phi''_{AB}(\omega)=\frac{1}{2i}[\phi_{AB}(z=\omega+i0^+)-\phi_{AB}(z=\omega-i0^+)]$$
$$=\frac{1}{2i}\int_{-\infty}^{+\infty}dt\,e^{i\omega t}\phi_{AB}(t)\ . \tag{9}$$

Although these functions are defined for real frequencies ω, they contain complete information about the full complex frequency functions since

$$\chi_{AB}(z)=\int_{-\infty}^{+\infty}\frac{d\omega}{\pi}\frac{\chi''_{AB}(\omega)}{z-\omega}\ , \tag{10}$$

$$\phi_{AB}(z)=\int_{-\infty}^{+\infty}\frac{d\omega}{\pi}\frac{\phi''_{AB}(\omega)}{z-\omega}\ . \tag{11}$$

The connection between S_{AB} and χ_{AB} is the fluctuation-dissipation theorem

$$2\chi''_{AB}(\omega)=(1-e^{-\beta\omega})S_{AB}(\omega)\ , \tag{12}$$

while the connection between $S_{AB}(\omega)$ and $\phi_{AB}(\omega)$ is

$$S_{AB}(\omega)=2\phi''_{AB}(\omega)/(1\pm\exp^{-\beta\omega})\ . \tag{13}$$

The correlation function $G_{AB}(\tau)$ is periodic (bosonic) or antiperiodic (fermionic) in τ with period β. From (1) and (7), it follows that

$$G_{AB}(\tau)=\frac{1}{\beta}\sum_{i\omega_n}e^{-i\omega_n}\phi_{AB}(z=i\omega_n\tau),\quad 0\le\tau\le\beta \tag{14}$$

where the ω_n are the Matsubara frequencies

$$\omega_n=\frac{\pi}{\beta}(2n+1),\quad n=0,\pm1,\pm2,\ldots$$

for fermion operators or

$$\omega_n=\frac{\pi}{\beta}2n,\quad n=0,\pm1,\pm2,\ldots$$

for boson operators. With (11) and the completion of the Fourier sums,

$$G_{AB}(\tau)=\int_{-\infty}^{+\infty}\frac{d\omega}{2\pi i}\frac{e^{-\tau\omega}}{1\pm e^{-\beta\omega}}[\phi_{AB}(z=\omega+i0^+)$$
$$-\phi_{AB}(z=\omega-i0^+)]$$

or with the use of (9)

$$G_{AB}(\tau)=\int_{-\infty}^{+\infty}\frac{d\omega}{\pi}\frac{e^{-\tau\omega}}{1\pm e^{-\beta\omega}}\phi''_{AB}(\omega)\ . \tag{15}$$

Thus the problem of extracting dynamical information from the imaginary-time correlation function is reduced to solving this integral equation. With this solution $S_{AB}(\omega)$ and $\chi''_{AB}(\omega)$ are found with the use of (12) and (13).

When $B=A^\dagger$,

$$\phi''_{AA^\dagger}(\omega)\ge 0,\quad S''_{AA^\dagger}(\omega)\ge 0,\quad \omega\chi''_{AA^\dagger}(\omega)\ge 0\ . \tag{16}$$

Hence, so is $G_{AA^\dagger}(\tau)$.

To be more concrete, we will choose $A=d_\sigma$, where d_σ is the operator that removes an electron from state d with spin σ, and define $\phi''_\sigma(\omega)$ as the corresponding spectral density. This quantity represents the probability of finding an electron with spin σ and energy ω in the state d, and $G_{AB}(\tau)$ is simply the single-particle Green's function associated with that state, $G_\sigma(\tau)=\langle d_\sigma(\tau)d_\sigma^\dagger\rangle$. Defining $A(\omega)=\sum_\sigma\phi_\sigma(\omega)/\pi$ and $G(\tau)=\sum_\sigma G_\sigma(\tau)$, then using (15), we can write

$$G(\tau)=\int_{-\infty}^{+\infty}d\omega\frac{e^{-\tau\omega}}{1+e^{-\beta\omega}}A(\omega)\ . \tag{17}$$

The solution of this equation, when the Hamiltonian H is that of the single-impurity Anderson model, will be the principal subject of this paper.

Another interesting case arises when $A=S^-$, where S^- is the spin lowering operator for some state. In this case,

$$\langle S^-(\tau)S^+(0)\rangle=\int_{-\infty}^{+\infty}\frac{d\omega}{\pi}\frac{\phi''_{-+}(\omega)e^{-\tau\omega}}{1-e^{-\beta\omega}}\ ,$$

where $\phi''_{-+}(\omega)$ is the associated spectral density. Such spin correlation functions are generally repressed in terms of the imaginary part of the dynamical susceptibili-

ty. Using (12) and (13), we can rewrite the above as

$$\langle S^-(\tau)S^+(0)\rangle = \int_{-\infty}^{+\infty} \frac{d\omega}{\pi} \frac{\chi''(\omega)e^{-\tau\omega}}{1-e^{-\beta\omega}} , \quad (18)$$

where $\chi''_{-+}(\omega)$ is the transverse magnetic susceptibility. For the Anderson model, we will use $S^-(\tau) = d^\dagger_\downarrow(\tau)d_\uparrow(\tau)$. The solution of this equation will be discussed elsewhere.[4]

III. MAXIMUM ENTROPY

For the analytic continuation problem of this paper, we are primarily concerned with inferring the spectral density function $A(\omega)$ from the imaginary-time Green's-function data $\bar{G}(\tau)$, which result from quantum Monte Carlo simulations. Given the data, what is our best estimate of $A(\omega)$ and how confident are we in our predictions? The answer to this question is not clear-cut since it depends on both the data and our prior knowledge about $A(\omega)$. For example, if physics told us that the spectral density function must have a particular functional form, then we need to consider only a very limited set of possibilities defined by a handful of parameters. Alternatively, if we did not have a good *a priori* reason to assume a functional form, then we must consider a much larger set of possibilities. Even without a functional form, we may know about the positivity of the spectral density function, its zeroth moment (normalization), or an asymptotic solution, and so on, which will restrict the set of allowed possibilities for $A(\omega)$. How, then, should we combine our prior knowledge with the evidence of the data to obtain our best estimate of $A(\omega)$ and a measure of its reliability?[20]

Cox[21] has shown that any method of inference that satisfies simple rules for logical and consistent reasoning must be equivalent to the use of ordinary probability theory. Accordingly, the conditional probability distribution function (PDF) $P[A|\bar{G},I]$ summarizes our interference about the spectral density function given the data $\bar{G}(\tau)$ and relevant background information I such as prior knowledge about $A(\omega)$. Since the numerical value of the probability assigned to a particular $A(\omega)$ is a measure of how much we believe that it is the true spectral density function, our best estimate is given by that $A(\omega)$ which maximizes $P[A|\bar{G},I]$. The width of the probability distribution function tells us the reliability of the estimate.

To compute $P[A|\bar{G},I]$, we use Bayes's theorem, which relates the PDF we require to one which we can calculate and to another which encodes our prior knowledge:

$$P[A|\bar{G},I] \propto P[\bar{G}|A,I]P[A|I] . \quad (19)$$

The term on the far right-hand-side $P[A|I]$ is called the *prior* PDF and represents our state of knowledge about $A(\omega)$ before we have the data. Our prior state of knowledge is modified by the data through the so-called *likelihood function* $P[\bar{G}|A,I]$, which encodes details about the nature of the simulation. The product of the prior PDF and the likelihood function yields the *posteriori PDF* we require and represents our state of knowledge about $A(\omega)$ after we have analyzed the data.

The likelihood function tells us how likely it is that we would have measured the data we actually did, if we were given an $A(\omega)$. In order to compute the likelihood function, therefore, it is essential (but not sufficient) that we should be able to calculate an ideal data set $G(\tau)$ given a spectral density function. For our case, the relevant transform is given by

$$G(\tau) = \int_{-\infty}^{+\infty} d\omega\, K(\tau,\omega)A(\omega) , \quad (20)$$

where the kernel

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

To calculate the likelihood function, we also need some knowledge about the statistical properties of the errors in the data. If we make the simplifying assumptions that the data are independent (so that one measurement does not affect another) and subject to additive Gaussian noise with a root-mean-square error σ_i, then the likelihood function takes the familiar form

$$P[\bar{G}|A,I] = \frac{e^{-\chi^2/2}}{\Pi_i \sqrt{2\pi\sigma_i^2}} ,$$

where χ^2 is the usual sum-of-squared-residuals misfit statistic:

$$\chi^2 = \sum_i \frac{[G_i - \bar{G}_i]^2}{\sigma_i^2} ,$$

where $G_i = G(\tau_i)$. The least-squares approach follows from the assumption that the prior PDF $P[A|I]$ is a constant (part of the normalization), and finds the most probable $A(\omega)$, $\hat{A}(\omega)$, by maximizing the likelihood PDF, which is the same as minimizing χ^2. The least-squares approach assumes that there is no prior knowledge about $A(\omega)$.

There is, however, an unrefutable piece of prior knowledge about $A(\omega)$: it is a positive and additive distribution. The appropriate prior for a positive and additive distribution is not immediately obvious, but many different types of arguments, including logical consistency, information theory, coding theory, and combinatorial arguments, lead to the entropic form[11,12]

$$P[A|I,m,\alpha] = \frac{e^{\alpha S[A,m]}}{Z_S(\alpha)} .$$

Here the prior information I assumes only that $A(\omega)$ is positive and additive, Z_S is a normalization constant, and S is the generalized Shannon-Jaynes entropy

$$S[A,m] = \int d\omega \{A(\omega) - m(\omega) - A(\omega)\ln[A(\omega)/m(\omega)]\} , \quad (21)$$

where $m(\omega)$ is a Lebesgue measure on the space of the distribution and α is a dimensional constant which is initially unknown. Multiplying this entropic prior with the likelihood function, we obtain (to within a normalization constant) the posterior PDF $P[A|\bar{G},I,m,\alpha]$. The function $m(\omega)$ is called the model. It sets both the zero and

the maximum of the entropy.

