

Article

# Bryan's Maximum Entropy Method—Diagnosis of a Flawed Argument and Its Remedy

Alexander Rothkopf 

Faculty of Science and Technology, University of Stavanger, 4021 Stavanger, Norway; alexander.rothkopf@uis.no

Received: 25 July 2020; Accepted: 15 September 2020; Published: 17 September 2020



**Abstract:** The Maximum Entropy Method (MEM) is a popular data analysis technique based on Bayesian inference, which has found various applications in the research literature. While the MEM itself is well-grounded in statistics, I argue that its state-of-the-art implementation, suggested originally by Bryan, artificially restricts its solution space. This restriction leads to a systematic error often unaccounted for in contemporary MEM studies. The goal of this paper is to carefully revisit Bryan's train of thought, point out its flaw in applying linear algebra arguments to an inherently nonlinear problem, and suggest possible ways to overcome it.

**Keywords:** Bayesian inference; inverse problems; maximum entropy method; Bayesian reconstruction method; singular value decomposition; systematic error

## 1. Introduction

Ill-posed inverse problems represent a major challenge to research in many disciplines. They arise, e.g., when one wishes to reconstruct the shape of an astronomical object after its light has passed through a turbulent atmosphere and an imperfect telescope or when we image the interior of a patient's skull during a computer tomography scan. In a more theoretical setting, extracting spectral functions from numerical simulations of strongly correlated quantum fields constitutes another example. The common difficulty among these tasks lies in the fact that we do not have direct access to the quantity of interest (from here on referred to as  $\rho$ ) but instead only to a distorted representation of it, measured in our experiment (from here on denoted by  $D$ ). Extracting  $\rho$  from  $D$ , in general, requires us to solve an inherently nonlinear optimization problem, which we construct and discuss in detail below.

Let us consider the common class of inverse problems, where the quantity of interest  $\rho$  and the measured data  $D$  are related via an integral convolution

$$\tilde{D}(\tau) = \int_{\omega_{\min}}^{\omega_{\max}} d\omega K(\tau, \omega) \rho(\omega), \quad 0 \leq \tau \leq \tau_{\max}, \quad (1)$$

with a kernel function  $K(\tau, \omega)$ . For the sake of simplicity let us assume (as is often the case in practice) that the function  $K$  is exactly known. The task at hand is to estimate the function  $\rho$  that underlies the observed  $D$ . The ill-posedness (and ill-conditioning) of this task is readily spotted if we acknowledge that our data comes in the form of  $N_\tau$  discrete estimates  $D_i = \tilde{D}(\tau_i) + \eta$  of the true function  $\tilde{D}$ , where  $\eta$  denotes a source of noise. In addition, we need to approximate the integral in some form for numerical treatment. In its simplest form, writing it as a sum over  $N_\omega$  bins we obtain

$$D_i = \sum_{l=1}^{N_\omega} \Delta\omega_l K_{il} \rho_l. \quad (2)$$

At this point, we are asked to estimate  $N_\omega \gg N_\tau$  optimal parameters  $\rho_l$  from  $N_\tau$  data points, which themselves carry uncertainty. A naive  $\chi^2$  fit of the  $\rho_l$ 's is of no use, since it would produce an

infinite number of degenerate solutions, which all reproduce the set of  $D_i$ 's within their error bars. Only if we introduce an appropriate regularization can the problem be made well-posed, and it is this regularization which in general introduces nonlinearities.

Bayesian inference represents one way to regularize the inversion task. It provides a systematic procedure for how additional (so called prior) knowledge can be incorporated to that effect. Bayes theorem

$$P[\rho|D, I] = \frac{P[D|\rho, I]P[\rho|I]}{P[D|I]}, \quad (3)$$

states that the *posterior* probability  $P[\rho|D, I]$  for some set of parameters  $\rho_l$  to be the true solution of the inversion problem is given by the product of the *likelihood* probability  $P[D|\rho, I]$  and the *prior* probability  $P[\rho|I]$ . The  $\rho$  independent normalization  $P[D|I]$  is often referred to as the *evidence*. Assuming that the noise  $\eta$  is Gaussian, we may write the likelihood as

$$P[D|\rho, I] = \exp[-\mathcal{L}], \quad \mathcal{L} = \frac{1}{2} \sum_{ij} (D_i - D_i^\rho) C_{ij}^{-1} (D_j - D_j^\rho), \quad (4)$$

where  $C_{ij}$  is the unbiased covariance matrix of the measured data with respect to the true mean and  $D_i^\rho$  refers to the synthetic data that one obtains by inserting the current set of  $\rho_l$  parameters into (2). A  $\chi^2$  fit would simply return one of the many degenerate extrema of  $\mathcal{L}$ , hence being referred to as a maximum likelihood fit.

The important ingredient of Bayes theorem is the presence of the prior probability, often expressed in terms of a regulator functional  $\mathcal{S}$

$$P[\rho|I] = \exp[\mathcal{S}]. \quad (5)$$

It is here where pertinent domain knowledge can be encoded. For the study of intensity profiles of astronomical objects and hadronic spectral functions for example, it is a-priori known that the values of  $\rho$  must be positive. Depending on which type of information one wishes to incorporate, the explicit form of  $\mathcal{S}$  will be different. It is customary to parameterize the shape of the prior distribution by two types of hyperparameters, the default model  $m(\omega)$  and a confidence function  $\alpha(\omega)$ . The discretized  $m_l = m(\omega_l)$  represents the maximum of the prior  $P[\rho_l|I(m, \alpha)]$  for each parameter  $\rho_l$  and each  $\alpha_l = \alpha(\omega_l)$  its corresponding spread.

Once both the likelihood and prior probability are set, we may search for the maximum a posteriori (MAP) point estimate of the  $\rho_l$ 's via

$$\left. \frac{\delta}{\delta \rho} P[\rho|D, I] \right|_{\rho=\rho^{\text{Bayes}}} = 0. \quad (6)$$

$\rho^{\text{Bayes}}$  constitutes the most probable parameters given our data and prior knowledge. (In the case that no data is provided, the Bayesian reconstruction will simply reproduce the default model  $m$ .) Note that since the exponential function is monotonous, instead of finding the extremum of  $P[\rho|D, I]$ , in practice, one often considers the extremum of  $\mathcal{L} - \mathcal{S}$  directly.

