6 Analytic Continuation of Quantum Monte Carlo Data

Erik Koch
Jülich Supercomputer Centre and
Institute for Advanced Simulation
Forschungszentrum Jülich

Contents

E. Pavarini, E. Koch, A. Lichtenstein, and D. Vollhardt (eds.)
Dynamical Mean-Field Theory of Correlated Electrons
Modeling and Simulation Vol. 12
Forschungszentrum Jülich, 2022, ISBN 978-3-95806-619-9
http://www.cond-mat.de/events/correl22

The analytic continuation of Monte Carlo data may appear as an exercise in achieving the unachievable. To understand why, let us consider the example of a fermionic finite-temperature Matsubara Green function $G(\tau)$. For imaginary times $\tau \in [0,\beta]$ it is related to the spectral function $\rho(\omega)$ by the integral equation

$$G(\tau) = -\frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\mathrm{e}^{-\omega\tau}}{1+\mathrm{e}^{-\beta\omega}}\,\rho(\omega)\,d\omega\,.$$

While calculating $G(\tau)$ from $\rho(\omega)$ is a straightforward integral, the inverse problem is hard. This is not because we have to solve a Fredholm equation of the first kind, the difficulty rather arises from the remarkable insensitivity of the imaginary-time data on changes in the spectral function. To illustrate this, we write the spectral function as a sum of delta-peaks $w_i\,\delta(\omega-\varepsilon_i)$, for which the imaginary-time Green function becomes a linear combination of exponentials

$$G(\tau) = -\frac{1}{2\pi}\sum_i w_i\big(1-n_{\mathrm{FD}}(\varepsilon_i)\big)\,\mathrm{e}^{-\varepsilon_i\tau} = -\frac{1}{2\pi}\sum_i w_i\,n_{\mathrm{FD}}(\varepsilon_i)\,\mathrm{e}^{+\varepsilon_i(\beta-\tau)},$$

where we have introduced the Fermi-Dirac distribution $n_{\mathrm{FD}}(\varepsilon) = 1/\big(\mathrm{e}^{+\beta\varepsilon}+1\big)$. While a peak at zero energy simply contributes a constant to $G(\tau)$, the contribution of peaks at large frequencies, $|\varepsilon| \gg 0$, is only noticeable close to $\tau = 0$ or β, while inside the interval $(0,\beta)$ it becomes exponentially small. To reconstruct the spectral function reliably over the entire ω-range, we thus need to know $G(\tau)$ very accurately very close to the boundaries of the interval $(0,\beta)$.
Numerical simulations can give, however, only a finite number of data points, $G(\tau_j)$. Obviously, this does not provide enough information to reconstruct a continuous spectral function: we expect that there are many different spectral functions $\rho(\omega)$ that reproduce a given set of data points $\{G(\tau_j)\}$. Such a problem without a well-defined solution is called ill posed [1]. If we insist on obtaining a unique result, we need to add constraints, e.g., by including additional information about what kind of solution we consider reasonable. In addition, Monte Carlo data are noisy. When reconstructing the spectral function, we thus need to take the accuracy of the data into account and quantify how reliable the result is, given the noise in the input. Both types of information, the estimate of the reliability of the data and our expectations about a reasonable solution of the inverse problem, can be handled using Bayesian reasoning [2].
In the following we will introduce the analytic properties that allow the continuation of Green and correlation functions. We then describe how to quantify the statistical errors in the numerical data and to set up the inverse problem. In the main part we use this to give an overview of methods to solve the inverse problem. The most straightforward approach simply performs a least-squares fit to the data points. We explain why this approach is ill posed and how it fails spectacularly. We then discuss the idea of regularization by introducing assumptions about a reasonable solution. This makes the problem well posed, but dependent on prior information. The effect of the prior information included in the regularizer can be quantified using Bayesian techniques. We discuss how they are used to argue for the different flavors of the Maximum Entropy method. Finally we introduce the average spectrum method which tries to avoid introducing prior information by calculating $\rho(\omega)$ as a functional integral over the space of all possible spectral functions.

1 Setting the stage

1.1 Analytic continuation

A system at finite temperature with time-independent Hamiltonian H is described as an ensemble of eigenstates, $H|n\rangle = E_n|n\rangle$, weighted by their Boltzmann factor. The expectation value of an operator A is thus given by

$$\langle A\rangle = \frac{\sum_n e^{-\beta E_n}\langle n|A|n\rangle}{\sum_n e^{-\beta E_n}} = \frac{1}{Z}\operatorname{Tr}\left(e^{-\beta H}A\right). \tag{1}$$

For a canonical ensemble the trace is over the N-electron Hilbert space. For a grand-canonical ensemble we get the same expression when measuring energies relative to the chemical potential, i.e., choosing $\mu = 0$, and taking the trace over the entire Fock space.
Time correlation functions can be calculated using the Heisenberg picture

$$\big\langle A(t_0{+}t)B(t_0)\big\rangle = \frac{1}{Z}\operatorname{Tr} e^{-\beta H}\, e^{\mathrm{i}H(t_0+t)}Ae^{-\mathrm{i}H(t_0+t)}\, e^{\mathrm{i}Ht_0}Be^{-\mathrm{i}Ht_0} = \big\langle A(t)B\big\rangle = \big\langle AB(-t)\big\rangle\,, \tag{2}$$

where the t_0-independence follows from the cyclic property of the trace $\operatorname{Tr} ABC = \operatorname{Tr} CAB$. Monte Carlo techniques are ideal to evaluate the high-dimensional sums needed to calculate such traces [3]. But since the time-evolution leads to complex coefficients, Monte Carlo sampling will have to fight with a serious phase-problem. This can be avoided using a Wick rotation, i.e., working in imaginary time. For this we need to analytically continue (2). This is straightforward: simply replace t in the analytic expression by the complex variable $\zeta = t{-}i\tau$ and determine for what values of ζ the result is well defined. This is most easily done using the spectral representation, i.e., evaluating the trace in the basis of eigenfunctions

$$\big\langle A(t{-}\mathrm{i}\tau)B\big\rangle = \frac{1}{Z}\operatorname{Tr}\mathrm{e}^{(\mathrm{i}t+\tau-\beta)H}A\,\mathrm{e}^{-(\mathrm{i}t+\tau)H}B = \frac{1}{Z}\sum_{n,m}\mathrm{e}^{(\mathrm{i}t+\tau-\beta)E_n}\,\mathrm{e}^{-(\mathrm{i}t+\tau)E_m}\langle n|A|m\rangle\langle m|B|n\rangle. \tag{3}$$

For systems with a finite number of states the sum is always analytic, while for systems whose spectrum is not bounded from above, we need $\beta \geq \tau \geq 0$ to maintain absolute convergence. Thus (2) can be analytically continued to a stripe below the real axis $\big\{\zeta \in \mathbb{C}\,\big|{-}\beta \leq \operatorname{Im}\zeta \leq 0\big\}$. We can then use quantum Monte Carlo to sample the function $C_{AB}(\tau) := \big\langle A(-\mathrm{i}\tau)B\big\rangle$ for $\tau \in [0,\beta]$. The analytic continuation back to the real axis is a bit less obvious, since QMC only gives us the function values, i.e., the left hand side of (3) for $t = 0$, but not the explicit functional form on the right hand side, for which we would have to know all eigenenergies and matrix elements. We can, however, define a spectral function that neatly contains all the required information by taking the Fourier transform

$$\int_{-\infty}^{\infty} dt\,\mathrm{e}^{\mathrm{i}\omega t}\big\langle A(t)B\big\rangle = \frac{2\pi}{Z}\sum_{n,m}\mathrm{e}^{-\beta E_n}\langle n|A|m\rangle\langle m|B|n\rangle\,\delta\big(\omega-(E_m{-}E_n)\big) =: \rho_{AB}(\omega) \tag{4}$$

in terms of which we can write (3) as

$$\big\langle A(t{-}\mathrm{i}\tau)B\big\rangle = \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega\,\mathrm{e}^{-(\mathrm{i}t+\tau)\omega}\rho_{AB}(\omega)\,. \tag{5}$$

For the special case $t = 0$ this gives us an integral equation directly relating $\rho_{AB}(\omega)$ to $C_{AB}(\tau)$

$$C_{AB}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega\, \mathrm{e}^{-\omega\tau}\rho_{AB}(\omega)\,, \tag{6}$$

which is, however, not suited for practical calculations since the integral kernel, $\exp(-\omega\tau)$, diverges for $\omega \to -\infty$. We can get around this problem by modifying the kernel, dividing it by a function that makes it finite, and correspondingly multiplying the spectral function to leave the integral unchanged

$$C(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega\, \frac{\mathrm{e}^{-\omega\tau}}{\mu(\omega)}\, \underbrace{\mu(\omega)\rho_{AB}(\omega)}_{=:\tilde\rho(\omega)}\,. \tag{7}$$

A suitable kernel modification would be $\mu(\omega) = 1 \pm \mathrm{e}^{-\beta\omega}$, which makes the kernel finite for $\omega \to -\infty$ as long as $\tau \leq \beta$, while keeping it finite for $\omega \to +\infty$. To analytically continue $C_{AB}(\tau) = \langle A(-\mathrm{i}\tau)B\rangle$ to the real axis we then solve the integral equation (with finite kernel)

$$C_{AB}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega\, \frac{\mathrm{e}^{-\omega\tau}}{1 \pm \mathrm{e}^{-\beta\omega}}\, \tilde\rho^{\pm}_{AB}(\omega) \tag{8}$$

for $\tilde\rho^{\pm}_{AB}(\omega)$ and use $\rho_{AB}(\omega) = \tilde\rho^{\pm}_{AB}(\omega)/(1\pm\mathrm{e}^{-\beta\omega})$ in (5) to calculate the analytical continuation on the real axis. For the plus sign $1/(1\pm\mathrm{e}^{-\beta\omega})$ is related to the Fermi-Dirac function $n_{\mathrm{FD}}(-\omega) = 1-n_{\mathrm{FD}}(\omega)$, while for minus to the Bose-Einstein function $-n_{\mathrm{BE}}(-\omega) = n_{\mathrm{BE}}(\omega)-1$.
It is reasonable to expect that $\tilde\rho^{\pm}_{AB}(\omega)$ is a spectral function in its own right. Reordering the spectral representation (4), we can write it as

$$\begin{aligned}
\tilde\rho^{\pm}_{AB}(\omega) &= \rho_{AB}(\omega) \pm \rho_{AB}(\omega)\,\mathrm{e}^{-\beta\omega} \\
&= \rho_{AB}(\omega) \pm \frac{2\pi}{Z}\sum_{n,m} \mathrm{e}^{-\beta E_n}\langle n|A|m\rangle\langle m|B|n\rangle\, \delta\big(\omega - (E_m{-}E_n)\big)\, \mathrm{e}^{-\beta(E_m-E_n)} \\
&= \rho_{AB}(\omega) \pm \rho_{BA}(-\omega)\,.
\end{aligned} \tag{9}$$

Comparing with (3) and (4) we see that $\tilde\rho^{\pm}_{AB}(\omega)$ is the spectral function of

$$\mathrm{i}G^{\pm}_{AB}(t) := \big\langle A(t)B\big\rangle \pm \big\langle B(-t)A\big\rangle = \big\langle A(t)B\big\rangle \pm \big\langle BA(t)\big\rangle = \big\langle [A(t),\, B]_{\pm}\big\rangle, \tag{10}$$

which, for $t > 0$, is the retarded correlation function $G^{R\pm}_{AB}(t) = \Theta(t)\, G^{\pm}_{AB}(t)$, with Θ the step function, $\Theta(t{>}0) = 1$ and $\Theta(t{<}0) = 0$. As discussed above, the first term can be analytically continued to $\big\{\zeta \in \mathbb{C}\,\big|{-\beta} \leq \operatorname{Im}\zeta \leq 0\big\}$, while the second term can be continued to the stripe of width β above the real axis. It is thus natural to define the Matsubara function

$$-\,G^{M\pm}_{AB}(\tau) := \big\langle\, \mathcal{T}^{\pm}_{\tau}\, A(-\mathrm{i}\tau)B(0)\big\rangle \tag{11}$$

with the imaginary-time ordering $\mathcal{T}^{\pm}_{\tau}\, A(-\mathrm{i}\tau)B(0) = \Theta(\tau)A(-\mathrm{i}\tau)B(0) \mp \Theta(-\tau)B(0)A(-\mathrm{i}\tau)$ taking care of selecting the appropriate analytic term for the given τ. This introduces a discontinuity at $\tau = 0$

$$G^{M\pm}_{AB}(0^+) - G^{M\pm}_{AB}(0^-) = -\big\langle [A,\, B]_{\pm}\big\rangle. \tag{12}$$

From the cyclic property of the trace in (3), it follows that the Matsubara functions for positive and negative τ are related (anti)symmetrically, i.e. for $\tau \in (0, \beta)$

$$G^{M\pm}_{AB}(\beta - \tau) = -\big\langle A(-\mathrm{i}(\beta-\tau))B\big\rangle = -\big\langle B(-\mathrm{i}\tau)A\big\rangle = -\big\langle BA(\,\mathrm{i}\tau)\big\rangle = \mp G^{M\pm}_{AB}(-\tau). \quad (13)$$

For $\tau \in (0, \beta]$ we obviously have (remember the sign introduced in (11)) $G^{M\pm}_{AB}(\tau) = -C_{AB}(\tau)$, so that from (8) we obtain

$$G^{M\pm}_{AB}(\tau) = -\frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega\, \frac{\mathrm{e}^{-\omega\tau}}{1 \pm \mathrm{e}^{-\beta\omega}}\, \tilde{\rho}^{\pm}_{AB}(\omega) \quad \text{for } \tau \in [0, \beta]. \quad (14)$$

It is convenient to choose the sign in the kernel modification to obtain a simple relation for the sum rule, which directly follows from the spectral representation, using $|n\rangle\langle n| = \mathbb{1}$

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega\, \tilde{\rho}^{\pm}_{AB}(\omega) = \big\langle [A,\, B]_{\pm}\big\rangle\,. \quad (15)$$

For observables and bosonic operators we thus choose the commutator, while for fermionic Green functions it is more convenient to choose the anticommutator.
For the special case $B = A^\dagger$ we find

$$\tilde{\rho}^{\pm}_{AA^\dagger}(\omega) = \frac{2\pi}{Z}\sum_{n,m}\big(\mathrm{e}^{-\beta E_n} \pm \mathrm{e}^{-\beta E_m}\big)\,\big|\langle n|A|m\rangle\big|^2\,\delta\big(\omega - (E_m - E_n)\big), \quad (16)$$

which is obviously non-negative for the fermionic case, for the bosonic sign choice it is non-negative for $\omega = (E_m - E_n) > 0$, non-positive for $\omega < 0$, and vanishes at least linearly at $\omega = 0$. We can thus define a non-negative function $\tilde{\rho}^{-}_{AB}(\omega)/\omega$ which is regular at $\omega = 0$

$$\lim_{\omega\to 0}\frac{\tilde{\rho}^{-}_{AB}(\omega)}{\omega} = \frac{2\pi\beta}{Z}\sum_{n,m}\mathrm{e}^{-\beta E_n}\,\big|\langle n|A|m\rangle\big|^2\,\delta(E_n - E_m) \quad (17)$$

so that we can rewrite (14) with non-negative functions as

$$G^{M+}_{AA^\dagger}(\tau) = -\frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega\, \frac{\mathrm{e}^{-\omega\tau}}{1 + \mathrm{e}^{-\beta\omega}}\, \tilde{\rho}^{+}_{AA^\dagger}(\omega) \quad (18)$$

$$G^{M-}_{AA^\dagger}(\tau) = -\frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega\, \frac{\omega\,\mathrm{e}^{-\omega\tau}}{1 - \mathrm{e}^{-\beta\omega}}\, \frac{\tilde{\rho}^{-}_{AA^\dagger}(\omega)}{\omega}\,, \quad (19)$$

which, when A is an annihilator, applies to the diagonal elements of Green functions.
When A is an observable, we see from (9) that $\tilde{\rho}^{-}_{AA}(\omega) = -\tilde{\rho}^{-}_{AA}(-\omega)$, so that we can restrict the integral to $\omega > 0$

$$G^{M-}_{AA}(\tau) = -\frac{1}{2\pi}\int_{0}^{\infty} d\omega\, \frac{\omega\big(\mathrm{e}^{-\omega\tau} + \mathrm{e}^{-\omega(\beta-\tau)}\big)}{1 - \mathrm{e}^{-\beta\omega}}\, \frac{\tilde{\rho}^{-}_{AA}(\omega)}{\omega} \quad \text{when } A \text{ hermitian.} \quad (20)$$

We could actually cancel the factor ω in the integrand since $\tilde{\rho}^{-}_{AA}(\omega \geq 0)$ is non-negative by itself, but when calculating susceptibilities it is common to keep it, since it shows the behavior for $\omega \to 0$, (17), more clearly.

