Connecting Tikhonov regularization to the maximum entropy method for the analytic continuation of quantum Monte Carlo data

Khaldoon Ghanem [1] and Erik Koch [2,3]
[1] *Quantinuum, Leopoldstrasse 180, 80804 Munich, Germany*
[2] *Jülich Supercomputer Centre, Forschungszentrum Jülich, 52425 Jülich, Germany*
[3] *JARA High-Performance Computing, 52425 Jülich, Germany*

(Received 11 December 2022; accepted 8 February 2023; published 17 February 2023)

Analytic continuation is an essential step in extracting information about the dynamical properties of physical systems from quantum Monte Carlo (QMC) simulations. Different methods for analytic continuation have been proposed and are still being developed. This paper explores a regularization method based on the repeated application of Tikhonov regularization under the discrepancy principle. The method can be readily implemented in any linear algebra package and gives results surprisingly close to the maximum entropy method (MaxEnt). We analyze the method in detail and demonstrate its connection to MaxEnt. In addition, we provide a straightforward method for estimating the noise level of QMC data, which is helpful for practical applications of the discrepancy principle when the noise level is not known reliably.

DOI: 10.1103/PhysRevB.107.085129

I. ANALYTIC CONTINUATION: AN ILL-POSED PROBLEM

From a mathematical perspective, the analytic continuation problem corresponds to solving a Fredholm integral equation of the first kind,

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

where $f(x)$ is the unknown spectrum, a non-negative integrable function. $K(y,x)$ is the kernel of the integral equation and is known analytically, while $g(y)$ is noisy data, typically obtained from QMC simulation at a finite number of points y_j.

To solve the analytic continuation numerically, the integral is discretized using a grid of n points x_i, giving a linear system of equations

$$\mathbf{g} = \mathbf{K}\mathbf{f}, \tag{2}$$

where the elements of the matrix $\mathbf{K}$ are the kernel values $K(y_j, x_i)$, $\mathbf{g}$ contains m measured data values $g(y_j)$, and f_i is the spectrum integral over the ith grid interval. The most naive and straightforward way of solving Eq. (2) is, as with any other linear system of equations, using the weighted least-squares method,

$$\mathbf{f}_{\mathrm{LS}} = \arg\min_{\mathbf{f}} \chi^2(\mathbf{f}), \tag{3}$$

which finds the spectrum minimizing the fit to the data:

$$\chi^2(\mathbf{f}) \coloneqq (\mathbf{g} - \mathbf{K}\,\mathbf{f})^{\mathsf{T}} \mathbf{C}^{-1} (\mathbf{g} - \mathbf{K}\,\mathbf{f}). \tag{4}$$

The fit is weighted by the inverse of $\mathbf{C}$, the covariance matrix of the noise on the data. By factorizing the covariance matrix into $\mathbf{C}^{-1} = \mathbf{T}^{\mathsf{T}}\mathbf{T}$, one can always replace the kernel matrix and data vector by the weighted ones $\mathbf{TK}$ and $\mathbf{Tg}$, respectively. Then the covariance matrix of the weighted data becomes the identity matrix, and one can use the ordinary least-squares method instead. In the following, we will always assume that such transformation has been applied to the kernel and the data despite using the same notations $\mathbf{K}$ and $\mathbf{g}$ to denote the weighted ones.

Using the least-squares solution for solving the analytic continuation problem gives generally bad results plagued by noise, as exemplified in Fig. 1. The reason is that the matrices in analytic continuation problems are highly ill-conditioned such that the inevitable small noise on the data leads to disastrous noise on the least-squares solution [1,2]. This can be seen more explicitly using the singular value decomposition (SVD) of the kernel matrix

$$\mathbf{K} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathsf{T}}, \tag{5}$$

where $\mathbf{S}$ is a diagonal matrix of size $m \times n$, and $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices of sizes $m \times m$ and $n \times n$, respectively. The columns of the matrix $\mathbf{U}$ form an orthonormal basis of the data space and are called the *data modes*, while the columns of matrix $\mathbf{V}$, which span the space of spectra, are called the *spectral modes*. The diagonal elements of $\mathbf{S}$ are the *singular values* and they are sorted in descending order. Using the SVD, the least-squares solution can be written as

$$\mathbf{f}_{\mathrm{LS}} = \sum_{i}^{\min(m,n)} \frac{\mathbf{u}_i^{\mathsf{T}}\mathbf{g}}{s_i}\,\mathbf{v}_i. \tag{6}$$

For matrices arising from analytic continuation problems, the singular values decay exponentially to zero (see Fig. 2). Dividing by these vanishing singular values hugely amplifies any small noise present in the data. This is the main problem with the least-squares solution.

The other source of ill-posedness is the incompleteness of the data, i.e., we only know the data at a finite number of

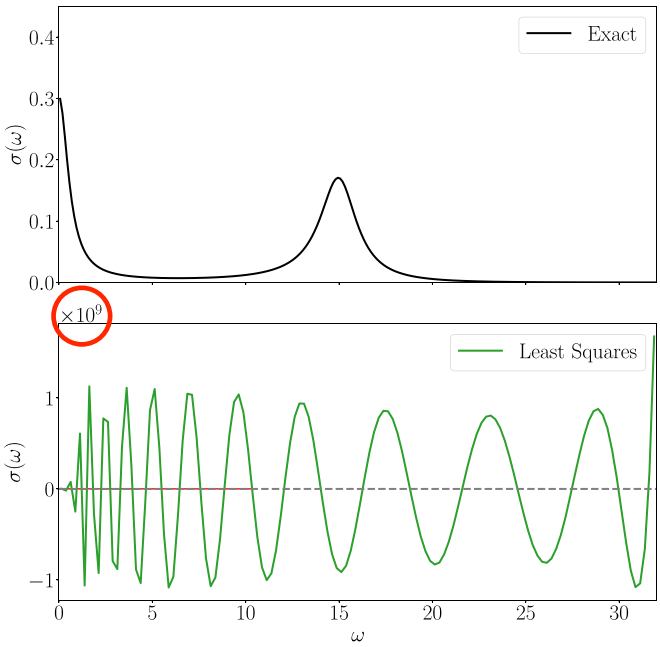

FIG. 1. Least-squares solution (bottom panel) for the analytic continuation of optical conductivity $\sigma(\omega)$ using noisy data of its correlation function. The exact correlation function is computed analytically from the exact optical conductivity (top panel) on the first $m = 60$ bosonic Matsubara frequencies with inverse temperature $\beta = 15$. The input data includes relative Gaussian noise with standard deviation 10^{-2}. This test case is an adaptation of the ones proposed by Ref. [29] and studied further in Refs. [12–14]. In the notation of the latter reference, the optical conductivity used here differs in the values of the following parameters: $\Gamma_e = 20, \epsilon_1 = 15$. We denote this data set as *test case 1*.

points $m < n$, where n is typically chosen large enough to resolve the desired features of the spectrum. Therefore, even for numerically exact data, if no regularization/additional information is provided, one can only ever hope to recover at most the first m modes of the spectrum.

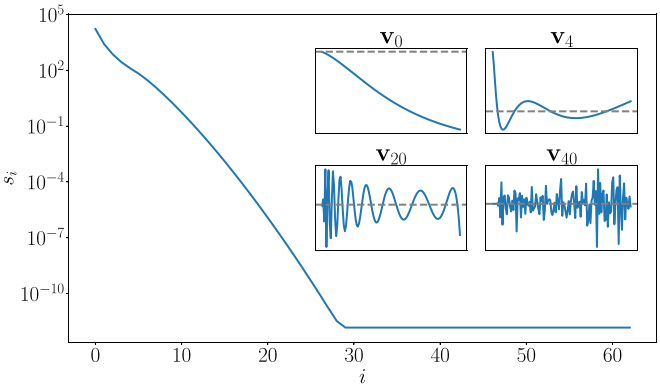

FIG. 2. Singular values of the (weighted) kernel of test case 1. The singular values decay exponentially until leveling off at a value determined by the machine epsilon. In the inset, we show some of the spectral modes. The leading spectral modes are smooth and slowly varying functions. As the mode index increases, the number of nodes increases and the modes become more oscillatory. Once the singular values reach numerical accuracy, the corresponding modes become numerically degenerate so the SVD routine returns arbitrary linear combinations of the exact modes.

Various methods have been developed and used to address the ill-posedness of analytic continuation, including the maximum entropy method (MaxEnt) [3–6], the average spectrum method [7–15], Padé approximation [16–19], stochastic optimization methods [20,21], machine learning methods [22–25], and genetic algorithms [26–28].

II. NOISE ESTIMATION

The SVD of the kernel matrix allows an accurate estimation of the overall scale of noise on QMC data. This can be valuable in practical situations where such an estimate is unavailable or as an important cross-check of the validity of the noise level estimate.

