Average spectrum method for analytic continuation: Efficient blocked-mode sampling and dependence on the discretization grid

Khaldoon Ghanem[1,2] and Erik Koch[1,3]
[1]*Jülich Supercomputer Centre, Forschungszentrum Jülich, 52425 Jülich, Germany*
[2]*Max-Planck-Institut für Festkörperforschung, 70569 Stuttgart, Germany*
[3]*JARA High-Performance Computing, 52425 Jülich, Germany*

(Received 4 December 2019; accepted 22 January 2020; published 10 February 2020)

The average spectrum method is a promising approach for the analytic continuation of imaginary time or frequency data to the real axis. It determines the analytic continuation of noisy data from a functional average over all admissible spectral functions, weighted by how well they fit the data. Its main advantage is the apparent lack of adjustable parameters and smoothness constraints, using instead the information on the statistical noise in the data. Its main disadvantage is the enormous computational cost of performing the functional integral. Here we introduce an efficient implementation, based on the singular value decomposition of the integral kernel, eliminating this problem. It allows us to analyze the behavior of the average spectrum method in detail. We find that the discretization of the real-frequency grid, on which the spectral function is represented, biases the results. The distribution of the grid points plays the role of a default model while the number of grid points acts as a regularization parameter. We give a quantitative explanation for this behavior, point out the crucial role of the default model and provide a practical method for choosing it, making the average spectrum method a reliable and efficient technique for analytic continuation.

DOI: 10.1103/PhysRevB.101.085111

I. INTRODUCTION

Strongly interacting quantum many-particle problems require nonperturbative solvers. Quantum Monte Carlo (QMC) approaches provide, in the absence of a sign problem, numerically exact results and are therefore widely used. Their key drawback is that they work well only for imaginary time or frequency. To make contact with experiment these data have to be analytically continued to obtain the spectral function $A(\omega)$ on the real-frequency axis. This requires solving an integral equation, presenting an ill-posed inverse problem. The standard approach to this problem for strongly correlated electron systems is the maximum entropy method (MaxEnt) described in Ref. [1], which is, with some variations, also used in Eliashberg theory [2] as well as in lattice QCD simulations [3].

The ill-posedness of the inverse problem implies that the spectral function $A(\omega)$ giving the best fit to the imaginary-axis data in a least-squares sense, while easily determined, is completely useless: it is dominated by rapid oscillations of diverging amplitude, arising from fitting the inevitable statistical noise in the QMC data. The standard approach for overcoming this problem is to impose smoothness on the solution, i.e., to regularize [4]. The maximum entropy method provides a regularization based on Bayesian arguments. It penalizes deviations of the spectral function from a default model, measured by the relative entropy of the two functions. While the nonlinearity of the entropy function makes optimization more difficult, it has the important advantage of ensuring the non-negativity of the spectral function. The method provides good results and is so efficient that it is the de facto standard for analytic continuation problems. Still there remains the problem of choosing an appropriate default model and regularization parameter, the latter giving rise to a number of different flavors of MaxEnt [5].

An alternative approach, the average spectrum method (ASM), that promises to avoid these ambiguities was proposed by White [6] and, independently, in Refs. [7,8]. The basic idea is of striking elegance: the spectral function is obtained as the average of all physically admissible spectral functions weighted with how well they fit the data given on the imaginary axis. Due to the ill-posedness of the inverse problem there are many spectral functions that differ drastically but fit the data equally well. Taking the average is thus expected to smooth out features that are not supported by the data, providing a regularization without the need for explicit parameters. The practical application of this conceptually appealing approach has, however, so far suffered from the computational cost of its implementations [6–10].

Here we introduce the blocked-mode sampling technique, which overcomes the main limitation of the average spectrum method: The commonly used recipe is to update the sampled spectral function at several points simultaneously, keeping a number of moments of $A(\omega)$ fixed [7,8]. Our more systematic approach introduces global moves, updating not individual components of $A(\omega)$, but changing it at all frequencies at once by an amount proportional to a singular mode of the kernel. This is very efficient when the global moves are not constrained too much by the non-negativity of $A(\omega)$. When the constraint limits these moves significantly it becomes more efficient to partition the frequency axis and perform global moves on the individual frequency blocks.

 ©2020 American Physical Society

Blocked-mode sampling makes the average spectrum method fast enough that we can systematically investigate how well it performs the analytic continuation. We find that the results depend on the way the real-frequency axis is discretized: the density function used for picking grid points acts as a default model, i.e., determines the result in the absence of data, while the number of grid points acts as a regularization parameter. That the ASM includes, via the parametrization of the real axis, a default model has already been noticed in Refs. [10,11], while in Ref. [12], it was observed that the results of the ASM are becoming more biased with increasing number of grid points. We find an explanation for this, which provides us with ways to undo the effect of a specific grid. Moreover, we develop a method for judging the reliability of the results of the average spectrum method, making it a reliable approach to analytic continuation.

II. AVERAGE SPECTRUM METHOD

The average spectrum method is designed to solve linear integral equations of the form

$$g(y) = \int K(y,x)\, f(x)\, dx \tag{1}$$

for $f(x)$. Calculating $g(y)$ given $f(x)$ merely involves a numerically stable integration. The inverse problem, on the other hand, is ill-conditioned since it is numerically hard to reconstruct sharp features in $f(x)$ that enter $g(y)$ only after being integrated over. That becomes harder the smoother the kernel $K(y,x)$ as a function of x. The problem is further complicated by the fact that $g(y)$ is usually determined by Monte Carlo methods, i.e., it is only known within the statistical errors of the simulation.

An important application is the determination of the spectral function $A(\omega)$ from the finite-temperature Green function at the fermionic Matsubara frequencies $\omega_m = (2m+1)\pi/\beta$

$$G(i\omega_m) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1}{i\omega_m - \omega} A(\omega)\, d\omega\,, \tag{2}$$

at imaginary times $[\tau \in (0,\beta)]$,

$$G(\tau) = -\frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-\omega\tau}}{1+e^{-\beta\omega}} A(\omega)\, d\omega\,, \tag{3}$$

or the coefficients $G_l = \sqrt{2l+1}\int_0^\beta P_l(2\tau/\beta - 1)\, G(\tau)\, d\tau$ of its expansion in Legendre polynomials $P_l(x)$ [13]

$$G_l = (-1)^{l+1}\frac{\sqrt{2l+1}}{2\pi}\frac{\beta}{2}\int_{-\infty}^{\infty} \frac{i_l^{(1)}(\beta\omega/2)}{\cosh(\beta\omega/2)} A(\omega)\, d\omega \tag{4}$$

where $i_l^{(1)}(x)$ are the modified spherical Bessel functions of first kind [14].

Another important application is the determination of the susceptibility $\chi''(\omega)$ from the correlation function at the bosonic Matsubara frequencies $\omega_m = 2m\pi/\beta$

$$\Pi(i\omega_m) = \frac{2}{\pi}\int_0^{\infty} \frac{\omega^2}{\omega_m^2+\omega^2}\frac{\chi''(\omega)}{\omega}\, d\omega\,, \tag{5}$$

imaginary times

$$\Pi(\tau) = \frac{1}{\pi}\int_0^{\infty} \omega\, \frac{e^{-\omega\tau}+e^{+\omega\tau}}{1-e^{-\beta\omega}}\frac{\chi''(\omega)}{\omega}\, d\omega\,, \tag{6}$$

or its Legendre expansion, which vanishes for odd l, while for even l

$$\Pi_l = \frac{\sqrt{2l+1}}{\pi}\beta\int_0^{\infty} \omega\, \frac{i_l^{(1)}(\beta\omega/2)}{\sinh(\beta\omega/2)}\frac{\chi''(\omega)}{\omega}\, d\omega. \tag{7}$$

In all these cases, the function $A(\omega)$ or $\chi''(\omega)/\omega$ to be determined is known to be non-negative.

In practice the QMC data is given as a discrete vector $\mathbf{g} = (g_1,\ldots,g_M)^\dagger$ of M data points. The mean over K samples is

$$\overline{\mathbf{g}} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{g}_k \tag{8}$$

and its statistical uncertainty, when the samples are uncorrelated, is characterized by the covariance matrix

$$\mathbf{C} = \frac{1}{K(K-1)}\sum_k (\mathbf{g}_k - \overline{\mathbf{g}})(\mathbf{g}_k - \overline{\mathbf{g}})^\dagger. \tag{9}$$

By the central limit theorem, the probability density of measuring $\overline{\mathbf{g}}$ instead of the exact result $\mathbf{g}_{\text{exact}}$ is proportional to $\exp(-(\overline{\mathbf{g}} - \mathbf{g}_{\text{exact}})^\dagger \mathbf{C}^{-1}(\overline{\mathbf{g}} - \mathbf{g}_{\text{exact}})/2)$.