The idea of introducing a regularization in order to meaningfully estimate the most probable set of parameters underlying observed data has a long history. As early as 1919 [1], it was proposed to combine what we here call the likelihood with a smoothing regulator. Let us have a look at three choices of regulators from the literature: the historic Tikhonov (TK) regulator [2] (1943–), the Shannon–Jaynes

entropy deployed in the Maximum Entropy Method (MEM) [3,4] (1986–), and the more recent Bayesian Reconstruction (BR) method [5] regulator (2013–), respectively,

$$\mathcal{S}_{\text{TK}} = - \sum_l \Delta\omega_l \alpha_l \frac{1}{2} (\rho_l - m_l)^2, \tag{7}$$

$$\mathcal{S}_{\text{MEM}} = \sum_l \Delta\omega_l \alpha_l (\rho_l - m_l - \rho_l \log[\frac{\rho_l}{m_l}]), \tag{8}$$

$$\mathcal{S}_{\text{BR}} = \sum_l \Delta\omega_l \alpha_l (1 - \frac{\rho_l}{m_l} + \log[\frac{\rho_l}{m_l}]). \tag{9}$$

Both  $\mathcal{S}_{\text{MEM}}$  and  $\mathcal{S}_{\text{BR}}$  are axiomatically constructed, incorporating the assumption of positivity of the function  $\rho$ . The assumption manifests itself via the presence of a logarithmic term that forces  $\rho$  to be positive-semidefinite in the former and positive-definite in the latter case. It is this logarithm that is responsible for the numerical optimization problem (6) to become genuinely nonlinear.

Note that all three functions are concave, which (as proven for example in [6]) guarantees that if an extremum of  $P[\rho|D, I]$  exists, it is unique—i.e., within the  $N_\omega$  dimensional solution space spanned by the discretized parameters  $\rho_l$ , in the case that a Bayesian solution exists, we will be able to locate it with standard numerical methods in a straightforward fashion.

## 2. Diagnosis of the Problem in Bryan’s MEM

In this section, we investigate the consequences of the choice of regularization on the determination of the most probable spectrum. Starting point is the fully linear Tikhonov regularization, on which a lot of the intuition in the treatment of inverse problems is built. We then continue to the Maximum Entropy Method, which amounts to a genuinely nonlinear regularization and point out how arguments, which were valid in the linear case, fail in the nonlinear context.

### 2.1. Tikhonov Regularization

The Tikhonov choice amounts to a Gaussian prior probability, which allows  $\rho$  to take on both positive and negative values. The default model  $m_l$  denotes the value for  $\rho_l$ , which was most probable before the arrival of the measured data  $D$  (e.g., from a previous experiment), and  $\alpha_l$  represents our confidence into the prior knowledge (e.g., the uncertainty of the previous experiment).

Since both (4) and (7) are at most quadratic in  $\rho_l$ , taking the derivative in (6) leads to a set of linear equations that need to be solved to compute the Bayesian optimal solution  $\rho^{\text{Bayes}}$ . It is this fully linear scenario from which most intuition is derived when it comes to the solution space of the inversion task. Indeed, we are led to the following relations

$$- \alpha_l (\rho_l - m_l) = \sum_i K_{il} \frac{\delta \mathcal{L}}{\delta D_i^\rho}, \quad -\hat{\alpha}(\vec{\rho} - \vec{m}) = \hat{K}^T \frac{\delta \vec{\mathcal{L}}}{\delta D^\rho}, \tag{10}$$

which can be written solely in terms of linear vector-matrix operations. Note that in this case  $\delta \mathcal{L} / \delta D^\rho$  contains the vector  $\vec{\rho}$  in a linear fashion. (10) invites us to parameterize the function  $\rho$  via its deviation from the default model

$$\vec{\rho} = \vec{m} + \vec{a}, \tag{11}$$

and to look for the optimal set of parameters  $a_l$ . Here, we may safely follow Bryan [7] and investigate the singular values of  $K^T = U \Sigma V^t$  with  $U$  being an  $N_\omega \times N_\omega$  special orthogonal matrix,  $\Sigma$  an  $N_\omega \times N_\tau$  matrix with  $N_\tau$  nonvanishing diagonal entries, corresponding to the singular values of  $K^T$  and  $V^t$  being an  $N_\tau \times N_\tau$  special orthogonal matrix. We are led to the expression

$$-\hat{\alpha}\vec{a} = \hat{U}\hat{\Sigma}\hat{V}^t \frac{\delta\vec{\mathcal{L}}}{\delta D^\rho}, \tag{12}$$

which tells us that in this case, the solution of the Tikhonov inversion lies in a functional space spanned by the first  $N_\tau$  columns of the matrix  $\hat{U}$  (usually referred to as the SVD or singular subspace) around the default model  $\vec{m}$ —i.e., we can parameterize  $\rho$  as

$$\vec{\rho} = \vec{m} + \sum_{k=1}^{N_\tau} c_k \vec{U}_k. \tag{13}$$

The point here is that if we add to this SVD space parametrization any further column of the matrix  $\hat{U}$ , it directly projects into the null space of  $K$  via  $\hat{K} \cdot \vec{U}_{k>N_\tau} = 0$ . In turn, such a column does not contribute to computing synthetic data via (2). As was pointed out in [8], in such a linear scenario, the SVD subspace is indeed all there is. If we add extra columns of  $\hat{U}$  to the parametrization of  $\rho$ , these do not change the likelihood. Thus, the corresponding parameter  $c_j$  of that column is not constrained by data and will come out as zero in the optimization procedure of (6), as it encodes a deviation from the default model, which is minimized by  $\mathcal{S}$ .

### 2.2. Maximum Entropy Method

Now that we have established that in a fully linear context the arguments based on the SVD subspace are indeed justified, let us continue to the Maximum Entropy Method, which deploys the Shannon–Jaynes entropy as regulator.  $\mathcal{S}_{MEM}$  encodes as prior information, e.g., the positivity of the function  $\rho$ , which manifests itself in the presence of the logarithm in (8). This logarithm however also entails that we are now dealing with an inherently nonlinear optimization problem in (6). Carrying out the functional derivative with respect to  $\rho$  on  $\mathcal{L} - \mathcal{S}$ , we are led to the following relation

$$-\alpha_l \log\left[\frac{\rho_l}{m_l}\right] = \sum_i K_{il} \frac{\delta\mathcal{L}}{\delta D_i^\rho}. \tag{14}$$

As suggested by Bryan [7], let us introduce the completely general (since  $\rho > 0$ ) and nonlinear redefinition of the parameters

$$\rho_l = m_l \exp[a_l]. \tag{15}$$

Inserting (15) into (14), we are led to an expression that is formally quite similar to the result obtained in the Tikhonov case

$$-\alpha_l a_l = \sum_i K_{il} \frac{\delta\mathcal{L}}{\delta D_i^\rho}, \quad -\hat{\alpha}\vec{a} = \hat{K}^T \frac{\delta\vec{\mathcal{L}}}{\delta D^\rho}. \tag{16}$$