1.2 Analytic properties of the integral equations

We can gain some insight into the integral equations (18) and (19) by realizing that they are intimately related to the Euler and Bernoulli polynomials [4]. Introducing the reduced variables $x = \beta\omega$ and $y = \tau/\beta \in [0,1]$ and the functions $f(x) = \tilde{\rho}^{\pm}(x/\beta)/\beta$ (scaled to conserve the sum rule) and $g(y) = G^{M\pm}(\beta y)$ we obtain, for the fermionic case

$$g(y) = -\frac{1}{2\pi}\int_{-\infty}^{\infty} dx\, \frac{\mathrm{e}^{-xy}}{1+\mathrm{e}^{-x}}\, f(x)\,, \tag{21}$$

from which we see that, for fixed kernel, the spectral function is spread out over an ever wider range as we go to lower temperatures. The scaled kernel of this equation is essentially the generating function of the Euler polynomials $E_n(s)$ on $s \in [0,1]$, which are defined by

$$\frac{2\mathrm{e}^{st}}{\mathrm{e}^{t}+1} = \sum_{n=0}^{\infty} E_n(s)\,\frac{t^n}{n!}\,. \tag{22}$$

With $s = \tau/\beta$ and $t = -\beta\omega$ we find from (18)

$$G^{M+}(\tau) = -\frac{1}{4\pi}\sum_{n=0}^{\infty} E_n(\tau/\beta)\;\frac{(-\beta)^n}{n!}\int_{-\infty}^{\infty} d\omega\,\omega^n\,\tilde{\rho}^{+}(\omega) \tag{23}$$

that the fermionic Matsubara function is a linear combination of Euler polynomials, where the expansion coefficients of $E_n(\tau/\beta)$ is proportional to the n-th moment of the spectral function. Since the Euler polynomials are not orthogonal, to determine the moments of $\tilde{\rho}$ from $G^{M+}(\tau)$, we first have to find the dual functions $E^n(s)$ with $\int_0^1 ds\, E^n(s)\, E_m(s) = \delta_{n,m}$. Integrating them with the generating function (22) we obtain

$$\int_0^1 ds\, E^n(s)\,\mathrm{e}^{st} = \frac{t^n}{n!}\,\frac{e^t+1}{2}\,, \tag{24}$$

which is solved by

$$E^n(s) = \frac{(-1)^n}{2\,n!}\Big(\delta^{(n)}(s-1) + \delta^{(n)}(s)\Big)\,, \tag{25}$$

where $\delta^{(n)}(s-a)$ is the n-th derivative of the delta function at $s = a$ (to make the evaluation for $a = 0$ and 1 unique, we consider the limit from inside the interval of integration). Integration by parts then produces $(-1)^n$ times the n-th derivative of the rest of the integrand at a. Using this in (23) and rewriting the Matsubara function at β as that at 0^-, eq. (13), we find that the discontinuity in the n-th derivative of the Matsubara function is proportional to the n-th moment of the spectral function

$$\frac{d^n G^{M+}(\beta)}{d\,\tau^n} + \frac{d^n G^{M+}(0)}{d\,\tau^n} = \frac{d^n G^{M+}(0^+)}{d\,\tau^n} - \frac{d^n G^{M+}(0^-)}{d\,\tau^n} = -\frac{(-1)^n}{2\pi}\int_{-\infty}^{\infty} d\omega\;\omega^n\,\tilde{\rho}^{+}(\omega)\,. \tag{26}$$

The higher moments contain the information about the spectral function at large frequencies. Extracting the derivatives from Monte Carlo data for $G(\tau)$ is difficult. Instead, they can be sampled directly: For $\tau > 0$ we have, (11),

$$-\,G^{M+}(\tau) = \big\langle A(-\mathrm{i}\tau)B\big\rangle = \frac{1}{Z}\,\mathrm{Tr}\,\mathrm{e}^{-\beta H}\mathrm{e}^{\tau H}A\mathrm{e}^{-\tau H}B\,. \tag{27}$$

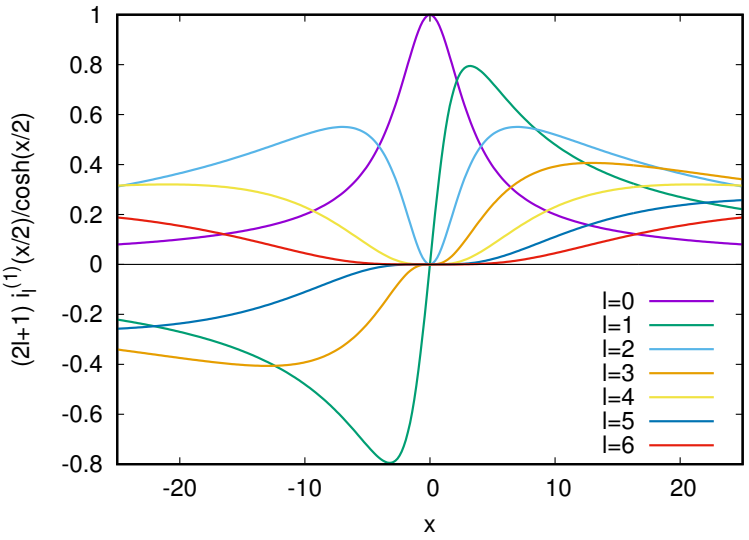


Fig. 1: *Dependence of the scaled Legendre kernel* $(2l{+}1)\, i_l^{(1)}(x/2)/\cosh(x/2)$ *on the order* l*. For* $l = 0$*,* G_l *contains information about the spectral function close to the Fermi level, while for increasing* l *it probes ever larger frequencies. As the Legendre polynomials themselves, the kernel is even/odd for even/odd* l*.*

Taking the derivative with respect to τ brings down the Hamiltonian to the left and the right of A, producing $\langle [H,\, A(-\mathrm{i}\tau)]\, B\rangle$. Repeated derivatives produce repeated commutators defined by $[H;\, A]_n := \big[H,\, [H;\, A]_{n-1}\big]$ and $[H;\, A]_0 := A$ as in the Baker-Campbell-Hausdorff formula. The moments can then be determined directly by sampling the expectation values

$$\Big\langle \big[[H;\, A]_n,\, B\big] \Big\rangle = -\frac{(-1)^n}{2\pi} \int_{-\infty}^{\infty} d\omega\; \omega^n\, \tilde{\rho}^+(\omega)\,. \tag{28}$$

Working with the Euler polynomials can become cumbersome due to their lack of orthogonality. This inconvenience can be overcome by expressing them in terms of orthogonal polynomials, e.g., shifted Legendre polynomials $P_l(2y-1)$. When the Matsubara function is expanded as [5]

$$G^{M+}(\tau) = \sum_{l=0}^{\infty} \frac{\sqrt{2l{+}1}}{\beta}\, G_l\, P_l(2\tau/\beta - 1) \quad \text{with} \quad G_l = \sqrt{2l{+}1} \int_0^\beta d\tau\, P_l(2\tau/\beta - 1)\, G^{M+}(\tau)$$

the expansion coefficients are related to the spectral function via (18) by

$$G_l = (-1)^{l+1} \sqrt{2l{+}1}\, \frac{\beta}{4\pi} \int_{-\infty}^{\infty} d\omega\, \frac{\mathrm{i}_l^{(1)}(\beta\omega/2)}{\cosh(\beta\omega/2)}\, \tilde{\rho}(\omega)\,, \tag{29}$$

where $\mathrm{i}_l^{(1)}(x)$ are the modified spherical Bessel functions of first kind. As shown in Fig. 1, for increasing l the integral kernel probes spectral features at higher and higher frequencies. From the derivatives of the recursion relation $(2l{+}1)P_l(x) = P'_{l+1}(x) - P'_{l-1}(x)$ and (26) we find that the n-th moment of the spectral function is given by a sum over all even or odd Legendre coefficients, starting at $l = n$

$$(-1)^{n+1}\, \frac{2}{n!} \sum_{k=0}^{\infty} \frac{\sqrt{4k+2n+1}}{\beta^{2k+n+1}}\, \frac{\big(2(k{+}n)\big)!}{\big(2(k{-}n)\big)!}\, G_{2k+n} = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega\; \omega^n\, \tilde{\rho}^+(\omega)\,. \tag{30}$$

For bosonic Matsubara functions we can obtain similar results using the Bernoulli polynomials $B_n(s)$ whose generating function is directly related to the bosonic kernel.

1.3 Preparing the data

Certainly the most important aspect of preparing Monte Carlo data for analytic continuation is the decision what data to sample. As we have seen, the information about spectral features further away from the chemical potential is concentrated in the Matsubara function extremely close to $\tau = 0$ and β. Reconstructing the spectral function from data given on a uniform τ-grid, we can therefore only expect to get reasonable results close to the chemical potential. Using, on the other hand, the derivatives of the Matsubara function at $\tau = 0$ and β gives us the moments of the spectral function, which, as we know, e.g., form the Lanczos method [6], accurately characterize the spectral function over the entire frequency range using just a few tens of the lowest moments.

The second concern is to properly characterize the statistical errors in the Monte Carlo data. Considering the integral equation

$$g(y) = \int K(y,x)\, f(x)\, dx\,, \tag{31}$$

the actual numerical data is not given as the function $g(y)$ but as vectors of M discrete data points $\boldsymbol{g} = (g_1, \dots, g_M)^\dagger$ representing $g(y)$. The mean over K independent samples is then

$$\bar{\boldsymbol{g}} = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{g}_k \tag{32}$$

with its statistical uncertainty being characterized by the $M \times M$ covariance matrix

$$\boldsymbol{C} = \frac{1}{K(K-1)} \sum_{k=1}^{K} \big(\boldsymbol{g}_k - \bar{\boldsymbol{g}}\big)\big(\boldsymbol{g}_k - \bar{\boldsymbol{g}}\big)^\dagger. \tag{33}$$

By the central limit theorem the probability density of measuring $\bar{\boldsymbol{g}}$ given the covariance matrix $\boldsymbol{C}$ instead of the exact result $\boldsymbol{g}_{\text{exact}}$ is then

$$p(\bar{\boldsymbol{g}}|\, \boldsymbol{g}_{\text{exact}}, \boldsymbol{C}) = \frac{1}{(2\pi)^{M/2} \det \boldsymbol{C}} \mathrm{e}^{-(\bar{\boldsymbol{g}}-\boldsymbol{g}_{\text{exact}})^\dagger \boldsymbol{C}^{-1} (\bar{\boldsymbol{g}}-\boldsymbol{g}_{\text{exact}})/2}. \tag{34}$$

This probability will play a central role in the reconstruction of the spectral function representing $\boldsymbol{g}_{\text{exact}}$. It is, therefore, crucial to have an accurate estimate of $\boldsymbol{C}$. Rewriting it as

$$\boldsymbol{C} = \frac{1}{K(K-1)} \sum_k (\boldsymbol{g}_k - \bar{\boldsymbol{g}})(\boldsymbol{g}_k - \bar{\boldsymbol{g}})^\dagger = \frac{1}{K(K-1)} \sum_k \boldsymbol{g}_k\, \boldsymbol{g}_k^\dagger - \frac{1}{K-1}\, \bar{\boldsymbol{g}}\, \bar{\boldsymbol{g}}^\dagger$$

and realizing that $\boldsymbol{g}\, \boldsymbol{g}^\dagger$ is the (scaled) projector onto $\boldsymbol{g}$, we see that the covariance matrix is a linear combination of K projectors to one-dimensional subspaces. We therefore need $K > M$ independent samples $\boldsymbol{g}_k$ in (33) to have a chance of obtaining a non-singular covariance matrix. Thus, reducing the discretization error requires taking more samples. The easiest way for obtaining independent samples are independent Monte Carlo runs, e.g., on a parallel computer. If we do not have enough CPUs available, we need to construct independent samples from a sequential run. This can be done, e.g., using the blocking technique described in appendix, A.1.

For the numerical solution of the integral equation (31) we also have to discretize $f(x)$ into a vector $\boldsymbol{f} = (f_1, \ldots, f_N)^\dagger$, e.g., by representing it as a piecewise constant function of value f_n on interval n. The integral equation then becomes a simple linear equation $\boldsymbol{g} = \boldsymbol{K}\boldsymbol{f}$, where the kernel matrix is obtained, e.g., from the Riemann sum [7]

$$g(y_m) = \sum_n K(y_m, x_n)\, w_n\, f(x_n)\,, \tag{35}$$

with w_n the width of interval n or, when the functions are expanded in an orthonormal set of functions $|\psi_m\rangle$ like in the Legendre expansion of the Green function, it is given by

$$g_m = \sum_n \int dy \int dx\, \overline{\psi_m(y)}\, K(y,x)\, \varphi_n(x)\; f_n = \sum_n \langle \psi_m | K | \varphi_n \rangle\, f_n\,. \tag{36}$$

Assuming $\boldsymbol{f}$ is the exact model, i.e., it gives the exact data, $\boldsymbol{K}\boldsymbol{f} = \boldsymbol{g}_{\text{exact}}$, it follows from (34)

$$p(\bar{\boldsymbol{g}}|\, \boldsymbol{f}, \boldsymbol{C}) \propto e^{-(\bar{\boldsymbol{g}}-\boldsymbol{K}\boldsymbol{f})^\dagger \boldsymbol{C}^{-1}(\bar{\boldsymbol{g}}-\boldsymbol{K}\boldsymbol{f})/2} \tag{37}$$

Factorizing the inverse covariance matrix, $\boldsymbol{C}^{-1} = \boldsymbol{T}^\dagger\boldsymbol{T}$, e.g., by Cholesky decomposition, we can absorb the explicit dependence on $\boldsymbol{C}$ by introducing $\tilde{\boldsymbol{g}} := \boldsymbol{T}\bar{\boldsymbol{g}}$ and $\tilde{\boldsymbol{K}} := \boldsymbol{T}\boldsymbol{K}$

$$\big(\bar{\boldsymbol{g}} - \boldsymbol{K}\boldsymbol{f}\big)^\dagger \boldsymbol{C}^{-1} \big(\bar{\boldsymbol{g}} - \boldsymbol{K}\boldsymbol{f}\big) = \big(\tilde{\boldsymbol{g}} - \tilde{\boldsymbol{T}}\boldsymbol{f}\big)^\dagger \big(\tilde{\boldsymbol{g}} - \tilde{\boldsymbol{T}}\boldsymbol{f}\big) = \big\|\tilde{\boldsymbol{g}} - \tilde{\boldsymbol{T}}\boldsymbol{f}\big\|^2. \tag{38}$$

The covariance of the transformed data $\tilde{\boldsymbol{g}}$ is then the unit matrix, i.e. the transformation produces uncorrelated data point $\tilde{g}_n$ that all have the same (unit) errorbar.

2 Optimization methods

After discretization of model $\boldsymbol{f}$ and data $\boldsymbol{g}$ and transformation to $\tilde{\boldsymbol{g}}$, analytic continuation is reduced to solving the linear system

$$\tilde{\boldsymbol{g}} = \tilde{\boldsymbol{K}}\boldsymbol{f}\,. \tag{39}$$

Nothing could be easier than that! When the number of data points M we are given equals the number of points N at which we want to know the model, the solution is unique, $\boldsymbol{f} = \tilde{\boldsymbol{K}}^{-1}\tilde{\boldsymbol{g}}$, as long as the kernel is not singular. When $M > N$ the model is overdetermined so that in general there will be no solution. Normally, however, we want to know the model at many more positions than we are given data points, $M < N$ so that the solution is underdetermined. A natural choice is then the $\boldsymbol{f}$ that gives the best fit to the data.