Given some function $f(x)$, it is straightforward to calculate the corresponding $g[f](y)$ by integration (1) and discretizing it to obtain $\mathbf{g}[f]$. Assuming that $f(x)$ is the exact model, the probability density for measuring $\overline{\mathbf{g}}$ given covariance $\mathbf{C}$ is

$$p(\overline{\mathbf{g}}|f,\mathbf{C}) \propto e^{-\frac{1}{2}(\overline{\mathbf{g}}-\mathbf{g}[f])^\dagger\,\mathbf{C}^{-1}\,(\overline{\mathbf{g}}-\mathbf{g}[f])} =: e^{-\frac{1}{2}\chi^2[f]}\,. \tag{10}$$

The idea of the average spectrum method is to average all functions $f(x)$ with the probability that they are the exact model, given the measured data $(\overline{\mathbf{g}},\mathbf{C})$, i.e., to perform the functional integral

$$f_{\text{ASM}}(\overline{\mathbf{g}},\mathbf{C};x) = \int \mathcal{D}f\; p(f|\overline{\mathbf{g}},\mathbf{C})\; f(x). \tag{11}$$

By the Bayes theorem, the posterior probability density is

$$p(f|\overline{\mathbf{g}},\mathbf{C}) = \frac{p(\overline{\mathbf{g}}|f,\mathbf{C})\, p(f)}{p(\overline{\mathbf{g}}|\mathbf{C})}\,, \tag{12}$$

where the likelihood is given by (10), $p(f)$ is the prior probability density, and $p(\overline{\mathbf{g}}|\mathbf{C}) = \int \mathcal{D}f\; p(\overline{\mathbf{g}}|f,\mathbf{C})\, p(f)$ is the normalization. For the spectral function and susceptibilities we know that f is non-negative. Setting the prior probability to zero for models that violate this constraint and constant otherwise, (11) becomes

$$f_{\text{ASM}}(\overline{\mathbf{g}},\mathbf{C};x) \propto \int_{f(x)\geqslant 0} \mathcal{D}f\; e^{-\frac{1}{2}\chi^2[f]}\; f(x)\,. \tag{13}$$

Estimating $f(x)$ just requires performing an integral over non-negative models while there is no need for any adjustable parameters. Instead, the regularization results exclusively from the uncertainty in the data as given by the covariance $\mathbf{C}$: the larger the statistical noise, the stronger the contribution of models that do not fit the data particularly well. We can thus expect that accurate data will give us spectra with sharp features, while for noisy data the spectra will contain less information, being more smoothed out by the averaging [6–8].

III. TEST CASES

For illustrating how the average spectrum method performs we use the test cases introduced in Ref. [15]: we try to reconstruct an optical conductivity given by

$$\sigma(\omega) = \frac{1}{1+(\omega/\Gamma_e)^6} \sum_{p=0,\pm1} \frac{W_{|p|}}{1+((\omega+\mathrm{sgn}(p)\varepsilon_{|p|})/\Gamma_{|p|})^2}, \tag{14}$$

where the overall factor with $\Gamma_e = 4$ cuts off $\sigma(\omega)$ for large frequencies and the terms in the sum give a (Drude) peak of weight $W_0 = 0.3$ and width $\Gamma_0 = 0.3$ (model 1) or 0.6 (model 2), and two symmetric peaks of weight $W_1 = 0.2$ and width $\Gamma_1 = 1.2$ centered at $\omega = \pm\varepsilon_1 = \pm 3$. The corresponding correlation function on the bosonic Matsubara frequencies $i\omega_m = 2\pi m i/\beta$

$$\Pi(i\omega_m) = \frac{2}{\pi}\int_0^\infty d\omega\, \frac{\omega^2}{\omega_m^2+\omega^2}\,\sigma(\omega) \tag{15}$$

can be calculated analytically. The input data for the analytic continuation is the imaginary-frequency correlation function $\Pi_m = \Pi(i\omega_m)(1+r_m)$ on the first 60 Matsubara frequencies $m = 0, \ldots, 59$ with Gaussian (relative) noise r_m of variance σ_Π, where $\sigma_\Pi = 0.01$ (noisy data) or 0.001 (accurate data). The inverse temperature is $\beta = 15$.

IV. BLOCKED-MODE SAMPLING

To evaluate the functional integral (13) numerically, we discretize $f(x)$. Introducing a grid of N intervals, we can, e.g., represent it as a piecewise constant function of value f_n on interval n: $\mathbf{f} = (f_1, \ldots, f_N)^\mathsf{T}$. The integral equation (1) then becomes a linear equation

$$\mathbf{g} = \mathbf{K}\mathbf{f} \tag{16}$$

and the functional $\chi^2[f]$ is approximated by

$$\chi^2(\mathbf{f}) = (\overline{\mathbf{g}} - \mathbf{K}\mathbf{f})^\dagger\, \mathbf{C}^{-1}\, (\overline{\mathbf{g}} - \mathbf{K}\mathbf{f}). \tag{17}$$

It is then easy to modify (17) such that the covariance matrix no longer appears explicitly. For this we factorize $\mathbf{C}^{-1} = \mathbf{T}^\dagger\mathbf{T}$, e.g., by Cholesky decomposition, to obtain

$$\chi^2(\mathbf{f}) = (\tilde{\mathbf{g}} - \tilde{\mathbf{K}}\mathbf{f})^\dagger\, (\tilde{\mathbf{g}} - \tilde{\mathbf{K}}\mathbf{f}) = \|\tilde{\mathbf{g}} - \tilde{\mathbf{K}}\mathbf{f}\|^2 \tag{18}$$

with $\tilde{\mathbf{g}} := \mathbf{T}\overline{\mathbf{g}}$ and $\tilde{\mathbf{K}} = \mathbf{T}\mathbf{K}$. The covariance $\tilde{\mathbf{C}}$ of the transformed data $\tilde{\mathbf{g}}$ is, by construction, the unit matrix.

The functional integral (13) is then estimated from

$$\mathbf{f}_{\mathrm{ASM}}(\tilde{\mathbf{g}}) \propto \prod_{n=1}^{N}\int_0^\infty df_n\, \mathbf{f}\, e^{-\frac{1}{2}\chi^2(\mathbf{f})}. \tag{19}$$

This N-dimensional integral can be evaluated by Monte Carlo techniques.

A. Component sampling

The straightforward method for evaluating (19) is to perform a random walk in the space of non-negative vectors $\mathbf{f}$, updating a single component, $f_n \to f_n'$, at a time. Detailed balance is fulfilled if we sample f_n' from the conditional distribution $\propto \exp(-\chi^2(\mathbf{f}; f_n')/2)$ with

$$\begin{aligned}\chi^2(\mathbf{f}; f_n') &= \|\underbrace{\tilde{\mathbf{g}} - \tilde{\mathbf{K}}\mathbf{f}}_{=:\tilde{\mathbf{r}}} - \tilde{\mathbf{K}}_n(f_n' - f_n)\|^2 \\ &= \tilde{\mathbf{K}}_n^\dagger\tilde{\mathbf{K}}_n\left(f_n' - f_n - \frac{\Re\tilde{\mathbf{K}}_n^\dagger\tilde{\mathbf{r}}}{\tilde{\mathbf{K}}_n^\dagger\tilde{\mathbf{K}}_n}\right)^2 + \tilde{\mathbf{r}}^\dagger\tilde{\mathbf{r}} - \frac{(\Re\tilde{\mathbf{K}}_n^\dagger\tilde{\mathbf{r}})^2}{\tilde{\mathbf{K}}_n^\dagger\tilde{\mathbf{K}}_n},\end{aligned} \tag{20}$$

where $\tilde{\mathbf{K}}_n$ is the nth column of $\tilde{\mathbf{K}}$. We thus have to sample f_n' from a univariate Gaussian of width $\sigma = 1/\|\tilde{\mathbf{K}}_n\|$ centered at $\mu = f_n + \Re\tilde{\mathbf{K}}_n^\dagger\tilde{\mathbf{r}}/\|\tilde{\mathbf{K}}_n\|^2$ and truncated to the non-negative values $f_n' \in [0, \infty)$. This can be done very efficiently [16].

Still, sampling components can be very slow because the width of the Gaussian is, in general, extremely small, i.e., the random walk performs only exceedingly small steps. This is evident when sampling spectral functions: we cannot change just a single f_n without violating the sum-rule. A common way out is to update several components simultaneously under the constraint that, e.g., a number of moments of $\mathbf{f}$ is conserved, and to use tempering techniques [6–10]. A simpler and more systematic way is to sample along the principal axes of the multivariate Gaussian $\exp(-\chi^2(\mathbf{f})/2)$, i.e., to change basis. This is illustrated in Fig. 1.

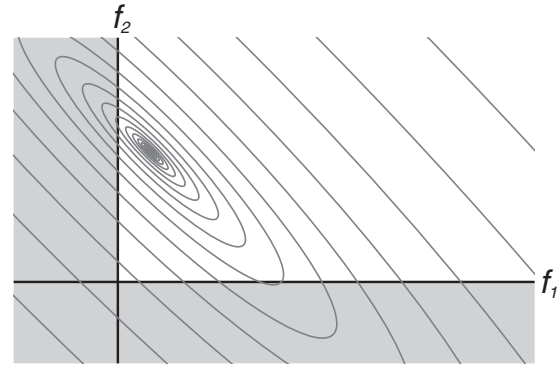

FIG. 1. Schematic contour plot of the Gaussian probability density $\exp(-\chi^2(\mathbf{f})/2)$ in the plane of two values f_1 and f_2. The unphysical region $\mathbf{f} < 0$ is shaded in gray. In components sampling the moves $f_i \to f_i'$ are proposed parallel to the coordinate axes, resulting in narrow Gaussians of widths that are of the order of $1/\max(\mathrm{d_i})$. In modes sampling, moves $e_i \to e_i'$ are proposed along the principal axes of the multivariate Gaussian, so that the moves in directions corresponding to small singular values can take large steps. Note that for ill-conditioned problems the singular values d_i vary over many orders of magnitude.

B. Mode sampling

To implement moves along the principal axes of χ^2, we use the singular value decomposition of the kernel $\tilde{\mathbf{K}} = \mathbf{U}\mathbf{D}\mathbf{V}^\mathsf{T}$, where $\mathbf{U}$ is a unitary matrix whose column vectors, $\mathbf{U}_m$, define a basis in the M-dimensional data space, $\mathbf{V}$ is a unitary matrix whose columns, $\mathbf{V}_n$, define a basis in the N-dimensional space of discretized models, and $\mathbf{D}$ is an $M \times N$ diagonal matrix with diagonal elements $d_1 \geqslant d_2 \geqslant \cdots \geqslant d_{\min(N,M)} \geqslant 0$. The singular values $d_n > 0$ determine how a mode in model space affects the data: $\tilde{\mathbf{K}}\mathbf{V}_n = d_n\mathbf{U}_n$,

while the zero modes $\mathbf{U}_n$ with $d_n = 0$ or $n > M$ do not affect the data. To simplify the notation we define $d_n := 0$ for $n = \min(N, M)+1, \ldots, \max(N, M)$.