While at first sight this relation is also amenable to be written as a linear relation for  $\vec{a}$ , it is actually fundamentally different from (10), since due to its nonlinear nature,  $\vec{a}$  enters  $\delta\mathcal{L}/\delta D^\rho$  via componentwise exponentiation. It is here, when attempting to make a statement about such a nonlinear relation with the tools of linear algebra, that we run into difficulties. What do I mean by that? Let us push ahead and introduce the SVD decomposition of the transpose Kernel as before,

$$-\hat{\alpha}\vec{a} = \hat{U}\hat{\Sigma}\hat{V}^t \frac{\delta\vec{\mathcal{L}}}{\delta D^\rho}. \tag{17}$$

At first sight, this relation seems to imply that the vector  $\vec{a}$ —which encodes the deviation from the default model (this time multiplicatively)—is restricted to the SVD subspace, spanned by the first  $N_\tau$  entries of the matrix  $\hat{U}$ . My claim (as put forward most recently in [9]) is that this conclusion is

false, since this linear-algebra argument is not applicable when working with (16). Let us continue to setup the corresponding SVD parametrization advocated, e.g., in [8]

$$\rho_l = m_l \exp \left[ \sum_{k=1}^{N_\tau} c_k U_{lk} \right]. \quad (18)$$

In contrast to (13), the SVD space is not all there is to (18) (see explicit computations in Appendix A). This we can see by taking the  $N_\tau + 1$  column of the matrix  $\hat{U}$ , exponentiating it componentwise, and applying it to the matrix  $K$ . In general, we get

$$\sum_l K_{il} \exp [U_{l(N_\tau+1)}] \neq 0. \quad (19)$$

This means that if we add additional columns of the matrix  $\hat{U}$  to the parametrization in (18), they do not automatically project into the null-space of the Kernel (see the explicit example in Appendix A of this manuscript) and thus will contribute to the likelihood. In turn, the corresponding parameter  $c_j$  related to that column will not automatically come out to be zero in the minimization procedure (6). Hence, we cannot a priori disregard its contribution and thus, the contribution of this direction of the search space, which is not part of the SVD subspace. We thus conclude that limiting the solution space in the MEM to the singular subspace amounts to an ad-hoc procedure, motivated by an incorrect application of linear-algebra arguments to a fully nonlinear optimization problem.

A representative example from the literature, where the nonlinear character of the parametrization of  $\rho$  is not taken into account, is the recent [8] (see Equations (7) and (12) in that manuscript). We emphasize that one does not apply a column of the matrix  $\hat{U}$  itself to the matrix  $K$  but the componentwise exponentiation of this column. This operation does not project into the null-space.

In those cases where we have only few pieces of reliable prior information and our datasets are limited, restricting to the SVD subspace may lead to significantly distorted results (as shown explicitly in [10]). On the other hand, its effect may be mild if the default model already encodes most of the relevant features of the final result and the number of datapoints is large, so that the SVD subspace is large enough to (accidentally) encompass the true Bayesian solution sought after in (6). Independent of the severity of the problem, the artificial restriction to the SVD subspace in Bryan's MEM is a source of systematic error, which needs to be accounted for when the MEM is deployed as precision tool for inverse problems.

Being liberated from the SVD subspace does not lead to any conceptual problems either. We have brought to the table  $N_\tau$  points of data, as well as  $N_\omega$  points of prior information in the form of the default model  $m$  (as well as its uncertainty  $\alpha$ ). This is enough information to determine the  $N_\omega$  parameters  $\rho_l$  uniquely, as proven in [6].

Recognizing that linear-algebra arguments fail in the MEM setup also helps us to understand some of the otherwise perplexing results found in the literature. If the singular subspace were all there is to the parametrization of  $\rho_l$  in (18), then it would not matter whether we use the first  $N_\tau$  columns of  $\hat{U}$  or just use the  $N_\tau$  columns of  $\hat{K}^T$  directly. Both encode the same target space, the difference being only that the columns of  $\hat{U}$  are orthonormal. However, as was clearly seen in Figure 28 of [11], using the SVD parametrization or the columns of  $\hat{K}^T$  leads to significantly different results in the reconstructed features of  $\rho$ . If the MEM were a truly linear problem, these two parameterizations gave exactly the same result. The finding that the results do not agree emphasizes that the MEM inversion is genuinely nonlinear and the restriction to the SVD subspace is ad hoc.

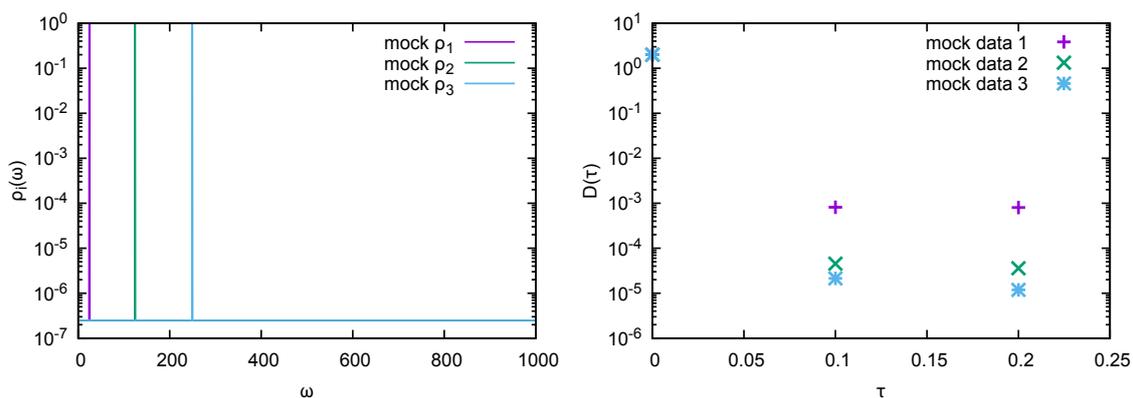
### 2.3. Numerical Evidence for the Inadequacy of the SVD Subspace

Let us construct an explicit example to illustrate the fact that the solution of the MEM reconstruction may lie outside of the SVD search space. Since Bryan's derivation of the SVD subspace proceeds independently of the particular form of the kernel  $K$ , the provided data  $D$ , and the choice of

the default model  $m$ , we are free to choose them at will. For our example, we consider a transform often encountered among inverse problems related to the physics of strongly correlated quantum systems.

One then has  $K(\tau, \omega) = 1/(\omega^2 + \tau^2)$  and we may set  $m = 1$ . With  $\alpha$  entering simply as scaling factor in (16), we do not consider it further in the following. Let me emphasize again that the arguments leading to (16) did not make any reference to the data we wish to reconstruct. Here, we will consider three datapoints that encode a single delta peak at  $\omega = \omega_0$ , embedded in a flat background.