2.1 Least squares and singular values

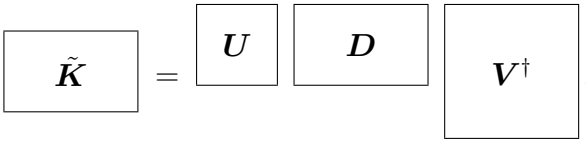
When we ask for a best-fit, we first have to define what we mean by that. Least-squares methods define "best" in terms of the Euclidian norm: minimize $\chi^2(\boldsymbol{f}) := \|\tilde{\boldsymbol{g}} - \tilde{\boldsymbol{T}}\boldsymbol{f}\|^2$. We can justify this choice using Bayesian reasoning: As we have noted in (37), the probability of measuring $\tilde{\boldsymbol{g}}$ when the true model is $\boldsymbol{f}$ is given by $p(\tilde{\boldsymbol{g}}|\, \boldsymbol{f}) = (2\pi)^{-M/2} \exp(-\chi^2(\boldsymbol{f})/2)$. We can invert

this relation using Bayes' theorem [2], $p(B|\,A)\,p(A) = p(A,B) = p(A|\,B)\,p(B)$, stating that the probability of outcome A and B can be written as the probability of B given A times the probability of A, or, equivalently, as the probability of A given B times that of B. For the relation between model and data this implies

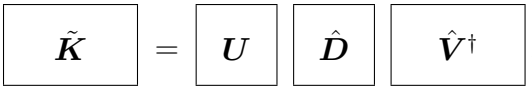
$$p(\boldsymbol{f}|\,\tilde{\boldsymbol{g}}) = \frac{p(\tilde{\boldsymbol{g}}|\,\boldsymbol{f})\,p(\boldsymbol{f})}{p(\tilde{\boldsymbol{g}})}\,. \tag{40}$$

The most probable model $\boldsymbol{f}$ given $\tilde{\boldsymbol{g}}$ thus maximizes $p(\tilde{\boldsymbol{g}}|\,\boldsymbol{f})\,p(\boldsymbol{f})$. In the absence of any further information about possible models it is reasonable to assume that $p(\boldsymbol{f})$ is the same for all $\boldsymbol{f}$, i.e., to use an "uninformative prior". A model that maximizes $p(\boldsymbol{f}|\,\tilde{\boldsymbol{g}})$ is then one that maximizes $\exp\big(-\chi^2(\boldsymbol{f})/2\big)$. It is called a "maximum likelihood estimator" and gives a best fit in the least-squares sense. Since the rank of the kernel matrix, $\operatorname{rank}\boldsymbol{K} \le \min(N,M)$, for $M < N$ the least-squares solution will not be unique: We can add any vector that is mapped by $\boldsymbol{K}$ into zero, without changing the fit. The least-squares problem is thus ill-posed. The usual way of making the solution unique is to ask in addition that $\boldsymbol{f}_{\mathrm{LS}}$ has vanishing overlap with any vector that is mapped to zero, i.e., $\boldsymbol{f}_{\mathrm{LS}}$ is orthogonal to the null space of $\tilde{\boldsymbol{K}}$.

A convenient tool for the theoretical analyzing least-squares problems is the singular value decomposition (SVD) of the matrix $\tilde{\boldsymbol{K}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\dagger$, where $\boldsymbol{U}$ is a unitary $M\times M$ matrix whose column vectors $|\boldsymbol{u}_m\rangle$ define an orthonormal basis in data space and $\boldsymbol{V}$ likewise is a unitary $N\times N$ matrix with columns $|\boldsymbol{v}_n\rangle$ spanning the space of models, while $\boldsymbol{D}$ is a diagonal $M\times N$ matrix with diagonal elements $d_1 \ge d_2 \ge \ldots \ge d_{\min(N,M)} \ge 0$. For the underdetermined case, $M < N$, the singular value decomposition can be pictured as



For the least-squares solution it is convenient to define the reduced singular value decomposition, where the null space of $\tilde{\boldsymbol{K}}$ is dropped in $\boldsymbol{V}$, pictorially,



The singular value decomposition provides a spectral representation of the kernel

$$\tilde{\boldsymbol{K}} = \sum_{i=1}^{\min(M,N)} |\boldsymbol{u}_i\rangle\, d_i\, \langle \boldsymbol{v}_i| \tag{41}$$

which allows us to write the residue vector for $M < N$ as

$$|\tilde{\boldsymbol{g}}\rangle - \tilde{\boldsymbol{K}}|\boldsymbol{f}\rangle = |\tilde{\boldsymbol{g}}\rangle - \sum_i |\boldsymbol{u}_i\rangle\, d_i\, \langle \boldsymbol{v}_i|\boldsymbol{f}\rangle = \sum_i |\boldsymbol{u}_i\rangle \Big(\langle \boldsymbol{u}_i|\tilde{\boldsymbol{g}}\rangle - d_i \langle \boldsymbol{v}_i|\boldsymbol{f}\rangle \Big) \tag{42}$$

so that the least-squares solution (for which the residue vanishes when $d_M > 0$) is

$$|\boldsymbol{f}_{\mathrm{LS}}\rangle = \sum_i \frac{\langle u_i|\tilde{\boldsymbol{g}}\rangle}{d_i}\, |\boldsymbol{v}_i\rangle \quad \text{or, equivalently,} \quad \boldsymbol{f}_{\mathrm{LS}} = \hat{\boldsymbol{V}}\hat{\boldsymbol{D}}^{-1}\boldsymbol{U}^\dagger\tilde{\boldsymbol{g}}\,. \tag{43}$$

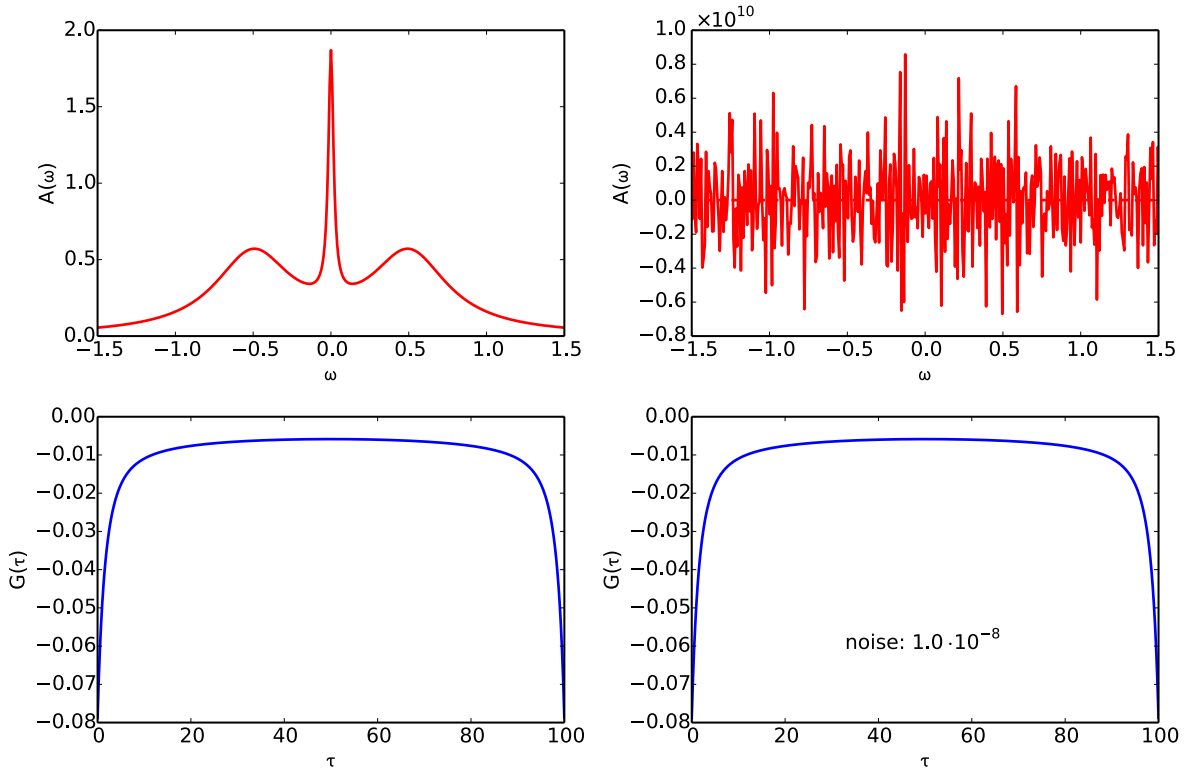


Fig. 2: *Least-squares solution for the analytical continuation of a fermionic imaginary-time Green function. The exact data (bottom left) is constructed from a simple model spectral function consisting of three Lorenz peaks (top left). We add noise of amplitude* 10^{-8} *to the data (bottom right). The least-squares solution given the noisy data is shown in the top right panel. It varies over ten orders of magnitude showing no resemblance at all to the original model.*

As simple and elegantly the least-squares solution can be constructed, as useless it is for the analytic continuation problem. This is illustrated in Fig. 2, showing that $\boldsymbol{f}_{\mathrm{LS}}$, despite giving a perfect fit to the data and, in particular, fulfilling the sum rule for $\sum_n f_n$, is completely dominated by numerical noise. What is the reason for this catastrophic failure? Making the noise in the data explicit, $\tilde{\boldsymbol{g}} = \tilde{\boldsymbol{g}}_{\mathrm{exact}} + \Delta\tilde{\boldsymbol{g}}$ we see that

$$|\boldsymbol{f}_{\mathrm{LS}}\rangle = |\boldsymbol{f}_{\mathrm{exact}}\rangle + \sum \frac{\langle \boldsymbol{u}_i | \Delta\tilde{\boldsymbol{g}}\rangle}{d_i} \, |\boldsymbol{v}_i\rangle \, . \tag{44}$$

When the kernel has close to vanishing singular values, the noise component is divided by a number close to numerical accuracy. This is, in fact, what we are seeing in Fig. 2: dividing noise of order 10^{-8} by the numerical epsilon of double precision numbers (of order 10^{-16}), we would expect the least-squares solution to vary over about eight orders of magnitude. We can verify this picture more quantitatively by looking at the singular values of the kernel matrix, shown in Fig. 3. The exponential decay of the singular values seen in this example actually is the hallmark of an ill-conditioned problem. It is a consequence of the orthogonality of the modes $|\boldsymbol{v}_i\rangle$: With increasing i they develop more and more nodes. Integrating over these oscillating modes with the positive fermionic Green function kernel means that the integral will decrease with the number of nodes. Once the singular value reaches machine precision, the singular modes become numerically degenerate. These modes contribute negligibly to the fit of the data, but cause the catastrophic numerical instability of the least-squares result.

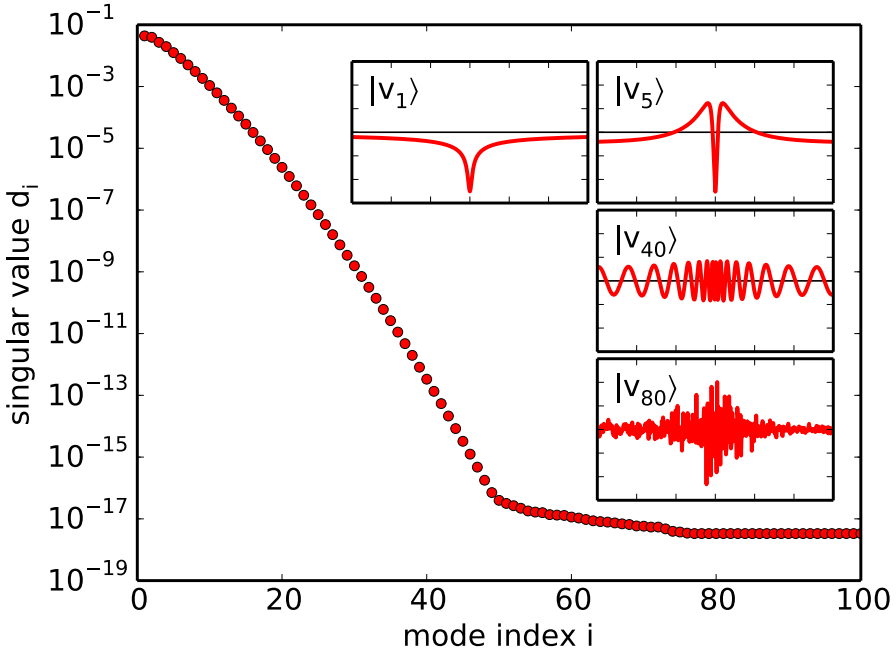


Fig. 3: *Singular values of the kernel matrix used in Fig. 2 on a logarithmic scale. The singular values decay exponentially until leveling off at a value determined by the numerical accuracy of the calculation. The insets show some of the singular modes $|\boldsymbol{v}_i\rangle$. With increasing mode index i, i.e., decreasing singular value, they have an increasing number of nodes. Once the singular value reaches the numerical accuracy, the singular modes become numerically degenerate so that the SVD routine returns arbitrary linear combinations as exemplified here for $|\boldsymbol{v}_{80}\rangle$.*

2.2 Non-negative least-squares

When motivating the least-squares approach using Bayesian reasoning, (40), one assumption was that we have no knowledge whatsoever about the possible models. When we are interested, e.g., in a diagonal spectral function, this is not quite true: We actually do know that $\boldsymbol{f}$ cannot be negative, cf. (18). To incorporate this information, the prior probability $p(\boldsymbol{f})$ should, in fact, vanish when $\boldsymbol{f}$ has a component $f_n < 0$. In other words, we really should maximize the likelihood over non-negative models only: $\max_{\boldsymbol{f}\geq 0} \exp(-\chi^2(\boldsymbol{f}))$. This approach is called non-negative least squares fitting (NNLS). A practical algorithm is discussed in A.2. It will, in general, not give a perfect fit, $\chi^2(\boldsymbol{f}_{\text{NNLS}}) > 0$, but what is not fitted is the part of the data that is incompatible with a non-negative model, i.e., pure noise.

As shown in Fig. 4, using non-negative least squares gives a dramatic improvement over the least-squares solution. Just incorporating the information about the non-negativity of the model reduces the oscillations in the result by nine orders of magnitude, bringing it into a reasonable range. This is because the amplitude of oscillating modes is now strongly limited by non-negativity. In fact, the constraints give the modes with small singular value or in the null space an important role: All modes except the first have nodes, so they can often not be included in the solution with their optimal value (43) without violating the constraint. Since the contribution of the modes with tiny singular value to the fit is tiny, they are free to arrange such that the modes with larger d_i can move closer to their optimum. Thus in NNLS the behavior of all modes is coupled, making the fit much more robust. Moreover, the non-negativity constraint makes the

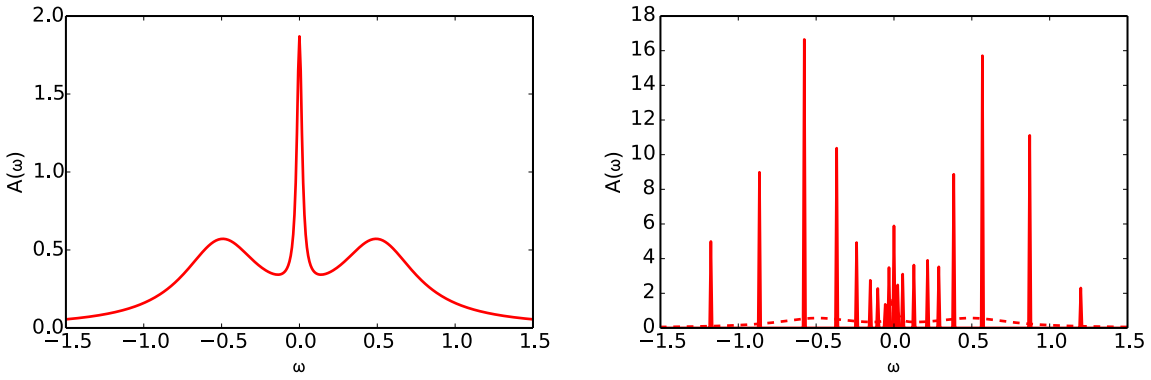


Fig. 4: *Non-negative least-squares solution of the same problem as in Fig. 2. Using our knowledge about the non-negativity of the spectral function gives a dramatic improvement, bringing the solution form a scale of the order of* $\pm 10^{10}$ *to a positive function with peaks of the order of* 10^2 *so that the original model shown on the left can actually also be seen in the plot on the right (dashed line). While the NNLS solution does show some resemblance to the original function it is far to spiky, even in the present case of exceedingly small noise (*$\sim 10^{-8}$*) in the data.*

problem well posed, i.e., giving a unique solution. Still, the spiky NNLS solutions indicate that we are still overfitting the noise in the data and this problem becomes stronger when considering data with noise levels larger than the $\sim 10^{-8}$ used for the example.

While the least-squares approaches do take information about the covariance of the data into account, via the modification of the kernel from $\boldsymbol{K}$ to $\tilde{\boldsymbol{K}}$, so that the data points that are given with higher accuracy have more weight, the results are completely independent of the absolute scale of $\boldsymbol{C}$: Multiplying it by a scalar σ^2 simply rescales all singular values of $\tilde{\boldsymbol{K}}$ by $1/\sigma$, which is compensated but the same rescaling of $\tilde{\boldsymbol{g}}$, leaving the solution unchanged. Thus the least-squares type solutions completely neglect the information about the overall noise in the data. This problem can be addressed when we include our intuition that the "true" solution should show some degree of smoothness. We then have to introduce a measure of smoothness, which puts an absolute scale in the fitting problem. This is the idea behind regularization approaches.

2.3 Linear regularization

To understand the failure of the least-squares methods better, we expand the noisy data and the fit in their respective singular modes $|\boldsymbol{u}_m\rangle$ and $|\boldsymbol{v}_n\rangle$. For the example of Fig. 2 this is shown in Fig. 5. It shows that, initially, the expansion parameters of $\boldsymbol{g}$ decrease somewhat faster with the mode index i than the singular values. Consequently the expansion of the least-squares solution also decrease with i. But once the $\langle \boldsymbol{u}_i|\boldsymbol{g}\rangle$ reach the level of the noise in the data, here $\sigma = 10^{-8}$, at $i \approx 30$, the expansion coefficients of the data remain constant while the singular values decrease further, leading to exponentially increasing contributions of the highly oscillating modes with large i that render the least-squares solution useless. The situation is quite similar for the non-negative least squares solution. The main difference being that the contributions of the modes with small or vanishing singular value are bounded $|\langle \boldsymbol{v}_n|\boldsymbol{f}_{\mathrm{NNLS}}\rangle| \lesssim 1$ by the non-negativity combined with the sum-rule for the model.