As a start, let us assume, as usual, that an estimate of the covariance matrix $\mathbf{C}$ already exists and that the data and kernel have been weighted by $\mathbf{T}$, the square root of its inverse. Consequently, the noise on the different components of the weighted data vector $\mathbf{g}$ is uncorrelated and has a unit variance. Since the matrix $\mathbf{U}$ is unitary, the noise ϵ_i present in the expansion coefficients of the data $\mathbf{u}_i^\mathsf{T}\mathbf{g}$ is also uncorrelated and has a unit variance. These noisy data coefficients are then related to the exact spectrum via the relation

$$\mathbf{u}_i^\mathsf{T}\mathbf{g} = s_i\, \mathbf{v}_i^\mathsf{T}\mathbf{f}_{\text{exact}} + \epsilon_i. \tag{7}$$

Given that the exact spectrum has a finite norm and that the singular values in analytic continuation decay exponentially, there is some index k, after which the exact data coefficients become negligible compared to the noise. For these indices, the measured data coefficients are practically plain noise,

$$\mathbf{u_i}^\mathsf{T}\mathbf{g} \approx \epsilon_i \quad : k < i \leqslant m, \tag{8}$$

and can be used to estimate the variance of the noise ϵ_i as

$$\sigma^2(\epsilon) \approx \frac{1}{m-k} \sum_{i=k+1}^{m} \left(\mathbf{u_i}^\mathsf{T}\mathbf{g}\right)^2, \tag{9}$$

where the formula for estimating population variance with a known mean of value zero has been employed. The expected value of this estimator is mostly independent of the index k as long as it is large enough that the corresponding exact data coefficients are well below the actual noise level. Estimating the earliest such index can be done by inspecting the noisy data coefficients and checking when they start to plateau at a certain level. That would be around $i = 10$ in Fig. 3. Using a larger value of k gives a similar estimate but with less accuracy. In practice, the index k can be automatically chosen as the numerical rank of $\mathbf{K}$, i.e., the index at which the singular values hit numerical accuracy. Since the number of data points is typically much larger than the numerical rank, this choice avoids checking for the plateau while it does not sacrifice much accuracy in estimating the noise.

When the covariance matrix $\mathbf{C}$ is properly scaled, we expect this value to be close to one. This is illustrated in Fig. 3 for test case 1, where the data coefficients decay exponentially till they reach the noise level $\sigma(\epsilon) = 1$ and fluctuate around it. However, when a covariance matrix with the wrong scaling is used, the aforementioned plateau of data coefficients will be

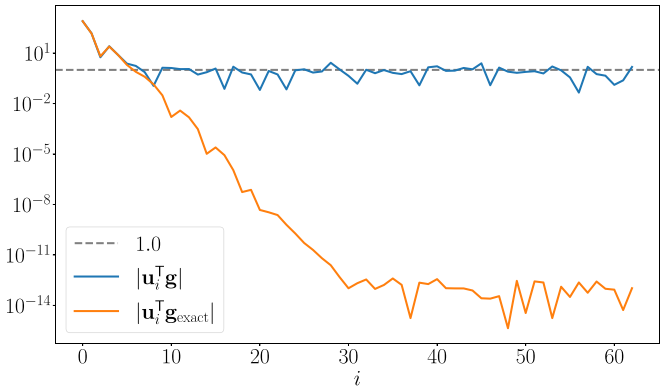

FIG. 3. Absolute values of the exact and noisy data coefficients of test case 1. While the exact coefficients decay to the machine epsilon, the noisy ones decay until they hit the noise level and then fluctuate around it. Here the noise level equals one because the data is weighted by the proper covariance matrix. Notice that large noisy coefficients are close to their exact values and that the deviation becomes significant only when their values drop to near the noise level.

scaled accordingly, and $\sigma(\epsilon)$ will deviate from the expected value of one. Values much larger than one indicate that the noise level has been underestimated, while values much lower than one indicate an overestimation of the noise level.

An important practical use case of the above formula is estimating the noise level of uncorrelated *relative* Gaussian noise. In this case, as an initial ansatz, one can use a diagonal covariance matrix whose diagonal elements are the squares of the data values. Equation (9) then provides an estimate of σ^2, which can be multiplied by the ansatz to obtain a properly scaled covariance matrix.

III. TIKHONOV REGULARIZATION

The expansion of the least-squares solution using SVD modes [cf. Eq. (6)] already suggests a direct remedy to the ill-posedness; namely, truncating the later modes, which are dominated by noise, while keeping the leading ones that are more stable. This is known as the truncated SVD solution. Tikhonov regularization [30,31] is a more refined method, where the noisy modes are turned off continuously, with each term in the least-squares solution multiplied by a filtering function $\phi(s;\alpha) \coloneqq s^2/(s^2+\alpha)$ that depends on its singular value s and an adjustable parameter α:

$$\mathbf{f}_{\text{Tikhonov}}(\alpha) = \sum_i^{\min(m,n)} \phi(s_i;\alpha)\,\frac{\mathbf{u}_i^\mathsf{T}\mathbf{g}}{s_i}\,\mathbf{v}_i. \tag{10}$$

Terms corresponding to very small singular values $s_i^2 \ll \alpha$ are practically removed, while ones corresponding to large singular values $s_i^2 \gg \alpha$ are hardly modified [32].

It can be shown that the above Tikhonov solution is the least-squares solution of an alternative problem with extended data and an extended kernel

$$\mathbf{f}_{\text{Tikhonov}}(\alpha) = \arg\min_{\mathbf{f}} \left\| \begin{pmatrix} \mathbf{K} \\ \sqrt{\alpha}\,\mathbf{I} \end{pmatrix} \mathbf{f} - \begin{pmatrix} \mathbf{g} \\ \mathbf{0} \end{pmatrix} \right\|^2, \tag{11}$$

where $\mathbf{I}$ is the unit matrix in the n-dimensional space of spectra. This formulation has a computational advantage for large-scale problems because it allows getting the Tikhonov solution using any linear solver without explicit computation of the SVD. Moreover, this least-squares problem can be written as the following minimization problem:

$$\mathbf{f}_{\text{Tikhonov}}(\alpha) = \arg\min_{\mathbf{f}} \chi^2(\mathbf{f}) + \alpha\|\mathbf{f}\|^2, \tag{12}$$

that aims to balance the fit to the data with the L_2-norm of the spectrum vector. The balance is controlled by the regularization parameter α. When α is very small, we approach the least-squares solution, which fits the data very well but has a very large L_2-norm. As α increases, more modes get filtered and the norm gets smaller while the fit gets worse. The smoothness typically associated with Tikhonov solutions comes from the fact that the leading modes are smoother than later ones for analytic continuation kernels (see, for example, the insets of Fig. 2).

While the aforementioned form of Tikhonov regularization is the most basic and widely used one in the inverse problem literature [33], it has two drawbacks for analytic continuation problems. The first is that the discretized L_2-norm is grid dependent because the spectral values $f_i \coloneqq w_i f(x_i)$ include the full weight of the grid interval at point x_i. Using a grid with n points and a grid density $\rho(x)$, these weights are defined as $w_i \coloneqq 1/[N\rho(x_i)]$ and the L_2- norm of the spectrum reads

$$\|\mathbf{f}\|^2 = \sum_i f_i^2 = \sum_i [w_i f(x_i)]^2 \approx \frac{1}{N}\int dx\,\frac{f^2(x)}{\rho(x)}. \tag{13}$$

This shows that the basic form of Tikhonov has an implicit dependence on the grid density [34]. We suggest replacing this implicit dependence with an explicit one on a default model $d(x)$. Let $d_i \coloneqq w_i d(x_i)$ be the integral of the default model over the ith grid interval, then we replace the usual L_2-norm $\sum_i f_i^2$ with the weighted L_2-norm $\sum_i f_i^2/d_i$. It can be easily verified that the weighted norm is indeed grid independent.

The second drawback is that the solution approaches zero in the limit of large regularization parameter α. In analytic continuation, however, we know that the spectrum must have a finite L_1-norm, so it would be desirable if the solution would approach some properly normalized spectrum in the limit of large α. We choose to center our regularization term at the default model $\mathbf{d}$ instead of zero.

In summary, we propose using the following form of Tikhonov regularization in analytic continuation problems:

$$\mathbf{f}_{\text{Tikhonov}}(\alpha,\mathbf{d}) = \arg\max_{\mathbf{f}} -\tfrac{1}{2}\chi^2(\mathbf{f}) + \alpha T(\mathbf{f}|\mathbf{d}), \tag{14}$$

where the Tikhonov penalty term is defined as

$$T(\mathbf{f}|\mathbf{d}) = -\frac{1}{2}\sum_i \frac{(f_i-d_i)^2}{d_i}. \tag{15}$$

It is worth noting that, like the original form, this formulation can be solved as an extended least-squares problem,

$$\mathbf{f}_{\text{Tikhonov}}(\alpha,\mathbf{d}) = \arg\min_{\mathbf{f}} \left\| \begin{pmatrix} \mathbf{K} \\ \sqrt{\alpha}\,\mathbf{D}^{-1} \end{pmatrix} \mathbf{f} - \begin{pmatrix} \mathbf{g} \\ \sqrt{\alpha}\mathbf{De} \end{pmatrix} \right\|^2, \tag{16}$$

with $\mathbf{D} = \mathrm{diag}(\mathbf{d})$ and $\mathbf{e} \coloneqq (1, 1, \ldots, 1)^\mathsf{T}$. Its solution can be similarly expressed in terms of the SVD of the rescaled kernel $\mathbf{K}\sqrt{\mathbf{D}}$ as shown in Appendix A.