To proceed further and deal explicitly with the extraneous factors of α and $m(\omega)$, we need to use two more results from probability theory. The first is essentially a restatement of Bayes's theorem:

$$P[a,b]=P[a|b]P[a]$$

and the second concerns marginalization, or the integrating out, of nuisance parameters:

$$P[a]=\int db\, P[a,b] .$$

The parameter α can be removed from the problem by using these relationships in the following manner:

$$P[A|\overline{G},m]=\int d\alpha\, P[A,\alpha|\overline{G},m]$$
$$=\int d\alpha\, P[A|\overline{G},m,\alpha]P[\alpha|\overline{G},m] ,$$

where we have omitted the conditioning on I as implicitly given throughout. So we eliminate α from the posterior PDF for $A(\omega)$ by integrating, with respect to α, the product of the posterior PDF derived above with the posterior PDF for α. If the number of data is large, then the posterior PDF for α is often sharply peaked, say around $\hat{\alpha}$, because we are trying to estimate a single parameter given many data. Therefore the integral over α is usually well approximated by fixing the value of α to be $\hat{\alpha}$:

$$P[A|\overline{G},m,\alpha]\approx P[A|\overline{G},m,\hat{\alpha}] .$$

To calculate the posterior PDF for α, we just use the same rules of probability theory as above:

$$P[\alpha|\overline{G},m]=\int \mathcal{D}A\, P[A,\alpha|\overline{G},m]$$
$$\propto \int \mathcal{D}A\, P[A,\alpha,\overline{G}|m]$$
$$\propto \int \mathcal{D}A\, P[\overline{G}|A]P[A|m,\alpha]P[\alpha] ,$$

where we have dropped irrelevant conditioning statements in the last line for simplicity (e.g., $P[\overline{G}|A,m,\alpha]=P[\overline{G}|A]$, because the data only depend on the spectral density functions and not the Lebesgue measure or α). The first term in the integral, $P[\overline{G}|A]$, is the likelihood function, the second term $P[A|\alpha,m]$ is the entropic prior, and the last term is the prior PDF for α. Skilling[12] has shown that the integral over the spectral density functions requires a measure $[A(\omega)]^{-1/2}$, which is equivalent to the entropy metric, to correctly define the volume element $\mathcal{D}A$. Any reasonable (i.e., not highly informative) choice for the prior PDF $P[\alpha]$ is soon overwhelmed by the evidence of the data to yield a sharply peaked posterior PDF for α. Following Skilling,[12] if we use $P[\alpha]=$const and carry out the integral over the spectral density functions in the Gaussian approximation (by expanding $\ln(P[A|\overline{G},m,\alpha])$ in a quadratic Taylor series about the optimal $A(\omega)$), we find that $\hat{\alpha}$ is given by

$$-2\hat{\alpha}S=\sum_i \frac{\lambda_i}{\hat{\alpha}+\lambda_i} , \qquad (22)$$

where λ_i are the eigenvalues of the Hessian operator $\partial^2\chi^2/\partial A(\omega)\partial A(\omega')$ viewed in the entropy metric. The sum on the right-hand side is often referred to as the number of good data, N_g.

To remove $m(\omega)$ from the posterior PDF for $A(\omega)$, we should, in principle, marginalize out the Lebesgue measure. In practice, this is far more difficult (if not impossible) than the marginalization of α, because $m(\omega)$ is not a single parameter. For example, we cannot generally assume that the posterior PDF for $m(\omega)$ will be sharply peaked or that the prior PDF we assign to $m(\omega)$ will not strongly affect its posterior PDF. Therefore we give the posterior PDF for $A(\omega)$ conditional upon our choice of $m(\omega)$: $P[A|\overline{G},m]$. Rather than being a failure, however, the explicit conditioning on $m(\omega)$ has the advantage that it enables us to incorporate our prior expectation about $A(\omega)$ into the problem in a natural way. This is because, in the absence of any data, the posterior PDF for $A(\omega)$ becomes directly proportional to the entropic prior with $m(\omega)$ being our best (prior) estimate for $A(\omega)$; this is the reason $m(\omega)$ is often called the *default model.* Although the choice of the default model is up to us and the results we give are conditional upon it, probability theory does allow us to choose quantitatively between different alternatives if they are available:

$$P[m|\overline{G}]=\int P[A,m|\overline{G}]\mathcal{D}A$$
$$\propto P[m]\int P[A|\overline{G},m]\mathcal{D}A .$$

So, the ratio of the posterior probabilities for two alternative default models $m_1(\omega)$ and $m_2(\omega)$ is given by ratio of prior probabilities (which we might take to be unity to be "fair") times the ratio of their evidences. We use the word "evidence" here, in common with its usage by Skilling,[22] to refer to the integral quantity which is the normalization factor in Bayes's theorem for the posterior PDF for $A(\omega)$.

Central to part of our original question, "What is our best estimate of the spectral density function," and to the computation of integrals over $A(\omega)$, calculated in the Gaussian approximation, is the optimal solution for $A(\omega)$, $\hat{A}(\omega)$. Formally, we need to find the maximum of the posterior PDF for $A(\omega)$, leading to the condition

$$\frac{\partial}{\partial A}\left[\hat{\alpha}S-\frac{\chi^2}{2}\right]\Bigg|_{\hat{A}}=0 ,$$

The solution is referred to as the maximum entropy reconstruction, or image. We leave an account of algorithmic issues related to finding this solution to the Appendix, and go on to a consideration of the other part of our original question, "What is the reliability of our best estimate of the spectral density function?"

Let us suppose that we wish to estimate a quantity B which is some function F of the spectral density function: $B=F[A(\omega)]$. For example, B might be given by the integral transform

$$B=\int d\omega\, P(\omega)A(\omega) . \qquad (23)$$

If $P(\omega)$ was a δ function at ω_0, then B would be equal to the value of the spectral density function at $\omega=\omega_0$. Our inference about B is given by the posterior PDF

$P[B|\bar{G},m,F]$, which is related to the posterior PDF for $A(\omega)$ we have derived above by

$$P[B|\bar{G},m,F]=\int \mathcal{D}A\, P[B,A|\bar{G},m,F]$$
$$=\int \mathcal{D}A\, P[B|A,F]P[A|\bar{G},m]\,,$$

again dropping irrelevant conditioning statements for simplicity. The PDF $P[B|A,F]$ is, of course, just a δ function: $\delta(B-F[A(\omega)])$. As usual, our best estimate of B, $\hat{B}$, is given by the maximum of the posterior PDF and will be equal to the value of B for the optimal spectral density function: $\hat{B}=F[\hat{A}(\omega)]$. The width of the PDF around $\hat{B}$ will give us the reliability or error bar δB. We will find that although the value of the inferred spectral density function at any particular frequency has a very large (if not infinite) error bar, because we cannot obtain microscopic information from macroscopic data, integrated features of $A(\omega)$ can be reliably determined.

Let us conclude this section with a few general remarks. We have presented here the Bayesian approach to the analytic continuation problem. Probability theory enables us to address both the question of the optimal solution and its reliability. It also reminds us that all answers are conditional upon our prior knowledge, but provides us with the machinery to quantitatively choose between alternative assumptions when they are available. If we do not "like" the answer, there must be additional prior knowledge or expectations that we have about the situation which we have not incorporated into the analysis; because all our assumptions are stated explicitly up front, in the conditioning statements, the causes of the shortcomings are easy to spot and rectify. The entropic prior, for example, encodes our unrefutable prior knowledge that the spectral density function is a positive and additive distribution, but we may also have reason to believe that $A(\omega)$ is locally "smooth." A slightly different formulation of the problem[22,23] enables us to combine legitimately local smoothness with entropy, and the probabilistic framework allows us to consider quantitatively the evidence for our inkling and to optimally choose the parameter(s) describing the corresponding spatial correlations.

It sometimes happens that there are uncertainties in our knowledge of the relationship between the data and the object of interest. Often this takes the form of an unknown constant γ, which scales the errors assigned to the data, $\sigma_i \to \gamma\sigma_i$, thereby affecting the likelihood function and the posterior PDF for $A(\omega)$. To estimate the value of this scaling factor, we just use the same rules of probability theory to compute the posterior PDF for γ:

$$P[\gamma|\bar{G},m]=\int \mathcal{D}A\, P[\gamma,A|\bar{G},m]$$
$$\propto P[\gamma]\int \mathcal{D}A\, P[A|\bar{G},m,\gamma]\,.$$

Again, since we are trying to estimate a single parameter from many data, the value of γ will be well determined by the data (independent of any reasonable assignment for the prior PDF for γ). Although this analysis is only strictly valid for the problem where there is such an unknown scaling constant for the errors, we frequently used it to check that the inferred value of γ was close to unity since $\gamma \gg 1$ would indicate serious systematic flaws in our calculation of the errors (and the assumptions that underlie it).

The model allows prior information about the solution to be incorporated in a consistent and convenient manner. In the calculations to be presented, we found that good perturbation theory provides good models. Sum rules and moments with the maximum-entropy method are also sources of models.[24,13] for instance, if

$$\int d\omega\, f(\omega)A(\omega)=\theta\,,$$

then a potentially useful model can be chosen by maximizing $-A\ln A$ subject to the above and normalization as constraints. The result is

$$m=pe^{-qf(\omega)}\,,$$

where

$$p\int d\omega\, e^{-qf(\omega)}=1,\quad p\int d\omega\, f(\omega)e^{-qf(\omega)}=\theta\,.$$

Extension to multiple sum rules is straightforward. In the absence of prior information the flat model, where A is a constant, independent of ω, is appropriate.

IV. QUANTUM MONTE CARLO

A. Anderson impurity model

As a test problem, we chose to find the spectral density of the symmetric, nondegenerate, single-impurity Anderson model[25]

$$H=\sum_{|k,\sigma} \varepsilon_k n_{k\sigma}+\sum_{k,\sigma} V_{kd}(c^\dagger_{k\sigma}d_\sigma+d^\dagger_\sigma c_{k\sigma})$$
$$+\varepsilon_d\sum_d d^\dagger_\sigma d_\sigma+Un_{d_\uparrow}n_{d_\downarrow}\,, \tag{24}$$

where ε_k is the band energy of an electron in state k, ε_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 electrostatic 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.

This model was originally proposed to describe the properties of dilute magnetic alloys. A central question was the development and persistence of a magnetic moment of an atom with a partially filled shell when alloyed into a metallic host. Experimental systems were found to have a number of anomalous transport properties, and several of these properties showed identical behavior (universality) if the temperature was scaled by a material-dependent constant called the Kondo temperature T_K. Universality within the Anderson single-impurity model was established by the renormalization-group calculations of Krishna-murthy, Wilkins, and Wilson.[26] Bickers, Cox, and Wilkins[27] have emphasized that for a dilute alloy various transport coefficients are moments of the inverse of the spectral density. For example, the electrical resistivity is

$$\frac{\rho(0)}{\rho(T)} \propto L_0(T) , \tag{25}$$

while the thermal conductivity is

$$\frac{\kappa(T)/T}{[\kappa(T)/T]_0} \propto L_2 - \frac{L_1^2}{L_0} , \tag{26}$$

where

$$L_n \propto \int_{-\infty}^{+\infty} \frac{\partial f}{\partial \omega} \omega^n A^{-1}(\omega) d\omega$$

and f is the Fermi-Dirac function. Since the derivative of f with respect to ω is a sharply peaked function about $\omega=0$, the transport coefficients are very dependent on the spectral density near the Fermi surface, and since these coefficients exhibit universality at low temperatures, the spectral density near $\omega \approx 0$ is expected to exhibit approximate universality when ω and T are scaled by T_K.

For the symmetric Anderson model, where $\varepsilon_d = -U/2$, the spectral density is an even function in ω. Its semiquantitative features are easily conjectured.[28] When both U and V_{kd} are zero, δ functions sit at $\omega \pm U/2$, corresponding to the probability of adding an electron to an occupied state or having a doubly occupied particle or hole state. When the hybridization is nonzero, the δ functions are broadened into Lorentzians (Friedel peaks) with a width $\Gamma = \pi V^2 N(0)$, where V is some measure of the effective hybridization and $N(0)$ is the density of state at the Fermi level. When $U \neq 0$ and the temperature T is near or below the Kondo temperature T_K, then a central, non-Lorentzian, resonance peak, whose width is proportional to T_K (the Kondo resonance), exists. As $T \to 0$, the Friedel sum rule[29] predicts that $A(0)$ approaches $1/\pi\Gamma$. From the point of view of the impurity, the presence of the conduction band is evidenced almost exclusively by the resonance broadening.