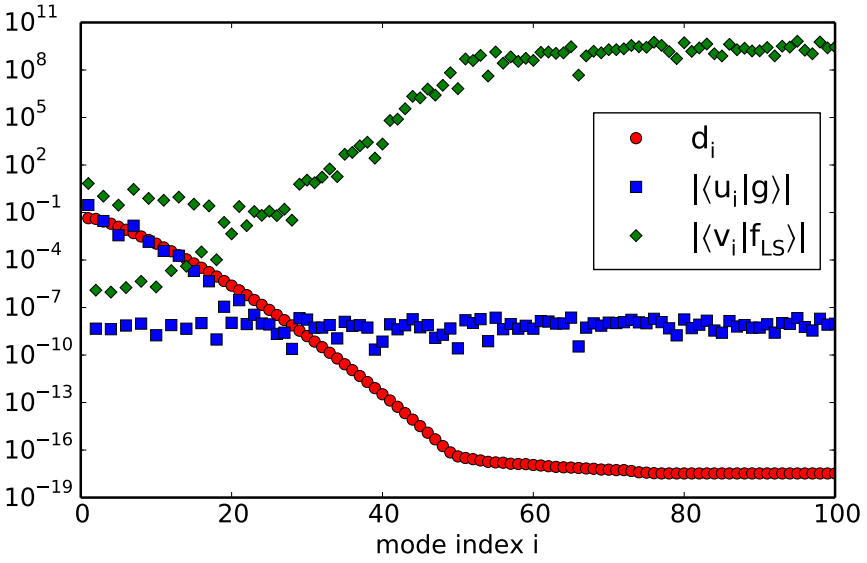


Fig. 5: *Picard plot for the example of Fig. 2. Since the model is symmetric, the expansion coefficients for the odd modes should vanish. For noisy data, instead of vanishing, the coefficients of odd modes are at the noise level. The even coefficients initially decay somewhat faster than the singular values so that the corresponding coefficients of the least-squares solution decrease with i. Once the $\langle \boldsymbol{u}_i|\boldsymbol{g}\rangle$ have decreased to the noise level, here 10^{-8}, they remain at that level while the singular values decrease further. This leads to exponentially increasing contributions of the corresponding modes to the least-squares solution.*

By maximizing the likelihood $\mathrm{e}^{-\chi^2(\boldsymbol{f})/2}$, with or without non-negativity constraint, we apparently overfit the noise that becomes most visible in the modes for which the singular value is below their contribution to the (noisy) data. The assumption behind this is that the exact solution cannot be dominated by the highly oscillating modes with vanishing singular value, i.e., that $\langle \boldsymbol{u}_i|\boldsymbol{g}_{\text{exact}}\rangle/d_i$, for large i decreases with the mode index. This is called the Picard condition. When it is not fulfilled, the reconstruction of the exact model is hopeless, since the relevant information is contained in vanishingly small coefficients $\langle \boldsymbol{u}_i|\boldsymbol{g}_{\text{exact}}\rangle$ that will be completely masked by the noise, cf. (44). When the exact model is not highly oscillating, the Picard condition holds and we have a chance of reconstructing the model from noisy data.

When the Picard condition is fulfilled we can get rid of a large part of the noise by suppressing the contribution of modes with singular value below the noise level in the data. This amounts to a least-squares fit with a truncated singular value decomposition, where the singular values beyond a limiting index are set to zero, $d_{i>i_{\text{trunc}}} := 0$.

A somewhat more refined method is to continuously switch off the small singular modes. This is called Tikhonov regularization. Introducing a regularization parameter α, the Tikhonov solution is given by

$$\boldsymbol{f}_{\text{T}}(\alpha) = \sum_{i=1}^{M} \frac{d_i}{d_i^2+\alpha^2}\langle \boldsymbol{u}_i|\tilde{\boldsymbol{g}}\rangle, \tag{45}$$

which in the limit $\alpha \to 0$ becomes the least-squares solution (43), while for $\alpha \to \infty$ the solution vanishes. For finite regularization parameter, modes with large singular value $d_i \gg \alpha$ are hardly affected, while the contribution of small singular values to $\boldsymbol{f}_{\text{T}}(\alpha)$ vanishes. To employ

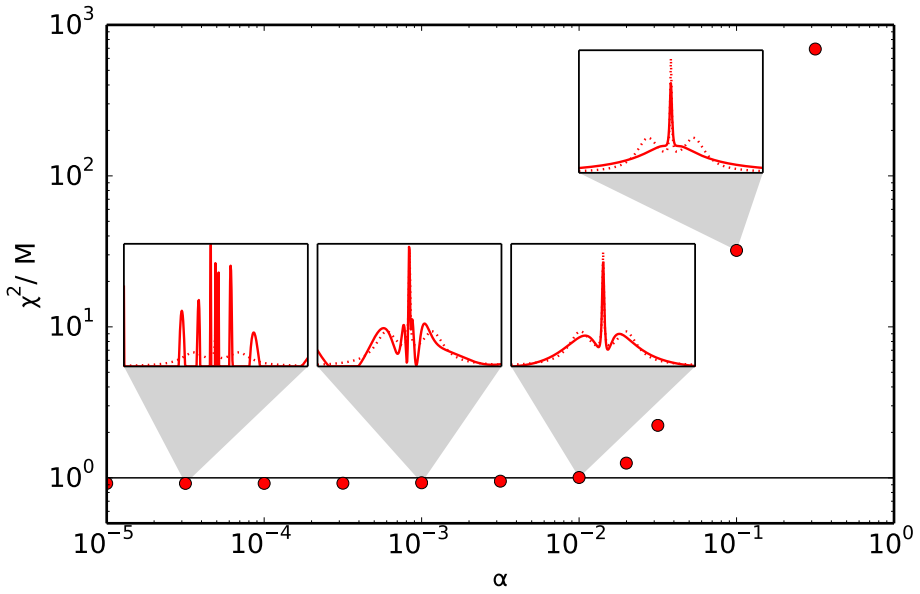


Fig. 6: *Non-negative Tikhonov regularization for the example of Fig. 2, but with noise level increased from* 10^{-8} *to* 10^{-4}*. The insets show the solutions* $\boldsymbol{f}_{\mathrm{T}}(\alpha)$ *at selected values of* α*. The dotted line shows the exact model for comparison. For small* α *the method overfits the noise, leading to strongly oscillating solutions, while the quality of the fit changes little. For large* α *the method underfits the data, leading to a rapid increase in* $\chi^2(\boldsymbol{f}_{\mathrm{T}}(\alpha)$ *and a loss of structure in the reconstructed models. The solid line indicates the expected noise in the data,* $\chi^2 = M$*, relevant for the discrepancy principle.*

Tikhonov regularization for non-negative models we need to formulate it as an optimization problem. Expanding in singular modes and completing the square, it can be written as

$$\begin{aligned}\|\tilde{\boldsymbol{K}}\boldsymbol{f}-\tilde{\boldsymbol{g}}\|^2+\alpha^2\|\boldsymbol{f}\|^2 &= \sum_{i=1}^{M}\left(\langle\boldsymbol{u}_i|\tilde{\boldsymbol{g}}\rangle - d_i\langle\boldsymbol{v}_i|\boldsymbol{f}\rangle\right)^2 + \alpha^2\sum_{n=1}^{N}\langle\boldsymbol{v}_i|\boldsymbol{f}\rangle^2 \\ &= \sum_{i=1}^{M}\left(\frac{\alpha^2\langle\boldsymbol{u}_i|\tilde{\boldsymbol{g}}\rangle}{d_i^2+\alpha^2} + \left(\frac{d_i\langle\boldsymbol{u}_i|\tilde{\boldsymbol{g}}\rangle}{\sqrt{d_i^2+\alpha^2}} + \sqrt{d_i^2+\alpha^2}\langle\boldsymbol{v}_i|\boldsymbol{f}\rangle\right)^2\right) + \alpha^2\sum_{i=M+1}^{N}\langle\boldsymbol{v}_i|\boldsymbol{f}\rangle^2\end{aligned} \tag{46}$$

which attains its minimum $\sum_i \alpha^2\langle\boldsymbol{u}_i|\tilde{\boldsymbol{g}}\rangle/(d_i^2+\alpha^2)$ for the unique solution (45). In Bayesian terms, (40), Tikhonov regularization chooses $p(\boldsymbol{f}) \propto \mathrm{e}^{-\alpha^2\|\boldsymbol{f}\|^2/2}$ as prior probability.
Alternatively, we can express Tikhonov regularization as a least-squares problem with an expanded kernel and data as

$$\min_{\boldsymbol{f}}\left(\|\tilde{\boldsymbol{K}}\boldsymbol{f}-\tilde{\boldsymbol{g}}\|^2+\alpha^2\|\boldsymbol{f}\|^2\right) = \min_{\boldsymbol{f}}\left\|\begin{pmatrix}\tilde{\boldsymbol{K}}\\ \alpha\mathbb{1}_N\end{pmatrix}\boldsymbol{f}-\begin{pmatrix}\tilde{\boldsymbol{g}}\\ 0_N\end{pmatrix}\right\|^2. \tag{47}$$

Performing the minimization over all models gives Tikhonov regularization, restricting the optimization to $\boldsymbol{f} \geq 0$ defines the non-negative Tikhonov regularization method.
The crucial question is how to choose the regularization parameter α. Fig. 6 shows the results of non-negative Tikhonov regularization for the example of Fig. 2 increasing, however, the noise level from 10^{-8} to 10^{-4} to make the problem not too easy. For small α the solutions show

strong oscillations, while the mean-square misfit, χ^2, increases only little with α. For large α the solution becomes featureless except for the peak at the Fermi level, which is already present in the leading singular mode, cf. Fig. 3, while $\chi^2(\boldsymbol{f}_\mathrm{T}(\alpha))$ rapidly gets worse. A compromise between overfitting of the noise in the data and smoothness of the model should be reached when α is chosen such that the deviation from the optimum fit $\chi^2(\boldsymbol{f}_\mathrm{T}(\alpha)) = \|\tilde{\boldsymbol{g}} - \tilde{\boldsymbol{K}}\boldsymbol{f}_\mathrm{T}(\alpha)\|^2$ equals the noise expected in M data points $\tilde{g}_m$ with unit covariance: $\chi^2 = M$. This criterion for choosing the regularization parameter is called the discrepancy principle [8]. We can formulate it as a constrained optimization problem with α^{-2} playing the role of the Lagrange parameter:

$$\min_{\boldsymbol{f}} \|\boldsymbol{f}\|^2 + \frac{1}{\alpha^2}\left(\left\|\tilde{\boldsymbol{g}} - \tilde{\boldsymbol{K}}\boldsymbol{f}\right\|^2 - M\right) \tag{48}$$

has the same variational equation as (46).

The regularization parameter α is the crucial ingredient of any regularization approach. Its role is to strike a balance between fitting the noisy data and keeping the solution smooth in some sense. While it is clear that with increasingly accurate data the chosen α should get smaller, there is no unique procedure for actually determining its value. The discrepancy principle is just one very reasonable way of choosing α but there is a plethora of other approaches, see [8] for a first overview. Likewise, the choice of the regularizer is not unique. Instead of $\|\boldsymbol{f}\|^2 = \langle\boldsymbol{f}|\mathbb{1}|\boldsymbol{f}\rangle$ we could choose any positive semidefinite $N\times N$ matrix $\boldsymbol{M}$ and use $\langle\boldsymbol{f}|\boldsymbol{M}|\boldsymbol{f}\rangle \geq 0$ instead. An obvious choice follows when we remember that $\boldsymbol{f}$ is the discretized version of the model function $f(x)$. As in (35), assuming a uniform x-grid, we can then write

$$\frac{1}{N}\|\boldsymbol{f}\|^2 = \frac{1}{N}\sum_{n=1}^{N}|f_n|^2 \approx \int dx\,|f(x)|^2. \tag{49}$$

Changing the integration variable from x to z, the integral and its Riemann sum in the new coordinates becomes

$$\int dx\,|f(x)|^2 = \int dz\,\frac{dx}{dz}\left|f\big(x(z)\big)\right|^2 \approx \frac{1}{N}\sum_{n=1}^{N}\frac{dx(z_n)}{dz}\left|f\big(x(z_n)\big)\right|^2 \tag{50}$$

so that Tikhonov regularization, $\boldsymbol{M} = \mathbb{1}$, on the old grid becomes regularization on the z-grid with a diagonal matrix $\boldsymbol{M}$ that contains the Jacobian factors $M_{nn} = \frac{dx(z_n)}{dz}$ on the diagonal.
Alternative choices of $\boldsymbol{M}$ impose smoothness by implementing finite-difference versions of the first or higher derivatives, choosing, e.g.,

$$\sum_{n=1}^{N-1}|f_n - f_{n+1}|^2 = \langle\boldsymbol{f}|\begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & & 0 & 0 & 0 & 0 \\ \vdots & & & & & & & & \vdots \\ 0 & 0 & 0 & 0 & & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & -1 & 1 \end{pmatrix}|\boldsymbol{f}\rangle. \tag{51}$$

A regularizer that penalizes the k-th derivative does not have full rank. For, e.g., the first derivative matrix in (51), all constant models give zero. In practice there is, however, no problem since the information about the low moments of the model are usually well contained in the modes with the largest singular values.

Going back to the Tikhonov regularizer, we might wonder why it actually has the effect of smoothing the solution. After all, $\|\boldsymbol{f}\|^2 = \sum_n |f_n|^2$ is local, i.e., does not depend on the change in neighboring values. So if we permuted the coordinate values $\{1, \ldots, N\}$ in an arbitrary way, the value of $\|\boldsymbol{f}\|^2$ would remain unchanged. The main reason why the identity regularizer $\boldsymbol{M} = \mathbb{1}$ leads to smooth models is that it reduces the effect of modes with small singular value. As we have seen in Fig. 3 these modes are highly oscillatory, while the modes that are least affected are the ones with few nodes that are relatively smooth. Still, even the leading mode is not entirely featureless. While a simple first derivative regularizer like (51) would reduce the contribution of such a mode, that is usually strongly supported by the data, Tikhonov will leave it largely unaffected. In that sense, Tikhonov regularization respects the variations in the important modes. To emulate this with a derivative regularizer would require to laboriously taylor an $\boldsymbol{M}$ suitable for every specific kernel $\tilde{\boldsymbol{K}}$.

There is also a second aspect. As we noted above, $\boldsymbol{f}_{\mathrm{T}}(\alpha \to \infty) = 0$. When we impose a sum rule, however, we force the solution to be finite and find from

$$\min_{\boldsymbol{f}} \left(\sum_n f_n^2 + \lambda_0 \Big(1 - \sum_n f_n \Big) \right) \tag{52}$$

that the Tikhonov regularizer prefers a flat solution, $f_n = 1/N$, or, in the case of a general diagonal matrix, $f_n \propto 1/M_{nn}$. These are the models resulting in the absence of data, i.e., for diverging variance resulting in vanishing $\tilde{\boldsymbol{K}}$ and $\tilde{\boldsymbol{g}}$, except for the 0-th moment sum-rule. They are called the default model of the regularizer.

We are, of course, not limited to bilinear regularizers of the type $\langle \boldsymbol{f}|M|\boldsymbol{f}\rangle$. An important non-linear regularizer is the entropy of the model. It is the basis of the maximum entropy approach.