IV. DISCREPANCY PRINCIPLE

Choosing the value of the regularization parameter α is an essential ingredient of any regularization method. Apart from the obvious criterion that α should be smaller for more accurate data, there is no unique procedure for actually determining its value. Any such procedure should strike a balance between fitting the noise and biasing the solution. A common method in the inverse problem literature is the discrepancy principle [35,36].

According to the discrepancy principle, a good spectrum would produce data such that the residual vector $\mathbf{r} \coloneqq \mathbf{g} - \mathbf{K}\mathbf{f}$ is dominated by noise. Therefore, we should choose α such that the norm of the residual $\|\mathbf{r}\|^2 = \chi^2(\mathbf{f})$ equals the expected norm of the noise vector. Assuming, as usual, that data and kernel have been reweighed with the square root of the noise covariance, the expected norm-squared of the noise vector follows the well-known chi-squared distribution. The mean value of this distribution equals the number of data points m, and its variance equals $2m$. To avoid accidental overfitting of noise, one may apply the discrepancy principle using a value (in terms of the standard deviation) somewhat larger than the mean. In this paper, however, we always use the mean value.

Interestingly, the Tikhonov solution using the discrepancy principle can be written in a form independent of any regularization parameter α as a maximization of the Tikhonov penalty,

$$\mathbf{f}_{\text{Tikhonov}}(\mathbf{d}) = \arg\max_{\mathbf{f}\in\mathcal{C}} T(\mathbf{f}|\mathbf{d}), \tag{17}$$

over the manifold $\mathcal{C}$ defined by the discrepancy principle:

$$\mathcal{C} \coloneqq \{\mathbf{f} \in \mathbb{R}^n : \chi^2(\mathbf{f}) = m\}. \tag{18}$$

Starting from some spectrum on the manifold $\mathcal{C}$, the Tikhonov solution can then be found by following the gradient of $T(\mathbf{f}|\mathbf{d})$, projected on $\mathcal{C}$:

$$\mathbf{a}^\perp = \left[\mathbf{I} - \frac{\mathbf{z}\,\mathbf{z}^\mathsf{T}}{\mathbf{z}^\mathsf{T}\mathbf{z}}\right]\mathbf{a}, \tag{19}$$

where $\mathbf{a} \coloneqq \nabla T$ is the gradient of the Tikhonov penalty with

$$a_i = -\frac{f_i - d_i}{d_i}, \tag{20}$$

and $\mathbf{z} \coloneqq -\frac{1}{2}\nabla\chi^2$ is the gradient of the fit function, i.e., the surface normal of $\mathcal{C}$ with

$$z_i = \mathbf{k}_i^\mathsf{T}[\mathbf{g} - \mathbf{K}\mathbf{f}], \tag{21}$$

where $\mathbf{k}_i$ is the ith column of the kernel matrix $\mathbf{K}$. At the optimal point, the projection vanishes and the gradient of T must be antiparallel to the fit gradient

$$\alpha\,\mathbf{a} = -\mathbf{z}, \tag{22}$$

which is nothing but the stationarity condition for Eq. (14). The optimal regularization parameter α thus reemerges as the ratio of the two gradients at the optimal point.

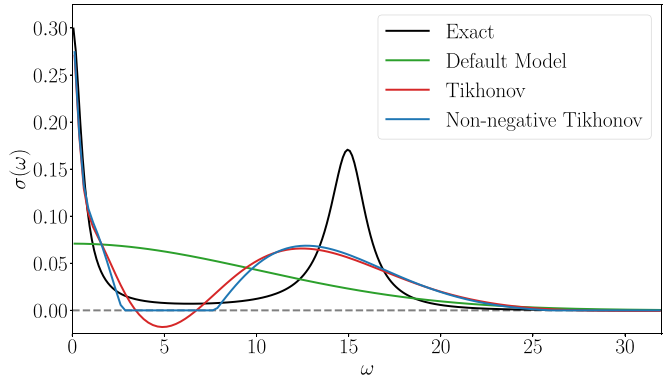

FIG. 4. Tikhonov solutions for test case 1 using a Gaussian default model centered at 0 with width 10. The values used for the regularization parameter α are determined by the discrepancy principle.

In practice, this constrained optimization problem is converted, using the method of Lagrange multiplier, into an unconstrained optimization of the objective function

$$\mathbf{f}_{\text{Tikhonov}}(\mathbf{d}) = \arg\max_{\mathbf{f},\beta} T(\mathbf{f}|\mathbf{d}) - \frac{\beta}{2}[\chi^2(\mathbf{f}) - m], \tag{23}$$

where the Lagrange multiplier β corresponds to the inverse of the regularization parameter α.

V. SELF-CONSISTENT TIKHONOV

Tikhonov regularization provides a simple and fast method to obtain a decent first impression of the analytic continuation solution. Its obvious disadvantage, however, is ignoring the non-negativity of the spectrum (see Fig. 4). One can enforce the non-negativity by explicitly restricting the optimization problem to non-negative spectra. This can be done straightforwardly by using the non-negative least squares method [37] with the extended kernel and data of Eq. (16). Nevertheless, enforcing the non-negativity in this artificial way does not improve the results as desired. As shown in Fig. 4, the non-negative Tikhonov solution looks like a clamped version of the original Tikhonov solution where the negative parts are set to zero, while the positive part stays roughly the same with minor adjustments to account for the truncated negative values.

Instead of enforcing the non-negativity constraint directly, one can reduce violations by increasing the regularization parameter α, which encourages the solution to be close to the non-negative default model. Under the discrepancy principle, the regularization parameter is determined implicitly and only has a large value if the default model fits the data well. This transforms the problem of satisfying non-negativity into one of improving the fit of the default model. In the limit, when the default model itself satisfies the discrepancy principle, it is its own Tikhonov solution, and thus non-negativity is guaranteed.

A simple way of improving the fit of a default model is by linearly mixing it with its Tikhonov solution under the discrepancy principle:

$$\mathbf{d} \leftarrow [1 - \mu]\,\mathbf{d} + \mu\,\mathbf{f}_{\text{Tikhonov}}(\mathbf{d}). \tag{24}$$

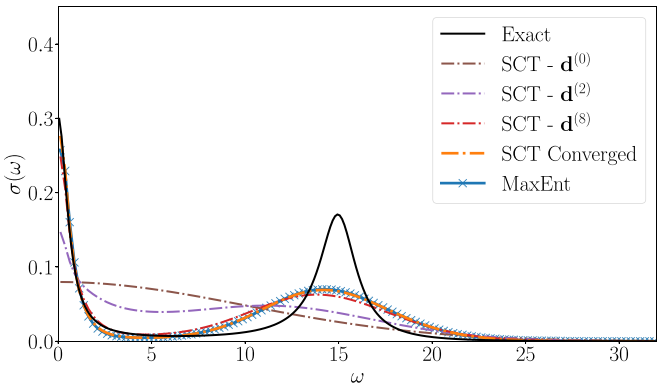

FIG. 5. Comparison of MaxEnt and default models produced by SCT at different iterations. The superscript of the default model represents its iteration number with $\mathbf{d}^{(0)}$ being the starting default model. For MaxEnt, the starting default model $\mathbf{d}^{(0)}$ was used, and the regularization parameter was determined by the discrepancy principle.

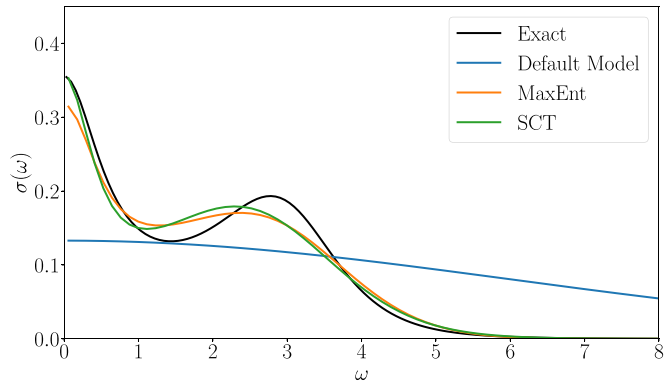

FIG. 6. Comparison of MaxEnt and SCT for a variant of test case 1. This case differs by the location of the second peak and the width of the envelope. In the notation of Ref. [14], the optical conductivity used here differs in the values of the following parameters: $\Gamma_e = 4, \epsilon_1 = 3$. We denote this data set as *test case 2*. The default model used here is a scaled Gaussian of width 6.

Assuming the fit of the starting default model is worse than m, this default model is guaranteed to have a better fit due to the convexity of the fit function χ^2. Additionally, if the starting default model is strictly positive, we can always choose the positive mixing parameter μ small enough such that this default model is also positive. The values of the mixing parameter that guarantee the positivity of this default model can be calculated explicitly from the values of the starting default model and its Tikhonov solution as

$$\mu < \min\left\{\frac{d_i}{d_i - f_i} : f_i < d_i\right\}. \tag{25}$$

These observations suggest an iterative approach to obtain an improved non-negative Tikhonov solution. In this approach, we keep linearly mixing the default model with its Tikhonov solution to obtain an improved default model until the difference between the default model and its Tikhonov solution becomes negligible. We call this method self-consistent Tikhonov (SCT).