Horvatić, Šokčević, and Zlatić[30] calculated the spectral density of the symmetric Anderson model using perturbation theory where the natural expansion parameter was $u = U/\pi\Gamma$. They expanded the self-energy as a function of u about an unperturbed state chosen to be the Hartree-Fock solution. The expansion is expected to be useful for $u < 1.0$ and is also expected to become exact as $\omega \to \infty$. For the symmetric model, the lowest order term is quadratic in u. This approximation to the proper self-energy is then used in Dyson's equation to calculate the correlation function from which the spectral density is found. Predicted are broadened peaks at $\pm U/2$ and the central peak at $\omega=0$; however, the central peak is not universal. We chose this perturbation theory as one of the models in our entropy expression. We found it superior to the flat model and better than one with three appropriately placed Lorentzians.

B. Algorithm

We calculated $G(\tau)$ for the Anderson model using the QMC algorithm of Hirsch and Fye.[7,31] This algorithm is particularly suited for impurity problems. For the symmetric model there are no "sign" problems that plague many quantum simulations and the algorithm is exceptionally stable at low temperatures. Its natural product is $G(\tau)$. Furthermore, the impurity can be embedded in an infinite medium to remove finite size effects. The conduction-band details enter only through the noninteracting problem, so it can be modeled in whatever fashion is convenient. In our calculations, we modeled the conduction band by a flat density of states and in relevant integrations took the bandwidth to infinity to remove the bandwidth as a parameter. The Coulomb parameter U and the resonant broadening Γ become the two parameters controlling the physics of our problem.

In the Hirsch-Fye algorithm, the problem is cast into a discrete path integral in imaginary time by first writing the partition function as

$$Z = \mathrm{Tr}\, e^{-\beta H}$$
$$= \mathrm{Tr} \prod_{l=1}^{L} e^{-\Delta\tau H} ,$$

and then using the Trotter approximation[32,33] to write

$$Z \simeq \mathrm{Tr} \prod_{l=1}^{L} e^{-\Delta\tau H_0} e^{-\Delta\tau H_1} + O(\Delta\tau^2) ,$$

where H_0 is the noninteracting part of the Hamiltonian, H_1 is the interacting (impurity) term, and $\Delta\tau = \beta/L$. The interacting term is reduced to a noninteracting contribution by the introduction of auxiliary Ising variables

$$e^{-\Delta\tau H_1} = e^{-\Delta\tau U[n_{d_\uparrow} n_{d_\downarrow} - (1/2)(n_{d_\uparrow} + n_{d_\downarrow})]}$$
$$= \tfrac{1}{2} \mathrm{Tr}_s e^{\lambda s (n_{d_\uparrow} - n_{d_\downarrow})} ,$$

where $\cosh\lambda = \exp(\Delta\tau U/2)$. Then taking the trace over the fermion degrees of freedom yields

$$Z = \mathrm{Tr}_s \prod_{\sigma=\pm 1} \det M_\sigma[s] , \tag{27}$$

where for a given configuration of Ising variables $M_\sigma[s]$ is an $L \times L$ matrix whose nonzero matrix elements are

$$[M_\sigma]_{\mu} = 1 ,$$

$$[M_\sigma]_{l,l+1} = -e^{-\Delta\tau K} e^{V_l^\sigma - 1}(1 - 2\delta_{1l}) ,$$

and $[M_\sigma]_{lm} = 0$ otherwise. K is a matrix representing the noninteracting (bilinear) part of the Hamiltonian and $V_l^\sigma = \sigma s(l)\lambda |d\rangle\langle d|$ is the (imaginary) time-dependent potential acting at only the impurity site.

Different Ising configurations give rise to different potentials V. The Green's function $g_\sigma = [M_\sigma]^{-1}$ for two different configurations are connected by a Dyson equation

$$g' = g + (g - I)(e^{V' - V} - I) g' .$$

As a matrix equation, because of the restricted range of the interaction, there is an element of g that connects only to the impurity state

$$g'_{dd} = g_{dd} + (g_{dd} - I)(e^{V' - V} - I) g'_{dd} . \tag{28}$$

If one starts with the Green's function g^0_{dd} of the noninteracting problem, the g_{dd} for any configuration can be

constructed by successive applications of this equation in which one Ising variable is added at each application.

At each time step l of the Monte Carlo procedure attempts to flip the value of s_l depending on whether the ratio of the determinants in (27) for two different configurations, $R = R_\uparrow R_\downarrow$ with

$$R_\sigma = 1 + [1 + g^\sigma_{dd}(l,l)][e^{V_l^\sigma[s'] - V_l^\sigma[s]} - 1] ,$$

is greater or less than a random number between 0 and 1. If the flip is accepted, then the function is updated by use of (28). The application of the Monte Carlo test to all L time steps is called a sweep. Averaging over a large number of sweeps generates an estimate for $G(\tau) \equiv \sum_\sigma \langle g^\sigma_{dd} \rangle$.

C. Specific points

At large positive and negative ω, the kernel (20) exponentially damps $A(\omega)$, causing the result of the integral, namely $G(\tau)$, to be insensitive to radically different high-frequency behaviors of $A(\omega)$. When measured by QMC, $G(\tau)$ is inherently noisy and necessarily incomplete. It is incomplete because the continuous variable τ can be sampled only for a finite number of discrete values of imaginary time τ_i, $G_i = G(\tau_i)$. The noise and incompleteness prompt for a least-squares solution to the problem, but the insensitivity, when coupled with the noise and incompleteness, translates into a terribly ill-posed problem, meaning an infinite number of $A(\omega)$ exists that can be consistent with the measured correlation function. Solving (20) means selecting by some criterion, a process called regularization, a "best" solution. Maximum entropy is our regularization procedure.

For quantum Monte Carlo data, the assumption of Gaussian-distributed errors is often adequate, but the assumption of independent errors is usually very poor. The definition of the χ^2 statistic must be generalized to the form:[1]

$$\chi^2 = \sum_{ij} (G_i - \bar{G}_i)[C^{-1}]_{ij}(G_j - \bar{G}_j) , \tag{29}$$

where C_{ij} is an element of the covariance matrix describing the correlations between the data. With angular brackets denoting a statistical average

$$C_{ij} = (\langle \bar{G}_i \bar{G}_j \rangle - \langle \bar{G}_i \rangle \langle \bar{G}_j \rangle)/(N-1) , \tag{30}$$

where N is the number of samples. The covariance matrix is symmetric, positive definite. Its diagonal elements are the familiar values of the variance of $\bar{G}_j$. We have found that C is poorly approximated by its diagonal elements, meaning that strong correlations between different values of $\bar{G}_j$ exist and the different $\bar{G}_j$ are not statistically independent. The importance of the off-diagonal terms has been ignored in all previous work.

By diagonalizing C with an orthogonal transformation S, we reexpress (29) in the resulting diagonal basis

$$\chi^2 = \sum_i \frac{(G'_i - \bar{G}'_i)^2}{\sigma_i^2} , \tag{31}$$

where

$$\bar{G}' = S^T \bar{G}, \quad G' = S^T G , \tag{32}$$

and σ_i^2 are the eigenvalues of C. We find that these eigenvalues can range over two to six orders of magnitude. This means that not all the $\bar{G}'_i$ values participate meaningfully in the least-squares problem.

V. TECHNICAL DETAILS

The proper extraction of $A(\omega)$ from $G(\tau)$ requires an estimate of errors in the data and an understanding of the propagation of these errors by maximum entropy. In this section, we discuss both issues.

A. Statistical control

From the QMC, computing an estimate of $\bar{G}_i$ is easy; determining the reliability of that estimate is more difficult. The general principle is to accumulate a sufficiently large number of statistically independent measurements so, by the central limit theorem, the sample variance of these measurements becomes a measure of the actual variance of measurements as their distribution becomes Gaussian. In practice, satisfying this principle is the motivation for the coarse-grained averaging procedure.[34] In this procedure, measurements of G_i are usually made after each Monte Carlo sweep. Since successive measurements will be correlated to some degree, a large number of successive measurements are collected into a bin. For each bin, an average G_i is computed. If the bin is sufficiently large, different bin averages become independent. Then for a large number of bins, the sample variance of the bin averages is a measure of the variance of the Gaussian distribution towards which these averages are tending. That a Gaussian distribution has been approached is rarely checked. We found it necessary to monitor the situation and take specific measures to promote Gaussian behavior.

For each τ_i we measured G_i only after every fourth sweep. Our flip acceptance rate was 40–50 %, which is generally considered nearly optimal. (This is produced by the algorithm; it is something we cannot control.) At this "hit" rate, the autocorrelation length[35] for each G_i was less than 1.5, so on this basis, we seem to be producing relatively independent measurements, but we found that different bin lengths would give different variances for the $\bar{G}_i$. Since we are trying to solve an ill-posed problem, these differences can change our results. We needed to produce variances that were insensitive to binning. The error estimation is further complicated by the strong correlations among bin-averaged values and the need to avoid storing large amounts of data. Instead of working in the diagonal basis defined by the covariance matrix, we chose to study the behavior of its diagonal elements. We found that when they were under statistical control, maximum entropy's estimate γ of error rescaling was close to unity, and if they were not under control, then γ was greater than unity.

We did three things with the bin averages. For a fixed number of bins, we assigned the data to a bin until the bin was full and then repeated the process until *all bins were* used (sequential binning). We also assigned the data to different bins until all bins were used and then repeated

the process until all bins were full (shuffled binning). In spot cases, we put the data into random order and then did the sequential binning. We found little difference between this binning and the shuffled binning. The average over all bins for these processes are the same; the variances are different, but if the bin size is large enough to produce statistically independent averages, then the differences in the variances cannot be statistically significant. We used the F test[36] to estimate the probability that the differences are significant. In general, the variances of only a few $\bar{G}_i$ values had greater than a 95% probability (more than three standard deviations) of being statistically significant. The τ values at which these large deviations occurred varied from simulation to simulation.

We also performed a χ^2 test[36] to measure how well histograms of our measurements compared to Gaussians with the same mean and variance. Again, in only few cases were the differences significant, and their location varied from simulation to simulation. A particularly simple, but the most revealing, figure of merit was the kurtosis.[36] It was generally negative, meaning our histograms are flat at the peak. In general, the kurtosis was just marginally within one standard deviation of what it should be if our bin averages were drawn from a Gaussian distribution. If the deviation was greater than one standard deviation, then our error estimates proved inadequate.

Overall, our data would be more accurately described as not being inconsistent with Gaussian behavior rather than being described as being Gaussian. The major point is greater effort than is normally done in a QMC calculation was required to achieve data that were not inconsistent with Gaussian sampling. Typically, we swept 1000 times to equilibrate the system before recording measurements. The number of measurements in a bin varied from 50 to 1000, and the number of bins varied from 100 to 200. The numbers used depended on the value of U, Γ, and T. The stated numbers are conservative and not optimal. When the $T > T_K$, the larger numbers were necessary, since at high temperatures, we have fewer time steps and hence fewer data values. Reducing the error associated with these values was necessary for good results. The choice of $\Delta\tau$ was also found to influence the quality of our data. Empirically, we found if $U\Gamma(\Delta\tau)^2 < 0.2$, then our error was statistical and not discretization.