Transforming to the new bases $\mathbf{h}:=\mathbf{U}^\dagger\tilde{\mathbf{g}}$ and $\mathbf{e}:=\mathbf{V}^\mathsf{T}\mathbf{f}$ diagonalizes the quadratic form

$$\chi^2(\mathbf{f}) = \|\mathbf{U}^\dagger\tilde{\mathbf{g}} - \mathbf{D}\mathbf{V}^\mathsf{T}\mathbf{f}\|^2 = \sum_{i=1}^{M}(h_i - d_i\, e_i)^2 \qquad (21)$$

and we can write (19) as $\mathbf{f}_{\mathrm{ASM}}(\tilde{\mathbf{g}}) = \mathbf{V}\mathbf{e}_{\mathrm{ASM}}(\mathbf{h})$, where the integral in the new basis factorizes

$$\mathbf{e}_{\mathrm{ASM}}(\mathbf{h})_i \propto \int_{\mathbf{f}\geqslant 0} de_i\, e_i\, \exp(-(d_i e_i - h_i)^2/2). \qquad (22)$$

For evaluating the integral, we perform a random walk, now updating one mode $e_i \to e_i'$ at a time. When the corresponding singular value does not vanish, we sample e_i' from a univariate Gaussian of width $\sigma = 1/d_i$ centered at h_i/d_i while for $d_i = 0$ we sample from a flat distribution. In both cases, the distribution is truncated to the interval for which $\mathbf{f}' \geqslant 0$.

Without the non-negativity constraint, the components of $\mathbf{e}_{\mathrm{ASM}}(\mathbf{h})$ for $d_i > 0$ would be given by h_i/d_i, resulting in a least-squares solution that, in general, would be completely dominated by the noise in data modes h_i with exceedingly small singular values. The coupling of the modes through the global condition $\mathbf{f} \geqslant 0$ is thus crucial for regularization.

We find the allowed values of e_i' from the condition $\mathbf{f}' = \mathbf{f} + (e_i' - e_i)\mathbf{V}_i \geqslant 0$, which, in terms of the components, is equivalent to $e_i' \geqslant e_i - f_n/V_{ni}$ for $V_{ni} > 0$ and correspondingly for $V_{ni} < 0$. Thus e_i' is constrained by

$$\max_n\left\{\frac{f_n}{V_{ni}}\middle| V_{ni} < 0\right\} \leqslant e_i - e_i' \leqslant \min_n\left\{\frac{f_n}{V_{ni}}\middle| V_{ni} > 0\right\}. \qquad (23)$$

Sampling modes e_i' is usually much more efficient than sampling components f_n': for modes with large singular value, the Gaussian is narrow so that the random walk quickly jumps close to the expected value h_i/d_i corresponding to the best fit, and then stays close to it. For modes with a small or a zero singular value, the distribution is very broad so that the random walk can take large steps, allowing for an efficient sampling of the degrees of freedom that are not strongly supported by the data.

Still, sampling may become quite inefficient when non-negativity restricts e_i' to a narrow interval. This will happen when $\mathbf{f}$ has regions where the f_n are very small. For a mode $\mathbf{V}_i$ that changes sign on such a region, e_i' cannot differ much from e_i without violating (23). Since the modes form a basis, there are many such modes. In particular, modes sampling can become quite slow when sampling spectral functions on grids with large cutoff. In the tail of the spectral function, where there are many small values f_n, it can be more efficient to sample the components f_n directly since they tend to change χ^2, Eq. (20), only little.

C. Blocked-mode sampling

The reason for the slow-down of modes sampling is that the narrow intervals originating from regions where the f_n are small also limit the changes in regions where they are large, i.e., where large steps could be taken. We can avoid this by decoupling such regions and sampling them separately. To do this, we split the kernel matrix $\mathbf{K}$ into blocks corresponding to the different regions, perform an SVD for each of them, and sample the resulting blocked modes. Now the non-negativity constraint (23) involves only components in the same region. Thus the intervals over which the blocked modes can be sampled will be larger than in modes sampling. On the other hand, the blocked modes no longer give the principal axes of the fit function χ^2 so that the Gaussians from which the modes are sampled will be more narrow than in modes sampling. When we choose the regions as just the individual grid points we are back to components sampling, where the intervals are semi-infinite $f_n \in [0, \infty)$, while the Gaussians become quite narrow.

FIG. 2. Example of the hierarchy of grid partitionings used in blocked-mode sampling. At the highest level (top), the grid on which $\mathbf{f}$ is represented forms a single block. Sampling on this block is modes sampling. At the level below, the grid is split into two blocks. If going to a lower level we split the blocks in half, there would always be a block boundary at the center of the grid. To avoid this, we shift the intervals at every other level by half their width. At the lowest level (bottom), the blocks are the individual intervals f_n. Sampling on these blocks is components sampling.

The idea of blocked-mode sampling is thus to exploit this trade-off between wide Gaussians and large intervals by interpolating between the limits of modes and components sampling. In practice, we use a hierarchy of partitionings of the grid as shown in Fig. 2 and sample in each step all blocks of a randomly chosen hierarchy level.

D. Efficiency

The computational complexity of the sampling methods per Monte Carlo step are comparable. For components sampling, calculating the Gaussian parameters, Eq. (20), for updating $f_n \to f_n'$ scales as $\mathcal{O}(MN)$ and there are N components to be updated. In modes sampling, the Gaussian parameters are given by the singular values, which are calculated only once, at the beginning of the simulation. Determining the constraint intervals, Eq. (23), takes $\mathcal{O}(N)$ operations, and there are N modes to be updated. In blocked-mode sampling, the singular value decompositions for all blocks are calculated once at the beginning. The computational cost of this is dominated by the SVD for the full block and scales as $\mathcal{O}(M\,N^2)$ when there are more grid than data points, $N > M$. Sampling a block of length N/B takes $\mathcal{O}(N/B)$ operations for determining the constraint intervals on the block plus $\mathcal{O}(MN(1 - 1/B))$ operations to calculate the contribution of the other blocks to the Gaussian parameters for each of the N/B modes in the block. There are B such blocks to be updated. Thus, the computational cost per Monte Carlo step is similar for all three approaches, so that their efficiency depends on how much the model $\mathbf{f}$ is changed per MC step.

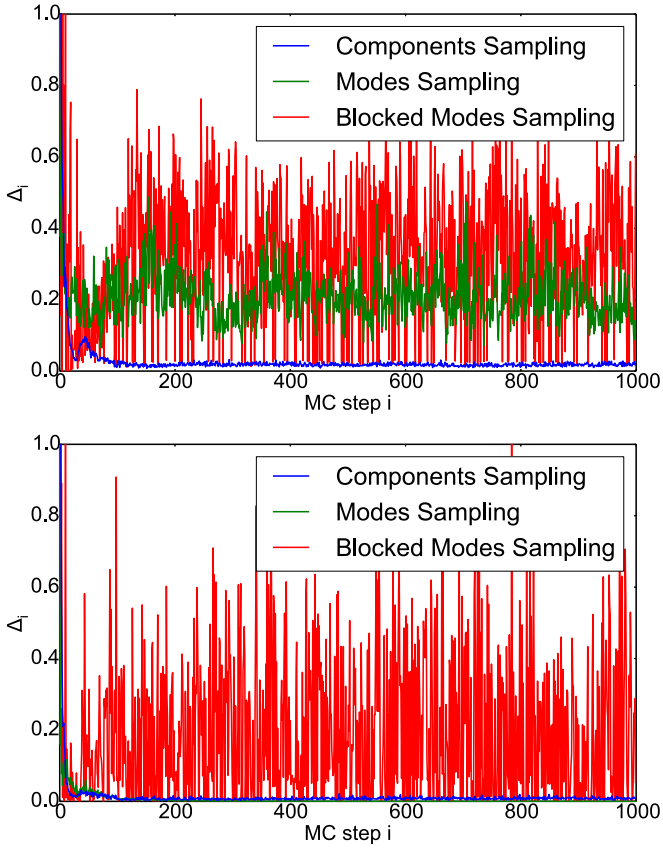

FIG. 3. Efficiency of different sampling methods comparing the changes in $\mathbf{f}$ between Monte Carlo steps, defined as $\Delta_i = \|\mathbf{f}^{(i+1)} - \mathbf{f}^{(i)}\|/\|\mathbf{f}^{(i)}\|$. For a grid with small cutoff (top), modes and blocked-mode sampling are of comparable efficiency, while the steps taken when sampling components are exceedingly small. When the cutoff is large enough so that the tail of the spectral function is represented on the grid (bottom) the steps taken when sampling modes also become extremely small. Only blocked-mode sampling always efficiently samples the space of models $\mathbf{f}$.

For a practical comparison of the different approaches, we apply them to the test cases, Sec. III. We try to recover the optical conductivity (14) from (15) for accurate data ($\sigma = 0.001$) generated from model 1 ($\Gamma_0 = 0.3$) on an equidistant frequency grid $\omega_n = 0.1(n-1/2)$ with small ($n = 1, \ldots, 40$) and large ($n = 1, \ldots, 120$) cutoff. The first cutoff is so small (about the width of the overall factor Γ_e) that the tail of the optical conductivity is hardly represented on the grid. The cutoff for the second grid is chosen such that it covers a large region of the tail where the model is going to zero.