Now, let us discretize the frequency domain between  $\omega_{\min} = 1/2000$  and  $\omega_{\max} = 1000$  with  $N_\omega = 2000$  points. Together with the choice of  $\tau_{\min} = 0$ ,  $\tau_{\max} = 0.2$ , and  $N_\tau = 3$ , this fully determines the kernel matrix  $K_{il}$  in (2). Three different mock functions  $\rho_i$  are considered, with  $\omega_0^{(1)} = 25$ ,  $\omega_0^{(2)} = 125$ , and  $\omega_0^{(3)} = 250$ , the background is assigned the magnitude  $1/N_\omega^2$  (see Figure 1 left).



**Figure 1.** The mock function  $\rho$  (left) and corresponding mock data (right) deployed in the explicit example discussed in the main text, which shows that the extremum of the posterior is not necessarily located within the singular value decomposition (SVD) subspace.

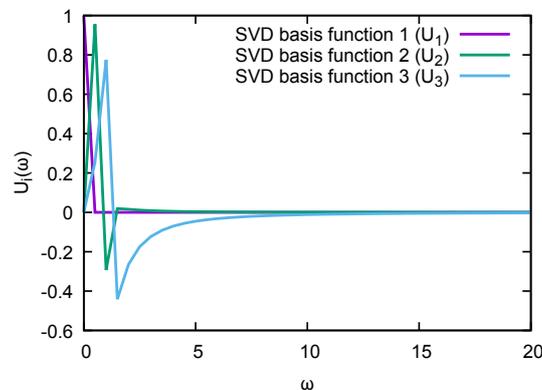
Bryan’s argument states that in the presence of three datapoints (see Figure 1 right), irrespective of the data encoded in those datapoints, the extremum of the posterior must lie in the space spanned by the exponentiation of the three first columns of the matrix  $\hat{U}$ , obtained from the SVD of the transpose kernel  $K^t$ . In Figure 2, its first three columns are explicitly plotted. Note that while they do show some peaked behavior around the origin, they quickly flatten off above  $\omega = 10$ . From this inspection by eye, it already follows that it will be very difficult to linearly combine  $U_1(\omega)$ ,  $U_2(\omega)$ , and  $U_3(\omega)$  into a sharply peaked function, especially for a peak located at  $\omega > 10$ .

Assuming that the data comes with constant relative error  $\Delta D/D = \kappa$ , let us find out how well we can reproduce it within Bryan’s search space. A minimization carried out by Mathematica (see the explicit code in Appendix B) tells us that  $\mathcal{L}_{\min}^1 \approx 10^6 \kappa^{-2}$ ,  $\mathcal{L}_{\min}^2 \approx 10^9 \kappa^{-2}$ , and  $\mathcal{L}_{\min}^3 \approx 10^{10} \kappa^{-2}$ —i.e., we clearly see that we are not able to reproduce the provided datapoints well (i.e.,  $\chi^2/\text{d.o.f} = \mathcal{L}/3 \gg 1$ ) and that as the deviation becomes more and more pronounced, the higher the delta peak is positioned along  $\omega$ . In the full search space on the other hand, we can always find a set of  $\rho_i$ ’s which bring the likelihood as close to zero as desired.

Minimizing the likelihood of course is not the whole story in a Bayesian analysis, which is why we have to take a look at the contributions from the regulator term  $\mathcal{S}$ . Remember that by definition, it is negative. We find that for all three mock functions  $\rho_i$ , Bryan’s best fit of the likelihood leads to values of  $\mathcal{S} \approx -0.24$ . On the other hand, a value of  $\mathcal{S} \approx -200$  is obtained in the full search space for the parameter set  $\rho_i$  which contains only one entry that is unity and all other entries being set to the background at  $1/N_\omega^2$ .

From the above inspection of the extremum of  $\mathcal{L}$  and the associated value of  $\mathcal{S}$  inside and outside of Bryan’s search space, we arrive at the following: Due to the very limited structure present in the SVD basis functions  $U_1(\omega)$ ,  $U_2(\omega)$ , and  $U_3(\omega)$ , it is in general not possible to obtain a good reconstruction of the input data. This in turn leads to a large minimal value of  $\mathcal{L}$  accessible within the SVD subspace.

Due to the fact that  $\mathcal{S}$  cannot compensate for a large value of  $\mathcal{L}$ , we have constructed an explicit example where at least one set of  $\rho_i$ 's (the one that brings  $\mathcal{L}$  close to zero in the full search space) leads to a smaller value of  $\mathcal{L} - \mathcal{S}$  and thus to a larger posterior probability than any of the  $\rho_i$ 's within the SVD subspace.



**Figure 2.** The SVD basis functions  $U_1(\omega)$ ,  $U_2(\omega)$ , and  $U_3(\omega)$ , which Bryan's argument suggests should capture the extremum of the posterior for all three mock functions  $\rho_i$  after exponentiation. Note that the SVD basis functions flatten out about  $\omega = 10$ . Their domain extends to  $\omega_{\max} = 1000$  but their value stays close to zero.

In other words, we have constructed a concrete example in which the MEM solution, given by the global extremum of the posterior, is not contained in the SVD subspace.

### 3. Remedy of the Problem

Having identified the shortcoming of Bryan's MEM as the ad-hoc restriction of the search space to the SVD subspace, we must consider how to overcome it. In this section, we present two possible routes to do so. In the first subsection, systematic modifications of the MEM, previously put forward by the author, are considered, which extend and generalize the SVD search space. In the second subsection, we present a collection of modern approaches to Bayesian inference developed by the research community which are not affected by limited search spaces. Particular attention is paid to the BR method recently codeveloped by the author.

#### 3.1. Staying within MEM

The most naive path to take is to simply add additional columns of the matrix  $\hat{U}$  to the nonlinear parametrization of (18), as proposed by the author in 2011 in [10] and successfully applied in practice in [12]. This procedure systematically approaches the full search space, in which the unique Bayesian solution is located (We are surprised by the statements made recently in [8], which claim that the numerics presented in [10] and thus also in [12] were unreliable. Both papers have been peer-reviewed and the underlying code has been freely available for many years on the author's website (ExtMEM link, <http://www.alexrothkopf.de/files/ExtMEMv3.tar.bz2>). Furthermore, the examples presented in [10] are explicitly parameterized and thus lend themselves to straightforward reproduction.).

Investigating in detail the shape of the SVD basis functions, it was found that they may not offer the same resolution for features located at different  $\omega$ 's. In practice, this means that one often needs to add a large number of additional SVD basis functions before the structures encoded in the data  $D_i$  are sufficiently resolved. In order to remedy this situation, we can exploit that we are not bound by the SVD subspace and instead, as suggested in [13], deploy for example Fourier basis functions in (18)—i.e., we replace the columns of  $\hat{U}$  by a linear combination of *sin* and *cos* terms. This proposal has been shown to lead to a resolution capacity of the MEM that is independent of the position of the encoded structures in  $\rho$  along  $\omega$  and has been put in practice, e.g., in [14].