B. Solution assurance

"How does one know that the spectral density $A(\omega)$ produced by a data analysis procedure is the correct one?" In a least-squares fitting of a model with a few parameters to the data, the best fit is judged to be the one with the smallest χ^2. The usual measure of a good fit is $\chi^2 \leq N_d$, where N_d is the number of independent data points. For fitting an entire spectral function, however, the number of free parameters is infinite, and an infinite number of parameters may satisfy $\chi^2 = 0$ or any other value of $\chi^2 \leq N_d$ one may choose. Such fits often exhibit structure which is a consequence of the statistical noise in the data, rather than the physics. "How does one distinguish this noise-induced structure from the real structure?"

The goal of any statistical regularization procedure is to provide criteria to select the best fit. Maximum entropy corresponds to seeking a fit which is closest to our prior knowledge, with negative entropy being an information measure of the distance between the solution and our prior knowledge. In the early development of the maximum-entropy method, the best fit was taken to be the solution that satisfied $\chi^2 = N_d$, with $1/\alpha$ being the Lagrange multiplier associated with the maximum-entropy constraint. This is known as the historic-maximum-entropy criterion for α.

In the more recent approach discussed in Sec. III, α is calculated by Bayesian logic. The best fit is the maximum-entropy solution for the most probable value for α. This approach is known as classic maximum entropy and results in $\chi^2 = N_d - N_g$, where N_g is the number of good eigenvalues of the likelihood function Eq. (22). For classic maximum entropy, χ^2 is thus less than N_d, the data are fit more closely than historic maximum entropy, and error estimates on the spectral density (see the Appendix) are larger. Because of the exponential character of the kernel that relates the spectral function to the Green's function (20), the eigenvalues of the likelihood function rapidly become very small so that N_g is a small number.[1] Typically, $N_g \sim 5-10$ for $N_d \sim 50-100$. As a consequence, the difference between the historic and classic criteria is small, but as we shall show, the difference can correspond to an order of magnitude or more change in α.

While maximum entropy is a statistical regularization procedure designed to minimize noise artifacts in the data, there are no guarantees. For example, if there is a broad feature in the spectrum, can it be resolved into two or three distinct overlapping features with improved resolution? How do we know that a particular feature in the spectrum is real? The answer to these and similar questions is given by the size of the error bars (see the Appendix) calculated for the features. If the error bars for a particular feature are small, then the likelihood of there being additional features or of that feature being a noise artifact is small; however, if the error bars are large, we can try to reduce the errors by improving either the data or the prior knowledge. For example, we can produce more data with larger number of Trotter steps L and smaller step sizes $\Delta\tau$, or reduce errors with a longer Monte Carlo run. We can improve our prior knowledge by selecting a more informative default model. Because of the exponential kernel of our analytic continuation problem, it is particularly helpful to input prior knowledge of the high-frequency part of the spectral function via the model.

Systematically increasing the probability of the data $P[\bar{G}|I]$ by improving the data and the default model generally allows one to distinguish real features from noise artifacts or test for unresolved features. The probability of the data $P[\bar{G}|I]$, or evidence, is the normalization constant in Bayes's theorem. Increasing $P[\bar{G}|I]$ reduces the size of the error bars.

VI. DETAILED CASE STUDY

The discussion in Sec. V underscores the point that the determination of the spectral density is not a "black-box" calculation, even with a procedure that has no adjustable parameters. In this section, we will provide a detailed case study of the determination of spectral densities using the maximum entropy method. In doing so, we will find it useful to utilize the distinction between classic and historic maximum entropy discussed briefly in the preceding section. Classic maximum entropy is the procedure described in Sec. III and the Appendix. It is based on the Bayesian approach to probability theory, with the parameter α being chosen (marginalized) by probability theory, resulting in $\chi^2=N_d-N_g$. Historic maximum entropy, on the other hand, refers to the traditional procedure of choosing the value of α such that $\chi^2=N_d$, where N_d is the number of data: it chooses the solution with the most entropy that "fits the data." Historic maximum entropy results in a larger χ^2 and larger α than classic maximum entropy, leading to solutions with less structure and correspondingly fewer artifacts. Historic maximum entropy places more weight on the prior knowledge, relative to the evidence in the data, than classic maximum entropy. This often results in visually more appealing fits, because the imperfections in the data are less able to filter through. However, a choice of α larger than that required by classic maximum entropy leads to overly optimistic estimates of the reliability of the fit.

The spectral function is to be inferred from quantum Monte Carlo data for the Green's function $G(\tau)$. Figure 1 shows the data set on which we shall focus initially. For the symmetric Anderson model, $A(\omega)$ is an even function, and $G(\tau)$, which equals $-G(\tau+\beta)$, is an even function in the interval $0\le\tau<\beta$ about $\tau=\beta/2$. The expansion parameter $u\equiv U/\pi\Gamma=2.5$ is much larger than

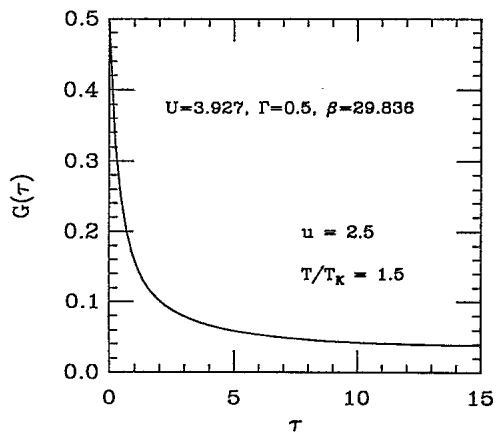

FIG. 1. Matsubara Green's function $G(\tau)$ plotted vs imaginary time τ for the single-impurity symmetric Anderson model. The parameters are $\beta\equiv 1/T$, the inverse temperature; Γ, the hybridization width between the impurity and the conduction electrons; U, the Coulomb repulsion between two electrons on the impurity; $u\equiv U/\pi\Gamma$, the expansion parameter for perturbative approaches to the Anderson Hamiltonian; and T_K, the corresponding Kondo temperature taken from the Wiegmann-Tsvelick Bethe ansatz solution of the Anderson model.

the domain of validity of Horvatić-Zlatić perturbation theory. It is well into the regime where we expect universal behavior. The ratio $T/T_K=1.5$ is too large to expect insight from the zero temperature Bethe ansatz[37] and renormalization-group[26] solutions. The large u in actuality corresponds to a large U, for which the quantum Monte Carlo algorithm, because of the Trotter approximation, will produce relatively large errors compared with data at smaller U. Consequently, this data set provides a good test for our methods.

We have found that the covariance matrix associated with $\bar{G}_i$ is dense, with elements going to zero only at $\tau=0$ as required by the sum rule for the symmetric model

$$G(0)=\int_{-\infty}^{+\infty}d\omega\frac{A(\omega)}{(1+e^{-\beta\omega})}$$
$$=0.5\ .$$

As discussed in Sec. IV C, the data can be considered to be independent only in an orthogonally transformed data space in which the covariance matrix is diagonalized. The eigenvalues of the covariance matrix for our test data set are shown in Fig. 2. The eigenvalues for all data sets typically spanned four to six orders of magnitude with the range depending greatly on the Anderson model parameters, as well as on the details of the quantum Monte Carlo run. Smaller $U\Gamma(\Delta\tau)^2$ and smaller β produced smaller eigenvalues. Since the Gaussian errors on independent data correspond to square roots of the eigenvalues and since the smallest eigenvalue for our test data set is approximately 10^{-10}, the errors are at the level of the fourth or fifth digit of the $G(\tau)$ data. Because of the large dynamic range of the eigenvalues of the covariance matrix, as well as the large range in the eigenvalues of the likelihood function,[1] we found it useful to run our maximum-entropy codes with 64-bit arithmetic. Failure to take this care in the data preparation and handling would often result in spurious structure in the spectral functions inferred.

A. Historic maximum entropy

We now proceed to the maximum-entropy analysis of the data in Fig. 1. We began by using as a default model

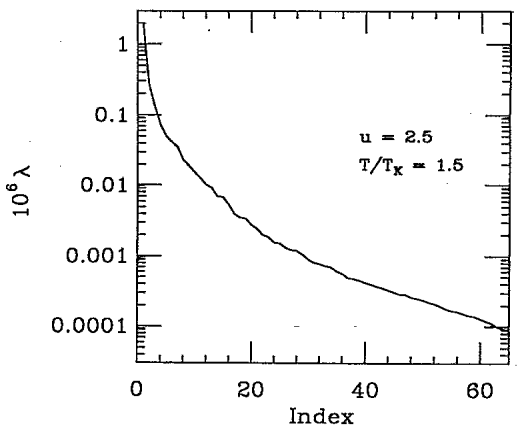

FIG. 2. Eigenvalues of the covariance matrix for the data in Fig. 1, in order of decreasing values.

a Lorentzian of width Γ centered at $\omega=0$. This function would be the correct spectral density only in the $U\rightarrow 0$ limit. It provides a very poor fit to the data, with $\chi^2/N_d=4.6\times10^4$, where N_d is the number of data values, but it does satisfy the $G(0)=0.5$ requirement from the spectral function sum rule.

We then found the maximum-entropy solution as a function of the statistical regularization parameter α. The calculation begins at $\alpha=\infty$ and iteratively reduces α. Figure 3 shows the entropy S and the χ^2/N_d of the maximum-entropy images as a function of α. At $\alpha=\infty$, the maximum-entropy image is equal to the default model, the entropy S is zero, and the χ^2/N_d is equal to its initial value for the default model. As α is reduced, the negative of the entropy, which is the measure of the distance between the maximum-entropy image and the default model, increases. At the same time the χ^2/N_d, which is the measure of the quality of fit to the data, decreases. When $\chi^2/N_d\leq 1.0$, we say that we have "fit the data," i.e., we have a feasible image.

Figure 4 shows some of the corresponding images plotted semilogarithmically with the frequency ω scaled by T_K^0 to emphasize the important low-frequency physics, especially the potential for universal behavior. Here T_K^0 is defined by $1/\chi(0)$, and it is related to the Kondo temperature defined by the resistivity by $T_K^0=\pi^2T_K/4$. The vertical axis is plotted as $\pi\Gamma A(\omega)$ for two reasons: this is the quantity which can be expected to show universal behavior, and the Friedel sum rule[29] requires it to equal 1 for $\omega\rightarrow 0$ and $T\rightarrow 0$. For large χ^2/N_d, the image is equal to the default model. As χ^2/N_d decreases, the spectral function deviates from the default model faster at low frequencies than at higher frequencies.

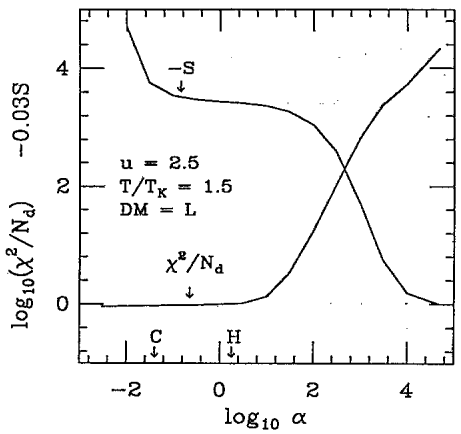

FIG. 3. Entropy S and χ^2 divided by the number of data N_d, as a function of the statistical regularization parameter α for the maximum-entropy analysis of the data in Fig. 1. The label $DM=L$ denotes that a Lorentzian of width Γ was used as the default model. The point labeled H indicates the stopping value of α in historic maximum entropy where $\chi^2/N_d=1.0$. The point labeled C marks the value of α for the classic-maximum-entropy stopping criterion.

