Solving the Bethe-Salpeter equation with exponential convergence

Markus Wallerberger,[1,*] Hiroshi Shinaoka,[2,*] and Anna Kauch[1]
[1]*Institute of Solid State Physics, TU Wien, 1040 Vienna, Austria*
[2]*Department of Physics, Saitama University, Saitama 338-8570, Japan*

(Received 10 December 2020; accepted 23 July 2021; published 20 August 2021)

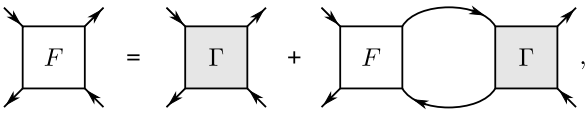
The Bethe-Salpeter equation plays a crucial role in understanding the physics of correlated fermions, relating to optical excitations in solids as well as resonances in high-energy physics. Yet, it is notoriously difficult to control numerically, typically requiring an effort that scales polynomially with energy scales and accuracy. This puts many interesting systems out of computational reach. Using the intermediate representation and sparse modeling for two-particle objects on the Matsubara axis, we develop an algorithm that solves the Bethe-Salpeter equation in $\mathcal{O}(L^8)$ time with $\mathcal{O}(L^4)$ memory, where L grows only logarithmically with inverse temperature, bandwidth, and desired accuracy. This opens the door for computations in hitherto inaccessible regimes. We benchmark the method on the Hubbard atom and on the multiorbital weak-coupling limit, where we observe the expected exponential convergence to the analytical results. We then showcase the method for a realistic impurity problem.

DOI: 10.1103/PhysRevResearch.3.033168

I. INTRODUCTION

The rapid increase in available computational resources has propelled material calculations into a new era. Computational methods based on the density functional theory are widely used for material design. One of the remaining grand challenges is computing dynamical response determining the functionalities of materials, especially those with strong electronic correlations. In theory, excitons, magnons, and other composite excitations can only be described by two-particle correlation functions (optical conductivity, susceptibilities). At the core of calculating these lie equations at the two-particle level, in particular the Bethe-Salpeter equation [1,2].

The Bethe-Salpeter equation (BSE) is a basic computational tool in a multitude of methods: from *ab initio* approaches [3–8] through many-body methods in condensed matter [9–21] but also high-energy [22,23] and nuclear physics [24,25]. It relates the sum F of all two-particle scattering channels to a smaller irreducible set of diagrams, Γ:



(1)

where F and Γ depend on three frequencies and four orbital indices [26] (cf. Sec. III B for details).

Solving the BSE is notoriously difficult since it scales unfavorably in the number of orbitals and frequencies. To reduce the computational effort most *ab initio* calculations at the two-particle level are performed with the static approximation, where only the zero-frequency component is taken [3,4,27–30], or with a reduced frequency dependence [6–8]. The fully dynamical calculations [10–14,31,32] are quickly stopped by the curse of dimensionality at only few orbitals or high temperatures. This renders many interesting regimes such as low-temperature phases of materials inaccessible.

Fully dynamical calculations mostly use a dense box of Matsubara frequencies. Solving the BSE then requires $\mathcal{O}(N^3)$ memory and, even worse, $\mathcal{O}(N^4)$ computational time, where N is the number of frequencies along each side of the box (see Fig. 1). The naïve truncation error decays as $\mathcal{O}(1/N)$. These exponents are problematic since N is proportional to the bandwidth $\omega_{\max}$ and to the inverse temperature $\beta = 1/T$ with a large prefactor due to substantial finite-size effects of the box. Chemical accuracy is then usually far out of reach and error control becomes difficult.

In this paper we offer a substantial step towards numerically low-cost evaluation of the BSE. Based on the previously developed concept of sparse modeling of the two-particle correlation functions [33,34], we directly solve the BSE in the sparse representation (intermediate representation, IR [35]). As illustrated in Fig. 1, the polynomial scaling of the maximal (red curves) and average (black curves) relative error with the linear system size (linear frequency box size N) is replaced by faster than exponential decay with the size of the IR representation L. Moreover, L grows only logarithmically with $\beta\omega_{\max}$ [36]. With this method we achieve not simply an improvement in scaling, but a paradigm shift: from brute force high-performance computing to data compression and systematic error control. The concept of sparse modeling is already widely adopted for *ab initio* calculations at the one-particle level [37–43]. Here, we pave the way to taking dynamical *ab initio* calculations with sparse modeling to the two-particle level.