Thanks to the availability of highly efficient numerical optimization algorithms, such as the limited-memory BFGS (Broyden–Fletcher–Goldfarb–Shanno) algorithm, it is nowadays easily possible to directly carry out the optimization task (6) in the full  $N_\omega$  dimensional search space, even if  $N_\omega \sim \mathcal{O}(10^3)$ . Together with the proof of uniqueness of the Bayesian solution and the related convexity of  $\mathcal{L} - \mathcal{S}$ , there does not exist any principal need to restrict to a low-dimensional subspace.

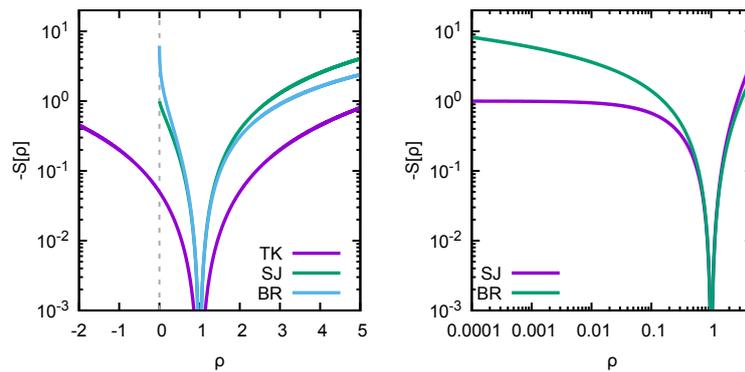
### 3.2. Beyond MEM

An active community is working on devising improved Bayesian approaches to inverse problems in the sciences. On the one hand, there exist works that go beyond the maximum a-posteriori approach and proceed towards sampling the posterior, such as the stochastic analytic continuation method [15], of which the MEM is but one special limit as shown in [16]. Together with the SOM method presented in [17], these stochastic methods have for example, been deployed in the study of nuclear matter at high temperatures in [18]. Recently, the community has seen heightened activity in exploring the use of neural networks for the solution of inverse problems, e.g., in [19–22].

Let me focus here on one recent Bayesian approach, the BR method, presented in [5] with regulator (9), which was designed with the particular one-dimensional inverse problem of (2) in mind. The motivation to develop this new method on the one hand arose from the observation that the specific form of the Shannon–Jaynes regulator of the MEM can pose a problem in finding the optimal Bayesian solution. Consider the (negative of the) integrands of (7)–(9), plotted for an arbitrary choice of  $\alpha = 0.1$  and  $m = 1$  in the left panel of Figure 3. By construction, all of them have an extremum at  $\rho = m$  and, as expected, only the Tikhonov regulator allows  $\rho$  to take on values smaller than zero. Both the MEM and BR regulator diverge as  $\rho \rightarrow \infty$ , but their behavior close to  $\rho = 0$  differs markedly. Let us have a closer look in the right panel of Figure 3. Plotted in log-log scale, we see that while the BR regulator diverges as  $\rho \rightarrow 0$ , the Shannon–Jaynes entropy just flattens off and intercepts the  $y$ -axis at a finite value. It is this flattening off that, in practice, can lead to very slow convergence of the deployed optimization algorithms, as the MEM wanders about in this flat direction.

The second reason was that the MEM originally arose in the context of two-dimensional astronomical image reconstruction and the assumptions that enter its construction make reference specifically to this two-dimensional nature of the inverse problems. Here, instead we are interested in a simple one-dimensional inverse problem, which is not directly related to at least one of the axioms underlying the MEM. As laid out in detail in [5], we started by replacing that axiom by a generic smoothness axiom and also introduced a scale invariance axiom to arrive at the BR method with its regulator given in (9). The form of this regulator differs in important aspects from the Shannon–Jaynes entropy. Note that it contains only ratios of the functions  $\rho$  and  $m$ . Since both quantities carry the same units, it implies that the value of the integral does not depend on the units assigned to them. In contrast, the Shannon–Jaynes integrand, and thus the integral, depend on the specific choice of units in  $\rho$ . In addition, the logarithmic term in  $S_{BR}$  is not multiplied with the function  $\rho$ . This changes the behavior of the integrand for  $\rho \rightarrow 0$ , making it diverge—i.e., the BR regulator avoids the flat direction encountered in  $S_{MEM}$  and thus shows much better convergence properties for functions  $\rho$ , which contain large ranges, where their values are parametrically much smaller than in the dominant contributions.

A straightforward implementation of the BR method in a general Bayesian context has recently been introduced in [23]. Using the MC-STAN Monte-Carlo library (<http://mc-stan.org/>), it has been shown how to sample the posterior distribution of the BR method in the full  $N_\omega$ -dimensional search space and thus how to access the full information encoded in it. The maximum a-posteriori solution considered in the literature so far only captures the maximum of this distribution. Not only does a full Bayesian implementation of the BR method allow for a self-consistent treatment of the hyperparameter  $\alpha$ , but it also provides the complete uncertainty budget from the spread of the posterior.



**Figure 3.** Comparison of the behavior of the integrand of the Tikhonov (TK), MEM, and Bayesian Reconstruction (BR) regulators for the choice  $\alpha = 0.1$  and  $m = 1$  using (left) a log-lin scale and (right) a log-log scale. Note that the flattening off of the Shannon–Jaynes entropy towards vanishing  $\rho$ .

#### 4. Summary and Conclusions

We have critically assessed in this paper the arguments underlying Bryan’s MEM, which we show to be flawed as they resort to methods of linear algebra when treating an inherently nonlinear optimization problem. Therefore, we conclude that even though the individual steps in the derivation of the SVD subspace are all correct, they do not apply to the problem at hand and their conclusions can be disproved with a direct counterexample. The counterexample we provided utilizes the fact that the componentwise exponentiated columns of the matrix  $\hat{U}$  do not project into the null-space of the Kernel when computing synthetic data. After establishing the fact that the restriction to the SVD subspace is an ad-hoc procedure, we discussed possible ways to overcome it, suggesting either to systematically extend the search space within the MEM or abandon the MEM in favor of one of the many modern Bayesian approaches developed over the past two decades.

In our ongoing work to improve methods for the Bayesian treatment of inverse problems, we focus on the development of more specific regulators. So far, the methods described above only make reference to very generic properties of the quantity  $\rho$ , such as smoothness and positivity. As a specific example, in the study of spectral functions in strongly-correlated quantum systems, the underlying first principles theory provides extra domain knowledge about admissible structures in  $\rho$  that so far is not systematically exploited. A further focus of our work is the extension of the separable regulators discussed in this work to fully correlated prior distributions that exploit cross-correlation between the individual parameters  $\rho_l$ .