In both cases, blocked-mode sampling updates $\mathbf{f}$ most efficiently so that we obtain uncorrelated samples after only a few Monte Carlo steps. In components sampling $\mathbf{f}$ is hardly changed in a MC update, so that very many steps are needed to obtain statistically independent samples. Modes sampling is as efficient as blocked-mode sampling when the model $\mathbf{f}$ does not go to zero. In case the tail is represented on the grid, however, it can become even less efficient than components sampling. Figure 3 shows that not every MC step in blocked-mode sampling results in a large change in $\mathbf{f}$. Since the level in the hierarchy of blockings (Fig. 2) is chosen randomly, there are steps where components or the modes of the full grid are updated. But most of the time a blocking in between these extremes is chosen, leading, on average, to an extremely rapid random walk in the space of models.

Since blocked-mode sampling moves so efficiently, it is not very important from which initial vector $\mathbf{f}^{(0)}$ the simulation is started. Still, a good choice is to start from non-negative least-squares (NNLS) solution [17] of Eq. (16), since this is gives the best fit under the constraint $\mathbf{f} \geqslant 0$. An even better starting point is obtained by choosing the NNLS solution of Eq. (16), after adding some noise to the data. This moves the initial vector $\mathbf{f}^{(0)}$ slightly away from the best fit solution, such that in effect we can immediately take data without having to warm-up the Monte Carlo run.

E. Linear constraints

Besides being non-negative, spectral functions can fulfill other constraints, e.g., the sum rule $\int d\omega A(\omega) = 2\pi$. After discretization, such linear constraints can be written as $\mathbf{C}\,\mathbf{f} = \mathbf{c}$. For C independent constraints, $\mathbf{C}$ is a $C \times N$ matrix. Using the reduced singular value decomposition $\mathbf{C} = \mathbf{U}_C\mathbf{D}_C\mathbf{V}_C^\mathsf{T}$, we see that the constraint is only active in the C-dimensional subspace that $\mathbf{P}_C = \mathbf{V}_C\mathbf{V}_C^\mathsf{T}$ projects to. Fixing $\mathbf{P}_C\mathbf{f}$ to fulfill the constraints, we can sample in the orthogonal space using the methods discussed above. In practice, we find that sum-rules are strongly represented in the data so that it is not really necessary to enforce them explicitly.

V. ROLE OF THE GRID

To implement the functional integral (13) numerically, we discretize the models $f(x)$ as a finite vector $\mathbf{f}$ representing $f(x)$ on a grid. We now analyze how the results depend on this discretization. As test cases we use again the optical conductivity described in Sec. III.

A. Uniform grid

The most natural choice is to represent $\sigma(\omega)$ on a uniform grid $\omega_n = \omega_0 + n\,\Delta\omega$. The number of grid points $n \in \{1, \ldots, N\}$ must be finite, so that such a grid necessarily has a cutoff. Since the optical conductivity quickly goes to zero for large frequencies, we would expect that once the cutoff is large enough so the tail of $\sigma(\omega)$ is well represented, the result should hardly change when increasing the cutoff further while keeping the step width $\Delta\omega$ fixed.

With our efficient blocked-mode sampling we can easily check this. For the optical conductivity test cases of Sec. III on grids with $\Delta\omega = 0.25$ and $N = 32$, 64, 128, and 256 frequency points, it is a matter of seconds on a modern laptop to obtain the average spectra with good statistical accuracy. The result for model 2 with noise $\sigma_\Pi = 0.001$ is shown in Fig. 4. To our great surprise, we find that the results change drastically: with increasing cutoff a set of pronounced spurious peaks develops. For the more noisy data, $\sigma_\Pi = 0.01$, the effect gets even stronger.

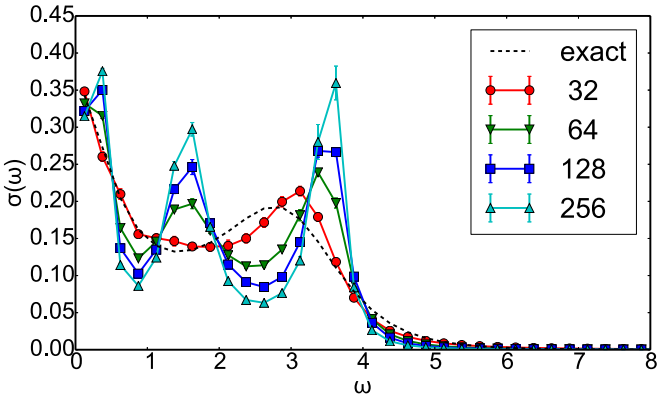

FIG. 4. Optical conductivity $\sigma(\omega)$ obtained by analytic continuation to uniform grids $\omega_n = (n-1/2)\Delta\omega$ for $n \in \{1, \ldots, N\}$ with fixed grid spacing $\Delta\omega = 0.25$ and increasing number of grid points $N = 32$, 64, 128, and 256, corresponding to a cutoff $\omega_{\max} \approx 8$, 16, 32, and 64. For comparison, the dashed line shows the optical conductivity from which the imaginary-frequency data for the analytic continuation was calculated. Even though all functions are essentially zero for $\omega \gtrsim 8$, the result depends very strongly on the length of the grid: the result of the analytic continuation develops spurious peaks that get sharper with increasing cutoff.

B. Nonuniform grids

To eliminate the cutoff for a finite grid on an infinite interval, we need to choose the grid points such that their spacing increases with their value. We can construct such a grid x_n on a general interval $x_{\min} \ldots x_{\max}$ using a positive and normalized function $\rho(x)$ that defines the density of the grid points. The cumulative distribution function $P(x) := \int_{x_{\min}}^{x} dx' \rho(x')$ is then a monotonous function mapping the interval $x_{\min} \ldots x_{\max}$ to [0,1]. Choosing a uniform discretization $z_n = (n-1/2)/N \in [0, 1]$ with $n \in \{1, \ldots, N\}$, we obtain a grid $x_n = P^{-1}(z_n)$. To get a more intuitive notation, we write the cumulative distribution function as $z(x) := P(x)$ and its inverse as $x(z) := P^{-1}(z)$. Then the x-grid $x_n = x(z_n)$ is given in terms of the uniform z grid. This mapping is illustrated in Fig. 5 for the interval $0 \ldots \infty$.

The following table lists a few useful nonuniform grids for the semi-infinite interval $[0, \infty)$. The names for the grids are derived from their density function. Note that our exponential grid is also known as logarithmic mesh, while our Lorentzian grid is sometimes called a conformal parametrization [18,19]. For the Gaussian grid, inverf is the inverse of the error function [14].

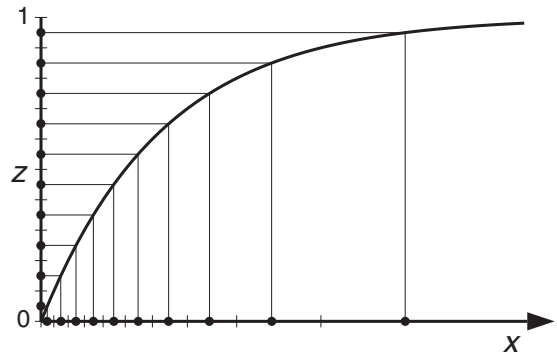

FIG. 5. Grid mapping from uniform grid on the interval [0,1] to a nonuniform grid on the half infinite interval $[0, \infty)$. Grid points are indicated as dots, limits of intervals by bars.

	$\rho(x) = \frac{dz}{dx}$	$x(z)$	$\frac{dx}{dz}$
Gaussian	$\frac{e^{-x^2/2\alpha^2}}{\sqrt{2\pi}\alpha/2}$	$\sqrt{2}\alpha\,\mathrm{inverf}(z)$	$\frac{\sqrt{2\pi}\alpha/2}{e^{-\mathrm{inverf}(z)^2}}$
exponential	$\frac{e^{-x/\beta}}{\beta}$	$-\beta\ln(1-z)$	$\frac{\beta}{1-z}$
Lorentzian	$\frac{2/\pi\gamma}{1+(x/\gamma)^2}$	$\gamma\tan(\pi z/2)$	$\frac{\pi\gamma/2}{\cos(\pi z/2)^2}$

The Gaussian and Lorentzian grids are easily extended to the interval $(-\infty, \infty)$ by replacing z by $2z-1$, giving $x_{\text{Gauß}}(z) = \sqrt{2}\alpha\,\mathrm{inverf}(2z-1)$ and $x_{\text{Lor}}(z) = -\gamma\cot(\pi z)$.

We express the integral equation in the new variable

$$g(y) = \int K(y,x)\,f(x)\,dx = \int_0^1 K(y,x(z))\,f(x(z))\frac{dx}{dz}\,dz.$$

To obtain a matrix equation as in (16) we write the integral as a Riemann sum [20]

$$g(y) \approx \frac{1}{N}\sum_{n=1}^{N} K(y,x_n)\,f(x_n)\,\frac{dx(z_n)}{dz}. \tag{24}$$

Since $w_n := (1/N)\,dx(z_n)/dz = 1/N\rho(x(z_n))$ is approximately the width of the interval $[x(z_{n-1/2}), x(z_{n+1/2})]$, we can interpret $\bar{f}_n := f(x_n)\,w_n$ as the integral of $f(x)$ over that interval.

Writing the matrix form of the integral equation as $\mathbf{g} = \mathbf{K}\bar{\mathbf{f}}$, we perform the integral [cf. (19)] over the $\bar{\mathbf{f}}$. The results for model 2 of Sec. III are shown in Fig. 6. We find that using nonuniform grids tends to give a dramatic improvement over the results for uniform grids with cutoff (Fig. 4). Still, results do depend on the choice of the grid, the more so the larger the noise in the data.

We can understand this by considering the limit where the data contains no information about the model except a sum rule $\sum \bar{f}_n = 1$ to keep the result finite. Then (19) becomes

$$\bar{\mathbf{f}}_{\text{ASM}} \propto \prod_{n=1}^{N}\int_0^\infty d\bar{f}_n\,\bar{\mathbf{f}}\;\delta\left(\sum_{n=1}^{N}\bar{f}_n - 1\right). \tag{25}$$

In this integral, all $\bar{f}_n$ play the same role, so that by symmetry all components of $\bar{\mathbf{f}}_{\text{ASM}}$ must be the same and, by the sum rule, equal to $1/N$. Consequently, in the absence of data except for a sum rule, the average spectrum is equal to $\mathbf{f}_{\text{ASM}}(x_n) = 1/N w_n = \rho(x_n)$. In that sense, the grid density acts as a default model.