[*]These authors contributed equally to this work.

(a)	Linear data size	Truncation error	Runtime scaling	Memory scaling
Sparse	$L = \mathcal{O}(\log(\beta\omega_{\max}))$	$\mathcal{O}(e^{-\alpha L})$	$N_{\text{orb}}^6 L^8$	$N_{\text{orb}}^4 L^4$
Dense	$N = \mathcal{O}(\beta\omega_{\max})$	$\mathcal{O}(N^{-1-\gamma})$	$N_{\text{orb}}^6 N^4$	$N_{\text{orb}}^4 N^3$

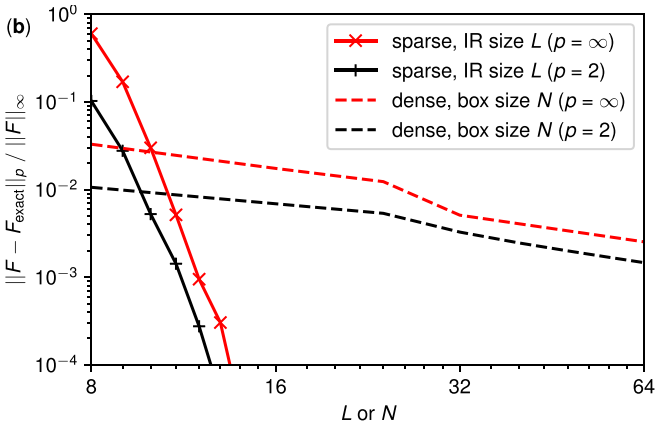


FIG. 1. Solving the BSE using sparse modeling with a cutoff L (this paper) compared with using a dense frequency box of linear size N. (a) Nominal truncation error and runtime/memory scaling, where N_{orb} is the number of orbitals, $\beta\omega_{\max}$ is the bandwidth in units of temperature, and γ is the power in inverse frequencies to which the asymptotics is known analytically. (b) Actual scaling of the error for the Hubbard atom (cf. Sec. VI A), where black and red lines are the average ($p = 2$) and maximum ($p = \infty$) deviation from the exact result, respectively.

In order to keep the paper self-contained and introduce the necessary concepts and notation, we review sparse modeling in Sec. II. In Sec. III, we will then show that these ideas carry over to diagrammatic equations such as the Bethe-Salpeter equation, which we rewrite in the compressed form. In Sec. IV, we then develop a method to also compress all intermediate results (summation over inner propagators) needed in solving the BSE (sparse convolution). The final step needed to obtain the solution is to solve the resulting least-squares problem, which we present in Sec. V. We then benchmark our method on two challenging systems: in Sec. VI, we compare sparse modeling to the analytic results for the Hubbard atom as well as for the weak-coupling limit of a multiorbital impurity. In Sec. VII, we then show the results for a realistic impurity problem. We conclude with a summary and outlook in Sec. VIII.

II. REVIEW: SPARSE MODELLING

We review sparse modeling before its use in solving the BSE. The concept of sparse modeling is based on a generic feature of imaginary frequency data: In transitioning numerical data from real to imaginary (Matsubara) frequencies, we lose information as features of correlation functions get smeared out. That information loss makes it difficult to reconstruct the original real-frequency signal, but at the same time allows imaginary frequency data to be compressed extremely efficiently [35,36,44].

The essence of this approach is twofold: (i) the truncated intermediate representation (IR) provides an efficient basis for representing Matsubara Green's functions and (ii) the basis coefficients can be inferred from Matsubara data at special sampling frequencies in a quick and stable fashion. In the following, we review the use of IR in representing one-particle and two-particle response functions, focusing specifically on the Green's function. For an extended review on the IR, we refer the reader to Refs. [45,46].