**Author Contributions:** The author has conceptualized and worked out the argumentative chain, as well as the explicit examples in the present study. He has written the manuscript and prepared all figures. The author has read and agreed to the published version of the manuscript.

**Funding:** The author acknowledges funding by the Research Council of Norway under the FRIPRO Young Research Talent grant 286883.

**Conflicts of Interest:** The author declares that he has no competing interests.

#### Abbreviations

The following abbreviations are used in this manuscript:

BFGS	Broyden–Fletcher–Goldfarb–Shanno
BR	Bayesian Reconstruction
MEM	Maximum Entropy Method
SJ	Shannon–Jaynes
SVD	singular value decomposition
TK	Tikhonov

## Appendix A. Explicit SVD Example of (Non-) Projection to the Null-Space

(\* We provide a concrete example showcasing the fact that after non-linear exponentiation, the columns of the SVD matrix U do not project into the Null space of the Kernel \*)

```
In[2]:= Nt = 10; Nw = 20;
```

```
In[3]:= dt = 1 / 10; dw = 1 / 2;
```

(\* The concrete form of the Kernel used here arises in the inverse Laplace transform, relevant e.g. in lattice QCD. \*)

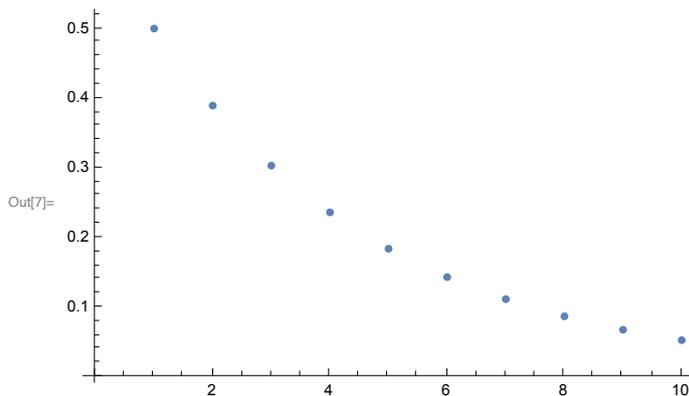
```
In[4]:= Kernel = Table[dw * Exp[-dt * t * dw * w], {t, 0, Nt - 1}, {w, 0, Nw - 1}];
```

```
In[5]:= Dimensions[Kernel]
```

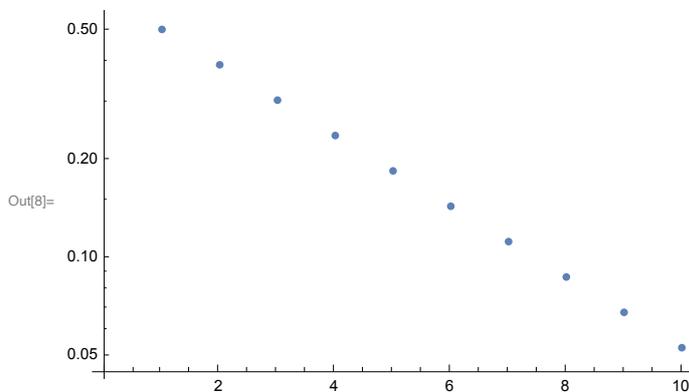
```
Out[5]:= {10, 20}
```

```
In[6]:= TestSpec = Table[If[w == 5, 1, 0], {w, 0, Nw - 1}];
```

```
In[7]:= ListPlot[Kernel.TestSpec]
```



```
In[8]:= ListLogPlot[Kernel.TestSpec]
```



(\* Here we carry out the SVD decomposition of the transposed Kernel \*)

```
In[9]:= {U, S, V} = SingularValueDecomposition[Transpose[N[Kernel, 20]]];
```

In[10]= **S**

```
Out[10]= { {3.3264299012192807701, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0},
            {0, 1.3217963119141028541, 0, 0, 0, 0, 0, 0, 0, 0, 0},
            {0, 0, 0.30345899410178106372, 0, 0, 0, 0, 0, 0, 0, 0},
            {0, 0, 0, 0.050903815971078317311, 0, 0, 0, 0, 0, 0, 0},
            {0, 0, 0, 0, 0.0064549951670843041285, 0, 0, 0, 0, 0, 0},
            {0, 0, 0, 0, 0, 0.00062805703054415921463, 0, 0, 0, 0, 0},
            {0, 0, 0, 0, 0, 0, 0.000046972404241219089401, 0, 0, 0, 0},
            {0, 0, 0, 0, 0, 0, 0, 2.6584704686839277841 × 10-6, 0, 0, 0},
            {0, 0, 0, 0, 0, 0, 0, 0, 1.0865307541628509085 × 10-7, 0, 0},
            {0, 0, 0, 0, 0, 0, 0, 0, 0, 2.8006775176021814106 × 10-9, 0},
            {0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0}, {0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0},
            {0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0}, {0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0},
            {0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0}, {0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0},
            {0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0}, {0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0},
            {0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0}, {0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0} }
```

(\* We first confirm the validity of the linear relations. I.e. any of the first  $N_{\tau}=10$  columns give finite results, when applied to the kernel \*)

In[11]= **Table** [

**Norm**[Transpose[V].Transpose[S].Transpose[U].Transpose[U][[k]]], {k, 1, Nt}]

```
Out[11]= {3.32642990121928077, 1.321796311914102854, 0.303458994101781064,
          0.050903815971078317, 0.006454995167084304, 0.00062805703054416,
          0.000046972404241219, 2.65847046868 × 10-6, 1.08653075416 × 10-7, 2.800677518 × 10-9}
```

(\* Then we confirm that the remaining columns are projected into the null space of the kernel \*)

In[12]= **Table**[Norm[Transpose[V].Transpose[S].Transpose[U].Transpose[U][[k]]],

{k, Nt + 1, Nw}]

```
Out[12]= {0, 0, 0, 0, 0, 0, 0, 0, 0, 0}
```

(\* Now let us check that the component-wise exponentiated columns are not projected into the null space \*)

In[13]= **Table**[Norm[Transpose[V].Transpose[S].Transpose[U].Exp[Transpose[U][[k]]]],

{k, Nt + 1, Nw}]

```
Out[13]= {14.38634619274251869, 14.39771209111371408,
          14.39697005512978974, 14.39028724651449921,
          14.39393010552004510, 14.39768153873137089, 14.38882643697154332,
          14.36025526323499610, 14.36506367520141366, 14.43479593707020195}
```