In the average spectra of Fig. 6, the effect of the grid is most clearly seen in the way the tail goes to zero. The Fredholm integral for the optical conductivity (15), e.g., depends, except for the sum rule given by $\Pi(0)$, only very weakly on the form of $\sigma(\omega)$ at large frequencies,

$$\Pi(i\omega_m) - \Pi(0) = -\frac{2}{\pi}\int_0^\infty d\omega\,\frac{\sigma(\omega)}{1+(\omega/\omega_m)^2}, \tag{26}$$

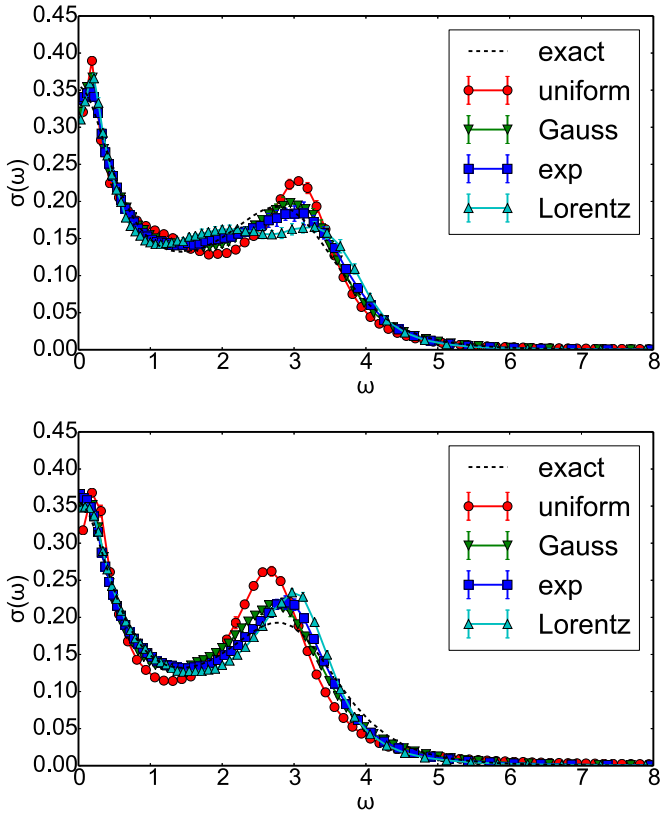

FIG. 6. Optical conductivity $\sigma(\omega)$ obtained by analytic continuation to different grids of $N = 64$ points. The uniform grid has a spacing $\Delta\omega = 0.125$, corresponding to a cutoff $\omega_{\max} \approx 8$. The width parameter of the Gaussian grid was chosen $\alpha = 4$, for the exponential $\beta = 3$, and for the Lorentzian $\gamma = 2.5$. The dashed line shows the exact result. Removing the cutoff by going to a nonuniform grid improves the result significantly, and the results depend much less on the chosen width parameter than on the cutoff. Still, the average spectra obtained for different grid densities differ by more than their error bars. This grid dependence becomes somewhat stronger for larger noise in the data [(top) $\sigma_\Pi = 0.001$ and (bottom) 0.01].

so that the data contains only little information about the shape of the tail. Indeed, as expected from (25), we find that for large ω the average spectrum vanishes as the chosen grid density.

It is important to realize that this behavior does not depend on our choice of including the width factor from (24) in the model vector $\bar{f}_n = f_n w_n$ or, $\bar{\mathbf{f}} = \mathbf{W}\mathbf{f}$, where $\mathbf{W} = \mathrm{diag}(\mathbf{w})$. If we include it, instead, in the kernel, the kernel matrix is modified $\bar{\mathbf{K}} := \mathbf{K}\mathbf{W}$, so that $\chi^2(\mathbf{f}) = \|\mathbf{g} - \bar{\mathbf{K}}\mathbf{f}\|^2 = \|\mathbf{g} - \mathbf{K}\bar{\mathbf{f}}\|^2 = \bar{\chi}^2(\bar{\mathbf{f}})$, and, by a change of variables

$$\begin{aligned}\bar{\mathbf{f}}_{\mathrm{ASM}} &= c_{\bar{\chi}^2} \prod_{n=1}^{N} \int_0^\infty d\bar{f}_n\, \bar{\mathbf{f}}\, e^{-\frac{1}{2}\bar{\chi}^2(\bar{\mathbf{f}})} \\ &= c_{\bar{\chi}^2} \prod_{n=1}^{N} w_n \prod_{n=1}^{N} \int_0^\infty df_n\, \mathbf{W}\mathbf{f}\, e^{-\frac{1}{2}\chi^2(\mathbf{f})} = \mathbf{W}\,\mathbf{f}_{\mathrm{ASM}}\,,\end{aligned} \tag{27}$$

where the constants w_n account for the change in normalization of the Gaussian after the change of variables: $c_{\chi^2} = c_{\bar{\chi}^2} \det(\mathbf{W}) = c_{\bar{\chi}^2} \prod_n w_n$.

To understand the grid dependence of $\bar{\mathbf{f}}_{\mathrm{ASM}}$ we can use a similar argument. Let $\bar{f}_n$ and $\tilde{f}_n$ be the models on two different grids, $\rho(x)$ and $\tilde{\rho}(x)$, that cover the same range, e.g., $x \in (0, \infty)$, and have the same number of grid points N. $\bar{f}_n$ is the integral of the model over the interval I_n centered around $x(z_n)$. Following (24), we may express it in terms of the $\tilde{\mathbf{f}}$ as a weighted sum of the $\tilde{f}_{\tilde{n}}$, defining a linear transformation $\bar{\mathbf{f}} = \tilde{\mathbf{W}}\tilde{\mathbf{f}}$. The situation is quite similar to (27), but with a crucial difference: In general $\tilde{\mathbf{W}}$ will not be diagonal, so that the transformation will change the limits of integration from $\bar{f}_n \geqslant 0$ for $\bar{f}_n$ to $\tilde{\mathbf{W}}\tilde{f}_n \geqslant 0$ for the integration over $\tilde{f}_n$ and consequently $\bar{\mathbf{f}}_{\mathrm{ASM}} \neq \tilde{\mathbf{W}}\tilde{\mathbf{f}}_{\mathrm{ASM}}$. Apparently, choosing different grids implies different definitions of what values of the model are allowed.

This becomes even more evident when we consider what happens when we refine the grid by halving each interval: instead of the original N values $\bar{f}_n$ on the original grid, we now have twice as many values $\tilde{f}_{\tilde{n}}$ representing the integral of the model over the halved intervals. The two sets are thus related by $\bar{f}_n = \tilde{f}_{2n-1} + \tilde{f}_{2n}$. Sampling the $\tilde{f}_{\tilde{n}} \geqslant 0$, we find that the probability of sampling a given value $\bar{f}_n$ is proportional to

$$\int_0^\infty d\tilde{f}_{2n-1} \int_0^\infty d\tilde{f}_{2n}\, \delta(\bar{f}_n - \tilde{f}_{2n-1} - \tilde{f}_{2n}) = \int_0^{\bar{f}_n} d\tilde{f}_{2n} = \bar{f}_n, \tag{28}$$

i.e., sampling the $\tilde{f}_{\tilde{n}}$ on the fine grid with a flat distribution implies sampling on the coarse grid with a distribution that is biased against small values of $\bar{f}_n$. In other words, the naive discretization of the functional integral (19) does not have a proper continuum limit. We, consequently, have to investigate the definition of a functional integral more carefully.

C. Functional integrals

We have just seen that the naive discretization of the functional integral, used so successfully in Feynman path integrals [21], does not work for averaging spectra. The problem is that sampling with a flat distribution on different grids gives incompatible results so that the discretized functional integral has no proper continuum limit [22]. We can, however, enforce such compatibility in (28) by introducing (separate) probability distributions for the $\bar{f}_n$ and the $\tilde{f}_{\tilde{n}}$ on the original and the halved intervals

$$\int_0^{\bar{f}_n} d\tilde{f}_{2n}\, \tilde{p}(\tilde{f}_{2n})\, \tilde{p}(\bar{f}_n - \tilde{f}_{2n}) = \bar{p}(\bar{f}_n). \tag{29}$$

In principle, the probability distributions on the two subintervals could be chosen independently, $\tilde{p}_{2n-1}$ and $\tilde{p}_{2n}$. To avoid any bias we assume, however, that the distribution only depends on the width but not the position of the interval. Thus $\tilde{p}_{2n-1} = \tilde{p}_{2n} =: \tilde{p}$, since each subinterval is half the width of the original interval.

The compatibility condition (29) means that the convolution of $\tilde{p}$ with itself equals $\bar{p}$ which, in terms of the Laplace transform

$$\mathcal{L}\{p\}(s) = \int_0^\infty dt\; p(t)\, e^{-st}, \tag{30}$$

is equivalent to $(\mathcal{L}\{\tilde{p}\})^2 = \mathcal{L}\{\bar{p}\}$. To find the compatible distribution on the fine grid given the distribution on the original grid, we just have to take the inverse transform of the square root of its Laplace transform: $\tilde{p} = \mathcal{L}^{-1}\{\sqrt{\mathcal{L}\{\bar{p}\}}\}$.