For the mixing parameter μ, we use half the maximum allowed value [cf. Eq. (25)]. Using this value implies that the updated default model has at least half its original value at any point. This mixing strategy works well for most cases but it can sometimes lead to slow convergence when the exact spectrum has values very close to zero (e.g., at the tail of a Gaussian peak). To accelerate the convergence of such cases, we put a lower limit on the mixing parameter μ. This may lead to a violation of the positivity of the default model, which can be directly reinforced by truncating values lower than some positive threshold. It should be emphasized that these limits are not strictly necessary but help accelerate convergence in pathological cases.

In Fig. 5, we plot a set of default models produced by SCT for test case 1 at different iterations. The default model gradually transforms and fits the data till it converges, with the converged solution satisfying the discrepancy principle. This solution represents a significant improvement over the original Tikhonov solution and its non-negative counterpart (see Fig. 4). Besides providing a smooth non-negative spectrum, the shape and width of the peaks are much better reproduced.

In the same figure, we also show the solution of the MaxEnt method using the same starting default model, $\mathbf{d}^{(0)}$, and a regularization parameter that is also determined by the discrepancy principle. Remarkably, the MaxEnt solution is indistinguishably close to SCT solution. By examining other test cases, we have always found that the solutions of MaxEnt and SCT are quite similar and in many cases virtually identical (see Fig. 6 for another example). The following sections will examine and clarify this surprising connection between MaxEnt and SCT. In this context, it is worth noting that MaxEnt has also been connected to the average spectrum method, a stochastic method for analytic continuation [15,38].

VI. MAXIMUM ENTROPY METHOD

Similarly to Tikhonov regularization, MaxEnt introduces a term that penalizes the mismatch between a spectrum and a default model [3–6]. The penalty term, known as Shannon entropy, is defined as

$$S(\mathbf{f}|\mathbf{d}) := \sum_{i=1}^{N}\left[f_i - d_i - f_i \ln\left(\frac{f_i}{d_i}\right)\right]. \tag{26}$$

It represents the expected amount of information in a spectrum $\mathbf{f}$ relative to the default model $\mathbf{d}$. This entropy is then optimized in MaxEnt simultaneously alongside the data fit:

$$\mathbf{f}_{\text{MaxEnt}}(\alpha, \mathbf{d}) = \arg\max -\tfrac{1}{2}\chi^2(\mathbf{f}) + \alpha S(\mathbf{f}|\mathbf{d}). \tag{27}$$

The fit and entropy trade-off is controlled via the regularization parameter α. When α is infinitesimally small, MaxEnt formally gives the non-negative least-squares solution, but as α increases, the solution gets smoother and closer to the default model.

There are different "flavors" of MaxEnt depending on how α is chosen [39]. The most relevant for our purpose is the one known as historic MaxEnt. In this method, α is chosen such that the fit χ^2 equals the number of the data points m. This choice is equivalent to the discrepancy principle when

the data and the kernel are transformed so the noise on the data becomes uncorrelated and has unit variance. Other commonly used methods for choosing α are the classic MaxEnt and Bryan's MaxEnt. Both methods derive a probability distribution over α using Bayesian theory and use either the maximum of this distribution (classic MaxEnt) or its average (Bryan's MaxEnt) as the final solution. In the rest of the paper, we will always assume that the discrepancy principle is applied, and thus, MaxEnt refers to the original way of choosing α, i.e.,

$$\mathbf{f}_{\text{MaxEnt}}(\mathbf{d}) = \arg\max_{\mathbf{f}\in\mathcal{C}} S(\mathbf{f}|\mathbf{d}), \tag{28}$$

where $\mathcal{C}$ is the manifold defined by the discrepancy principle in Eq. (18).

The Shannon entropy is directly related to the Tikhonov regularization term, $T(\mathbf{f}|\mathbf{d})$ being the entropy expanded to second order in $\Delta_i := f_i - d_i$:

$$\begin{aligned} S(\mathbf{f}|\mathbf{d}) &= \sum_i \Delta_i - (\Delta_i + d_i)\ln\left(1+\frac{\Delta_i}{d_i}\right) \\ &\approx \sum_i \Delta_i - \frac{\Delta_i^2}{d_i} - d_i\left(\frac{\Delta_i}{d_i} - \frac{\Delta_i^2}{2d_i^2}\right) \\ &= T(\mathbf{f}|\mathbf{d}). \end{aligned} \tag{29}$$

This means that the Tikhonov method can be considered an approximation to MaxEnt. The quality of this approximation depends on how close the starting default model $\mathbf{d}$ is to the hypersurface defined by the discrepancy principle $\mathcal{C}$. When the default model satisfies the discrepancy principle, then MaxEnt and Tikhonov give the same solution—the default model itself. As the fit of the default model deteriorates, it gets further away from that hypersurface, and the maxima of the penalty terms S and T in $\mathcal{C}$ start to diverge. A more quantitative analysis of the difference between MaxEnt and Tikhonov solutions is given in Appendix B.

VII. MAXENT FAMILY OF EQUIVALENT DEFAULT MODELS

Analogously to the discussion in Sec. IV about optimizing the Tikhonov penalty, maximizing the Shannon entropy under the discrepancy constraint can also be achieved by following its gradient, projected on $\mathcal{C}$,

$$\mathbf{b}^{\perp} = \left[\mathbf{I} - \frac{\mathbf{z}\,\mathbf{z}^{\mathsf{T}}}{\mathbf{z}^{\mathsf{T}}\mathbf{z}}\right]\mathbf{b}, \tag{30}$$

where $\mathbf{b} := \nabla S$ is the gradient of the entropy with

$$b_i = -\ln\left(\frac{f_i}{d_i}\right). \tag{31}$$

At the MaxEnt solution $\mathbf{f}^\star$, the gradient of Shannon entropy and the gradient of the fit function must be antiparallel:

$$\alpha\mathbf{b}^\star = -\mathbf{z}^\star. \tag{32}$$

This gives rise to the following self-consistent system of equations satisfied by any MaxEnt solution:

$$f_i^\star = d_i \exp\left(\frac{z_i^\star}{\alpha}\right), \tag{33}$$

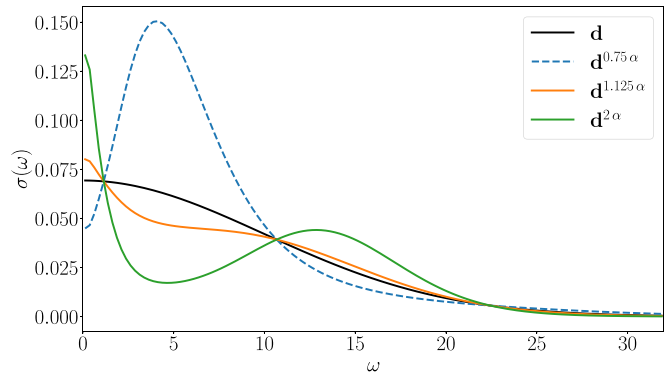

FIG. 7. Different default models equivalent to $\mathbf{d}$ for test case 1. The value of α is determined via the discrepancy principle.

where the fit gradient of the MaxEnt solution $\mathbf{z}^\star$ depends on the solution itself. By rearranging this equation, it becomes clear that the same MaxEnt solution can be obtained using a whole family of other equivalent default models $\mathbf{d}$ and their corresponding regularization parameters α. This family can be constructed explicitly using the MaxEnt solution and its fit gradient:

$$d_i := f_i^\star \exp\left(-\frac{z_i^\star}{\alpha}\right). \tag{34}$$

Alternatively, given a default model $\mathbf{d}$ with regularization parameter α, we can construct an entire family of default models $\mathbf{d}^{\alpha'}$ that result in the same MaxEnt solution $\mathbf{f}^\star$:

$$d_i^{\alpha'} = d_i \exp\left[-z_i^\star\left(\frac{1}{\alpha'} - \frac{1}{\alpha}\right)\right]. \tag{35}$$

Note that $\lim_{\alpha'\to\infty}\mathbf{d}^{\alpha'} = \mathbf{f}^\star$. In Fig. 7, we show a set of equivalent default models for test case 1.