(\* Thus we have found a clear counterexample to the claim that adding to the non-linear parametrization of the spectral function in Bryan's MEM one of the  $N_{\tau}+i$  columns of the matrix U does not affect the reconstruction outcome. \*)

## Appendix B. Explicit MEM Example with a Solution Outside of the SVD Subspace

(\* We construct an explicit example of an MEM inverse problem, where the global extremum of the posterior does not reside within the SVD subspace \*)

```
In[15]:= Nt = 3; Nw = 2000;
```

```
In[16]:= dt = 1 / 10; dw = 1 / 2;
```

(\* The concrete form of the Kernel used here arises e.g. in the reconstruction of spectral functions of strongly correlated quantum systems from Matsubara frequency correlators. \*)

```
In[17]:= Kernel =
```

```
Table[dw * 1 / ( (dt * t) ^ 2 + (dw * (w + 1 / Nw)) ^ 2), {t, 0, Nt - 1}, {w, 0, Nw - 1}];
```

```
In[18]:= Dimensions[Kernel]
```

```
Out[18]:= {3, 2000}
```

```
In[19]:= TestSpec1 = Table[If[w == 50, 1, 1 / Nw ^ 2], {w, 0, Nw - 1}];
```

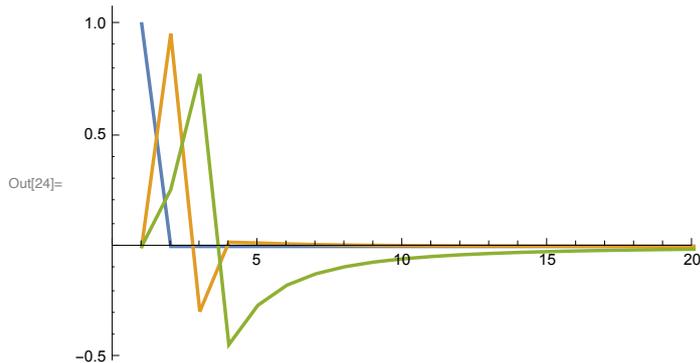
```
In[20]:= TestSpec2 = Table[If[w == 250, 1, 1 / Nw ^ 2], {w, 0, Nw - 1}];
```

```
In[21]:= TestSpec3 = Table[If[w == 500, 1, 1 / Nw ^ 2], {w, 0, Nw - 1}];
```

(\* Explicit SVD with high numerical precision \*)

```
In[22]:= {U, S, V} = SingularValueDecomposition[Transpose[N[Kernel, 80]]];
```

```
In[24]:= ListLinePlot[{U[[1]], U[[2]], U[[3]]}, PlotRange -> {{0, 20}, All},  
PlotStyle -> {{Thick, Joined -> True}, {Thick, Joined -> True}}]
```



(\* Finding the minimum of L in the SVD subspace \*)

```
In[25]:= IdealData1 = Kernel.TestSpec1;
```

```
In[26]:= IdealData2 = Kernel.TestSpec2;
```

```
In[27]:= IdealData3 = Kernel.TestSpec3;
```

```
In[28]= MyMinSVD1 = FindMinimum[ (1 / 2) * Norm[
      (IdealData1 - Kernel.( Exp[a * U[[1]] + b * U[[2]] + c * U[[3]])) / IdealData1]^2,
      {{a, 1}, {b, 1}, {c, 1}}, PrecisionGoal -> 30, AccuracyGoal -> 30,
      WorkingPrecision -> 30, MaxIterations -> 5000]
```

```
Out[28]= {3.42128398306572180310416688213 × 106, {a -> -21.1437298365263458094788852912,
      b -> -2.78909016024624830050728020634, c -> -1.58971804336547244848102541463}}
```

```
In[29]= MyMinSVD2 = FindMinimum[ (1 / 2) Norm[
      (IdealData2 - Kernel.( Exp[a * U[[1]] + b * U[[2]] + c * U[[3]])) / IdealData2]^2,
      {{a, 1}, {b, 1}, {c, 1}}, PrecisionGoal -> 30, AccuracyGoal -> 30,
      WorkingPrecision -> 30, MaxIterations -> 5000]
```

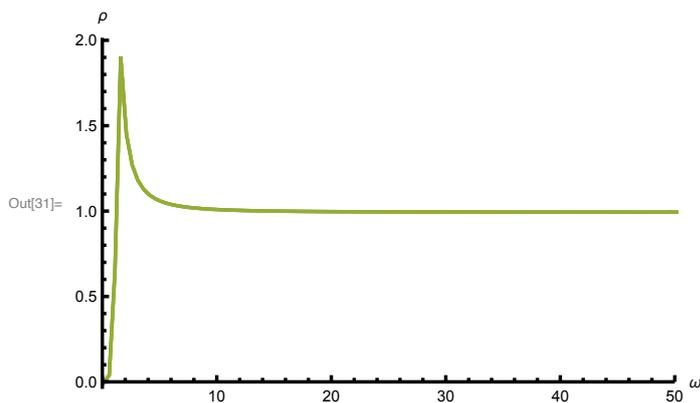
```
Out[29]= {1.40988511953296931150634141038 × 109, {a -> -21.0125682530657542502962215752,
      b -> -2.78037836336990724569212067212, c -> -1.58559045740992587533119137236}}
```

```
In[30]= MyMinSVD3 = FindMinimum[ (1 / 2) Norm[
      (IdealData3 - Kernel.( Exp[a * U[[1]] + b * U[[2]] + c * U[[3]])) / IdealData3]^2,
      {a, b, c}, PrecisionGoal -> 30, AccuracyGoal -> 30,
      WorkingPrecision -> 30, MaxIterations -> 5000]
```

```
Out[30]= {1.03146604416698341240440273339 × 1010, {a -> -20.7847018679437685122657396700,
      b -> -2.76823465886073358007383383406, c -> -1.57986469803312296678195012825}}
```

(\* Best fit solution of  $\rho^*$ )

```
In[31]= ListLinePlot[{Chop[((Exp[a * U[[1]] + b * U[[2]] + c * U[[3]]) / MyMinSVD1[[2]]]),
      Chop[((Exp[a * U[[1]] + b * U[[2]] + c * U[[3]]) / MyMinSVD2[[2]]]),
      Chop[((Exp[a * U[[1]] + b * U[[2]] + c * U[[3]]) / MyMinSVD3[[2]])], DataRange ->
      {0, 1000}, PlotStyle -> {{Thick, Joined -> True}, {Thick, Joined -> True}},
      AxesStyle -> Thick, AxesLabel -> {" $\omega$ ", " $\rho$ "}, PlotRange -> {{0, 50}, All}]
```