We want the distribution on the original grid to resemble a flat distribution. An obvious choice is to simply introduce a cutoff: $\bar{p}_c(\bar{f}) = (\Theta(\bar{f}) - \Theta(\bar{f}-\bar{c}))/\bar{c}$, where $\Theta(x)$ is the step function that vanishes for $t<0$ and is one for $t>0$. The square root of its Laplace transform is $\sqrt{1-e^{-\bar{c}s}}/\sqrt{\bar{c}s}$. Expanding the numerator for $s>0$ in $e^{-\bar{c}s}$ and using that $\mathcal{L}\{\Theta(t-a)/\sqrt{t-a}\}(s) = e^{-as}\,\Gamma(\frac{1}{2})/\sqrt{s}$, where $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\,dx$ is the Gamma function, we find

$$\mathcal{L}^{-1}\{\sqrt{\mathcal{L}\{\bar{p}_c\}}\}(\tilde{f}) = \frac{1}{\sqrt{\pi\bar{c}}}\left(\frac{\Theta(\tilde{f})}{\sqrt{\tilde{f}}} - \frac{1}{2}\frac{\Theta(\tilde{f}-\bar{c})}{\sqrt{\tilde{f}-\bar{c}}} - \cdots\right), \tag{31}$$

which is negative due to the divergences at integer multiples of the cutoff $\bar{c}$. Thus, for flat distributions $\bar{p}_c(\bar{f})$ with cutoff there exist no compatible distributions $\tilde{p}_c(\tilde{f})$ on the halved intervals. They are called indivisible [22].

Alternatively, we can start from an exponential $\bar{p}_e(\bar{f}) = \lambda e^{-\lambda\bar{f}}$, which for $\lambda \searrow 0$ approaches a flat distribution. Its Laplace transform is $\mathcal{L}\{\bar{p}_e\}(s) = \lambda/(s+\lambda)$. Using $\mathcal{L}\{e^{-at}/\sqrt{\pi t}\}(s) = (s+a)^{-1/2}$ we see that $\tilde{p}_e(\tilde{f}) = e^{-\lambda\tilde{f}}/\sqrt{\pi\tilde{f}/\lambda}$. Thus the exponential distribution is divisible. In fact, from $\mathcal{L}\{f(t)\,e^{-\lambda t}\}(s) = \mathcal{L}\{f(t)\}(s+\lambda)$ and

$$\mathcal{L}\{t^{\tilde{w}-1}\}(s) = s^{-\tilde{w}}\int_0^\infty x^{\tilde{w}-1}\,e^{-x}\,dx = s^{-\tilde{w}}\,\Gamma(\tilde{w}), \tag{32}$$

it follows that it can be divided into any number, n, of intervals of width $\tilde{w} = 1/n$, i.e., it is infinitely divisible. Note that $\tilde{w}$ is the width of the subinterval in units of the width of the original interval. The process of subdivision is consistent: halving the small intervals produces a distribution $\tilde{\tilde{p}} = \mathcal{L}^{-1}\{\sqrt{\mathcal{L}\{\tilde{p}\}}\}$, which, by $\mathcal{L}\{\tilde{p}\} = \sqrt{\mathcal{L}\{\bar{p}\}}$, is equal to $\mathcal{L}^{-1}\{\sqrt[4]{\mathcal{L}\{\bar{p}\}}\}$, so that the continuum limit of the functional integral is well defined. Of course, we are not restricted to subintervals of equal width. For

$$\begin{aligned} p_{w,\lambda}(f) &= \mathcal{L}^{-1}\{\sqrt[1/w]{\mathcal{L}\{\lambda e^{-\lambda f}\}}\}(f) \\ &= \frac{\lambda^w}{\Gamma(w)}f^{w-1}\,e^{-\lambda f} = \frac{f^{w-1}\,e^{-\lambda f}}{\int_0^\infty x^{w-1}\,e^{-\lambda x}\,dx}, \end{aligned} \tag{33}$$

which is a gamma distribution with shape parameter w and scale λ, we find the generalized compatibility relation

$$\int_0^{\bar{f}} d\tilde{f}\; p_{\tilde{w},\lambda}(\tilde{f})\, p_{\bar{w}-\tilde{w},\lambda}(\bar{f}-\tilde{f}) = p_{\bar{w},\lambda}(\bar{f}), \tag{34}$$

where the scale λ remains unchanged, while the shape parameter changes with the width of the interval.

Using gamma distributions, we can now write down a discretization of the functional integral with a well defined continuum limit. For a particular grid of N points and density $\rho(x)$, we start with the naive discretization (19), i.e., we sample the $\bar{f}_n$ from a flat distribution

$$\bar{\mathbf{f}}_{\text{ASM}} = \lim_{\lambda\to 0} c_{\bar{\chi}^2}\prod_{n=1}^{N}\int_0^\infty d\bar{f}_n\,\frac{p_{1,\lambda}(\bar{f}_n)}{\lambda}\,\bar{\mathbf{f}}\,e^{-\frac{1}{2}\bar{\chi}^2(\mathbf{f})}, \tag{35}$$

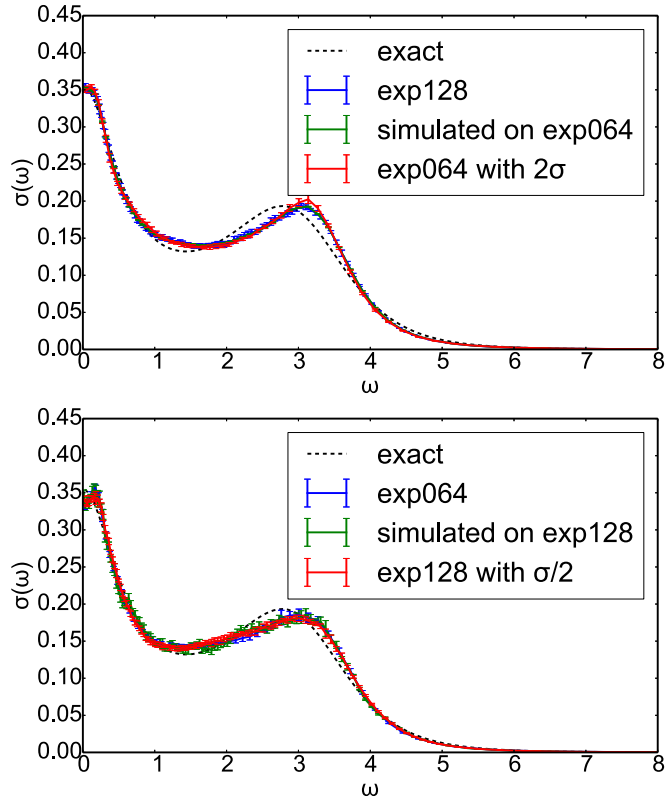

FIG. 7. Simulating one finite grid on another. For the same problem as in the upper panel of Fig. 6, we compare the result for an exponential grid with $\beta = 3$ and $N = 128$ points with the simulation on the same exponential but with $\tilde{N} = 64$ (top) and vice versa (bottom). The results agree within error bars. In addition, runs on the $\tilde{N}$ grid are shown, where the noise in the data is scaled by $N/\tilde{N}$. We do not plot the large symbols distinguishing the curves as they would obscure the near perfect agreement.

where convergence in the limit $\lambda \to 0$ is guaranteed by the Gaussian. On a different grid of $\tilde{N}$ points with grid density $\tilde{\rho}(x)$ we then have to sample the $\tilde{f}_{\tilde{n}}$ from a gamma distribution, where the shape parameter is the width of the interval of grid $[\tilde{N}, \tilde{\rho}(x)]$ in units of the width of the corresponding interval on grid $[N, \rho(x)]$. Approximating the width of an interval containing $\tilde{x}$ by $1/N\rho(\tilde{x})$ as in (24), we obtain

$$\tilde{\mathbf{f}}_{\text{ASM}} \sim \prod_{\tilde{n}=1}^{\tilde{N}}\int_0^\infty d\tilde{f}_{\tilde{n}}\,\tilde{f}_{\tilde{n}}^{\frac{N\rho(\tilde{x}_{\tilde{n}})}{\tilde{N}\tilde{\rho}(\tilde{x}_{\tilde{n}})}-1}\,\tilde{\mathbf{f}}\,e^{-\frac{1}{2}\tilde{\chi}^2(\tilde{\mathbf{f}})}, \tag{36}$$

which for $\tilde{N} \to \infty$ has a well defined continuum limit, i.e., defines a specific functional integration.

We can actually use (36) to simulate on grid $(\tilde{N}, \tilde{\rho}(x))$ the result we would obtain sampling with a flat distribution on a different grid $(N, \rho(x))$. This is illustrated in Fig. 7. Note that for $\tilde{N}\tilde{\rho}(\tilde{x}_{\tilde{n}}) > N\rho(\tilde{x}_{\tilde{n}})$ the reweighting factor in (36) diverges for small $\tilde{f}_{\tilde{n}}$ (but still giving a probability distribution). In the limit $\tilde{N} \to \infty$ individual samples $\tilde{\mathbf{f}}$ will therefore be zero almost everywhere except for finite values on a few intervals, i.e., they will look like a collection of discrete peaks [22]. This atomicity property of the gamma distributions makes sampling coarse grids on finer ones somewhat noisy.