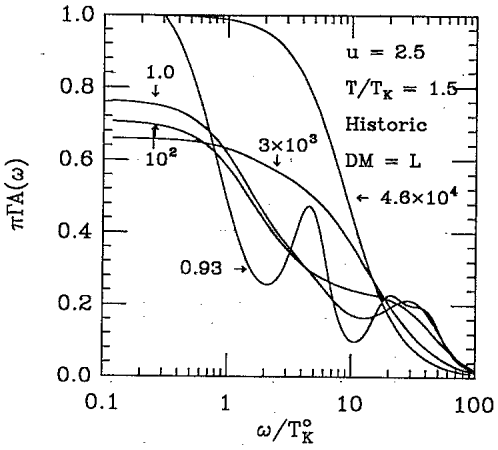

FIG. 4. Maximum-entropy images of the single-impurity Anderson model spectral function $A(\omega)$ plotted vs ω/T_K^0 for various values of χ^2/N_d as indicated. Here, the Wilson Kondo temperature is defined by $T_K^0\equiv 1/\chi(T)$ at $T=0$ where $\chi(T)$ is the static magnetic susceptibility. The Wilson temperature characterizes the low-temperature energy scale of the Anderson model and is related to the Kondo temperature by $T_K^0\equiv\pi^2T_K/4$. The curve with $\chi^2/N_d=1.0$ is the historic-maximum-entropy result.

Until the advent of classic maximum entropy, the image for $\chi^2/N_d=1.0$ was regarded as the "best image." The corresponding choice for α is termed the *historic-maximum-entropy stopping criterion,* and its value is indicated by the H in Fig. 3. This stopping criterion is based on the belief that it provides the feasible image (i.e., it barely fits the data) which is closest to the prior knowledge represented by the default model. It contains the least input from the noisy, incomplete quantum Monte Carlo data. Since we chose a Lorentzian default model which is smooth, we obtain an image with the least structure which is consistent with the data. Indeed, pushing χ^2/N_d down to 0.93 results in considerably more structure of doubtful validity. We would say that we are *overfitting* the data and that the image is *ringing.* In fact, this image is physically wrong since it violates the Friedel sum rule, unitarity limit bound that $\pi\Gamma A(\omega=0)\leq 1$. This image points to an important distinction between maximum entropy and least-squares approaches to the analytic continuation problem. Traditional least-squares approaches would argue that images with smaller χ^2/N_d are better. Indeed, one might prefer to fit the data exactly, as is done in simple Padé methods,[7] but the corresponding spectral function would show spurious structure corresponding to fitting the statistical noise in the QMC simulation rather than the physics. Getting a smooth spectral function would require better statistics on the quantum Monte Carlo data, perhaps an order of magnitude better than required by the maximum-entropy approach.

In fact, there exist an infinite set of possible spectral functions that fit the data for any chosen value of χ^2/N_d, and most of them are undesirable according to physical criteria. A popular modification to the least-squares procedure is to add a smoothing term,[9,10] but then one is left

with an arbitrary choice of smoothing parameter. In addition, these terms arbitrarily cause sharp features in the spectral function to be smoothed even if the sharpness is justified by the data.

Figure 5 compares the historic maximum entropy spectral function with both the Horvatić-Zlatić (HZ) perturbation theory result and our best result obtained by the classic maximum-entropy procedure discussed in Sec. VI B. We obtain the Friedel peak at $\omega = U/2$, even though it was absent in the Lorentzian default model, showing that such high-frequency information was contained in the quantum Monte Carlo data. The difference between the HZ and the quantum Monte Carlo maximum-entropy (QMC-ME) results at low ω/T_K^0 is statistically and physically significant. As we shall show, it represents the difference between the correct universal behavior of the spectral function at low frequencies and the nonuniversal behavior of perturbation theory at finite order in u.

One may be concerned about the maximum-entropy image going unstable between $\chi^2/N_d = 1.0$ and 0.93. As shown in Fig. 3, the statistical regularization parameter α, which controls the degree of fluctuation of the image about the default model, is changing rapidly for very small changes in χ^2/N_d. This is untrue for most applications of maximum entropy to data analysis problems, but it is a consequence of the very small number of well-determined *good* eigenvectors N_g of the likelihood function.[1]

One may be especially concerned that quantities of physical interest, such as the resistivity, might depend strongly on the choice of χ^2/N_d. We found that quantities like these, which are sensitive mainly to the well-determined low-frequency properties of the spectral function [Eqs. (25) and (26)], are fortunately insensitive to the choice of χ^2/N_d.[3-5]

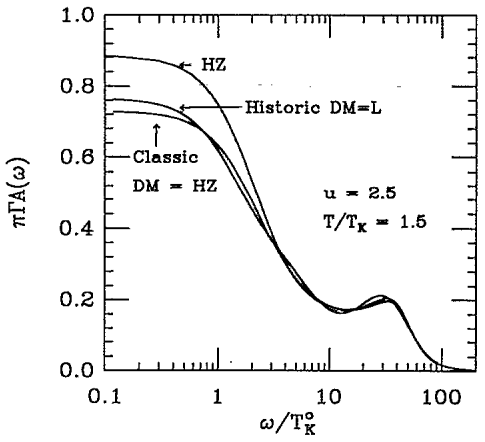

FIG. 5. HZ labels the prediction of Hovatić-Zlatić perturbation theory for the spectral function in which the self-energy is calculated to second order in the expansion parameter $u \equiv U/\pi\Gamma$. We find this is accurate for $u \leq 1.25$. The curve labeled *historic* is the maximum-entropy image for the stopping criterion of $\chi^2/N_d = 1.0$ and a Lorentzian ($DM = L$) default model. The curve labeled *classic* is our best result for the spectral function. The peak centered at $\omega = 0$ of width approximately T_K^0 is the Kondo resonance. The peaks centered at $\omega \pm U/2$ of width Γ are the Friedel peaks corresponding to the removal or addition of an electron to the impurity state.

B. Classic maximum entropy

While we could have analyzed all of our quantum Monte Carlo data satisfactorily using the historic-maximum-entropy stopping criterion, recent developments in maximum-entropy methods remove the ambiguity about the stopping criterion, i.e., the choice of α (or equivalently χ^2/N_d). In addition, they provide a fully probabilistic approach to data analysis that can provide error estimates on integrated functions of the spectral function, such as the resistivity, and answer questions of solution assurance, i.e., "Is the ringing structure in the $\chi^2/N_d = 0.93$ image statistically significant?" In Sec. III, we described the theory of classic maximum entropy.[12] We now illustrate it.

We now analyze our test-data set with classic maximum entropy, using two different default models. One is the Lorentzian ($DM = L$) model used in the VI A. This model is shown in Fig. 4. The other is a more informative model, the Horvatić-Zlatić perturbation second-order perturbation theory ($DM = HZ$) extrapolated beyond its domain of validity ($u \leq 1.0$) to $u = 2.5$. This model is shown in Fig. 5. While the HZ theory cannot be expected to describe quantitatively the low-frequency behavior, it can be expected to describe correctly the high-frequency behavior. In this sense, the HZ model inputs more prior knowledge than the Lorentzian model.

In classic maximum entropy, the stopping criterion is determined by Eq. (22), $-2\alpha S = N_g$, where N_g is the number of good measurements. We find that S/N_g is almost constant, which allows us to discuss the behavior of the classic-maximum-entropy solution in terms of α. In the classic-maximum-entropy approach, one calculates the probability of α, given the data and the prior knowledge (default model), and chooses to maximize this probability. This maximum occurs at $-2\alpha S/N_g = 1.0$.

Figure 6 shows the logarithm of the probability $P[\bar{G}|\alpha, m]$ as a function of α. For the Lorentzian default model ($DM = L$), this probability is sharply peaked at a particular α marked by the C symbol. In contrast, the historic-maximum-entropy value of α, indicated by the H symbol, is much larger than the classic value and is far from the peak in the probability function by orders of magnitude. For the HZ model, the posterior probability is a shallow function of α, and the classic and historic values for α are very close to one another. Since the probability distribution for α is so shallow for the HZ model, one might argue that taking the optimal value of α is a poor approximation for the integral over α required by the probability theory (described in Sec. III). The need for an explicit marginalization over α is one motivation for the alternate algorithm[38] to solve the maximum-entropy equations discussed in the Appendix.

Figure 7 shows χ^2/N_d plotted against α for the two default models. This plot should be compared to Fig. 3. For the Lorentzian default model ($DM = L$), the classic value of the stopping criterion (marked by C) is much

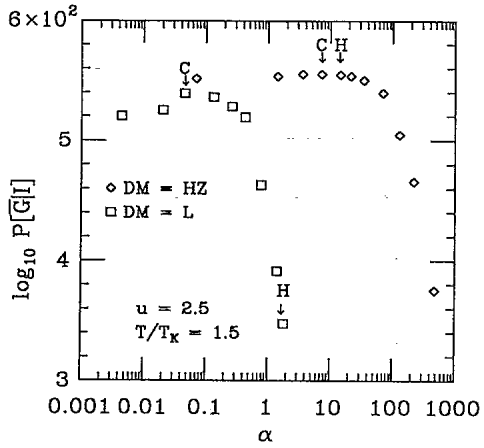

FIG. 6. Bayesian probability of the data $P[\bar{G}|I]$ plotted vs α. $DM=L$ labels the results for the Lorentzian default model, while $DM=HZ$ labels the results for the HZ default model. The stopping criterion for classic maximum entropy is at the peak of this probability distribution, whereas the stopping criterion for historic maximum entropy is an α such that $\chi^2/N_d=1.0$. H labels the position of the historic values for the stopping criterion, while C labels the classic values.

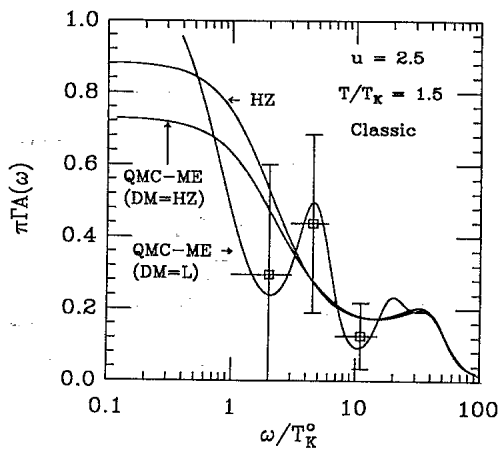

FIG. 8. Classic-maximum-entropy images of the spectral function for two different default models. The quantum Monte Carlo classic-maximum-entropy images of the spectral function are labeled by QMC-ME for the Horvatić-Zlatić (HZ) and Lorentzian (L) default models. The squares mark the average image for the Lorentzian default model over the ω/T_K^0 regions indicated by the horizontal bars. The vertical error bars indicate that the ringing obtained by classic maximum entropy for the Lorentzian model is not statistically significant. HZ labels the Horvatić-Zlatić second-order perturbation theory prediction for $u=2.5$.

smaller than the Historic value (marked by H). The value of χ^2/N_d at the classic-maximum-entropy stopping criterion is less than unity and is smaller for the Lorentzian than for the HZ model. Since the number of good measurements N_g is small for the analytic continuation problem (typically 5–10), χ^2/N_d is only slightly less than 1. Since the HZ model is closer to the truth than the Lorentzan model, the number of good measurements is smaller and the χ^2 is larger. Thus the better the model, the less is the tendency for classic maximum entropy to overfit the data. Indeed, the classic-maximum-entropy image for the Lorentzian default model is the same as the $\chi^2/N_d=0.93$ *ringing* image shown in Fig. 4, as shown again in Fig. 8. However, since classic maximum entropy

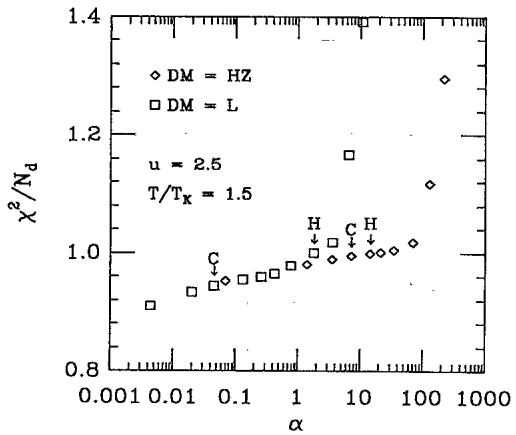

FIG. 7. χ^2/N_d plotted vs α for the maximum-entropy images obtained using the Horvatić-Zlatić (HZ) and Lorentzian (L) default models.

is fully based on probability theory, whereas historic maximum entropy is not, it tells us that the image obtained with the HZ model is several orders of magnitude more probable than that obtained with the Lorentzian model. It also provides error estimates for the structure in the image. The squares show the averages of the image over the ω/T_K^0 regions indicated by the horizontal bars. The vertical bars indicate the statistical error estimates on these averages. One sees that the structure in the ringing image is not statistically significant.