2.4 Maximum entropy

Maximum entropy methods differ from Tikhonov-type regularization in the assumptions they make about the solutions. While Tikhonov is based on the Picard condition giving preference to the modes with large singular value, maximum entropy favors models that contain as little information as possible. This is measured by the information entropy, see A.3 for details. Using the generalized entropy (89)

$$H(\boldsymbol{f}; \boldsymbol{\rho}) = -\sum_n \left(f_n \ln \frac{f_n}{\rho_n} - f_n + \rho_n \right) \tag{53}$$

as regularizer that should be maximized, we have to solve the non-linear optimization problem

$$\min_{\boldsymbol{f}} \Big(\chi^2(\boldsymbol{f})/2 - \alpha H(\boldsymbol{f}; \boldsymbol{\rho}) \Big), \tag{54}$$

where we use α instead of α^2 as in (46) to conform with the conventions used, e.g., in [9]. A convenient property of the non-linear entropy regularizer is that it automatically ensures the positivity of the solution since the gradient

$$-\frac{\partial H(\boldsymbol{f};\boldsymbol{\rho})}{\partial f_n} = \ln\frac{f_n}{\rho_n} \tag{55}$$

diverges for $f_n \to 0$ while the gradient of the fit function

$$\frac{\partial \frac{1}{2}\chi^2(\boldsymbol{f})}{\partial f_n} = \frac{\partial}{\partial f_n}\frac{1}{2}\sum_m \Big(\tilde{g}_m - \sum_{n'}\tilde{K}_{mn'}f_{n'}\Big)^2 = -\sum_m \tilde{K}^\dagger_{nm}\Big(\tilde{g}_m - \sum_{n'}\tilde{K}_{mn'}f_{n'}\Big) \tag{56}$$

is always finite, so that any solution of the variational equations f_n has the same sign as ρ_n. Since it is easy to also calculate the Hessians

$$\frac{\partial^2 \frac{1}{2}\chi^2(\boldsymbol{f})}{\partial f_{n'}\partial f_n} = \sum_m \tilde{K}^\dagger_{nm}\tilde{K}_{mn'} \quad \text{and} \quad -\frac{\partial^2 H(\boldsymbol{f};\boldsymbol{\rho})}{\partial f_{n'}\partial f_n} = \frac{1}{f_n}\,\delta_{nn'} \tag{57}$$

it is straightforward to solve the non-linear minimization problem using, e.g., the Levenberg-Marquardt method. The only slight complication being that finite steps in the iteration might change the sign of some component f_n.

In the absence of data, minimizing (54), i.e, setting (55) to zero, we find $\boldsymbol{f} = \boldsymbol{\rho}$. Thus $\boldsymbol{\rho}$ is the default model and, as in (88) or (50) it can be related to the choice of the grid. Even when we have decided on a default model, we still have to determine the value of the regularization parameter. Choosing it according to the discrepancy principle is called historic MaxEnt [9].

Other flavors of the maximum entropy method determine the regularization parameter using Bayesian methods. For this we write the entropy regularizer as a prior probability

$$p(\boldsymbol{f}\,|\boldsymbol{\rho},\alpha) \propto \mathrm{e}^{+\alpha H(\boldsymbol{f};\boldsymbol{\rho})} \tag{58}$$

so that the minimization (54) becomes equal to maximizing the posterior probability, cf. (40),

$$p(\boldsymbol{f}\,|\tilde{\boldsymbol{g}},\boldsymbol{\rho},\alpha) = \frac{p(\tilde{\boldsymbol{g}}\,|\boldsymbol{f},\boldsymbol{\rho},\alpha)\,p(\boldsymbol{f}\,|\boldsymbol{\rho},\alpha)}{p(\tilde{\boldsymbol{g}})} \propto \mathrm{e}^{-\chi^2(\boldsymbol{f})/2+\alpha H(\boldsymbol{f};\boldsymbol{\rho})}\,, \tag{59}$$

where we have used that the QMC data $\tilde{\boldsymbol{g}}$ is actually independent of our choice of regularization parameter and default model. The Bayesian approach to determining the regularization parameter uses the posterior probability of α

$$p(\alpha\,|\tilde{\boldsymbol{g}},\boldsymbol{\rho}) = \int\prod_n \frac{df_n}{\sqrt{f_n}}\; p(\boldsymbol{f},\alpha\,|\tilde{\boldsymbol{g}},\boldsymbol{\rho}) = \int\prod_n 2d\sqrt{f_n}\;\frac{p(\boldsymbol{f}\,|\tilde{\boldsymbol{g}},\boldsymbol{\rho},\alpha)\,p(\tilde{\boldsymbol{g}},\boldsymbol{\rho},\alpha)}{p(\tilde{\boldsymbol{g}},\boldsymbol{\rho})} \tag{60}$$

obtained from marginalizing out $\boldsymbol{f}$, i.e., integrating over the space of models $\boldsymbol{f}$. The peculiar choice of the integration measure, $2d\sqrt{f_n}$, is discussed in [9]. It naturally appears in the expression for the entropy when using Stirling's approximation to one order higher than in the derivation given in A.3, which rather suggests that the factor $1/\sqrt{f_n}$ should be considered part of the entropic prior and not the integration measure.

For historic MaxEnt it was not necessary to know the normalized probabilities (58) and (59). When, however, we want to compare the probabilities of different renormalization parameters, we need to determine the normalization of the distributions that depend on α by, again, integrating over $\boldsymbol{f}$

$$Z_{\chi^2}(\tilde{\boldsymbol{g}}) = \int_{-\infty}^{\infty} \prod_n df_n \, \mathrm{e}^{-\chi^2(\boldsymbol{f})} = (2\pi)^{M/2} \tag{61}$$

$$Z_H(\boldsymbol{\rho}, \alpha) := \int \prod_n df_n \, \frac{\mathrm{e}^{+\alpha H(\boldsymbol{f},\boldsymbol{\rho})}}{\prod_{n'} \sqrt{f_{n'}}} \tag{62}$$

Since the likelihood is a Gaussian, the integral is straightforward, cf. (34). The normalization of the entropic prior is more difficult. In MaxEnt such functional integrals are approximated by Gaussian integrals obtained from expanding the exponent to second order about its maximum. As already discussed above, the entropy term is maximized when the model equals the default model. The second-order expansion (57) is thus given by the diagonal matrix $-\delta_{nn'}/\rho_n$, i.e., the entropy becomes just a Tikhonov regularizer with general diagonal matrix. Also expanding $1/\sqrt{f_n} \approx \big(1-(f_n-\rho_n)/2\rho_n\big)/\sqrt{\rho_n}$, the normalization of the entropy prior thus is approximated by that of the simple Tikhonov prior without non-negativity constraint [9]

$$Z_H(\boldsymbol{\rho}, \alpha) \approx \prod_n \int_{-\infty}^{\infty} df_n \, \frac{\mathrm{e}^{-\frac{\alpha(f_n-\rho_n)^2}{2\rho_n}}}{\sqrt{\rho_n}} \left(1 - \frac{f_n - \rho_n}{2\rho_n}\right) = \left(\frac{2\pi}{\alpha}\right)^{N/2}. \tag{63}$$

To calculate $p(\alpha \,|\tilde{\boldsymbol{g}}, \boldsymbol{\rho})$, (60), we still have to choose a prior probability $p(\tilde{\boldsymbol{g}}, \boldsymbol{\rho}, \alpha)$. From the discrepancy principle it seems reasonable that it should be independent of the data normalized to have a unit covariance matrix. If we also assume that α is independent of the default model, we only have to choose $p(\alpha)$. Assuming that the prior is scale invariant

$$p(\alpha)\, d\alpha \stackrel{!}{=} p(s\alpha)\, d(s\alpha) \tag{64}$$

one obtains the Jeffreys prior $p(\alpha) \propto 1/\alpha$ [2], which might not be the most appropriate choice, since the scale of the regularization parameter is fixed by the noise in the data, which we know. Using all this for calculating the posterior probability of the renormalization parameter, there are two different flavors of how α enters the analytical continuation: Historic MaxEnt chooses the α that maximizes $p(\alpha \,|\tilde{\boldsymbol{g}}, \boldsymbol{\rho})$ to determine $\boldsymbol{f}_{\text{historic}} = \boldsymbol{f}(\alpha_{\max})$. Bryan's method no longer insists on picking a specific value of α. The approach rather determines the model as the average over all regularization parameters, weighted with their posterior probability

$$\boldsymbol{f}_{\text{Bryan}} = \int_0^{\infty} d\alpha \, \boldsymbol{f}(\alpha)\, p(\alpha \,|\tilde{\boldsymbol{g}}, \rho)\,. \tag{65}$$

It might seem that the MaxEnt approaches could be improved by actually performing the integrals over model space exactly rather than using simple Gaussian approximations that even violate the non-negativity of the models, which is one of the precious priors that we are sure

of. But doing these integrals is fiendishly hard. A second drawback of MaxEnt, or rather of all regularization approaches, is the need to deal with a regularization parameter, introducing the need for making assumptions about its behavior, its prior probability and the like, for which there is no apparent solution. If only we could efficiently integrate over model space there might be a way of eliminating all these complication arising from the need to regularize. Instead of looking for a solution that maximizes some posterior probability, we could ask for the average over all possible models, weighted with their likelihood. This approach which is free of explicit regularization parameters is the average spectrum method.

3 Average spectrum method

The average spectrum methods is an appealing alternative to the optimization approaches. It was probably first proposed by White [10] and reinvented several times after. The basic idea is of striking elegance: The spectral function is obtained as the average over all physically admissible functions, weighted by how well they fit the data

$$f_{\mathrm{ASM}}(x) := (2\pi)^{-M/2} \int_{f(x)\geq 0} \mathcal{D}f \; f(x)\, \mathrm{e}^{-\chi^2[f]/2}. \tag{66}$$

Due to the ill-conditioning of the inverse problem there are very many functions that differ drastically but essentially fit the data equally well. Taking the average, we can thus expect that the spectral features not supported by the data will be smoothed out, providing a regularization without the need for explicit parameters. So far the practical application of this conceptually appealing approach has, however, been hampered by the immense computational cost of numerically implementing the functional integration.

It is worth emphasisizing that the non-negativity constraint is essential. An unconstrained integration over the Gaussian in (66) actually produces a least-squares solution. Since the width of the likelihood increases with the inverse of the singular value this would, of course, be numerically very inefficient, and since the width diverges for the modes in the null space of the kernel, their contribution will never converge to a definite value, reflecting that the problem is underdetermined.

When we discretize the model function $f(x)$ as discussed in Sec. 1.3, the functional integral becomes

$$\boldsymbol{f}_{\mathrm{ASM}} \propto \prod_{n=1}^{N} \int_0^\infty df_n \, \boldsymbol{f} \, \mathrm{e}^{-\chi^2(\boldsymbol{f})/2}, \tag{67}$$

where the N-dimensional integral can be evaluated by Monte Carlo techniques. The most straightforward approach is to perform a random walk in the space of non-negative vectors $\boldsymbol{f}$, updating a single component, $f_n \to f'_n$, at a time. Detailed balance is fulfilled when we sample the new component f'_n for the conditional distribution $\propto \mathrm{e}^{-\chi^2(\boldsymbol{f};f'_n)/2}$ with

$$\chi^2(\boldsymbol{f}; f'_n) := \big\| \underbrace{\tilde{\boldsymbol{g}} - \tilde{\boldsymbol{K}}\boldsymbol{f} + \tilde{\boldsymbol{K}}_n f_n}_{=:\tilde{\boldsymbol{g}}_n} - \tilde{\boldsymbol{K}}_n f'_n \big\|^2 = \tilde{\boldsymbol{K}}_n^\dagger \tilde{\boldsymbol{K}}_n \left(f'_n - \tilde{\boldsymbol{K}}_n^\dagger \, \tilde{\boldsymbol{g}}_n / \tilde{\boldsymbol{K}}_n^\dagger \tilde{\boldsymbol{K}}_n \right)^2, \tag{68}$$

where $\tilde{\boldsymbol{K}}_n$ is the n-th column of $\tilde{\boldsymbol{K}}$. We thus have to sample f'_n from a univariate Gaussian of width $\sigma = 1/\|\tilde{\boldsymbol{K}}_n\|$ centered at $\mu = \tilde{\boldsymbol{K}}_n^\dagger \tilde{\boldsymbol{g}}_n / \|\tilde{\boldsymbol{K}}_n\|^2$ and truncated to the non-negative values $f'_n \in [0, \infty)$. This can be done very efficiently, e.g., as described in A.4.

Still, sampling components can be very slow because the width of the Gaussian (68) is, in general set by the inverse of the largest singular value, i.e., the random walk performs only exceedingly small steps. This is even more evident when sampling spectral functions, where we cannot change just a single f_n without violating the sum-rule (15). A way around is to introduce global moves along the principal axes of χ^2, i.e., along the singular modes.

Transforming to the new bases $\boldsymbol{h} := \boldsymbol{U}^\dagger \tilde{\boldsymbol{g}}$ in data and $\boldsymbol{e} := \boldsymbol{V}^\dagger \boldsymbol{f}$ in model space diagonalizes

$$\chi^2(\boldsymbol{f}) = \left\| \boldsymbol{U}^\dagger \tilde{\boldsymbol{g}} - \boldsymbol{D}\boldsymbol{V}^\dagger \boldsymbol{f} \right\|^2 = \sum_{i=1}^{M} (h_i - d_i e_i)^2 \tag{69}$$

so that in the new basis the integral (67) factorizes into Gaussian integrals

$$e_i^{\mathrm{ASM}} \propto \int_{\boldsymbol{f} \geq 0} de_i \, e_i \, \mathrm{e}^{-(h_i - d_i e_i)^2/2}. \tag{70}$$

Without non-negativity constraint the integrals would be independent and result in a least-squares solution. With the constraint they are coupled via their range of integration. Updating modes $e_i \to e'_i$ is restricted by the condition $\boldsymbol{f}' = \boldsymbol{f} + (e'_i - e_i)\boldsymbol{V}_i \geq 0$, where $\boldsymbol{V}_i$ is the i-th column vector of $\boldsymbol{V}$. This is equivalent to $e'_i \geq e_i - f_n / V_{ni}$ for $V_{ni} > 0$ and correspondingly for $V_{ni} < 0$ so that e'_i is constrained to

$$\max \left\{ \frac{f_n}{V_{ni}} \middle| V_{ni} < 0 \right\} \leq e_i - e'_i \leq \min \left\{ \frac{f_n}{V_{ni}} \middle| V_{ni} > 0 \right\}. \tag{71}$$

Sampling modes is usually much more efficient than sampling components: For modes with large singular value the Gaussian is narrow so that the random walk quickly jumps close to the optimal value h_i/d_i and then takes small steps around there. For modes with small or zero singular value the distribution is very broad so that the random walk can take large steps, allowing for an efficient sampling. Still, sampling may become inefficient when non-negativity restricts a mode to a very narrow interval. This will happen when $\boldsymbol{f}$ has regions where it becomes very small, e.g., in the tail of the spectral function. Then the scale for the step size is not given by the singular value but rather by the width of the interval (71). Also this problem can be overcome by using a real space renormalization group technique, introducing blocks of varying size in which modes are sampled. This way the method can interpolate efficiently between sampling components, i.e., blocks of size 1, and sampling modes, i.e., blocks of size N. Details of the method and its performance are given in [11, 12].

Using this approach, we find that the results of the average spectrum method actually depend on the choice of the discretization (67). This is not a problem of the particular method, but a general feature of the functional integral and would also affect, e.g., MaxEnt were it to do the normalization and marginalization integrals exactly, see e.g. Sec. 6.2 of [2]. We find that the choice of the coordinates for the discretization grid plays the role of a default model, while the number of grid points N acts as a regularization parameter.