Still, we are left with the problem of how to choose the grid $[N, \rho(x)]$ used in (35), which determines the functional

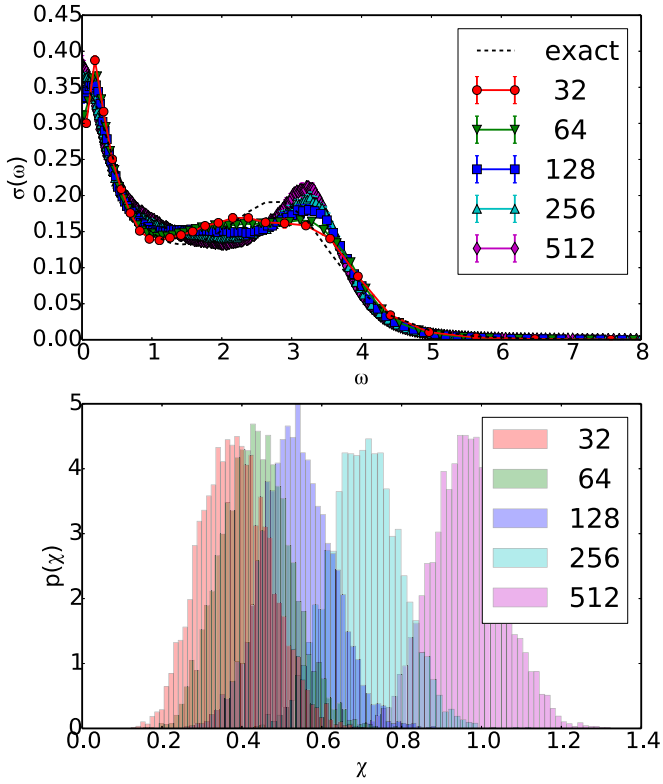

FIG. 8. Dependence of the average spectrum on the number N of grid points for the same problem as in Fig. 7 on a Lorentzian with $\gamma = 2.5$. For larger N, the average spectra get worse. The reason becomes clear from the histograms in the bottom panel, showing the contribution of spectra with a given fit χ to $\tilde{\mathbf{f}}_{\mathrm{ASM}}$: with increasing N the histogram moves to the right, i.e., worse fits.

measure. Our first impulse might be to choose N as large as possible as to minimize discretization errors. As shown in Fig. 8, however, for larger N the average spectra tend to develop spurious structures. To understand the origin of this counterintuitive behavior, we analyze what models actually contribute to the average $\mathbf{f}_{\mathrm{ASM}}$. Figure 8 shows that with increasing N, the average spectrum is eventually dominated by models that fit the data less and less well. We can understand this qualitatively by realizing that N is the number of degrees of freedom in a model. So increasing N allows for a larger variety of different models. Still, for any given N there is only a single model that gives the best fit χ_{NNLS}. Thus the density of models with worse fit increases with N, explaining the drift of the histogram towards larger χ.

We can make a more rigorous argument and gain further insights by using the reweighting approach. Let us assume that we are calculating (35) on a very fine grid $[N, \rho(x)]$. We can simulate the result on a much coarser grid of the same density $[\tilde{N}, \rho(x)]$ with $\tilde{N} \ll N$. Imposing the sum rule $\sum_{\tilde{n}} \tilde{f}_{\tilde{n}} = \tilde{F}$, (36) becomes

$$\tilde{\mathbf{f}}_{\mathrm{ASM}} \sim \prod_{\tilde{n}=1}^{\tilde{N}} \int_0^\infty d\tilde{f}_{\tilde{n}}\, \tilde{f}_{\tilde{n}}^{\tilde{w}_{\tilde{n}}-1}\, \delta\left(\tilde{F} - \sum_{\tilde{n}} \tilde{f}_{\tilde{n}}\right) \tilde{\mathbf{f}}\, e^{-\frac{1}{2}\tilde{\chi}^2(\tilde{\mathbf{f}})}$$

with $\tilde{w}_{\tilde{n}} := N/\tilde{N} \gg 1$. The models are thus sampled from a Dirichlet distribution

$$p_D(\tilde{w}_1, \ldots, \tilde{w}_{\tilde{N}}; \tilde{f}_1, \ldots, \tilde{f}_{\tilde{N}}) = \frac{\Gamma\left(\sum \tilde{w}_{\tilde{n}}\right)}{\tilde{F}^{\sum \tilde{w}_{\tilde{n}}-1} \prod \Gamma(\tilde{w}_{\tilde{n}})} \prod \tilde{f}_{\tilde{n}}^{\tilde{w}_{\tilde{n}}-1} \tag{37}$$

with fixed $\tilde{F}$, where the normalization constant

$$\int_0^{\tilde{F}} d\tilde{f}_1 \tilde{f}_1^{\tilde{w}_1-1} \int_0^{\tilde{F}-\tilde{f}_1} d\tilde{f}_2 \tilde{f}_2^{\tilde{w}_2-1} \ldots \int_0^{\tilde{F}-\sum_{\tilde{n}=1}^{\tilde{N}-2}\tilde{f}_{\tilde{n}}} d\tilde{f}_{\tilde{N}-1} \times \tilde{f}_{\tilde{N}-1}^{\tilde{w}_{\tilde{N}-1}-1} \left(\tilde{F} - \sum_{\tilde{n}=1}^{\tilde{N}-1} \tilde{f}_{\tilde{n}}\right)^{\tilde{w}_{\tilde{N}}-1}$$

follows from Euler's Beta integral [14] for $\alpha, \beta > 0$

$$\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\, dt = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \tag{38}$$

For $\tilde{w}_{\tilde{n}} \gg 1$, using Stirling's formula $\ln \Gamma(z) \approx z \ln z - z$, we obtain

$$\ln p_D \approx \sum \left(\tilde{w}_{\tilde{n}} \ln \frac{\sum_{\tilde{n}} \tilde{w}_{\tilde{n}} \frac{\tilde{f}_{\tilde{n}}}{\tilde{F}}}{\tilde{w}_{\tilde{n}}}\right) = -\frac{N}{\tilde{F}} \sum_{\tilde{n}} \tilde{f} \ln \frac{\tilde{f}}{\tilde{f}_{\tilde{n}}} \tag{39}$$

which is proportional to the entropy $S(\tilde{f}|\tilde{\mathbf{f}})$ of $\tilde{f} := \tilde{F}/\tilde{N}$ relative to $\tilde{f}_{\tilde{n}}$. Hence

$$\tilde{\mathbf{f}}_{\mathrm{ASM}} \sim \prod_{\tilde{n}=1}^{\tilde{N}} \int_0^\infty d\tilde{f}_{\tilde{n}}\, \delta\left(\tilde{F} - \sum_{\tilde{n}} \tilde{f}_{\tilde{n}}\right) \tilde{\mathbf{f}}\, e^{\frac{N}{\tilde{F}} S(\tilde{f}|\tilde{\mathbf{f}}) - \frac{1}{2}\tilde{\chi}^2(\tilde{\mathbf{f}})}. \tag{40}$$

For $N \to \infty$, the entropy term will dominate χ^2 so that the integrals of the model over the intervals, $\tilde{\mathbf{f}}_{\mathrm{ASM}}$, will tend to a constant, independent of the data. The situation is quite analogous to that discussed for (25): Sampling on a very dense grid gives a model proportional to the grid density, which, again, acts as a default model.

In fact, the prior on the models in (40) is strikingly similar to the maximum entropy prior, which, however, uses the entropy of the model relative to the default model. The two relative entropies are closely related, with the MaxEnt entropy $-\sum \tilde{f}_{\tilde{n}} \ln \tilde{f}_{\tilde{n}}/\tilde{f}$ penalizing models deviating from the default somewhat less than the average-spectrum entropy $-\sum \tilde{f} \ln \tilde{f}/\tilde{f}_{\tilde{n}}$.

While the grid density acts as a default model, the number N of grid points plays the role of a regularization parameter: going from N to N' grid points changes the prefactor of the entropy term relative to that of the fit function by N'/N. In (40), we can reach the same effect by staying with the N grid points but scaling the fit function by N/N', i.e., scaling the overall variance in the data. Fig. 7 shows that this is a simple, efficient, and remarkably accurate way of simulating grids with the same density but different number of points. This explains why the idea of rescaling the noise of Monte Carlo data is widely used in practice [7,8,10,12]. Moreover, it beautifully confirms the intuition underlying the idea of the average spectrum method stated after Eq. (13): the noise in the data leads, via the averaging of spectra, to a smoothing of the model, and the larger the noise, the larger this regularizing effect.

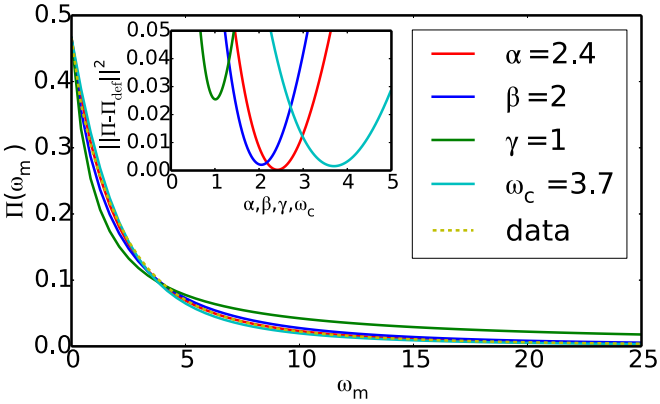

FIG. 9. Determining a reasonable default model by fitting the grid density $\rho(x)$ (Gaussian of standard deviation α, exponential with rate β, or Lorentzian of width γ, uniform with cutoff ω_c) to the data (dotted line) for the same problem as in Fig. 8. The inset shows $\sum (\Pi(\omega_m) - \Pi_{\text{def}}(\omega_m))^2$ as a function of the grid parameter, highlighting the importance of choosing a reasonable default model.

VI. PRACTICAL METHOD

To make the average spectrum approach a practical method, we have to understand how to choose the regularization. As we have seen in Fig. 8, results can depend strongly on the number of grid points. Most striking about this dependence is that with increasing regularization the average spectra are not smoothed but rather develop increasingly sharp features—the opposite of what one would expect from a regularization! To understand this, we look at how the default model fits the imaginary-axis data. For this we need to relate the grid density to the default model. We can, e.g., write the default optical conductivity as $\sigma_{\text{def}}(\omega) := \pi\,\Pi(0)\rho(\omega)/2$ from which we can calculate the values on the imaginary axis as

$$\Pi_{\text{def}}(\omega_m) = \Pi(0)\left(1 - \int_0^\infty \frac{\rho(\omega)}{1+(\omega/\omega_m)^2}\,d\omega\right). \qquad (41)$$

The deviation of Π_{def} from the actual data tells us how compatible the default model is with the data. This is shown in Fig. 9. We find that the grid density used in Fig. 8, a Lorentzian of width $\gamma = 2.5$, does not even remotely represent the imaginary-axis data. The situation is even worse for the uniform grids of Fig. 4, which, with increasing cutoff, become more and more inconsistent with the data. In all these cases the default model does not resemble the data on the imaginary axis at all. This misfit has dramatic consequences, since the information we try to extract from the data is hidden in the tiny details on the imaginary axis—the very reason why analytic continuation is so ill-conditioned. Regularizing towards a grossly wrong default model then forces the models to develop unphysical features in order to somehow achieve a decent fit nevertheless.