For the informative Horvatić-Zlatić model ($DM=HZ$), the classic value of the stopping criterion is very close to the historic value. The χ^2/N_d is larger than the value for the Lorentzian model. The corresponding image, shown in Fig. 8, is smooth, and the error bars on integrated features in the image (not shown) are much smaller.

The quality of the classic-maximum-entropy images also depends, of course, on the quality of the data. Figure 9 shows images obtained for $u=1.25$ quantum Monte Carlo data, which have approximately an order of magnitude smaller statistical error than the $u=2.5$ data set we have been considering so far. This value of u is also the largest where we might expect Horvatić-Zlatić perturbation theory to be reliable. The figure shows in this case that the classic-maximum-entropy image obtained with the Lorentzian model ($DM=L$) is stable and is essentially identical to Horvatić-Zlatić perturbation theory. However, it is still possible to drive the maximum-entropy image into instability by using an even less informative model, which we chose to be the flat model ($DM=F$) where $m(\omega)=0.1$. This model does not

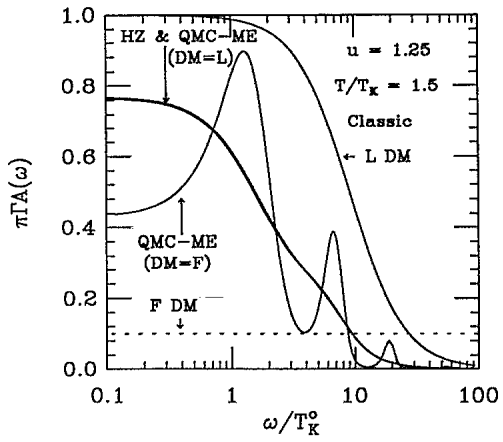

FIG. 9. Classic-maximum-entropy images of the spectral function for $u=1.25$, around the limit of validity of second-order Horvatić-Zlatić (HZ) perturbation theory. The line labeled HZ and QMC-ME($DM=L$) is actually two curves: one is the prediction of perturbation theory, and the other, the classic-maximum-entropy image obtained with a Lorentzian model (shown as the curve labeled *L DM*). The curve labeled QMC-ME($DM=F$) is the classic-maximum-entropy image obtained with a flat model (shown as the dashed line labeled *F DM*), which is even less informative than the Lorentzian model. For example, the Lorentzian model satisfies the $G(0)=0.5$ sum rule, whereas the flat model does not. The image obtained with the flat model is ringing, i.e., overfitting the data.

even satisfy the $G(0)=0.5$ sum-rule constraint, and the corresponding image is ringing. Again, as in Fig. 8, classic-maximum-entropy error estimates would show that this structure is not statistically significant.

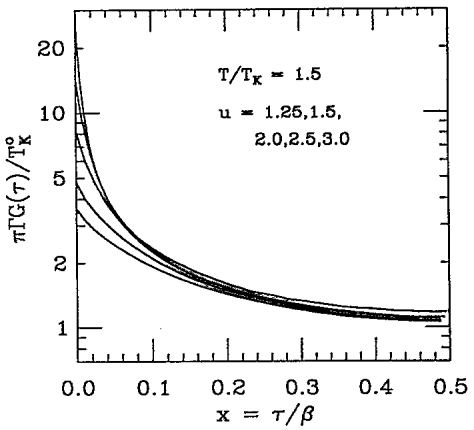

FIG. 10. Testing the quantum Monte Carlo data for universality (independence of u). The scaled Green's function obtained from quantum Monte Carlo $\pi\Gamma G(\tau)/T_K^0$ is plotted vs $x\equiv\tau/\beta$ for fixed $T/T_K=1.5$ and varying $u\equiv U/\pi\Gamma$. Larger intercepts at $x=0$ correspond to larger values of u. The convergence of the curves at large x for different u indicates that the underlying spectral function $\pi\Gamma A(\omega)$ is a universal function for small ω/T_K^0.

C. Universality

As discussed in Sec. IV A, universality is the hypothesis that low-frequency properties should only depend on T_K and be independent of $u\equiv U/\pi\Gamma$. Specifically, we expect at low frequencies $\Gamma A(\omega)$ to depend only on ω/T_K and T/T_K. For fixed T/T_K this would predict that $\Gamma G(\tau)/T_K^0$ would be a function only of τ/β. Figure 10 shows a plot of a sequence of quantum Monte Carlo simulations with fixed $T/T_K=1.5$ and varying u. The predicted scaling relation appears to hold at large τ/β, but not at small τ/β. Since large τ corresponds to the low-frequency part of the spectral function and small τ picks up high-frequency information, universality (independence from u) is present in the low-frequency properties of the spectral function but not in the high-frequency properties. Indeed, as illustrated in Fig. 5, the low-frequency peak is the Kondo resonance, which we expect to show universal behavior, while at high frequencies the spectral function exhibits the $\omega=U/2$ peak, which is nonuniversal.

While the raw data, when properly scaled, appear to show evidence for universal behavior, the default model we shall use is not universal. Figure 11 shows the Horvatić-Zlatić perturbation theory truncated at second order in u for the same sequence of T/T_K and u values. Increasing u increases the value of the intercept at $\omega=0$. Of course, in principle, one could calculate higher-order terms in the HZ expansion in u which would improve the convergence toward a universal behavior at large u, but this has not yet been done. Instead, we use the HZ theory as a default model in order to reduce the variance of the maximum-entropy data analysis. Figure 12 shows the spectral functions obtained by using classic maximum entropy and the HZ model. Within statistical error, the spectral function is universal for $\omega/T_K^0\leq 10$.

Figure 13 shows the corresponding results for the resistivity ratio, calculated by imputing the spectral function

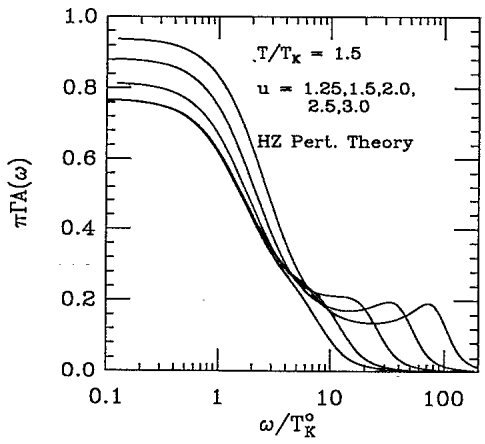

FIG. 11. Predictions of HZ perturbation theory up to second order in u, at fixed $T/T_K=1.5$ and at a variety of u indicated. To second order the perturbation theory is not universal at low ω/T_K^0, although in principle higher-order terms in u may be calculated which would extend its range of validity.

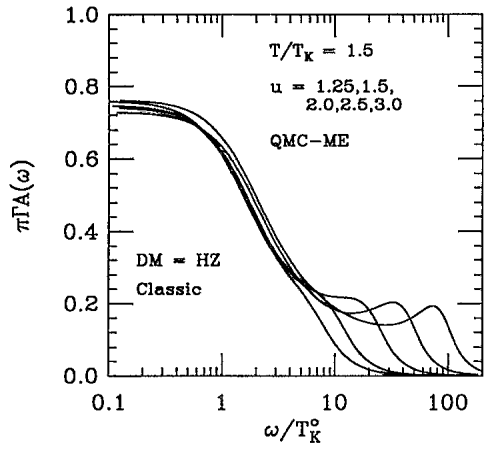

FIG. 12. Quantum Monte Carlo and classic-maximum-entropy images of the spectral function at fixed $T/T_K=1.5$ and a variety of u, obtained using the Horvatić-Zlatić perturbation theory as a default model and automatic noise scaling. The Kondo resonance in the spectral function centered at $\omega=0$ is a universal function for $\omega/T_K^0 \leq 10$, within statistical error. The Friedel peak centered at $\omega=U/2$ of width Γ is nonuniversal.

into Eq. (25). The HZ resistivity ratio is distinctly nonuniversal. Our QMC-ME calculated using the HZ default model are universal within one standard deviation statistical error. The error bars are much smaller than the distance between the QMC-ME and HZ results. Also shown in this figure are the resistivity ratios obtained using the Lorentzian default model. The error bars with the Lorentzian are much too large to reach any conclusions about the validity of universality. In this sense, the use of an informative default model is essential to the proof of universality.

The error bars are provided by the ability of classic maximum entropy to provide error estimates on any integrated property of the spectral function (see the Appendix). We have found these error bars to be reliable in the traditional sense: if we repeat the QMC simulation with different random number seeds, the values for the resistivity ratios are typically equal to within one standard deviation; if we repeat the QMC simulation with smaller statistical error (more CPU time), the resistivity ratio has smaller errors, and the result usually lies within the error bars of the poorer statistics runs; and if we increase the number of binned measurements, the error bars scale with the inverse of the square root of the number of measurements. One may also notice the trend toward larger errors with increasing u. This has two origins: the quantum Monte Carlo calculations have increasing statistical errors with increasing u as measured by the eigenvalues of the covariance matrix as shown in Fig. 2, and the Horvatić-Zlatić perturbation theory becomes a less informative model with increasing u.

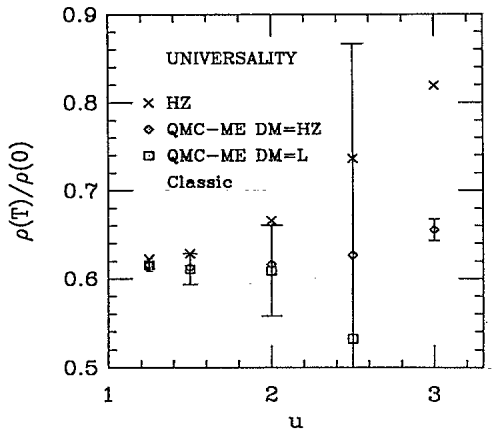

FIG. 13. Resistivity ratios $\rho(T)/\rho(0)$ at fixed T/T_K as a function of u. The second-order HZ perturbation theory prediction is distinctly nonuniversal. The quantum Monte Carlo, classic-maximum-entropy (QMC-ME) resistivity ratio obtained with the HZ default model is universal within one standard deviation statistical error. The resistivity ratio obtained with the Lorentzian ($DM=L$) model has error bars which are much too large to prove universality.