Besides establishing the existence of equivalent default models, Eq. (34) can be used to study the stability of MaxEnt solution with respect to perturbations to these default models. The partial derivatives of the default model with respect to variations in MaxEnt solution $\mathbf{f}^\star$ and regularization parameter α read

$$\frac{\partial d_i}{\partial f_j^\star} = \frac{d_i}{f_i^\star}\delta_{i,j} + \frac{d_i}{\alpha}\,\mathbf{k}_i^{\mathsf{T}}\mathbf{k}_j, \tag{36}$$

$$\frac{\partial d_i}{\partial \alpha} = \frac{z_i^\star}{\alpha^2} d_i. \tag{37}$$

Therefore, an infinitesimal change in the MaxEnt solution $d\mathbf{f}^\star$ and an infinitesimal change in the regularization parameter $d\alpha$ induce the following relative change in the default model:

$$\boldsymbol{\delta} := \mathbf{D}^{-1}\,d\mathbf{d} = \mathbf{L}\,d\mathbf{f}^\star + \frac{d\alpha}{\alpha^2}\,\mathbf{z}^\star, \tag{38}$$

where $\mathbf{D} = \mathrm{diag}(\mathbf{d})$ and $\mathbf{L}$ is the scaled Hessian of the MaxEnt objective function

$$-\alpha\mathbf{L} := -(\mathbf{K}^{\mathsf{T}}\mathbf{K} + \alpha\mathbf{F}^\star), \tag{39}$$

with $\mathbf{F}^\star := \mathrm{diag}(\mathbf{f}^\star)$. Inverting Eq. (38) gives the changes in the MaxEnt solution in terms of perturbations to its default model. Under the discrepancy principle, the change in the regularization parameter is fixed by the constraint

$\mathbf{z}^{\star\mathsf{T}} d\mathbf{f}^\star = 0$ (ensuring that $d\mathbf{f}^\star$ has no component perpendicular to $\mathcal{C}$) to the value

$$d\alpha = \alpha^2 \frac{\mathbf{z}^{\star\mathsf{T}}\mathbf{L}^{-1}\boldsymbol{\delta}}{\mathbf{z}^{\star\mathsf{T}}\mathbf{L}^{-1}\mathbf{z}^\star}, \tag{40}$$

and the corresponding change in the MaxEnt solution is

$$d\mathbf{f}^\star = \mathbf{L}^{-1}\left[\mathbf{I} - \frac{\mathbf{z}^\star\mathbf{z}^{\star\mathsf{T}}\,\mathbf{L}^{-1}}{\mathbf{z}^{\star\mathsf{T}}\mathbf{L}^{-1}\mathbf{z}^\star}\right]\boldsymbol{\delta} =: \mathbf{L}^{-1}\,\boldsymbol{\delta}^\perp, \tag{41}$$

where $\boldsymbol{\delta}^\perp$ is the part of the vector $\boldsymbol{\delta}$ perpendicular to the surface normal $\mathbf{z}^\star$ under the inner product defined by the matrix $\mathbf{L}^{-1}$.

Relative changes in the default model along the direction of $\mathbf{z}^\star$ give an equivalent default model and thus have no effect on the MaxEnt solution. To assess the effect of changes in the default model along orthogonal directions, we need to look into the spectral decomposition of the matrix $\mathbf{L}$. The eigenvectors of $\mathbf{L}$ match the spectral modes of the rescaled kernel $\mathbf{K}' \coloneqq \mathbf{K}\sqrt{\mathbf{F}^\star}$ and the eigenvalues of the former λ_i are related to the singular values of the later s'_i as

$$\lambda_i = \frac{\alpha + s'^2_i}{\alpha}. \tag{42}$$

We now distinguish two limiting cases depending on the direction of the vector $\boldsymbol{\delta}^\perp$. When $s'^2_i \gg \alpha$, then $\lambda_i^{-1} \approx \alpha/s'^2_i$. Therefore, changes along the leading modes have little effect on the MaxEnt solution, and the effect is smaller the further away the default model is from $\mathcal{C}$. On the other hand, when $s'^2_i \ll \alpha$, then $\lambda_i^{-1} \approx 1$. Therefore, changes along the trailing modes are directly reflected in the MaxEnt solution. Assuming that a MaxEnt solution is smooth, the leading modes of $\mathbf{K}'$ are smooth and slowly varying functions while the trailing ones are highly oscillating. These results then confirm and elucidate the common wisdom that slowly varying details of the default model have little to no effect on MaxEnt solutions, while sharp features tend to introduce strong biases. Finally, note that having more accurate data scales up the singular values s'_i, and thus the MaxEnt solution becomes less sensitive to changes in the default model, as one would intuitively anticipate.

VIII. CONNECTING SCT TO MAXENT

Let $\mathbf{d}^{(t)}$ be the default model at step t of SCT and $\mathbf{f}^{(t)}$ and $\mathbf{z}^{(t)}$ be the corresponding Tikhonov solution and its fit gradient. By combining Eq. (20) with Eq. (22), we see that the Tikhonov solutions satisfy the following self-consistent equation (analogous to Eq. (33) of MaxEnt):

$$f_i^{(t)} = d_i^{(t)}\left[1 + \frac{\mathbf{z}^{(t)}}{\alpha^{(t)}}\right]. \tag{43}$$

Using mixing parameters $\mu^{(t)}$, the default models at subsequent iterations are then related by

$$d_i^{(t+1)} = (1-\mu^{(t)})d_i^{(t)} + \mu^{(t)} f_i^{(t)} = d_i^{(t)}\left[1 + \frac{\mu^{(t)}}{\alpha^{(t)}}\,\mathbf{z}^{(t)}\right]. \tag{44}$$

Applying this relation recursively and assuming very small $\mu^{(t)}/\alpha^{(t)}$, we get the following exponential form for the default models produced by SCT:

$$d_i^{(t)} = d_i^{(0)} \exp\left[\sum_{\tau=0}^{t-1} \frac{\mu^{(\tau)}}{\alpha^{(\tau)}}\,\mathbf{z}^{(\tau)}\right] = d_i^{(0)} \exp\left[\frac{\tilde{\mathbf{z}}^{(t)}}{\tilde{\alpha}^{(t)}}\right], \tag{45}$$

where in the last equation we defined the effective fit gradients $\tilde{\mathbf{z}}^{(t)}$ and the effective regularization parameters $\tilde{\alpha}^{(t)}$ as

$$\tilde{\mathbf{z}}^{(t)} \coloneqq \tilde{\alpha}^{(t)} \sum_{\tau=0}^{t-1} \frac{\mu^{(\tau)}}{\alpha^{(\tau)}}\mathbf{z}^{(\tau)}, \quad \frac{1}{\tilde{\alpha}^{(t)}} \coloneqq \sum_{\tau=0}^{t-1} \frac{\mu^{(\tau)}}{\alpha^{(\tau)}}. \tag{46}$$

Comparing the default models generated by SCT [cf. Eq. (45)] with the MaxEnt family of equivalent default models [cf. Eq. (35)], it is clear that the two have the same functional form and would match if the effective fit gradients $\tilde{\mathbf{z}}^{(t)}$ match the MaxEnt fit gradient $\mathbf{z}^\star$.

Indeed, the effective gradients of SCT provide an excellent approximation to the MaxEnt gradient. In Fig. 8, we plot the normalized overlap between the two at different iterations of SCT. The starting effective gradient is nothing but the original Tikhonov gradient, which already has a very good overlap of 0.83. This is to be expected since, as discussed in the previous section, Tikhonov provides an approximation to MaxEnt. As the SCT procedure iterates, the effective gradient not only maintains the good initial overlap, but the overlap improves until it saturates at about 0.99 when the procedure converges. Interestingly, the overlap with the bare gradients $\mathbf{z}^{(t)}$, i.e., the gradients of Tikhonov solutions at different iterations, does not necessarily increase. The plot shows that the bare overlap actually drops after a couple of iterations. We observed cases where the bare overlap even drops below its starting value (see Fig. 9). Nevertheless, in all cases we investigated, the effective gradients always had a monotonically increasing overlap with the MaxEnt gradient. An argument for this behavior of the fit gradients is detailed in Appendix C.

These results demonstrate that the set of default models produced by SCT provides an approximation to the MaxEnt family of equivalent default models, and thus solving the MaxEnt problem with any one of them gives a solution that is

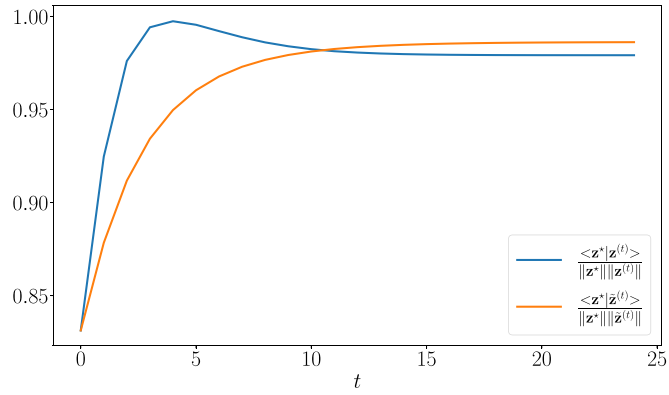

FIG. 8. Normalized overlap between the MaxEnt fit gradient $\mathbf{z}^\star$ and the fit gradients produced at different SCT iterations (denoted as t) in test case 1. Both the bare gradients $\mathbf{z}^{(t)}$ (gradients of Tikhonov solutions) and the effective gradients $\tilde{\mathbf{z}}^{(t)}$ are shown. The overlaps and norms are calculated using the inner product $\langle \mathbf{x}, \mathbf{y}\rangle \coloneqq \mathbf{x}^\mathsf{T}\mathbf{L}^{-1}\mathbf{y}$, where $\mathbf{L}$ is the scaled Hessian of MaxEnt objective function defined in Eq. (39).