The problem completely disappears when using a reasonable default model. An example is shown in Fig. 10, using a Gaussian grid with $\alpha = 2.4$. As we read off from Fig. 9, this default model is compatible with the data and we see that with increasing regularization the resulting spectra become smoother. Moreover, this smoothing is not very strong so that the results are remarkably robust under changes in the number

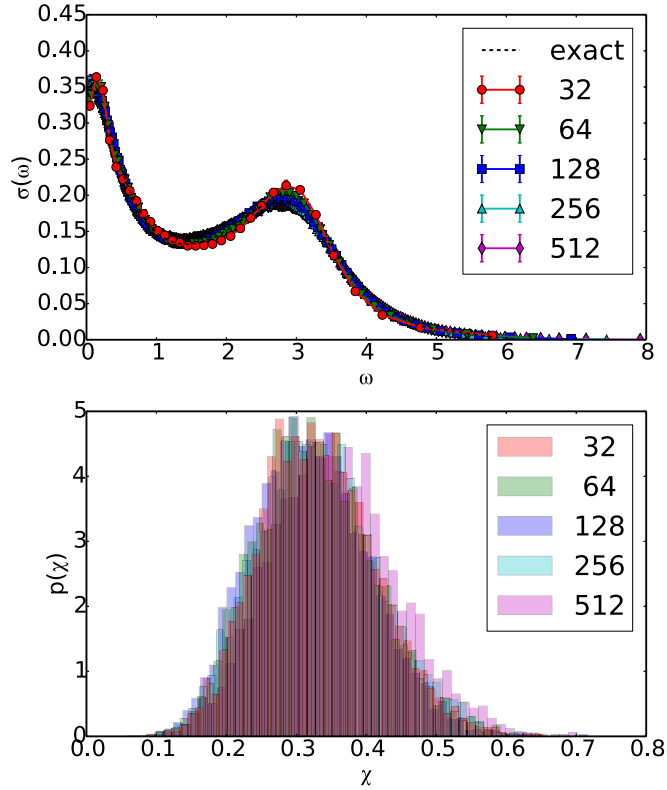

FIG. 10. Dependence of the average spectrum on the number N of grid points for the same problem as in Fig. 8 on an optimized Gaussian grid with $\alpha = 2.4$. The average spectra are largely independent of N and the histograms show a consistently good fit.

of grid points. It thus turns out that the choice of the default model is much more important than that of the regularization parameter.

In this respect, a flat default model with cutoff is a particularly unfortunate choice. As we see from Fig. 9, for a cutoff $\omega_c \approx 3.7$, we actually obtain quite a reasonable default model so that we would expect robust average spectra. Such a grid, however, has no points in the tail of the model. If we want to resolve the model at higher frequencies we need to "improve" the cutoff, necessarily giving increasingly poor default models that are responsible for the disastrous results obtained in Fig. 4.

VII. CONCLUSIONS

We have seen that the average spectrum method is not the parameter free method suggested by the deceptively written functional integral (13): We have to choose a grid density $\rho(x)$, which acts as a default model, and a number N of grid points, which acts as a regularization parameter. The reason for this is that the naive discretization (19) does not converge to a well defined functional integral. Instead we have to sample the components of the models we are integrating over from distributions that are consistent for different discretizations. For general non-negative functions these are gamma distributions (33), when, in addition, the functions fulfill a sum-rule they are Dirichlet distributions (37). This raises, of course, the question why the naive discretization

does work for path integrals. In the Feynman approach, the integrand itself already fulfills the consistency relation giving rise to a complex Wiener measure [23], so that the appropriate functional measure is inherent in the path integrand. This is not the case for the functional integral (13), requiring us to explicitly specify the functional measure by singling out a specific grid on which to evaluate (19). Using the corresponding family of gamma or Dirichlet distributions, we can then take the continuum limit.

We find that approaching this limit we sample models $\tilde{\mathbf{f}}$ with a prior given by the entropy of the flat distribution on that grid relative to $\tilde{\mathbf{f}}$, making the grid density act as a default model, while the number of grid points acts as the regularization parameter. The similarity with the maximum-entropy method (MaxEnt) is obvious. Of course, the entropies differ, but this only means that MaxEnt regularizes large deviations from the default model somewhat less. More importantly, MaxEnt determines the model from maximizing rather than averaging. This appears to avoid having to specify a functional measure. However in the derivation of MaxEnt marginalizations over the model space do, in fact, require functional integrals. To quote Ref. [24], p. 137: "This shortcoming has been missed earlier due to a deceptive side-effect of the Gaussian approximation made in the calculation, and because the quantitative answers from the analysis were generally sensible in practice."

To make the average spectrum method a practical technique for analytic continuation we need reliable recipes for choosing grid density and number of grid points. As we have demonstrated, the results can depend quite strongly on these choices. A badly chosen default model will bias the results towards models that give an extremely bad fit to the imaginary-axis data. In such cases, we obtain utterly unreasonable results: with increasing regularization the result develops stronger and stronger features. Interestingly this is particularly true for flat default models with cutoff, which are by their very nature ill suited for analytic continuation. A good default model should, instead, not only be featureless but also be overall consistent with the data. For such default models, the features in the results will be suppressed with increasing regularization—as it should be. In fact, then results become fairly independent of the actual choice of the regularization parameter over a wide range, highlighting the importance of the default model rather than the regularization parameter.

Finally, a practical method must be efficient. This has so far been the cardinal problem of the average spectrum method. We have described an optimized implementation, without which we could not have analyzed the method in such detail. While we have discussed here only one specific test case, more can be found in Ref. [25].

In addition, we make an efficient web-based implementation freely available at [26].

[1] M. Jarrell and J. E. Gubernatis, Phys. Rep. **269**, 133 (1996).
[2] A. Reymbaut, D. Bergeron, and A.-M. S. Tremblay, Phys. Rev. B **92**, 060509(R) (2015).
[3] Y. Burnier and A. Rothkopf, Phys. Rev. Lett. **111**, 182003 (2013).
[4] P. C. Hansen, *Discrete Inverse Problems* (SIAM, Philadelphia, 2010).
[5] M. Jarrell, in *Correlated Electrons: From Models to Materials*, edited by E. Pavarini, E. Koch, F. Anders, and M. Jarrell (Forschungszentrum Jülich, Jülich, 2012).
[6] S. R. White, in *Computer Simulation Studies in Condensed Matter Physics III*, edited by D. P. Landau, K. K. Mon, and B.-B. Schüttler (Springer, Heidelberg, 1991), pp. 145–153.
[7] A. W. Sandvik, Phys. Rev. B **57**, 10287 (1998).
[8] K. Vafayi and O. Gunnarsson, Phys. Rev. B **76**, 035115 (2007).
[9] O. F. Syljuåsen, Phys. Rev. B **78**, 174429 (2008).
[10] S. Fuchs, T. Pruschke, and M. Jarrell, Phys. Rev. E **81**, 056701 (2010).
[11] K. S. D. Beach, Identifying the maximum entropy method as a special limit of stochastic analytic continuation, arXiv:cond-mat/0403055.
[12] A. W. Sandvik, Phys. Rev. E **94**, 063308 (2016).
[13] L. Boehnke, H. Hafermann, M. Ferrero, F. Lechermann, and O. Parcollet, Phys. Rev. B **84**, 075145 (2011).
[14] DLMF, *NIST Digital Library of Mathematical Functions*, http://dlmf.nist.gov/, Release 1.0.17 of 2017-12-22, edited by F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, and B. V. Saunders.
[15] O. Gunnarsson, M. W. Haverkort, and G. Sangiovanni, Phys. Rev. B **82**, 165125 (2010).
[16] C. P. Robert, Statistics Computing **5**, 121 (1995).
[17] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems* (SIAM, Philadelphia, 1974).
[18] I. S. Krivenko and A. N. Rubtsov, Analytic continuation of quantum Monte Carlo data: Optimal stochastic regularization approach, arXiv:cond-mat/0612233.
[19] L.-F. Arsenault, R. Neuberg, L. A. Hannah, and A. J. Millis, Inverse Probl. **33**, 115007 (2017).
[20] J. Waldvogel, in *Approximation and Computation*, Springer Optimization and Its Applications, Vol. 42, edited by W. Gautschi, G. Mastroianni, and T. Rassias (Springer, New York, NY, 2010), pp. 267–282.
[21] L. S. Schulman, *Techniques and Applications of Path Integration* (Dover Publications, Mineola, NY, 2005).
[22] J. Skilling and S. Sibisi, Priors on measures, in *Maximum Entropy and Bayesian Methods*, edited by K. M. Hanson and R. N. Silver (Kluwer, Dordrecht, 1996), pp. 261–270.
[23] I. M. Gel'fand and A. M. Yaglom, J. Math. Phys. **1**, 48 (1960).
[24] C. S. Sivia and J. Skilling, *Data Analysis: A Bayesian Tutorial*, 2nd ed. (Oxford University Press, Oxford, 2006).
[25] K. Ghanem, Stochastic analytic continuation: A Bayesian approach, Ph.D. thesis, RWTH Aachen University, 2017.
[26] www.spektra.app.