(\* Check the corresponding values of S \*)

```
In[38]= MinLikelihoodSpec1 = (Exp[a * U[[1]] + b * U[[2]] + c * U[[3]]) / MyMinSVD1[[2]];
```

```
In[39]= MinLikelihoodSpec2 = (Exp[a * U[[1]] + b * U[[2]] + c * U[[3]]) / MyMinSVD2[[2]];
```

```
In[40]= MinLikelihoodSpec3 = (Exp[a * U[[1]] + b * U[[2]] + c * U[[3]]) / MyMinSVD3[[2]];
```

```

In[41]:= Sum[0.1 * (MinLikelihoodSpec1[[k]] - 1 -
      MinLikelihoodSpec1[[k]] * Log[MinLikelihoodSpec1[[k]]]), {k, 1, Nw}]
Out[41]= -0.237298

In[42]:= Sum[0.1 * (MinLikelihoodSpec2[[k]] - 1 -
      MinLikelihoodSpec2[[k]] * Log[MinLikelihoodSpec2[[k]]]), {k, 1, Nw}]
Out[42]= -0.236848

In[43]:= Sum[0.1 * (MinLikelihoodSpec3[[k]] - 1 -
      MinLikelihoodSpec3[[k]] * Log[MinLikelihoodSpec3[[k]]]), {k, 1, Nw}]
Out[43]= -0.236224

In[44]:= MinLikelihoodSpecFull = Table[If[k == 25, 1, 1 / 2000 ^ 2], {k, 0, Nw}];

In[45]:= Sum[0.1 * (MinLikelihoodSpecFull[[k]] - 1 -
      MinLikelihoodSpecFull[[k]] * Log[MinLikelihoodSpecFull[[k]]]), {k, 1, Nw}]
Out[45]= -199.899

```

## References

- Whittaker, E.T. On a New Method of Graduation. *Proc. Edinb. Math. Soc.* **1922**, *41*, 63–75. [\[CrossRef\]](#)
- Tikhonov, A.N. On the stability of inverse problems. *Dokl. Akad. Nauk SSSR* **1943**, *39*, 195–198.
- Narayan, R.; Nityananda, R. Maximum Entropy Image Restoration in Astronomy. *Annu. Rev. Astron. Astrophys.* **1986**, *24*, 127–170. [\[CrossRef\]](#)
- Jarrell, M.; Gubernatis, J.E. Bayesian inference and the analytic continuation of imaginary-time quantum Monte Carlo data. *Phys. Rep.* **1996**, *269*, 133–195. [\[CrossRef\]](#)
- Burnier, Y.; Rothkopf, A. Bayesian Approach to Spectral Function Reconstruction for Euclidean Quantum Field Theories. *Phys. Rev. Lett.* **2013**, *111*, 182003. [\[CrossRef\]](#)
- Asakawa, M.; Nakahara, Y.; Hatsuda, T. Maximum entropy analysis of the spectral functions in lattice QCD. *Prog. Part. Nucl. Phys.* **2001**, *46*, 459–508. [\[CrossRef\]](#)
- Bryan, R.K. Maximum entropy analysis of oversampled data problems. *Eur. Biophys. J.* **1990**, *18*, 165–174. [\[CrossRef\]](#)
- Asakawa, M. Comment on “Heavy Quarkonium in Extreme Conditions”. *arXiv* **2020**, arXiv:2001.10205.
- Rothkopf, A. Heavy quarkonium in extreme conditions. *Phys. Rep.* **2020**. [\[CrossRef\]](#)
- Rothkopf, A. Improved maximum entropy analysis with an extended search space. *J. Comput. Phys.* **2013**, *238*, 106–114. [\[CrossRef\]](#)
- Jakovác, A.; Petreczky, P.; Petrov, K.; Velytsky, A. Quarkonium correlators and spectral functions at zero and finite temperature. *Phys. Rev. D* **2007**, *75*, 014506. [\[CrossRef\]](#)
- Rothkopf, A.; Hatsuda, T.; Sasaki, S. Complex Heavy-Quark Potential at Finite Temperature from Lattice QCD. *Phys. Rev. Lett.* **2012**, *108*, 162001. [\[CrossRef\]](#)
- Rothkopf, A. Improved Maximum Entropy Method with an Extended Search Space. In Proceedings of the 30th International Symposium on Lattice Field Theory (LATTICE 2012), Cairns, Australia, 24–29 June 2012; p. 100. [\[CrossRef\]](#)
- Kelly, A.; Rothkopf, A.; Skullerud, J.I. Bayesian study of relativistic open and hidden charm in anisotropic lattice QCD. *Phys. Rev. D* **2018**, *97*, 114509. [\[CrossRef\]](#)
- Sandvik, A.W. Stochastic method for analytic continuation of quantum Monte Carlo data. *Phys. Rev. B* **1998**, *57*, 10287–10290. [\[CrossRef\]](#)
- Beach, K.S.D. Identifying the maximum entropy method as a special limit of stochastic analytic continuation. *arXiv* **2004**, arXiv:Cond-mat/0403055.

17. Mishchenko, A.S.; Prokof'ev, N.V.; Sakamoto, A.; Svistunov, B.V. Diagrammatic quantum Monte Carlo study of the Fröhlich polaron. *Phys. Rev. B* **2000**, *62*, 6317–6336. [[CrossRef](#)]
18. Ding, H.T.; Kaczmarek, O.; Mukherjee, S.; Ohno, H.; Shu, H.T. Stochastic reconstructions of spectral functions: Application to lattice QCD. *Phys. Rev. D* **2018**, *97*, 094503. [[CrossRef](#)]
19. Yoon, H.; Sim, J.H.; Han, M.J. Analytic continuation via domain knowledge free machine learning. *Phys. Rev. B* **2018**, *98*, 245101. [[CrossRef](#)]
20. Fournier, R.; Wang, L.; Yazyev, O.V.; Wu, Q. Artificial Neural Network Approach to the Analytic Continuation Problem. *Phys. Rev. Lett.* **2020**, *124*, 056401. [[CrossRef](#)]
21. Kades, L.; Pawłowski, J.M.; Rothkopf, A.; Scherzer, M.; Urban, J.M.; Wetzel, S.J.; Wink, N.; Ziegler, F. Spectral Reconstruction with Deep Neural Networks. *arXiv* **2019**, arXiv:1905.04305.
22. Offler, S.; Aarts, G.; Allton, C.; Glesaaen, J.; Jäger, B.; Kim, S.; Lombardo, M.P.; Ryan, S.M.; Skullerud, J.I. News from bottomonium spectral functions in thermal QCD. *arXiv* **2019**, arXiv:1912.12900.
23. Rothkopf, A. Bayesian techniques and applications to QCD. In Proceedings of the 13th Conference on Quark Confinement and the Hadron Spectrum (Confinement XIII), Maynooth, Ireland, 31 July–6 August 2018; p. 26.



© 2020 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).