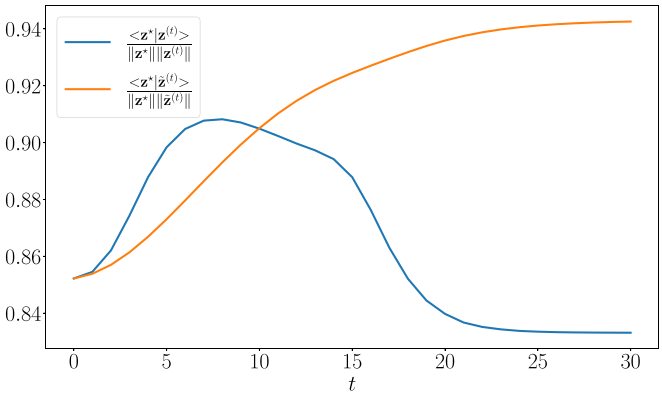

FIG. 9. Normalized overlap between the MaxEnt fit gradient $\mathbf{z}^\star$ and the fit gradients produced at different SCT iterations (denoted as t) in test case 2.

close to the solution of the original MaxEnt problem. At convergence, the default model of SCT satisfies the discrepancy principle, and thus, it is trivially the solution of its own MaxEnt problem and a good approximation of the original MaxEnt solution. In Appendix D, we give an alternative perspective in which SCT can be seen as an approximate and simplified variant of Newton's method for obtaining the MaxEnt solution.

IX. SUMMARY AND DISCUSSION

In this paper, we used SVD to derive a generally applicable method for estimating the noise level on QMC data. Having a reliable error estimate is crucial when using the discrepancy principle/historic MaxEnt. We then introduced a particular form of the Tikhonov regularization that is more suitable for analytic continuation problems. Besides solving the implicit grid dependence and normalization issues, this form is closely connected to Shannon entropy. A quadratic approximation of the entropy around its default model gives precisely the introduced Tikhonov penalty term. This form allows approximating the MaxEnt solution using the Tikhonov method when the default model already has a good fit to the data (i.e., in the limit of large regularization parameter). In the typical cases where the default model does not fit the data well, we showed that an iterative procedure where the default model is repeatedly mixed with its Tikhonov solution still gives similar results to MaxEnt. We investigated the connection between the two methods, which revealed that the same MaxEnt solution could be produced by a whole family of equivalent default models. This family is approximately traced by the self-consistent Tikhonov procedure.

SCT provides a simple and efficient alternative to MaxEnt that could be easily implemented using any linear algebra library. While we have not observed a numerical advantage of SCT over MaxEnt, an important difference between the methods is that MaxEnt has the non-negativity of the spectral function built in, while SCT is not constrained *per se* but implements non-negativity by keeping the default model from going negative. We expect this added flexibility to be useful for the analytic continuation of matrix-valued Green functions, where MaxEnt is trickier to implement [40,41]. This requires generalizing the Tikhonov penalty term to handle matrix-valued spectra $\mathcal{F}_i$ as follows: $T(\mathcal{F}|\mathcal{D}) \coloneqq -\sum_i (\mathcal{F}_i - \mathcal{D}_i)^\mathsf{T} \mathcal{D}_i^{-1} (\mathcal{F}_i - \mathcal{D}_i)/2$, where $\mathcal{D}_i$ is now a default model of positive-definite matrices. Positive definiteness of the solution can then be similarly achieved by enforcing positive definiteness of the default model. Whether such a procedure gives better results than other methods remains to be investigated.

APPENDIX A: TIKHONOV SOLUTION USING SVD

The minimization problem of Tikhonov in Eq. (14) can be written as the following least-squares problem with an extended kernel matrix and extended data vector:

$$\mathbf{f}_{\text{Tikhonov}}(\alpha, \mathbf{d}) = \arg\min_{\mathbf{f}} \left\| \begin{pmatrix} \mathbf{K} \\ \sqrt{\alpha\, \mathbf{D}^{-1}} \end{pmatrix} \mathbf{f} - \begin{pmatrix} \mathbf{g} \\ \sqrt{\alpha \mathbf{D}}\mathbf{e} \end{pmatrix} \right\|^2, \tag{A1}$$

where $\mathbf{D} = \text{diag}(\mathbf{d})$ and $\mathbf{e} = (1, 1, \ldots, 1)^\mathsf{T}$. The normal equation of this least-squares problem reads

$$(\mathbf{K}^\mathsf{T}\mathbf{K} + \alpha \mathbf{D}^{-1})\mathbf{f} = \mathbf{K}^\mathsf{T}\mathbf{g} + \alpha \mathbf{e}, \tag{A2}$$

$$\Leftrightarrow [\tilde{\mathbf{K}}^\mathsf{T}\tilde{\mathbf{K}} + \alpha \mathbf{I}](\sqrt{\mathbf{D}^{-1}}\,\mathbf{f}) = \tilde{\mathbf{K}}^\mathsf{T}\mathbf{g} + \alpha\sqrt{\mathbf{D}}\mathbf{e}, \tag{A3}$$

where a rescaled kernel matrix $\tilde{\mathbf{K}}$ is defined as $\tilde{\mathbf{K}} \coloneqq \mathbf{K}\sqrt{\mathbf{D}}$. Using SVD of the rescaled matrix $\tilde{\mathbf{K}} = \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^\mathsf{T}$, the normal equation in the mode space reads

$$[\tilde{\mathbf{S}}^\mathsf{T}\tilde{\mathbf{S}} + \alpha \mathbf{I}]\tilde{\mathbf{V}}^\mathsf{T}(\sqrt{\mathbf{D}^{-1}}\,\mathbf{f}) = \tilde{\mathbf{S}}\tilde{\mathbf{U}}^\mathsf{T}\mathbf{g} + \alpha\tilde{\mathbf{V}}^\mathsf{T}\sqrt{\mathbf{D}}\mathbf{e}. \tag{A4}$$

The Tikhonov solution can then be expressed in terms of the rescaled modes of the rescaled matrix $\mathbf{V}' \coloneqq \sqrt{\mathbf{D}}\tilde{\mathbf{V}}$ as

$$\mathbf{f}_{\text{Tikhonov}} = \sum_i \frac{\tilde{s}_i^2}{\tilde{s}_i^2 + \alpha} \frac{\tilde{\mathbf{u}}_i^\mathsf{T}\mathbf{g}}{\tilde{s}_i}\mathbf{v}'_i + \alpha \sum_i \frac{\mathbf{v}_i'^\mathsf{T}\mathbf{e}}{\tilde{s}_i^2 + \alpha}\mathbf{v}'_i. \tag{A5}$$

The first term is similar to the expansion of the original grid-dependent Tikhonov in Eq. (10), while the second term comes from centering the regularization term around the default model.

Unlike the spectral modes $\mathbf{v_i}$ in Eq. (10), however, the modes $\mathbf{v}'_\mathbf{i}$ are not orthonormal under the standard inner product. They are instead orthonormal under the modified inner product:

$$\langle \mathbf{x}, \mathbf{y} \rangle \coloneqq \mathbf{x}^\mathsf{T}\mathbf{D}^{-1}\mathbf{y}. \tag{A6}$$

Moreover, these vectors can be seen as the modes of the original kernel matrix, with the orthogonality being defined under this modified inner product. This view holds since

$$\mathbf{K}\mathbf{v}'_i = \tilde{\mathbf{K}}\tilde{\mathbf{v}}_i = \tilde{s}_i\tilde{\mathbf{u}}_i \tag{A7}$$

and

$$\langle \mathbf{v}'_i, \mathbf{v}'_j \rangle = \delta_{i,j}. \tag{A8}$$

Equation (A5) can then be seen as a direct expansion of the Tikhonov solution in terms of the spectral modes of the kernel matrix

$$\mathbf{f}_{\text{Tikhonov}} = \sum_i \langle \mathbf{v}'_i, \mathbf{f}_{\text{Tikhonov}} \rangle\, \mathbf{v}'_i, \tag{A9}$$

with

$$\langle \mathbf{v}'_i, \mathbf{f}_{\text{Tikhonov}} \rangle = \frac{1}{\tilde{s}_i^2 + \alpha} \left[\tilde{s}_i^2 \, \frac{\tilde{\mathbf{u}}_i^\mathsf{T} \mathbf{g}}{\tilde{s}_i} + \alpha \, \langle \mathbf{v}'_i, \mathbf{d} \rangle \right]. \tag{A10}$$

Note how each component of the Tikhonov solution is an interpolation between the components of the least-squares spectrum and the default model. Different components, however, are mixed differently (each according to its singular value), and thus the overall Tikhonov solution is generally not a simple interpolation of the two spectra.