A. One-particle response functions

A one-particle response function can be written as [47]

$$\hat{G}^\alpha(\mathrm{i}\omega) = \int_0^\beta d\tau\, e^{\mathrm{i}\omega\tau} \langle T_\tau A^\alpha(\tau) B^\alpha(0)\rangle, \tag{2}$$

where τ denotes imaginary (Euclidean) time, T_τ indicates time ordering, $\alpha \in \{\mathrm{B}, \mathrm{F}\}$ denotes fermionic or bosonic statistics, A and B are operators, and $\mathrm{i}\omega$ is a fermionic (bosonic) Matsubara frequency for $\alpha = \mathrm{F}$ ($\alpha = \mathrm{B}$). Equation (2) and the associated spectral function $\rho(\omega)$ in real frequency ω are connected through analytic continuation [47]:

$$\hat{G}^\alpha(\mathrm{i}\omega) = \int_{-\omega_{\max}}^{\omega_{\max}} d\omega' \underbrace{\frac{\omega'^{\delta_{\alpha,\mathrm{B}}}}{\mathrm{i}\omega - \omega'}}_{K^\alpha(\mathrm{i}\omega,\omega')} \rho^\alpha(\omega'), \tag{3}$$

where we require the spectral function to be of the form $\rho(\omega) = \sum_i \rho_i \delta(\omega - \omega_i)$ with $\omega_i \in [-\omega_{\max}, \omega_{\max}]$. (Note that the bosonic kernel has been regularized using an additional factor ω', consistent with analytic continuation practice.)

The kernel in a Fredholm integral equation of the first kind, K^α admits a singular value expansion [48]:

$$K^\alpha(\mathrm{i}\omega, \omega') = \sum_{l=0}^{\infty} \hat{U}_l^\alpha(\mathrm{i}\omega) S_l^\alpha V_l^\alpha(\omega'), \tag{4}$$

where S_l^α are the singular values in strictly decreasing order, $S_0 > S_1 > \cdots > 0$, $\hat{U}_l^\alpha$ are the left singular functions, which form an orthonormal set on the Matsubara frequencies, and V_l^α are the right singular functions, which form an orthonormal set on the real frequencies [49]. The singular functions $\hat{U}_l^\alpha$ and V_l^α are the so-called IR basis functions [35].

We can use the left singular functions of the kernel as representation for the Green's function [35]:

$$\hat{G}^\alpha(\mathrm{i}\omega) = \sum_{l=0}^{L-1} \hat{U}_l^\alpha(\mathrm{i}\omega) G_l^\alpha + \epsilon_L, \tag{5}$$

where $G_l^\alpha = S_l \sum_i \rho_i V_l^\alpha(\omega_i)$ is a basis coefficient and ϵ_L is an error term associated with truncating the series. One can prove that S_l drop faster than any power with l [50], and there is numerical evidence for faster-than-exponential decay [36]. One empirically finds logarithmic growth of singular values with respect to increasing the energy cutoff [36], $S_L/S_0 = \mathcal{O}(\log(\beta\omega_{\max}))$, and also that the right singular functions are bounded, i.e., there exists a $V_{\max}$ such that $|V_l(\omega)| < V_{\max}$. This implies that the truncated representation (5) converges faster than exponentially $\epsilon_L = \mathcal{O}(S_L)$ and is substantially more compact than a polynomial representation.

In order to efficiently extract the basis coefficients G_l from $\hat{G}(\mathrm{i}\omega)$, we exploit the fact that the roots of the singular functions are similar in structure to the roots of orthogonal

polynomials [51]. Hence, we choose a set of sampling frequencies close to the sign changes of $\hat{U}_L(\mathrm{i}\omega)$:

$$\mathcal{W}^\alpha = \{\mathrm{i}\omega_1^\alpha, \mathrm{i}\omega_2^\alpha, \ldots, \mathrm{i}\omega_L^\alpha\}, \tag{6}$$

and turn Eq. (5) into an ordinary least-squares fit [37]:

$$G_l^\alpha = \arg\min_{G_l} \sum_{\mathrm{i}\omega\in\mathcal{W}^\alpha} \left| \hat{G}^\alpha(\mathrm{i}\omega) - \sum_{l=0}^{L-1} \hat{U}_l^\alpha(\mathrm{i}\omega) G_l^\alpha \right|^2. \tag{7}$$

An example of sampling frequencies is shown later in Fig. 4.

One empirically observes that the $|\mathcal{W}| \times L$ fitting matrix $U_{nl} = \hat{U}_l^\alpha(\mathrm{i}\omega_n^\alpha)$ is well-conditioned provided that $\mathcal{W}$ was chosen as in Eq. (6) [37]. This in turn implies the fitting error is consistent with the overall truncation error ϵ_L. Similar rules can be found for imaginary time.