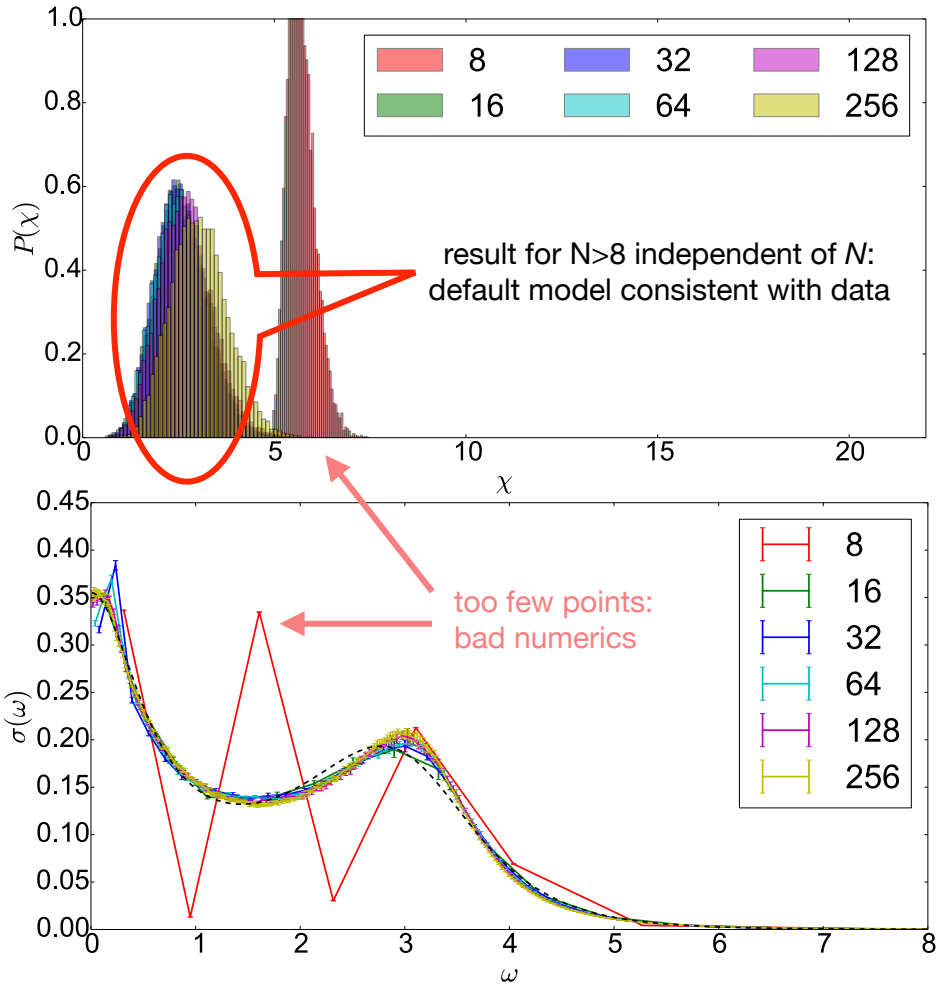


Fig. 7: *Average spectrum method on a Gaussian grid determined from NNLS for an optical conductivity reconstructed from Matsubara data [11]. The resulting spectral shown below are rather insensitive to the choice of the regularization parameter, the number of grid points N, except for $N = 8$ where the coarse grid leads to discretization errors when evaluating the Fredholm integral. The exact model is shown as the dashed line for comparison. The top panel shows the histograms of χ taken during the Monte Carlo sampling of the functional integral. The small variations in the histograms indicate a robust choice of the grid.*

The reason for this is that the notion of sampling uniformly, i.e., with a flat prior, is tied to the choice of a specific grid. This is most easily understood when we consider what happens when we double the number of grid points. On the original grid we sample $f \in [0, \infty)$. On the denser grid we represent f over the large interval by two values $\hat{f}_1$ and $\hat{f}_2$ over intervals of half the width, so that $f = \hat{f}_1 + \hat{f}_2$. If we sample the $\hat{f}_i \in [0, \infty)$ with a flat prior, $p(\hat{f}_i) = \text{const.}$, this implies a probability distribution for f

$$p(f) = \int_0^\infty d\hat{f}_1\, p(\hat{f}_1) \int_0^\infty d\hat{f}_2\, p(\hat{f}_2)\; \delta\big(\hat{f}_1 + \hat{f}_2 - f\big) \propto \int_0^f d\hat{f}_1 = f \tag{72}$$

which is *not* flat. Properly defining the flat prior as $p(\hat{f}_i) = \lim_{\lambda\to\infty} \mathrm{e}^{-\hat{f}_i/\lambda}/\lambda$, $p(f)$ becomes a gamma distribution. More generally, we find that sampling with a flat distribution on a particular grid defines a measure for the functional integral (66) represented by a gamma process [11].

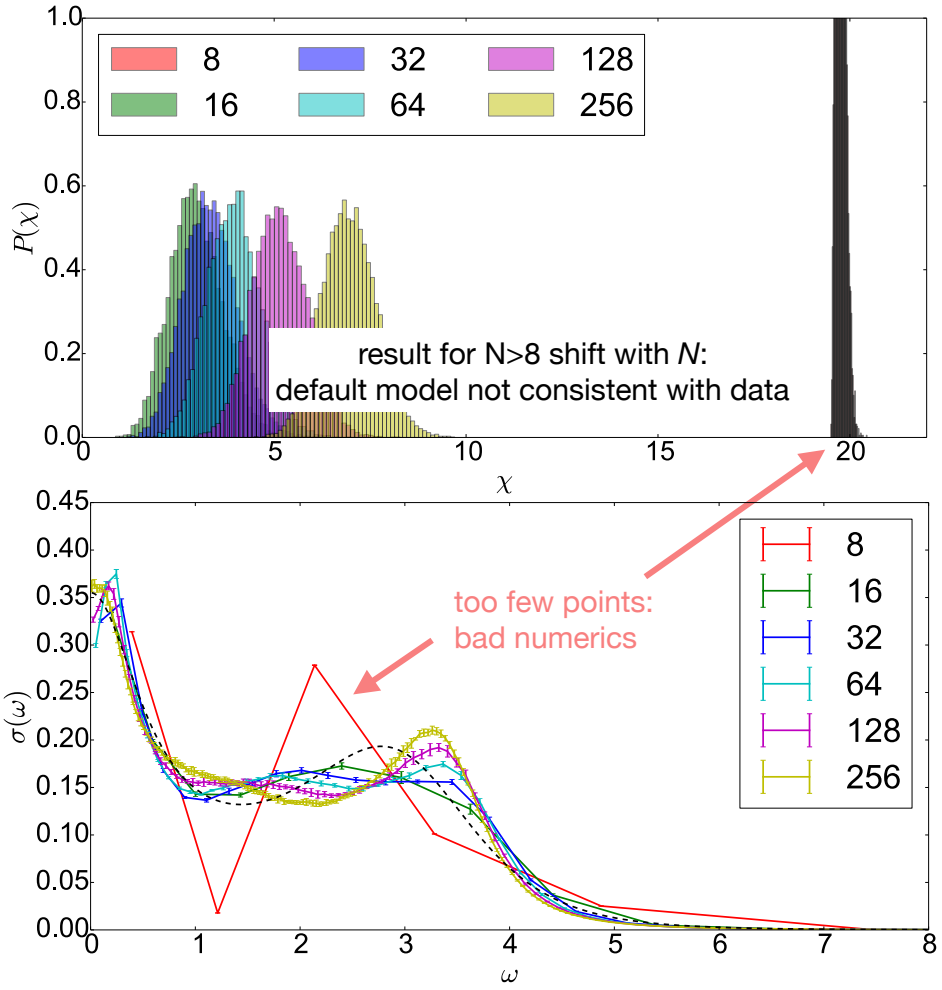


Fig. 8: *Average spectrum method on a Lorentzian grid of the same width as the Gaussian in Fig. 7. The resulting spectral functions shown below are now quite sensitive to the choice of the regularization parameter, the number of grid points N. In the top panel we see from the histograms of χ that the fit sizeably deteriorates with increasing N, indicating problems with the choice of the grid.*

Still, we can give a practical recipe for determining the regularization parameters and checking the quality of the results. To find the grid type (default model), we use non-negative least-squares to determine the width of the spectrum. This implies a Gaussian grid of particular width. We then vary the number of grid points to check how the results change with increasing N. When N is too small, the result will be inaccurate because of discretization errors in evaluating the integral (31) entering χ^2. When N becomes to large there is a rapidly increasing number of vectors $\boldsymbol{f}$ that, despite having a small weight $\mathrm{e}^{-\chi^2(\boldsymbol{f})/2}$, contribute to the average due to their sheer number. In between there will be a region, where the results are fairly independent of the actual choice of N. This is shown in Fig. 7. When the grid is not chosen well, as in Fig. 8, where the grid uses a Lorentzian density of the same width as the Gaussian in the previous figure, results vary strongly with N.

This approach gives already reliable and robust results. When we have to deal with particularly difficult cases, we can use Bayesian techniques to make the method even more robust by sampling over different grids, albeit at an increased computational cost [11].

4 Conclusions

As for so many problems, there is no magic solution to the problem of analytic continuation. Any method can only reconstruct what is in the data and must substitute missing information ideally by exact prior knowledge or, otherwise, by mere assumptions about the solution. The most important aspect of analytic continuation is thus encountered already before the solution of the inverse problem is even started. Depending on what features of the model we are interested in, we have to decide where to measure the data. If we want, e.g., to reconstruct the spectral function far from the Fermi level, it does not help to just have highly accurate values for the Green function when they are not close enough to $\tau = 0$ or β to give information about the discontinuity in its derivatives.

Moreover, it is deceiving to just look at the single result returned by a regularization approach. There is not "the" solution, rather every method produces an expected solution with its uncertainty quantified by a non-intuitive $N \times N$ covariance matrix. This is, however, rarely analyzed because it is hard to calculate and difficult to interpret. Still, there are approaches to estimate the error in observables derived by integrating over the spectrum. They are nicely discussed in [9] and should be used wherever possible.

We have presented the approaches to the analytic continuation problem in the order of increasing sophistication and accuracy — and numerical cost. The analysis of QMC data should ideally follow this progression until the desired information about the spectral function has been reliably obtained. A Picard plot will give a first impression of how much information is actually contained in the data and what can be expected from a straightforward linear regularization. Despite the uncontrolled approximations in the practical flavors of MaxEnt, the approach has developed into the standard approach for analytical continuation. The average spectrum method, that is now numerically competitive, provides an appealing alternative since it makes all assumptions via the choice of the discretization explicit, while being numerically exact.

The most important lesson is that results of analytic continuation must not be overinterpreted. When the results depend on the details of the method, they rather reflect the choices made by the approach than the data. Thus before interpreting details of the spectral function, we have to make sure that they are robust under (reasonable) variations in the regularization parameters. The discrepancy principle and the fit histogram are practical methods for doing this.

A Technical appendices

A.1 Blocking method for correlated data

Let us assume we have an ergodic Markov chain Monte Carlo method, e.g., using Metropolis sampling, that generates a set of K data points $m_1, \ldots, m_K$ drawn from a probability distribution $p(m)\,dm$ and we are interested in the mean value $\mu = \int dm\, p(m)\, m$. The obvious estimate for μ is the average $\bar{m} = \sum_{k=1}^{K} m_k/K$. It will, of course, be different for different Monte Carlo runs, but, by the central limit theorem, for large K the averages $\bar{m}$ of different runs will tend to be distributed as a Gaussian centered at μ with variance

$$\sigma^2(\bar{m}) = \langle \bar{m}^2 \rangle - \langle \bar{m} \rangle^2 = \frac{1}{K^2} \sum_{k,l=0}^{K} \Big(\langle m_k m_l \rangle - \langle m_k \rangle \langle m_l \rangle \Big), \tag{73}$$

where $\langle \cdot \rangle$ is the average over all possible Monte Carlo runs producing K data points. How can we estimate $\sigma^2(\bar{m})$ from the simulation data of a single run? Splitting the double sum

$$\sigma^2(\bar{m}) = \frac{1}{K} \underbrace{\frac{1}{K} \sum_{k=1}^{K} \Big(\langle m_k^2 \rangle - \mu^2 \Big)}_{= \langle m^2 \rangle - \mu^2 =: s_0} + \frac{1}{K^2} \sum_{k \neq l} \Big(\langle m_k m_l \rangle - \mu^2 \Big) \tag{74}$$

we see that for uncorrelated data, $\langle m_k m_l \rangle = \langle m_k \rangle \langle m_l \rangle$ for $k \neq l$, the variance is given by s_0/K. But in general, samples obtained from Markov chain Monte Carlo will be positively correlated, so that $\sigma^2(\bar{m}) \geq s_0/K$. We can eliminate this correlation using an elegant renormalization group technique [13]. For this we consider the transformation of the original data set $m_1, \ldots, m_K$ of K samples (assuming K is even) into half as many data points, obtained by averaging

$$\hat{m}_{\hat{k}} := \frac{m_{2\hat{k}-1} + m_{2\hat{k}}}{2}\,. \tag{75}$$

Obviously, the average of the new data points $\sum_{\hat{k}=1}^{K/2} \hat{m}_{\hat{k}}/(K/2)$ is still $\bar{m}$ and thus must have the same distribution as the averages of the original data. Consequently, $\sigma^2(\bar{m})$ must remain invariant under the blocking transformation (75). Looking at the uncorrelated part of the variance for the blocked data $\hat{m}_{\hat{k}}$ and remembering that the ensemble average $\langle m_k^2 \rangle$ is independent of k, we see that

$$\hat{s}_0 = \frac{1}{K/2} \sum_{\hat{k}=1}^{K/2} \left(\frac{\langle m_{2\hat{k}-1}^2 + 2 m_{2\hat{k}-1} m_{2\hat{k}} + m_{2\hat{k}}^2 \rangle}{4} - \mu^2 \right) = \frac{s_0}{2} + \frac{1}{2\,K/2} \sum_{\hat{k}=1}^{K/2} \big(\langle m_{2\hat{k}-1} m_{2\hat{k}} \rangle - \mu^2 \big) \tag{76}$$

contains part of the correlations not contained in s_0. Therefore $\hat{s}_0/(K/2) \geq s_0/K$. Under repeated blocking transformations the uncorrelated part of the variance will thus increase. When it reaches a plateau, i.e., a fixed-point under the blocking transformation, it becomes equal to $\sigma^2(\bar{m})$ and the blocked data has become uncorrelated.

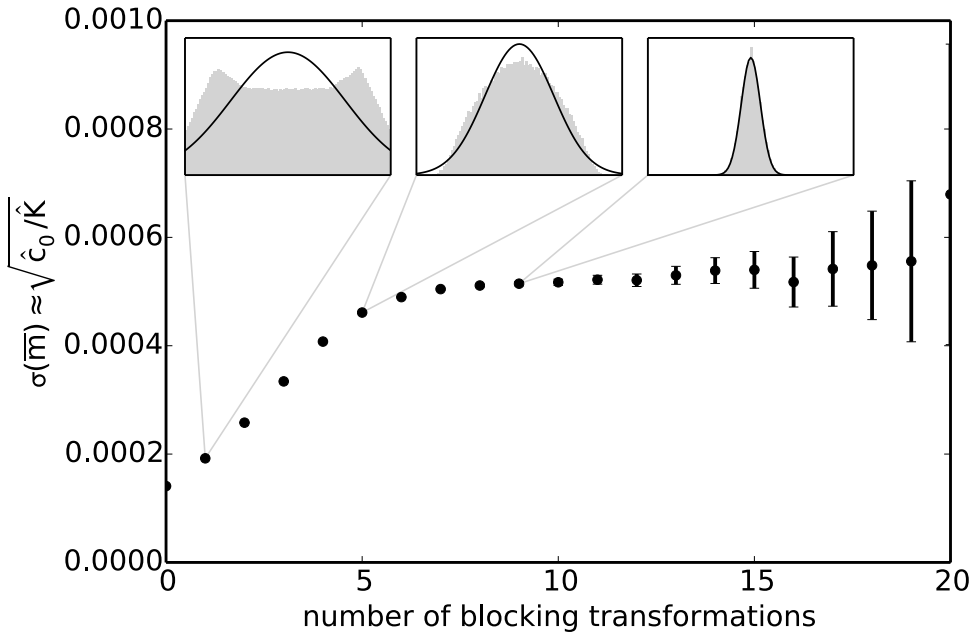


Fig. 9: *Estimate of the standard deviation of the average of correlated data obtained with the blocking method. Initially the variance is severely underestimated but the estimate increases with each blocking step until a plateau is reached, at which point the blocked data has become uncorrelated. By that time the distribution of the $\hat{m}_{\hat{k}}$ has become Gaussian of width $\hat{c}_0$, as shown in the insets. Eventually the number of blocked samples, $\hat{K} = K/2^n$, become so small that the estimates become unreliable.*

We can try to estimate the ensemble average s_0 from the data from one specific simulation run as $\sum_{k=1}^{K}(m_k^2 - \bar{m}^2)/K$. Taking the ensemble average and comparing to s_0

$$\frac{1}{K}\sum_{k=1}^{K}\left(\langle m_k^2\rangle - \langle\bar{m}^2\rangle\right) = \frac{1}{K}\sum_{k=1}^{K}\Big(\big(\langle m_k^2\rangle - \langle\bar{m}\rangle^2\big) - \big(\langle\bar{m}^2\rangle - \langle\bar{m}\rangle^2\big)\Big) = s_0\left(1-\frac{1}{K}\right)$$

we find that the unbiased estimator actually is

$$s_0 \approx c_0 := \frac{1}{K-1}\sum_{k=1}^{K}\big(m_k^2 - \bar{m}^2\big) \quad\Rightarrow\quad \sigma^2(\bar{m}) \approx \frac{1}{K(K-1)}\sum_{k=1}^{K}\big(m_k^2 - \bar{m}^2\big). \tag{77}$$

In an actual implementation of the blocking method, we repeatedly block the data and calculate the corresponding estimator of the uncorrelated variance $\hat{s}_0/\hat{K}$. An example is shown in Fig. 9. As expected, $\hat{c}_0/\hat{K}$ increases with each blocking step until it reaches a plateau. There the blocked data $\hat{m}_{\hat{k}}$ are uncorrelated and, by the central limit theorem, approach Gaussian variables of variance $\sigma^2(\hat{m}) = \hat{K}\sigma^2(\bar{m})$. For such variables the variance of the variance $\sigma^2(\bar{m})$ is given by $\langle(\hat{c}_0/\hat{K})^2\rangle - \langle\hat{c}_0/\hat{K}\rangle^2 = 2\sigma^4(\bar{m})/(\hat{K}-1)$, which provides us with the errorbars. Since the number of blocked data points is halved in each step, eventually the blocked sample becomes very small and $\hat{c}_0/\hat{K}$ starts to fluctuate, also indicated by rapidly increasing errorbars. We can then identify the plateau by checking when $\hat{s}_0/\hat{K}$ does not change between blocking steps within its error bar.