APPENDIX B: DIFFERENCE BETWEEN MaxEnt AND TIKHONOV

We can quantify the difference between the Tikhonov and MaxEnt solutions of the same default model and regularization parameter as the following:

$$\boldsymbol{\Delta}^\star \coloneqq \mathbf{f}_{\text{Tikhonov}} - \mathbf{f}^\star = \mathbf{H}^{-1} \, \boldsymbol{\nabla} T^{\mathbf{f}^\star}, \tag{B1}$$

where $\mathbf{H}$ is minus the Hessian of the Tikhonov objective function of Eq. (14),

$$\mathbf{H} \coloneqq \alpha \, \mathbf{D}^{-1} + \mathbf{K}^\mathsf{T} \mathbf{K}, \tag{B2}$$

and $\boldsymbol{\nabla} T^{\mathbf{f}^\star}$ is its gradient at the MaxEnt solution:

$$\begin{aligned} \nabla T_i^{\mathbf{f}^\star} &= z_i^\star + \alpha \, a_i^\star = z_i^\star - \alpha \, \frac{f_i^\star - d_i}{d_i} \\ &= z_i^\star - \alpha \left[\exp\left(\frac{z_i^\star}{\alpha} \right) - 1 \right] \\ &= -\frac{1}{2\alpha} z_i^{\star 2} + O(\alpha^{-2}). \end{aligned} \tag{B3}$$

Therefore, the gradient scales linearly with the inverse of α. To analyze how the difference $\boldsymbol{\Delta}^\star$ scales, we look at the spectral decomposition of the Hessian matrix $\mathbf{H}$. Its eigenvectors are the same as the spectral modes of the rescaled matrix $\tilde{\mathbf{K}} = \mathbf{K}\sqrt{\mathbf{D}}$, and its eigenvalues h_i are related to the singular values of $\tilde{\mathbf{K}}$ as follows:

$$h_i = \alpha + \tilde{s}_i^2. \tag{B4}$$

The ith component of the difference then scales as $1/(\alpha^2 + \alpha\tilde{s}_i^2)$, and thus, the difference between Tikhonov and MaxEnt vanishes quadratically in the limit of strong regularization. Note that the components of the gradient along the leading spectral modes (i.e., the smooth components with large singular values) get suppressed more than the trailing ones (i.e., the oscillating components with small singular values).

APPENDIX C: DYNAMICS OF SCT

Let $\mathbf{z}'$ be the fit gradient of a Tikhonov solution. When mixing the default model with the Tikhonov solution, the relative change in the default model is proportional to this fit gradient, namely, $\boldsymbol{\delta} = \mu/\alpha \, \mathbf{z}'$. The part of $\mathbf{z}'$ along the MaxEnt fit gradient $\mathbf{z}^\star$ gives an equivalent default model and thus does not affect the MaxEnt solution. Let $d\mathbf{z}'$ denote the part of $\mathbf{z}'$ perpendicular to $\mathbf{z}^\star$ under the inner product defined by $\mathbf{L}^{-1}$, i.e.,

$$d\mathbf{z}' \coloneqq \mathbf{z}' - \frac{\mathbf{z}^{\star\mathsf{T}} \mathbf{L}^{-1} \mathbf{z}'}{\mathbf{z}^{\star\mathsf{T}} \mathbf{L}^{-1} \mathbf{z}^\star} \, \mathbf{z}^\star. \tag{C1}$$

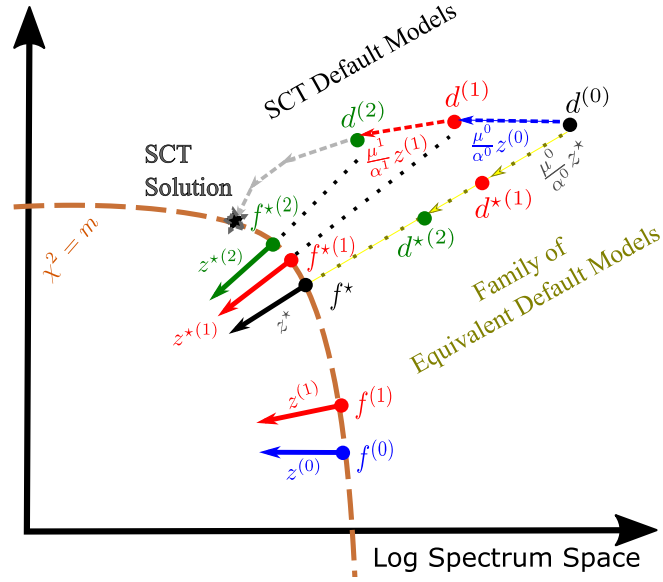

FIG. 10. Schematic diagram illustrating how the default models and their MaxEnt and Tikhonov solutions evolve with the SCT iterations. The diagram is depicted in the logarithmic space of spectra. Note that Tikhonov solutions are assumed here to be strictly positive, although, in general, they may have either sign.

Then the relevant relative change in the default model is $\boldsymbol{\delta}^\perp = \mu/\alpha \, d\mathbf{z}'$. From Eq. (41), we see that the corresponding change in the fit gradient of the MaxEnt solution reads

$$\begin{aligned} d\mathbf{z}^\star &= -\mathbf{K}^\mathsf{T}\mathbf{K} \, d\mathbf{f}^\star = -\mathbf{K}^\mathsf{T}\mathbf{K} \, \mathbf{L}^{-1} \, \boldsymbol{\delta}^\perp \\ &= -\frac{\mu}{\alpha} \, \mathbf{K}^\mathsf{T}\mathbf{K} \, \mathbf{L}^{-1} \, d\mathbf{z}'. \end{aligned} \tag{C2}$$

Given that the matrices $\mathbf{L}^{-1}$ and $\mathbf{K}^\mathsf{T}\mathbf{K}$ are positive semidefinite, the overlap between $d\mathbf{z}^\star$ and $d\mathbf{z}'$ is nonpositive, i.e., the fit gradient of the MaxEnt solution moves opposite to the change in the fit gradient that induced it. Since Tikhonov solutions generally follow the MaxEnt solutions, these Tikhonov gradients would be closer to the original MaxEnt gradient than the previous ones. This explains why the bare fit gradient vectors in SCT initially move closer to the original MaxEnt gradient vector (Figs. 8 and 9). However, the MaxEnt solutions using SCT default models keep drifting away in the same direction, so the Tikhonov solutions and their fit gradients would eventually also start moving away from the original MaxEnt. The effective fit gradient, on the other hand, is an average of these bare gradients and thus can be closer to the original MaxEnt than any of its summands. This happens when the bare gradients circulate around the original MaxEnt gradient, which is the case in SCT.

The dynamics described above is depicted schematically in Fig. 10. In this diagram, we represent the log spectra as points, so the family of equivalent default models $\mathbf{d}^{\star(1)}, \mathbf{d}^{\star(2)}, \ldots$ all lie on a straight line between the initial default model $\mathbf{d}^{(0)}$ and its MaxEnt solution $\mathbf{f}^\star$. This line is specified by the fit gradient vector $\mathbf{z}^\star$. In SCT, $\mathbf{z}^\star$ is replaced by $\mathbf{z}^{(t)}$, the bare fit gradients at the Tikhonov solutions $\mathbf{f}^{(t)}$, leading to a set of alternative default models $\mathbf{d}^{(t)}$ that approximates the equivalent family $\mathbf{d}^{\star(t)}$. Each approximate default model $\mathbf{d}^{(t)}$ has its

own MaxEnt solution $\mathbf{f}^{\star(t)}$ and Tikhonov solution $\mathbf{f}^{(t)}$. In this two-dimensional case, according to Eq. (C2), the Tikhonov solution $\mathbf{f}^{(t)}$ and the MaxEnt solution at the next iteration $\mathbf{f}^{\star(t+1)}$ must be on opposite sides of the MaxEnt solution $\mathbf{f}^{\star(t)}$. Therefore, the fit gradients of Tikhonov $\mathbf{z}^{(t)}$ would initially get closer to $\mathbf{z}^{\star}$ before moving away. Also, note how the effective fit gradients (i.e., consecutive weighted averages of $\mathbf{z}^{(t)}$) get monotonically closer to $\mathbf{z}^{\star}$. This is the result of $\mathbf{z}^{(t)}$ moving from one side of $\mathbf{z}^{\star}$ to the other, and the weights $\mu^{(t)}/\alpha^{(t)}$ getting lower for higher iterations.

APPENDIX D: SCT AS RESET NEWTON METHOD

Another perspective on SCT is seeing it as a variant of Newton's method for optimization. Assuming that the optimal regularization parameter for satisfying the discrepancy principle is somehow known in advance, solving the MaxEnt problem of Eq. (28) reduces to optimizing the MaxEnt objective function of Eq. (27). Using the default model $\mathbf{d}$ as an initial guess, an improved solution can be obtained using Newton's method as

$$\mathbf{d}' = \mathbf{d} + \gamma \mathbf{H}^{-1}\, \nabla S^{\mathbf{d}}, \tag{D1}$$

where γ is a small step size and $\mathbf{H}$ is minus the Hessian of the objective function at the default model [which coincides with minus the Tikhonov Hessian in Eq. (B2)] and $\nabla S^{\mathbf{d}}$ is its gradient, also evaluated at the default model.