D. Automatic noise scaling

For Fig. 12, we used an additional feature of classic maximum entropy, automatic noise scaling, which was explained in Sec. III. Specifically, one chooses the noise scaling to enforce the condition $N_d=\chi^2+N_g$. This feature is useful because of the difficulty in estimating the errors in the quantum Monte Carlo data. In practice, we have found that the noise scaling varied from 1.0 to 1.6 for most data sets, with 1.1–1.2 being typical values. Generally, data sets having the smallest $(\Delta\tau)^2\Gamma U$ had the best statistics. Conversely, large noise scalings (≥ 2) appeared to be a sure sign of pathology in the QMC data. We would also get large noise scalings when $(\Delta\tau)^2\Gamma U \geq 0.2$ indicating a breakdown of the Trotter approximation. Large noise scalings would also occur when the number of binned measurements was much too small. However, we have not yet developed a systematic analysis of the correlation between the noise scaling and the behavior of the QMC algorithm.

Without the automatic noise scaling feature, we would have had a tendency to overfit the data, and we would have obtained ringing images for many of our data sets with large error estimates on the resistivity ratio. In cases where large noise scalings occurred, we would rerun the QMC simulations until the noise scaling was less than 1.5. It proved expensive to do this for large U or at very small T, which limited the range of Anderson model parameters we could reliably calculate.

VII. CONCLUSIONS AND DISCUSSION

Using the classic method of maximum entropy, we show how to extract, with no adjustable parameters, spectral and transport properties from the imaginary-time correlation function data generated by QMC simulations of the nondegenerate, symmetric, single-impurity Anderson model. The physical quantities obtained were found to be universal functions when the frequency and temperature are scaled by the Kondo temperature. This universality is a striking feature, not found in perturba-

tion theory, but found in experiment, and provides the benchmark for our claim of successful extractions. We suggest that our techniques are not restricted to the Hamiltonian and quantum Monte Carlo algorithm used here, but that maximum entropy and these techniques lay the general groundwork for the extraction of dynamical information from imaginary-time data generated by other quantum Monte Carlo simulations.

We showed that the quality of classic-maximum-entropy images depends on both the quality of our prior knowledge and the quality of the data. Given sufficiently high-quality data (defined in terms of maximum completeness, minimal statistical errors, and minimal systematic errors), the default model is irrelevant, and the correct result will always be obtained. Of course, one might say that such a statement is true of any analytic continuation procedure. Conversely, in the absence of data maximum entropy simply returns the default model. The advantage of maximum entropy over other analytic continuation procedures comes when we consider real data, which is always more incomplete and subject to statistical errors than we would like. Our experience is that, compared with other analytic continuation procedures,[1,40] maximum entropy can pull more information out of the data available and can reduce the cost required to generate adequate data by orders of magnitude.

While the error estimates provided by classic maximum entropy are a great aid in solution assurance, there are no guarantees. Ultimately, a decision on whether a good solution has been found depends on physical judgment, experience with the use of the maximum-entropy procedure, and experience with the data generation process. Does the image found make physical sense? Does the structure obtained correspond to the typical energy and length scales in the problem? If not, additional prior knowledge is needed to constrain the problem or to choose the model. Does the image depend strongly on small changes in χ^2/N_d? If so, it is unlikely that details of the structure are statistically significant. This can be checked by computing the errors on the image. Needed are either higher quality data or a more informative model. Are the error rescalings large? If so, there may be a problem in the process of estimating the statistical errors on the data. The measurements may not be Gaussian distributed or there may be roundoff errors. Is the model returned as the image with negligible error estimates? Although one may have been wise enough to know the result of the calculation beforehand, in this instance it is more probable that the quality of the data was too poor to provide anything new. To test for this possibility, one should try a different default model and see whether the image tracks it. A strong dependence of the image on the model is a sure sign of poor data quality. Does the image depend systematically on supposedly irrelevant parameters of the quantum Monte Carlo run, e.g., the Trotter $\Delta\tau$ parameter? If so, systematic errors in the data generation process, which are outside the domain of maximum entropy, may exist. We have encountered all of these pathologies at various stages in the development of the maximum-entropy approach to extracting the physical properties of the symmetric, single-impurity Anderson model.

The reader may have noticed that classic maximum entropy has a tendency to overfit the data. In particular, consider the equivalence of the $\chi^2/N_d=0.93$ curve in Fig. 4 and the classic-maximum-entropy result for the Lorentzian default model in Fig. 8. To get a stable (i.e., nonringing) result from classic maximum entropy, we had to input more prior knowledge in the form of a more informative default model from HZ perturbation theory. However, we might be strongly tempted to avoid the need for either inputing more prior knowledge or obtaining higher-quality quantum Monte Carlo data simply by relaxing α. After all, the historic-maximum-entropy image obtained with the Lorentzian default model is essentially equivalent to the classic-maximum-entropy image obtained with the HZ default model, as is shown in Fig. 5. While this behavior is true generally and it is easy to adjust α, we pay a price in having to introduce an arbitrary choice of stopping criterion and the loss in the ability to make reliable error estimates. However, we admit to a certain level of *cheating* in the sense that we have often compared the classic maximum-entropy image to the historic-maximum-entropy image in order to assess whether classic maximum entropy may be overfitting the data. In this case, the historic maximum-entropy image may provide a better hint of what the correct answer may be. Such comparisons point to the need for improving either our prior knowledge or our QMC data for the classic-maximum-entropy analysis.

Such overfitting by classic maximum entropy is a general phenomenon, not unique to the analytic continuation problem. Gull and Skilling[12] have recently shown that an additional hypothesis that the image is *locally smooth* is required to overcome this tendency. They have introduced a Bayesian approach to calculating *intrinsic correlation functions* which minimizes this overfitting.[22,24] Unlike the attempts to introduce smoothing in least-squares approaches,[9,10] where the choice of smoothing parameter is *ad hoc*, in the Bayesian approach the smoothing parameter is determined from the data and so there are no adjustable parameters. In addition, this method provides for meaningful error estimates for individual points in the spectral function, whereas classic maximum entropy only provides finite error estimates on integrated properties of the spectral function. For details, we refer the reader to their articles, for we have not yet implemented this method for the Anderson model. It was unnecessary given the high quality of our Horvatić-Zlatić default model, but we expect this recent development to be important in cases where informative default models are difficult to obtain.

ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy. We thank the Ohio Supercomputing Center for computational time. R.N.S. and D.S.S. also thank OBES/DMS for support of the Manuel Lujan, Jr. Neutron Scattering Center.

APPENDIX

We summarize two numerical algorithms for solving the maximum-entropy equations.[40] We closely followed the suggestions of Bryan and Skilling.[38,41] The algorithms maximize $Q=\alpha S-L$ by solving $\nabla Q=0$ with respect to the image $A(\omega)$ and the parameter α where

$$S=\sum_i [A_i-m_i-A_i\ln(A_i/m_i)] ,$$

$$L=\tfrac{1}{2}\sum_i \frac{(\bar{G}_i-\sum_j K_{ij}A_i)^2}{\sigma_i^2} ,$$

$\bar{G}_i$ is the data with standard errors σ_i, K_{ij} is the kernel which relates G_i and A_i, and $A_i=A(\omega_i)\Delta\omega_i$ with $\Delta\omega_i$ being the appropriate integration weight associated with a discrete frequency ω_i.

Both algorithms proceed by a Newton-Raphson search, but they differ in the space in which the search is conducted. Most of the results presented in this paper were analyzed by an iterative search in the image space $\{A_i\}$. We will describe this method first. Recently, we implemented a method which searches in the singular space of the kernel K_{ij}.[38] This second algorithm will also be described.

At each step in the iterative, image space algorithm, the current estimate of the image A_i is treated as a vector $\mathbf{A}$ in an r-dimensional space. The new image is generated from the old one by a Newton-Raphson step

$$\delta\mathbf{A}=-(\nabla\nabla Q)^{-1}\cdot\nabla Q$$
$$=-(\alpha\nabla\nabla S-\nabla\nabla L)^{-1}\cdot\nabla Q .$$

Since $-\nabla\nabla S=(1/A_i)\delta_{ij}$, the natural metric for the search in the image space is $g_{ij}=(1/A_i)\delta_{ij}$. With this metric surprisingly few directions are needed to approximately span the space defined by ∇Q. The algorithm proceeds to maximum Q by searching along the directions generated from a binomial expansion of the matrix $(\alpha\nabla\nabla S-\nabla\nabla L)^{-1}$. In our case, we found three directions to be sufficient:

$$\mathbf{e}_1=A(\nabla S) ,$$
$$\mathbf{e}_2=A(\nabla L) ,$$
$$\mathbf{e}_3=A\{\nabla\nabla L\cdot[A(\nabla S)/|\nabla S|-A(\nabla L)/|\nabla L|]\} .$$

Here $A(B)$ is a shorthanded notation for the vector A_iB_i. Because of the metric, each vector ∇S and ∇L and each matrix $\nabla\nabla S$ and $\nabla\nabla L$ are multiplied by the $g^{ij}=A_i\delta_{ij}=g_{ij}^{-1}$ to bring them into the same space as the image. Since $g_{ij}=-\nabla\nabla S$, the directions $A[\nabla\nabla S\cdot A(\nabla S)]\propto A(\nabla S)$ and $A[\nabla\nabla S\cdot A(\nabla L)]\propto A(\nabla L)$ are not independent search directions. Q is maximized in the three-dimensional space using a quadratic approximation

$$Q(x)\approx Q_0+\sum_\mu \mathbf{e}_\mu^\dagger\cdot\nabla Q x^\mu+\tfrac{1}{2}\sum_{\mu\nu}\mathbf{e}_\mu^\dagger\cdot\nabla\nabla Q\cdot\mathbf{e}_\nu x^\mu x^\nu .$$

The location of the maximum $x^\mu_{\max}$ yields the new image

$$A^{\text{new}}=A+\sum_\mu x^\mu_{\max}e_\mu .$$

This procedure is repeated until it converges to a final value of the image $\mathbf{A}(\alpha)$ for a given value of α.

The parameter α is determined by maximizing the posterior probability of α

$$P[\alpha|\bar{G},m]\propto\frac{Z_Q}{Z_SZ_L}P[\alpha]$$

with respect to $\ln\alpha$. The prior probability $P[\alpha]$ is assumed to be relatively flat as a function of $\ln\alpha$ and its variation with $\ln\alpha$ is sometimes neglected. The partition functions Z_Q, Z_S, and Z_L are defined so that the metric g modifies their differential measure. For example,

$$Z_Q=\int\frac{d^rA}{\Pi_i A_i^{1/2}}e^Q ,$$

with Z_S and Z_L being defined similarly. Z_L is independent of α. The integral for Z_S is done by expanding S to second order in a Taylor series and then approximating the integral using the method of steepest descents. With this approximation,[12]

$$Z_S=\frac{r}{2}\ln\left[\frac{\alpha}{2\pi}\right] .$$

The integral for Z_Q is done in a similar manner. Here the matrix $\Lambda_{ij}=A_i^{1/2}(\nabla\nabla L)_{ij}A_j^{1/2}$ with eigenvalues λ_i is defined. The maximum of the resulting ratio is then found by solving

$$\frac{\partial(Z_Q/Z_L)}{\partial\ln(\alpha)}=0 ,$$

with the result [Eq. (22)]

$$-2\alpha S=\sum_j \lambda_j/(\alpha+\lambda_j) .$$

This equation is solved for a new value of α, and the procedure described above is reused to find a new image. These steps are repeated until they converge to a final image and α.