A.2 Non-negative least-squares algorithm (NNLS)

The model $\boldsymbol{f}$ fitting a given data vector $\boldsymbol{g}$ best in the least-squares sense minimizes the norm of the residual vector $\chi^2(\boldsymbol{f}) = \|\boldsymbol{K}\boldsymbol{f} - \boldsymbol{g}\|^2$. At the minimum $\boldsymbol{f}_{\mathrm{LS}}$ the gradient $\boldsymbol{w}$ vanishes

$$w_n(\boldsymbol{f}_{\mathrm{LS}}) := \frac{1}{2}\,\frac{\partial\,\chi^2(\boldsymbol{f})}{\partial f_n}\bigg|_{\mathrm{LS}} = \mathrm{Re}\left(\boldsymbol{K}^\dagger(\boldsymbol{K}\boldsymbol{f}_{\mathrm{LS}}-\boldsymbol{g})\right)_n = 0 \quad \forall n\,. \tag{78}$$

Since χ^2 is a non-negative quadratic form in $\boldsymbol{f}$ stationary points must be minima

$$\chi^2(\boldsymbol{f}_{\mathrm{LS}}+\boldsymbol{\delta}) = \chi^2(\boldsymbol{f}_{\mathrm{LS}}) + 2\boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{w}(\boldsymbol{f}_{\mathrm{LS}}) + \|\boldsymbol{K}\boldsymbol{\delta}\|^2 \geq \chi^2(\boldsymbol{f}_{\mathrm{LS}})\,. \tag{79}$$

The least-squares fit can be found from the singular value decomposition (SVD) $\boldsymbol{K}^\dagger = \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\dagger$

$$\boldsymbol{f}_{\mathrm{LS}} = \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}^\dagger\boldsymbol{g}\,, \tag{80}$$

where the diagonal matrix $\boldsymbol{D}$ has dimension $K = \mathrm{rank}(\boldsymbol{K})$ while the matrix $\boldsymbol{U}$ is $M\times K$ and $\boldsymbol{V}$ is $N\times K$-dimensional. In terms of the SVD the gradient (78) is given by

$$\boldsymbol{w}(\boldsymbol{f}) = \mathrm{Re}\,\boldsymbol{V}\boldsymbol{D}(\boldsymbol{D}\boldsymbol{V}^{\mathsf{T}}\boldsymbol{f} - \boldsymbol{U}^\dagger\boldsymbol{g})\,. \tag{81}$$

This way of calculating the gradient is numerically more stable than calculating it directly in terms of $\boldsymbol{K}$. It also immediately shows that the gradient vanishes for $\boldsymbol{f}_{\mathrm{LS}}$.
Finding the best fit, $\min\|\boldsymbol{K}\boldsymbol{f} - \boldsymbol{g}\|^2$, under the constraint $\boldsymbol{f} \geq 0$ (non-negative least-squares, NNLS) is more complicated. When all components of the unconstrained solution are non-negative, $(\boldsymbol{f}_{\mathrm{LS}})_n \geq 0$, it is obviously also the solution of the constrained problem. When there are components $(\boldsymbol{f}_{\mathrm{LS}})_n < 0$ we might expect that the constrained fit assumes its minimum on the boundary, $(\boldsymbol{f}_{\mathrm{NNLS}})_n = 0$, where the gradient is positive $w_n > 0$. These are the Karush-Kuhn-Tucker conditions [14]:

$$f_n > 0 \quad \text{and} \quad w_n = 0 \qquad \text{or} \qquad f_n = 0 \quad \text{and} \quad w_n \geq 0\,. \tag{82}$$

We distinguish the two cases by defining the two sets $\mathcal{P} = \{n|\, f_n > 0\}$ and $\mathcal{Z} = \{n|\, f_n = 0\}$ which partition the set of indices, $\mathcal{P} \cup \mathcal{Z} = \{1, \ldots, N\}$.
When $\boldsymbol{f}_{\mathrm{KT}}$ fulfills the Karush-Kuhn-Tucker conditions, it minimizes $\chi^2(\boldsymbol{f})$ on $\boldsymbol{f} \geq 0$. To see this we consider a vector $\boldsymbol{f}_{\mathrm{KT}} + \boldsymbol{\delta}$ with $\delta_z \geq 0$ so that $\boldsymbol{f}_{\mathrm{KT}} + \boldsymbol{\delta} \geq 0$. Then

$$\chi^2(\boldsymbol{f}_{\mathrm{KT}}+\boldsymbol{\delta}) = \chi^2(\boldsymbol{f}_{\mathrm{KT}}) + 2\boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{w}(\boldsymbol{f}_{\mathrm{KT}}) + \|\boldsymbol{K}\boldsymbol{\delta}\|^2 \geq \chi^2(\boldsymbol{f}_{\mathrm{KT}}) \tag{83}$$

since $\boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{w} = \sum_n \delta_n w_n = \sum_{n\in\mathcal{P}} \delta_n w_n + \sum_{n\in\mathcal{Z}} \delta_n w_n \geq 0$, where the first sum vanishes because of the gradient, while in the second sum both factors in each term are non-negative. Conversely, when $\boldsymbol{f}_{\mathrm{NNLS}}$ solves the non-negative least-squares problem it must fulfill the Karush-Kuhn-Tucker condition, otherwise an infinitesimal change (respecting non-negativity) of a component violating it could lower χ^2. Thus, to solve the NNLS problem we just have to find a vector that fulfills the Karush-Kuhn-Tucker conditions. For this we can simply go through all possible partitionings of the indices $\{1, \ldots, N\} = \mathcal{P} \cup \mathcal{Z}$. For a given partitioning we determine the

least-squares solution on the indices in $\mathcal{P}$, i.e., we minimize $\|\boldsymbol{K}\boldsymbol{P}_\mathcal{P}\boldsymbol{f}-\boldsymbol{g}\|^2$, where $\boldsymbol{P}_\mathcal{P}$ is the projector to the space spanned by the components in $\mathcal{P}$. This makes sure that the gradients for these components vanish, $w_p=0$, while $f_z=0$. If also $f_p\geq 0$ for all $p\in\mathcal{P}$ and $w_z\geq 0$ for all $z\in\mathcal{Z}$, we have found the NNLS solution, otherwise we try the next partitioning. The only problem is that there are 2^N partitionings of the N indices (each index can be either in $\mathcal{P}$ or $\mathcal{Z}$).

A practical algorithm [14] considers possible partitionings in a much more efficient way. We start from some partition for which $\boldsymbol{f}_\mathcal{P}>0$ and $\boldsymbol{w}_\mathcal{P}=0$, e.g., an empty positive set, $\mathcal{P}=\{\}$ and $\boldsymbol{f}_{\{\}}=0$. Given a set $\mathcal{P}$ and the corresponding $\boldsymbol{f}_\mathcal{P}$ for which the Karush-Kuhn-Tucker condition is not yet fulfilled, we add the component i with the most negative gradient. Least-squares fitting on the expanded set $\mathcal{P}'{=}\mathcal{P}\cup\{i\}$ will produce an improved fit $\chi^2(\boldsymbol{f}_{\mathcal{P}'}){<}\chi^2(\boldsymbol{f}_\mathcal{P})$: because of the negative gradient, the new component will not stay at zero but rather take a positive value. In case $\boldsymbol{f}_{\mathcal{P}'}\geq 0$, we calculate the new gradient. If it is non-negative, we have found the Karush-Kuhn-Tucker solution, otherwise we repeat the procedure. Each iteration will produce a non-negative solution with improved fit, so that we will converge to the minimum of $\chi^2(\boldsymbol{f})$ under the constraint $\boldsymbol{f}\geq 0$.

In general, however, the least squares solution $\boldsymbol{f}_{\mathcal{P}'}$ will have negative components. In this case, we can find a mixing with the previous fit $\boldsymbol{f}_\alpha=(1{-}\alpha)\boldsymbol{f}_\mathcal{P}{+}\alpha\boldsymbol{f}_{\mathcal{P}'}$ with $\alpha\in(0,1)$, that brings the most negative component of $\boldsymbol{f}_{\mathcal{P}'}$ to zero. Since $\chi^2(\boldsymbol{f}_\alpha)\leq(1{-}\alpha)\chi^2(\boldsymbol{f}_{\mathcal{P}'})+\alpha\chi^2(\boldsymbol{f}_\mathcal{P})<\chi^2(\boldsymbol{f}_\mathcal{P})$ the fit will still be improved. We remove the components where $\boldsymbol{f}_\alpha$ vanishes from $\mathcal{P}'$ and perform a least-squares fit on the new set, repeating the procedure until we get a non-negative least squares solution. This must happen after a finite number of iterations, since in each step at least one element is removed from the positive set while the resulting $\boldsymbol{f}_\alpha\geq 0$ keeps improving the fit $\chi^2(\boldsymbol{f}_\alpha)<\chi^2(\boldsymbol{f}_\mathcal{P})$. Thus we can continue the outer loop with $\boldsymbol{f}_\alpha$, calculating the new gradient and adding the component where it is most negative to the positive set.

Since each step produces a $\boldsymbol{f}\geq 0$ with improved fit, the algorithm does not visit any partitioning twice and will thus always converge. At worst it may take 2^N steps, but in practice the down-hill search produces a solution after trying less than hundred partitionings. Numerically, the most delicate part is the calculation of the gradient, which should be stabilized using a factorization of the kernel. Obviously, checking the Karush-Kuhn-Tucker condition for the gradient must take the numerical accuracy into account. Moreover, the implementation may not converge, when, after adding the component with the most negative gradient, the least-squares fit gives that component a negative value. This can only happen as a consequence of numerical errors. In this case we rather include the component with the second most negative gradient in $\mathcal{P}$.

Note that the non-negative least squares solution $\boldsymbol{f}_\mathrm{NNLS}$ is unique, unlike the least-squares solution (80), to which we can add any multiple of a vector with zero singular value without changing the fit. While the least-squares problem is thus ill-posed when there are vectors that do not contribute to the fit, these vectors play a crucial role in non-negative least squares fitting: They take values such that the modes that are important for the fit can approach their optimal value as closely as possible without violating the constraint. Thus NNLS is well posed.

function NNLS($\boldsymbol{K}$, $\boldsymbol{g}$)
  $\boldsymbol{f} \leftarrow 0$
  $\mathcal{Z} \leftarrow \{1, \ldots, N\}$ ▷ below we use the abbreviation $\mathcal{P} = \{1, \ldots, N\} \setminus \mathcal{Z}$
  loop
    $\boldsymbol{w} \leftarrow \boldsymbol{K}^{\dagger}(\boldsymbol{K}\boldsymbol{f} - \boldsymbol{g})$ ▷ for robust calculation use, e.g., SVD
    if $\boldsymbol{w}[\mathcal{Z}] \geq 0$ **then return** $\boldsymbol{f}$
    end if
    $i \leftarrow \operatorname{argmin}(\boldsymbol{w}[\mathcal{Z}])$ ▷ find component with most negative gradient
    $\mathcal{Z} \leftarrow \mathcal{Z} \setminus \{i\}$
    loop
      $\boldsymbol{f}' \leftarrow LS(\boldsymbol{K}\boldsymbol{P}\boldsymbol{f}, \boldsymbol{g})$ ▷ LS solution on components $\mathcal{P}$
        ▷ $\boldsymbol{f}_i > 0$, if not: numerical error in gradient! Do not pick i again this round
      if $\boldsymbol{f}'[\mathcal{P}] > 0$ **then**
        $\boldsymbol{f} \leftarrow \boldsymbol{f}'$
        break
      end if
      $\alpha \leftarrow \min\left\{ \frac{\boldsymbol{f}_i}{\boldsymbol{f}_i - \boldsymbol{f}'_i} \middle| i \in \mathcal{P} \wedge \boldsymbol{f}'_i \geq 0 \right\}$
      $\boldsymbol{f} \leftarrow (1-\alpha)\boldsymbol{f} + \alpha\boldsymbol{f}'$ ▷ now $\boldsymbol{f} \geq 0$ and $\boldsymbol{f}_i = 0$ for $i = \operatorname{argmin}$
      $\mathcal{P} \leftarrow \mathcal{P} \setminus \{i | \boldsymbol{f}_i = 0\}$
    end loop
  end loop
end function

Fig. 10: *Function that returns the non-negative least-squares solution $\boldsymbol{f} \geq 0$ of $\boldsymbol{g} = \boldsymbol{K}\boldsymbol{f}$.*

A.3 Shannon entropy

When developing *The Mathematical Theory of Communication*, Claude Shannon introduced the bit as the amount of information needed to decide between two equally probable events [15]. Receiving an unlikely (surprising) message should convey more information than receiving a likely one, and the information contained in two independent messages should be the sum of the information carried by each individually. These axioms lead to $-\log_2 p_i$ as the information contained in receiving a message of probability p_i. Summing over a set of M possible messages of probabilities p_i and weighting the information contained in them by their probability defines the average information or entropy of an information source

$$H(\{p_i\}) = -\sum p_i \log_2 p_i \,. \tag{84}$$

It gives a lower limit to the number of bits needed for encoding the N messages. The maximum number of bits, $\log_2 N$, is needed when we know least about which message to expect, i.e., when all probabilities are the same. In the opposite limit, when one of the messages is certain, we need not encode it at all. Thus the entropy of an information source measures our ignorance before receiving one of the possible messages.

Changing the base of the logarithm, $\log_b(x) = \log_2(x)/\log_2(b)$, for $b > 1$ simply multiplies the entropy by a positive constant, i.e., changes the units in which we measure information. For convenience, we use the natural logarithm, $\ln$, working in natural units, 1 nat $\approx$ 1.44 bits.

An alternative derivation [2] of (84) starts by considering microstates representing the probabilities as $p_i = n_i/M$ by placing N (distinguishable) objects into M bins. Since we can place any of the N objects into any of the M bins, there are M^N such states. The number of different ways of placing the objects into bins and obtaining the same set of $\{n_i\}$, i.e., the same macrostate, is also easily determined: We can pick any n_1 of the N objects and put them into the first bin. Then we can pick any n_2 of the remaining $N - n_1$ objects and put them into the second bin. The probability of realizing a macrostate $\{n_1, \dots, n_M\}$ is thus

$$\frac{1}{M^N}\binom{N}{n_1}\binom{N-n_1}{n_2}\binom{N-n_1-n_2}{n_3}\cdots\binom{n_M}{n_M} = \frac{1}{M^N}\,\frac{N!}{n_1!\,n_2!\cdots n_M!}\,.$$

Taking the logarithm and using the Stirling approximation $\ln n! \approx n\ln n - n$ we find

$$-N\ln M + N\ln N - N - \sum_{i=1}^{M}\big(n_i\ln n_i - n_i\big) = N\left(\ln\frac{1}{M} - \sum_i p_i \ln p_i\right),$$

which is proportional to the $H(\{p_i\})$ minus the entropy of a flat distribution $\{1/M\}$.
Subtracting the entropy of the flat distribution becomes crucial when taking the limit of a continuous probability distribution: encoding an infinite number of messages will, in general, take an infinite number of bits. Subtracting $-\log 1/M$ keeps the limit $M \to \infty$ finite. To see this, we discretize a continuous distribution $p(x)$ on an equidistant grid of M points, $p_i = p(x_i)\Delta x$ with $\Delta x = (x_{\max} - x_{\min})/M = (\int dx)/M$, and take the limit of the Riemann sum

$$-\sum_{i=1}^{M} p_i \ln\left(\frac{p_i}{1/M}\right) = -\sum_{i=1}^{M}\Delta x\, p(x_i)\ln\left(\frac{p(x_i)\Delta x}{1/M}\right) \to -\int dx\, p(x)\ln\left(\frac{p(x)}{1/\int dx}\right).$$

This defines the entropy of a distribution $p(x)$

$$H[p] = -\int dx\, p(x)\ln\left(\frac{p(x)}{1/\int dx}\right). \tag{85}$$

We can find the $p(x)$ that maximizes this functional from the variational principle. Remembering that the functional derivatives are defined by the expansion

$$H[p+\delta p] = H[p] + \int dx\,\frac{\delta H[p]}{\delta p(x)}\,\delta p(x) + \frac{1}{2}\int dx\,dx'\,\frac{\delta^2 H[p]}{\delta p(x')\delta p(x)}\,\delta p(x')\,\delta p(x) + \mathcal{O}^3(\delta p)$$

we read off the first variation from

$$H[p+\delta p] = -\int dx\,(p+\delta p)\big(\underbrace{\ln(p+\delta p)}_{=\ln p+\ln(1+\frac{\delta p}{p})=\ln p+\frac{\delta p}{p}+\mathcal{O}^2} + \ln\textstyle\int dx\big) = H[p] - \int dx\,\underbrace{\left(1+\ln\frac{p(x)}{1/\int dx}\right)}_{=-\frac{\delta H[p]}{\delta p(x)}}\delta p + \mathcal{O}^2,$$

where we have used $\ln(1+x) = x - x^2/2 + \cdots$. In second order we find $\delta^2 H[p]/\delta p(x')\,\delta p(x) = -2\delta(x-x')/p(x) \le 0$ so that the stationary points are maxima. Imposing normalization of the 0-th moment via a Lagrange parameter, the variational equation becomes

$$0 = \frac{\delta}{\delta p(x)}\,H[p] + \lambda_0\big(1 - \textstyle\int dx\, p(x)\big) = -1 - \ln p(x) - \ln\int dx - \lambda_0$$

which is solved by the constant distribution, where $\lambda_0 = -1$ is fixed by normalization

$$p(x) = \frac{1}{\int dx}\, \mathrm{e}^{-(1+\lambda_0)} = \frac{1}{\int dx}\,. \tag{86}$$