The vector $\mathbf{x} \coloneqq \mathbf{H}^{-1}\nabla S^{\mathbf{d}}$ is the solution of

$$\begin{aligned} &[\mathbf{K}^{\mathsf{T}}\mathbf{K} + \alpha\, \mathbf{D}^{-1}]\mathbf{x} = \mathbf{K}^{\mathsf{T}}[\mathbf{g} - \mathbf{K}\mathbf{d}] \\ \Leftrightarrow\ &[\mathbf{K}^{\mathsf{T}}\mathbf{K} + \alpha\, \mathbf{D}^{-1}][\mathbf{x} + \mathbf{d}] = \mathbf{K}^{\mathsf{T}}\mathbf{g} + \alpha\mathbf{e}. \end{aligned} \tag{D2}$$

Comparing with Eq. (A2), we see that $\mathbf{x} + \mathbf{d}$ equals the Tikhonov solution, and thus Newton's update formula can be written as

$$\mathbf{d}' = \mathbf{d} + \gamma\big(\mathbf{f}_{\text{Tikhonov}} - \mathbf{d}\big), \tag{D3}$$

which is precisely the mixing formula used in SCT. Note that the entropy has no contribution to the gradient vector $\nabla S^{\mathbf{d}}$ at the starting default model. However, at later steps there is an additional term $-\alpha^{(t)}\ln(d_i^{(t)}/d_i^{(0)})$. SCT ignores this term; thus, SCT is equivalent to Newton's method, where the default model is always reset to its most recent solution.

Interestingly, the missing entropy contributions can be expressed in terms of the effective fit gradients:

$$-\alpha^{(t)}\ln\left(\frac{d_i^{(t)}}{d_i^{(0)}}\right) = -\frac{\alpha^{(t)}}{\tilde{\alpha}^{(t)}}\tilde{\mathbf{z}}^{(t)}. \tag{D4}$$

Therefore, we can recover the full Newton's method as a variant of the SCT method where the data is modified at each step to take into account the residuals of the previous Tikhonov solutions.

[1] P. C. Hansen, Numerical tools for analysis and solution of Fredholm integral equations of the first kind, Inverse Probl. **8**, 849 (1992).

[2] P. C. Hansen, *Discrete Inverse Problems* (SIAM, Philadelphia, 2010).

[3] R. N. Silver, D. S. Sivia, and J. E. Gubernatis, Maximum-entropy method for analytic continuation of quantum Monte Carlo data, Phys. Rev. B **41**, 2380 (1990).

[4] M. Jarrell and J. E. Gubernatis, Bayesian inference and the analytical continuation of imaginary-time quantum Monte Carlo data, Phys. Rep. **269**, 133 (1996).

[5] O. Gunnarsson, M. W. Haverkort, and G. Sangiovanni, Analytical continuation of imaginary axis data using maximum entropy, Phys. Rev. B **81**, 155107 (2010).

[6] D. Bergeron and A. M. S. Tremblay, Algorithms for optimized maximum entropy and diagnostic tools for analytic continuation, Phys. Rev. E **94**, 023303 (2016).

[7] S. R. White, The average spectrum method for the analytic continuation of imaginary-time data, 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.

[8] A. W. Sandvik, Stochastic method for analytic continuation of quantum Monte Carlo data, Phys. Rev. B **57**, 10287 (1998).

[9] O. F. Syljuåsen, Using the average spectrum method to extract dynamics from quantum Monte Carlo simulations, Phys. Rev. B **78**, 174429 (2008).

[10] S. Fuchs, T. Pruschke, and M. Jarrell, Analytical continuation of quantum Monte Carlo data by stochastic analytical inference, Phys. Rev. E **81**, 056701 (2010).

[11] A. W. Sandvik, Constrained sampling method for analytic continuation, Phys. Rev. E **94**, 063308 (2016).

[12] K. Ghanem, Stochastic analytic continuation: A Bayesian approach, Ph.D. thesis, RWTH Aachen University, 2017.

[13] K. Ghanem and E. Koch, Average spectrum method for analytic continuation: Efficient blocked-mode sampling and dependence on the discretization grid, Phys. Rev. B **101**, 085111 (2020).

[14] K. Ghanem and E. Koch, Extending the average spectrum method: Grid point sampling and density averaging, Phys. Rev. B **102**, 035114 (2020).

[15] H. Shao and A. W. Sandvik, Progress on stochastic analytic continuation of quantum Monte Carlo data, Phys. Rep. **1003**, 1 (2023).

[16] H. J. Vidberg and J. W. Serene, Solving the Eliashberg equations by means of N-point Padé approximants, J. Low Temp. Phys. **29**, 179 (1977).

[17] K. S. D. Beach, R. J. Gooding, and F. Marsiglio, Reliable padé analytical continuation method based on a high-accuracy symbolic computation algorithm, Phys. Rev. B **61**, 5147 (2000).

[18] A. Östlin, L. Chioncel, and L. Vitos, One-particle spectral function and analytic continuation for many-body implementation in the exact muffin-tin orbitals method, Phys. Rev. B **86**, 235107 (2012).

[19] J. Schött, I. L. M. Locht, E. Lundin, O. Grånäs, O. Eriksson, and I. Di Marco, Analytic continuation by averaging Padé approximants, Phys. Rev. B **93**, 075104 (2016).

[20] A. S. Mishchenko, Stochastic optimization method for analytic continuation, in *Correlated Electrons: From Models to Materials*, edited by E. Pavarini, E. Koch, F. Anders, and M. Jarrell (Forschungszentrum Jülich, Jülich, 2012).

[21] F. Bao, Y. Tang, M. Summers, G. Zhang, C. Webster, V. Scarola, and T. A. Maier, Fast and efficient stochastic optimization for analytic continuation, Phys. Rev. B **94**, 125149 (2016).

[22] L.-F. Arsenault, A. Lopez-Bezanilla, O. A. von Lilienfeld, and A. J. Millis, Machine learning for many-body physics: The case of the Anderson impurity model, Phys. Rev. B **90**, 155136 (2014).

[23] L.-F. Arsenault, R. Neuberg, L. A. Hannah, and A. J. Andrew J. Millis, Projected regression method for solving Fredholm integral equations arising in the analytic continuation problem of quantum physics, Inverse Probl. **33**, 115007 (2017).

[24] H. Yoon, J.-H. Sim, and M. J. Han, Analytic continuation via domain knowledge free machine learning, Phys. Rev. B **98**, 245101 (2018).

[25] R. Fournier, L. Wang, O. V. Yazyev, and Q. S. Wu, Artificial Neural Network Approach to the Analytic Continuation Problem, Phys. Rev. Lett. **124**, 056401 (2020).

[26] E. Vitali, M. Rossi, L. Reatto, and D. E. Galli, Ab initio low-energy dynamics of superfluid and solid ^{4}He, Phys. Rev. B **82**, 174510 (2010).

[27] G. Bertaina, D. E. Galli, and E. Vitali, Statistical and computational intelligence approach to analytic continuation in quantum Monte Carlo, Adv. Phys.: X **2**, 302 (2017).

[28] N. S. Nichols, P. Sokol, and A. Del Maestro, Parameter-free differential evolution algorithm for the analytic continuation of imaginary time correlation functions, Phys. Rev. E **106**, 025312 (2022).

[29] O. Gunnarsson, M. W. Haverkort, and G. Sangiovanni, Analytic continuation of imaginary axis data for optical conductivity, Phys. Rev. B **82**, 165125 (2010).

[30] D. L. Phillips, A technique for the numerical solution of certain integral equations of the first kind, J. ACM **9**, 84 (1962).

[31] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems* (V. H. Winston & Sons, Washington, D.C.; John Wiley & Sons, New York, 1977).

[32] In the inverse-problem literature, it is common for Tikhonov regularization parameter α to appear squared. We choose to deviate from that convention to make the correspondence with the regularization parameter of MaxEnt more seamless.

[33] The most general form of Tikhonov is obtained by replacing the L_2-norm with a bilinear function $\|\mathbf{f}-\mathbf{f}_0\|^2_{\mathbf{M}}$, where $\mathbf{M}$ is some positive-definite matrix and $\mathbf{f}_0$ is an arbitrary vector that acts as a default model.

[34] One can obtain trivial grid independence by including one square root of the grid weights in the spectrum vector and the other square root in the kernel matrix. In this case, the discretized L_2-norm of $\mathbf{f}$ corresponds to the continuous l_2-norm of $f(x)$. However, using this form implies a specific choice of the measure on x that is equivalent to fixing the grid density $\rho(x)$ to be uniform.

[35] C. Groetsch, *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*, Chapman & Hall/CRC Research Notes in Mathematics Series (Pitman Advanced Pub. Program, Boston, 1984).

[36] V. A. Morozov, Criteria for selection of regularization parameter, in *Methods for Solving Incorrectly Posed Problems* (Springer, New York, 1984), pp. 32–64.

[37] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems* (SIAM, Philadelphia, 1995).

[38] K. S. D. Beach, Identifying the maximum entropy method as a special limit of stochastic analytic continuation, arXiv:cond-mat/0403055.

[39] M. Jarrell, The maximum entropy method: Analytic continuation of QMC data, in *Correlated Electrons: From Models to Materials*, edited by E. Pavarini, E. Koch, F. Anders, and M. Jarrell (Forschungszentrum Jülich, Jülich, 2012).

[40] G. J. Kraberger, R. Triebl, M. Zingl, and M. Aichhorn, Maximum entropy formalism for the analytic continuation of matrix-valued Green's functions, Phys. Rev. B **96**, 155128 (2017).

[41] J. Fei, C.-N. Yeh, D. Zgid, and E. Gull, Analytical continuation of matrix-valued functions: Carathéodory formalism, Phys. Rev. B **104**, 165111 (2021).