Recently, Bryan[38] introduced an algorithm that is more efficient for oversampled problems where N_g is small. Such problems can arise when the kernel is exponential in nature as in the present case. With this method, the kernel is decomposed with a singular value decomposition $\mathbf{K}=\mathbf{V}\mathbf{\Sigma}\mathbf{U}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices, and $\mathbf{\Sigma}$ is the diagonal matrix which contains the singular values. The search then proceeds by a Newton-Raphson iteration, very similar to that described above; however, here the search space is composed of the column space of $\mathbf{U}$, plus the remaining independent direction $\mathbf{A}(\nabla S)$. Specifically, the image $\mathbf{A}$ is related to the search vector $\mathbf{u}$ through the relation

$$A_i/m_i=\exp\left[\sum_j U_{ij}u_j\right] .$$

For problems involving exponential kernels $\mathbf{K}$, this method may be superior to the image-space methods for two reasons. First, the most important search directions

in the column space of **U** are associated with the largest of the singular values (the diagonal elements of Σ). For an exponential kernel, the singular values will fall off exponentially fast, so that in the most extreme cases, only one direction (one column of **U**) may be important. When rotated back into image space, this direction could, in principle, be orthogonal to all three of the search directions used in the image-space code. Second, the singular space is always much smaller than the image space. Thus this algorithm has a much smaller space to search for the maximum of Q, and preferentially searches in the most important directions and is usually more efficient for problems involving a singular kernel than the image-space algorithm.

In both algorithms, we find the image **A** (or **u**) which maximizes $Q=\alpha S-L$, with α chosen so that it maximizes the posterior probability $P[\alpha|\overline{G},m]$. However, for the analytic continuation problems we have so far studied, $P[\alpha|\overline{G},m]$ is often not sharply peaked; rather, it can be a broad distribution, heavily skewed to large values of α (Fig. 6). In such cases, the mode and mean of the distribution differ, and it is proper to integrate the image over the posterior probability to obtain the average image,

$$\langle \mathbf{A} \rangle = \int d\alpha\, P[\alpha|\overline{G},m]\, \mathbf{A}(\alpha) .$$

Here $\mathbf{A}(\alpha)$ is the image that maximizes Q for fixed α. In the cases, we have studied, this mean image has a slightly less tendency to ring than the mode image; however, as discussed in the manuscript, this could be because the mean value of α is larger than the mode value.

For fixed α, once we have the image which maximizes Q, we can estimate the error δB of an integrated function B of the image

$$B=\sum_i A_i P_i .$$

The function $P_i=P(\omega_i)$ may, for instance, equal ω_i^n and hence B would be the nth moment of the image.

At its maximum, Q is well described by a quadratic expansion in the image space

$$Q \approx Q_0+\delta\mathbf{A}^\dagger\cdot\nabla Q+\tfrac{1}{2}\delta\mathbf{A}^\dagger\cdot\nabla\nabla Q\cdot\delta\mathbf{A} .$$

Then using the method of steepest descents to evaluate the integral, one may show that

$$\begin{aligned}\mathbf{E}&=\langle \delta\mathbf{A}\delta\mathbf{A}^\dagger\rangle \\ &=\frac{1}{Z_Q}\int\frac{d^rA}{\Pi_i A_i^{1/2}}\delta\mathbf{A}\delta\mathbf{A}^\dagger e^Q \\ &\approx -(\nabla\nabla Q)^{-1} .\end{aligned}$$

The matrix **E** is the covariance of the image **A**. In order to propagate this error to B, we need to work in a representation where this covariance is diagonal. Thus we find an orthogonal matrix **O** such that $\mathbf{O}^{-1}\mathbf{E}\mathbf{O}=\mathbf{D}$ is diagonal, then we define a vector **d** such that

$$\mathbf{d}=\mathbf{O}^{-1}\mathbf{A}, \quad B(\alpha)=\sum_{ij} P_i O_{ij} d_j .$$

The covariance for **d** is the matrix **D**, which is diagonal, so that the errors of d_i and d_j are uncorrelated, thus $\delta d_n^2=D_n$:

$$\begin{aligned}\delta B(\alpha)^2&=\sum_n\left[\frac{\partial B(\alpha)}{\partial d_n}\right]^2\delta d_n^2 \\ &=\sum_n\left[\sum_i P_i O_{in}\right]^2 D_n .\end{aligned}$$

If we want an estimate of the error for the mean image $\langle \mathbf{A}\rangle$, as described above, then we must integrate over the posterior probability distribution of α

$$\delta B^2=\int d\alpha\, P[\alpha|\overline{G},m][\delta B(\alpha)^2-B(\alpha)^2] + \left[\int d\alpha\, P[\alpha|\overline{G},m]B(\alpha)\right]^2 .$$

[1] R. N. Silver, D. S. Sivia, and J. E. Gubernatis, Phys. Rev. B **41**, 2380 (1990); and in *Quantum Simulations of Condensed Matter Phenomena,* edited by J. E. Gubernatis and J. D. Doll (World Scientific, Singapore, 1990), p. 340.

[2] R. N. Silver, D. S. Sivia, J. E. Gubernatis, and M. Jarrell, Phys. Rev. Lett. **65**, 496 (1990).

[3] M. Jarrell, J. E. Gubernatis, R. N. Silver, and D. S. Sivia, Phys. Rev. B **43**, 1206 (1991).

[4] M. Jarrell, J. E. Gubernatis, and R. N. Silver, Phys. Rev. B **44**, 5347 (1991).

[5] R. N. Silver, J. E. Gubernatis, D. S. Sivia, and M. Jarrell, in *Condensed Matter Theories 6,* edited by S. Fantoni and A. Fabrocini (Plenum, New York, 1991), pp. 189–202.

[6] A brief review is given by B. J. Berne and D. Thirumalai, Annu. Rev. Chem. Phys. **47**, 401 (1986). More recent works include J. D. Doll, R. D. Coalson, and D. L. Freeman, J. Chem. Phys. **87**, 1641 (1987); and J. D. Doll, D. L. Freeman, and T. L. Beck, Adv. Chem. Phys. **78**, 61 (1990), and references therein.

[7] J. E. Hirsch, in *Quantum Monte Carlo Methods,* edited by M. Suzuki (Springer-Verlag, New York, 1987), p. 205.

[8] H.-B. Schüttler and D. J. Scalapino, Phys. Rev. Lett. **55**, 1204 (1985); Phys. Rev. B **34**, 4744 (1986).

[9] S. R. White, D. J. Scalapino, R. L. Sugar, and N. E. Bickers, Phys. Rev. Lett. **63**, 1523 (1989).

[10] M. Jarrell and O. Biham, Phys. Rev. Lett. **63**, 2504 (1989).

[11] For tutorial introductions, see E. T. Jaynes, in *Maximum Entropy and Bayesian Methods,* edited by J. H. Justice (Cambridge University Press, Cambridge, England, 1986); S. F. Gull, in *Maximum Entropy and Bayesian Methods in Science and Engineering,* edited by G. J. Erickson and C. R. Smith (Kluwer Academic, Dordrecht, 1988); D. S. Sivia, Los Alamos Sci. **19**, 180 (1990), and references therein.

[12] J. Skilling, in *Maximum Entropy and Bayesian Methods,* edited by J. Skilling (Kluwer Academic, Dordrecht, 1989), p. 45; S. F. Gull, *ibid.*, p. 53.

[13] J. Deisz, M. Jarrell, and D. L. Cox, Phys. Rev. B **42**, 4869 (1990).

[14]M. Jarrell, D. S. Sivia, and B. Patton, Phys. Rev. B **42**, 4804 (1990).
[15]S. R. white (unpublished).
[16]J. Deisz, K.-H. Luk, M. Jarrell, and D. L. Cox (unpublished).
[17]S. W. Lovesey, *Condensed Matter Physics: Dynamic Correlations* (Benjamin/Cummings, Reading, MA, 1980).
[18]R. Kubo, M. Toda, and N. Hashitsume, *Statistical Physics II* (Springer-Verlag, New York, 1978).
[19]S. Doniach and E. H. Sondheimer, *Green's Functions for Solid State Physicists* (Benjamin, Reading, MA, 1974).
[20]General references to this section of the paper are given in Refs. 11 and 12.
[21]R. T. Cox, Am. J. Phys. **14**, 1 (1946).
[22]J. Skilling and S. Sibisi, in Neutron Scattering Data Analysis 1990, edited by M. W. Johnson, IOP Conf. Proc. No. 107 (Institute of Physics and Physical Society, London, 1990), p. 1.
[23]M. K. Charter, in *Maximum Entropy and Bayesian Methods,* edited by P. F. Fougère (Kluwer Academic, Dordrecht, 1990), p. 325.
[24]A. Papoulis, *Probability and Statistics* (Prentice-Hall, New York, 1990), p. 422.
[25]P. W. Anderson, Phys. Rev. **124**, 41 (1961).
[26]H. R. Krishna-murthy, J. W. Wilkins, and K. G. Wilson, Phys. Rev. B **21**, 1021 (1980); **21**, 1044 (1980).
[27]N. E. Bickers, D. L. Cox, and J. W. Wilkins, Phys. Rev. B **36**, 2036 (1987).
[28]V. Zlatić, G. Grünier, and N. Rivier, Solid State Commun. **14**, 639 (1974).
[29]D. C. Langreth, Phys. Rev. **130**, 516 (1964).
[30]B. Horvatić, D. Šokčević, and V. Zlatić, Phys. Rev. B **36**, 675 (1987).
[31]J. E. Hirsch and R. M. Fye, Phys. Rev. Lett. **56**, 2521 (1986).
[32]H. F. Trotter, Proc. Am. Math. Soc. **10**, 545 (1959).
[33]M. Suzuki, Commun. Math. Phys. **51**, 183 (1976).
[34]W. W. Wood, in *Physics of Simple Liquids,* edited by N. V. Temperley, J. S. Rowlinson, and G. S. Rushbrooke (North-Holland, Amsterdam, 1968), Chap. 5.
[35]H. Müller-Krumbhaar and K. Binder, J. Stat. Phys. **8**, 1 (1973).
[36]W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes* (Cambridge University Press, Cambridge, England, 1986), Chap. 13.
[37]A. M. Tsvelick and P. B. Wiegmann, Adv. Phys. **34**, 453 (1983); P. B. Wiegmann and A. M. Tsvelick, J. Phys. C **16**, 2281 (1983).
[38]R. K. Bryan, Eur. Biophys. J. **18**, 165 (1990).
[39]S. R. White (private communication).
[40]Most of the results presented in this paper were obtained by using MEMSYS3, produced by Maximum Entropy Data Consultants, Ltd. We have implemented, and now frequently use, however, the algorithms described in the Appendix.
[41]J. Skilling and R. K. Bryan, Mon. Not. R. Astron. Soc. **211**, 111 (1984).