Inserting into (85) we find $H[1/\int dx] = 0$, as it must by construction.
Likewise, we can ask which distribution maximizes the entropy, when we know in addition its first moment $\mu = \int dx\, x\, p(x)$. The variational equation then contains two Lagrange parameters

$$0 = \frac{\delta}{\delta p(x)}\, H[p] + \lambda_0\big(1 - \int dx\, p(x)\big) + \lambda_1\big(\mu - \int dx\, x\, p(x)\big) = -1 - \ln \frac{p(x)}{1\int dx} - \lambda_0 - \lambda_1\, x$$

and we obtain a Boltzmann distribution

$$p_\mu(x) = \frac{1}{\int dx}\, \mathrm{e}^{-(1+\lambda_0+\lambda_1 x)}, \tag{87}$$

where λ_0 and λ_1 are fixed by solving the system $\int dx\, p_\mu(x) = 1$ and $\int dx\, x\, p_\mu(x) = \mu$. Likewise, when we also know the variance of $p(x)$, maximizing the entropy results in a Gaussian.
Given the entropy functional, it is natural to ask what happens under a change of variable. Remembering that density functions transform as $p(x)\, dx = p(z)\, dz$, we find

$$H[p] = -\int \frac{dx}{dz}\, dz\;\; p(z)\frac{dz}{dx}\;\; \ln\left(\frac{p(z)\frac{dz}{dx}}{1/\int dx}\right) = -\int dz\, p(z) \ln\left(\frac{p(z)}{\rho(z)}\right),$$

where we introduced $\rho(z)\, dz = dx/\int dx$. It reflects how the intervals on x change under the under transformation to z. When we define $\rho(x) = 1/\int dx$, we see that the form of the entropy functional is invariant under coordinate transformations

$$H[p\,|\,\rho] = -\int dx\, p(x) \ln \frac{p(x)}{\rho(x)}\,. \tag{88}$$

This is the relative entropy or Kullback-Leibler divergence. From $\ln x \leq x-1$ it follows that $H[p] \leq 0$. By construction, the maximum is attained for $p(x) = \rho(x)$. The relative entropy describes the average information contained in the distribution $p(x)$ when what we expected was the distribution $\rho(x)$. The prior $\rho(x)$ plays the role of a density of states: from the functional derivative of the relative entropy $\delta H[p]/\delta p(x) = -1 - \ln(p(x)/\rho(x))$ we see that the solutions of the variational equations for $p(x)$ derived above become proportional to $\rho(x)$.
For convenience we might want to allow non-normalized densities of states $\tilde\rho(x)$ and correspondingly drop the normalization constraint for $\tilde p(x)$. If we write [16]

$$\tilde H[\tilde p\,|\,\tilde\rho] = \int dx \left(\tilde p(x) - \tilde\rho(x) - \tilde p(x) \ln \frac{\tilde p(x)}{\tilde\rho(x)}\right) \tag{89}$$

we obtain from the variational equation $\delta\tilde H/\delta\tilde p = 0$ (without normalization constraint) the solution $\tilde p(x) = \tilde\rho(x)$ with $\tilde H[\tilde\rho\,|\,\tilde\rho] = 0$ as for (88).
We note that a flat prior $\rho(x) = \mathrm{const.}$ is bound to the choice of variable: Given any $\rho(z)$ we can always transform to $x(z) \propto \int^z \rho(z')\, dz'$ to obtain $\rho(x) = \rho(z)\, dz/dx = \mathrm{const.}$, where x must be restricted to a finite interval to be normalizable.

A.4 Sampling from a truncated normal distribution

It is straightforward to generate a random variable x with a normal probability distribution

$$p_n(x) = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-x^2/2} \tag{90}$$

using, e.g., the Box-Muller method [17]. When taking a constraint into account, we need, however, variables with a normal distribution restricted to some interval, $x \in [a, b]$.

$\boldsymbol{x \geq a}$: When x is restricted to be larger than some value a, a straightforward approach is to sample normally distributed values x until we find an $x > a$. The probability for finding such an x is, on average, just the integral over the Gaussian

$$\bar{P}_n(a) = \int_a^\infty dx\, p_u(x) = I(a)\,. \tag{91}$$

This is easily written in terms of the complementary error function $\mathrm{erfc}(z) = 2/\sqrt{\pi}\int_z^\infty dt\,\mathrm{e}^{-t^2}$. For $a > 0$

$$I(a \geq 0) = \frac{1}{2}\,\mathrm{erfc}(a/\sqrt{2})\,, \tag{92}$$

while for $a < 0$

$$I(a \leq 0) = 1 - I(-a)\,. \tag{93}$$

The average acceptance probability (91) is shown in figure 11. For a to the left of the peak of the normal distribution it is very likely that a proposed random variable x is larger than a and thus is accepted. For $a > 0$ this probability is, however, rapidly decreasing to zero, meaning that we would have to propose very many normally distributed variables x until we find one that is larger than a. This is very inefficient, so for $a > 0$ we need a better approach. Following [18], we generate random variables $x \geq a$ with an exponential probability distribution

$$p_{\mathrm{exp}}(x) = \alpha\,\mathrm{e}^{-\alpha(x-a)}\,. \tag{94}$$

These are easily obtained as $x = a - \ln(u/\alpha)/\alpha$ from uniformly distributed random numbers $u \in [0, 1)$. To transform these exponentially distributed random numbers $x \geq a$ into the desired normally distributed random numbers we use the rejection method [17], accepting x with probability proportional to the ratio $p_n(x)/p_{\mathrm{exp}}(x)$ of the desired and the proposed probability distribution functions. To obtain a probability, we introduce a prefactor to make sure that for no $x \geq a$ the ratio exceed one. Completing the square in the exponential we find

$$p_{\mathrm{acc}}(x; A) = \frac{1}{A}\,\frac{\mathrm{e}^{-x^2/2}}{\mathrm{e}^{-\alpha(x-a)}} = \frac{1}{A}\,\underbrace{\mathrm{e}^{-(x-\alpha)^2/2}}_{\leq 1}\,\mathrm{e}^{\alpha^2/2-\alpha a} \overset{!}{\leq} 1\,. \tag{95}$$

For $x \geq a$ the choice $A = \mathrm{e}^{\alpha^2/2-\alpha a}$ maximizes the acceptance probability, which becomes

$$p_{\mathrm{acc}}(x) = \mathrm{e}^{-(x-\alpha)^2/2}\,. \tag{96}$$

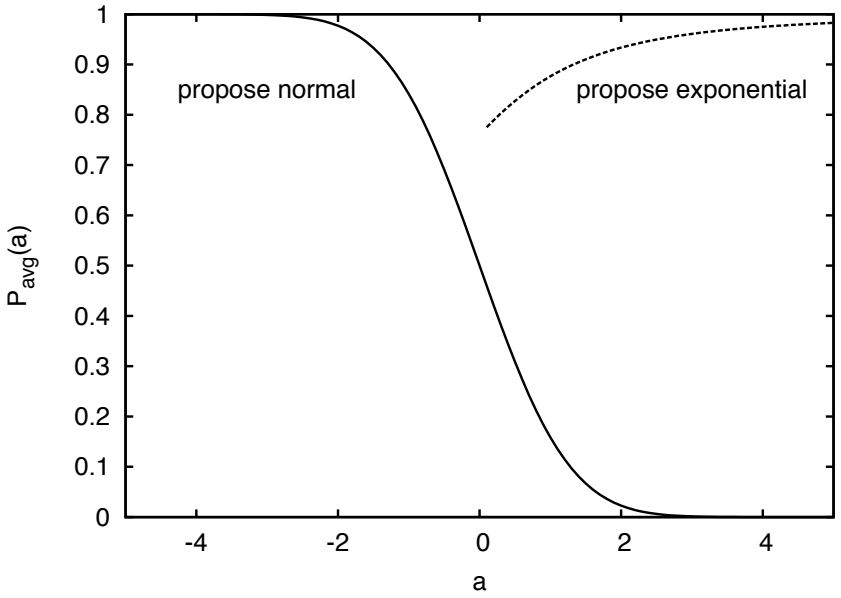


Fig. 11: *Efficiency of the methods for sampling from a normally distributed variable x restricted to the interval $[a, \infty)$: For $a < 0$ the average acceptance probability for an unrestricted normally distributed random variable ($\bar{P}_n(a)$, full line) is larger than $1/2$, while for $a > 0$ it rapidly approaches zero. For positive a we therefore propose exponentially distributed random numbers ($\bar{P}_{\exp}(a \geq 0)$, dotted line), for which the average acceptance probability is at least $\sqrt{\pi/2\mathrm{e}} \approx 0.76$.*

The corresponding average acceptance probability is then the integral of the product of the probability for proposing a value x times the probability for accepting it

$$\bar{P}_{\exp}(a \geq 0) = \int_a^\infty dx\, p_{\exp}(x)\, p_{\mathrm{acc}}(x) = \sqrt{2\pi}\, \alpha\, \mathrm{e}^{-\alpha^2/2+\alpha a}\, I(a) \tag{97}$$

In this expression α is still a free parameter, which we choose to maximize the average acceptance. Solving the variational equation we obtain

$$\alpha = \frac{a + \sqrt{a^2+4}}{2}\,. \tag{98}$$

We note that for $a \geq 0$, $\bar{P}_{\exp}(a)$ has the same form as (91), differing, however, by a prefactor $\gamma_{\exp} = \sqrt{2\pi}\, \alpha\, \mathrm{e}^{-\alpha^2/2+\alpha a}$ which grows faster than the complementary error function decays. Therefore, as can be seen from Fig. 11, for $a > 0$ this method is dramatically more efficient than sampling from an unbounded uniform distribution. Thus for $a < 0$ we choose the method with $\gamma_n = 1$, while for $a \geq 0$ we choose $\gamma_{\exp}$, obtaining an average acceptance probability $\bar{P}_\gamma(a) = \gamma(a)\, I(a)$.

$\boldsymbol{x \geq b}$: When we need to sample a normal variable constrained from above we can use the same methods as above, sampling $-x \geq -b$, with average acceptance probability $\bar{P}_\gamma(-b)$.

$\boldsymbol{a \leq x \leq b}$: When the random variable is constrained to a finite interval, an obvious approach is to first sample an $x \in [a, \infty)$ and to accept it if $x \leq b$. The average acceptance probability is then $\bar{P}_\gamma(a)$ for proposing an $x \in [a, \infty)$, times $\int_a^b dx\, p_n(x) / \int_a^\infty dx\, p_n(x)$ for accepting it, i.e.,

$$\bar{P}_\gamma(a, b) = \gamma\big(I(a) - I(b)\big)\,. \tag{99}$$

For large intervals this will be an efficient approach, while for narrow intervals the acceptance will go to zero. In this case it becomes more efficient to propose x uniformly distributed on $[a,b]$ and accept them with probability

$$p_{\rm acc} = \mathrm{e}^{-(x^2-m^2)/2}\,, \tag{100}$$

where m is the coordinate at which the normal distribution assumes its maximum value on $[a,b]$, ensuring that $p_{\rm acc} \le 1$. When $0 \in [a,b]$ then $m=0$, otherwise $m = \min(|a|,|b|)$. The average acceptance probability for this approach is

$$\bar{P}_u(a,b) = \int_a^b dx\, \frac{1}{b-a}\,\mathrm{e}^{-(x^2-m^2)/2} = \frac{\mathrm{e}^{m^2/2}}{b-a}\,\sqrt{2\pi}\big(I(a)-I(b)\big). \tag{101}$$

Again, (101) differs from (99) only by its prefactor γ_u, which increases as the width of the interval $b-a$ becomes smaller.
For a given interval $[a,b]$ we then choose the most efficient method:

- $a<0<b$: Since $a<0$ we have to choose between normal sampling with $\gamma_n = 1$ and uniform sampling with $\gamma_u = \sqrt{2\pi}/(b-a)$. For $\gamma_n = \gamma_u$ both methods have the same average acceptance probability. Solving this gives us the critical width $w_0 = \sqrt{2\pi}$. For intervals $b-a<w_0$ we thus use uniform sampling with γ_u, otherwise γ_n. The worst case is $\bar{P}_{\gamma=1}(0,w_0) = I(0)-I(w_0) = \mathrm{erf}(\sqrt{\pi})/2 \approx 0.49$.

- $0<a<b$: Since $a>0$ we choose between exponential sampling with $\gamma_{\rm exp} = \sqrt{2\pi}\,\alpha\,\mathrm{e}^{-\alpha^2/2+\alpha a}$ and $\gamma_u = \sqrt{2\pi}\,\mathrm{e}^{a^2/2}/(b-a)$. Solving $\gamma_{\rm exp} = \gamma_u$ gives the critical width $w_>(a) = e^{(\alpha-a)^2/2}/\alpha$. For intervals $b-a<w_>(a)$ we use uniform sampling with γ_u, otherwise $\gamma_{\rm exp}$. The worst case is $\bar{P}_\gamma(0,w_>(0)) = \gamma_{\rm exp}\big(I(0)-I(\sqrt{e})\big) = \sqrt{\pi/2e}\,\mathrm{erf}(\sqrt{\mathrm{e}/2}) \approx 0.68$.

- $a<b<0$: We sample $-x$ in the interval from $-b$ to $-a$ as described above.

Overall, we can thus sample from a truncated normal distribution with an average acceptance larger than $\mathrm{erf}(\sqrt{\pi})/2 \approx 0.49$.
The generalization to sampling from a Gaussian distribution $\exp(-(x-\mu)^2/2\sigma^2)/\sqrt{2\pi}\sigma$ of variance σ centered at μ restricted to $x\in[a,b]$ is then straightforward: use $x{=}\sigma\tilde{x}{+}\mu$, where $\tilde{x}$ is sampled, as described above, from a normal distribution (90) on the interval $[(a{-}\mu)/\sigma, (b{-}\mu)/\sigma]$.

References

[1] J. Hadamard, Princeton Univ. Bull. **13**, 49–52 (1902): ...ces problèmes se présentait comme parfaitement bien posé, je veux dire comme *possible* et *déterminé*.

[2] D.S. Sivia and J. Skilling: *Data Analysis: A Bayesian Tutorial* (Oxford Univ. Press, 2006)

[3] A. Flesch, E. Gorelov, E. Koch, and E. Pavarini, Phys. Rev. B **87**, 195141 (2013)

[4] A. Erdélyi (ed.): *Higher Transcendental Functions* (McGraw-Hill, New York, 1953)

[5] L. Boehnke, H. Hafermann, M. Ferrero, F. Lechermann, and O. Parcollet, Phys. Rev. B **84**, 075145 (2011)

[6] E. Koch: *The Lanczos Method* in E. Pavarini, E. Koch, D. Vollhardt, A. Lichtenstein (eds.): *The LDA+DMFT approach to strongly correlated materials* Modeling and Simulation Vol. 1 (Forschungszentrum Jülich, 2011)

[7] J. Waldvogel inn: W. Gautschi, G. Mastroianni, T. Rassias (eds): *Approximation and Computation*, Springer Optimization and Its Applications, Vol. 42 (Springer, New York, 2010) p. 267

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

[9] M. Jarrell: *The Maximum Entropy Method: Analytic Continuation of QMC Data* in E. Pavarini, E. Koch, F. Anders, and M. Jarrell (eds.): *Correlated Electrons: From Models to Materials* Modeling and Simulation Vol. 2 (Forschungszentrum Jülich, 2012)

[10] S.R. White in D.P. Landau, K.K. Mon, and B.B. Schüttler: *Computer Simulation Studies in Condensed Matter Physics III* (Springer, 1991) p. 145–153

[11] K. Ghanem: *Stochastic Analytic Continuation: A Bayesian Approach* PhD Thesis (RWTH Aachen, 2017)

[12] K. Ghanem and E. Koch, Phys. Rev. B **101**, 085111 (2020); *ibid.* **102**, 035114 (2020)

[13] H. Flyvbjerg and H.G. Petersen, J. Chem. Phys. **91**, 461(1989)

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

[15] C. Shannon, Bell Syst. Tech. J. **27**, 379 and 623 (1948)

[16] J. Skilling, in J. Skilling (ed.): *Maximum Entropy and Bayesian Methods* (Kluwer, Dordrecht, 1989) p. 45–52

[17] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery: *Numerical Recipes* (Cambridge University Press, 2007)

[18] C.P. Robert Statistics and Computing **5**, 121 (1995)