B. Two-particle response functions

We now turn to the two-particle Green's function:

$$\begin{aligned}\hat{G}&(\mathrm{i}\nu_1, \mathrm{i}\nu_2, \mathrm{i}\nu_3, \mathrm{i}\nu_4) \\ &= \int_0^\beta d\tau_1\, d\tau_2\, d\tau_3\, d\tau_4 \\ &\quad \times \mathrm{e}^{\mathrm{i}\nu_1\tau_1 - \mathrm{i}\nu_2\tau_2 + \mathrm{i}\nu_3\tau_3 - \mathrm{i}\nu_4\tau_4} \langle T_\tau A(\tau_1) B(\tau_2) C(\tau_3) D(\tau_4)\rangle, \end{aligned} \tag{8}$$

where $A, \ldots, D$ are now fermionic operators and, consequently, $\mathrm{i}\nu_1$ to $\mathrm{i}\nu_4$ are fermionic Matsubara frequencies.

The Lehmann representation of Eq. (8) can be cast in the following form [33]:

$$\begin{aligned}\hat{G}&(\mathrm{i}\nu_1, \mathrm{i}\nu_2, \mathrm{i}\nu_3, \mathrm{i}\nu_4) \\ &= \sum_{r=1}^{12} \sum_{\nu\nu'\omega} T_r(\mathrm{i}\nu_1, \ldots, \mathrm{i}\nu_4; \mathrm{i}\nu, \mathrm{i}\nu', \mathrm{i}\omega) \\ &\quad \times \int d^3\omega K^{\mathrm{F}}(\mathrm{i}\nu, \omega_1) K^{\mathrm{F}}(\mathrm{i}\nu', \omega_2) K^{\bar{\mathrm{B}}}(\mathrm{i}\omega, \omega_3) \\ &\quad \times \rho_r(\omega_1, \omega_2, \omega_3), \end{aligned} \tag{9}$$

where $\mathrm{i}\nu, \mathrm{i}\nu'$ are fermionic Matsubara frequencies, $\mathrm{i}\omega$ is a bosonic Matsubara frequency, $\omega_1, \omega_2, \omega_3$ are real frequencies, and T_r is a frequency translation tensor defined in Table I. This translation is necessary in order to have a spectral function of the form

$$\rho_r(\omega_1, \omega_2, \omega_3) = \sum_{ijk} A_{ijk}^{(r)} \delta(\omega_1 - \omega_i)\delta(\omega_2 - \omega_j)\delta(\omega_3 - \omega_k), \tag{10}$$

because the two-particle Green's function cannot be made compact in any single frequency convention due to permutations induced by time ordering in Eq. (8). Therefore, we require the sum over $r = 1, \ldots, 12$ different representations in Eq. (9), unlike Eq. (3), where only one representation is sufficient.

Entering Eq. (9) is a product of the one-particle kernels K^{F} and K^{B} from Eq. (3). However, K^{B} must be augmented in order to ensure proper decay of the expansion coefficients:

$$K^{\bar{\mathrm{B}}}(\mathrm{i}\omega, \omega') = \frac{\omega'}{\mathrm{i}\omega - \omega'} + S_0^{\bar{\mathrm{B}}}\delta_{\mathrm{i}\omega,0} + \frac{S_1^{\bar{\mathrm{B}}}(1 - \delta_{\mathrm{i}\omega,0})}{\mathrm{i}\omega}, \tag{11}$$

where $S_0^{\bar{\mathrm{B}}}, S_1^{\bar{\mathrm{B}}}$ are arbitrary prefactors (we shift the remaining singular values by $S_l^{\mathrm{B}} \to S_{l+2}^{\bar{\mathrm{B}}}$). A δ function at $\mathrm{i}\omega = 0$ is not included in the unaugmented kernel (3) as it cannot be resolved by it; however, terms like those are indeed present in the two-particle Green's function.

Since the one-particle kernels can be truncated, Eq. (9) implies that there also exists a truncated, compact representation for the two-particle Green's function analogous to Eq. (5):

$$\begin{aligned}\hat{G}(\mathrm{i}\nu_1, \ldots, \mathrm{i}\nu_4) &= \sum_{r=1}^{12} \sum_{\nu\nu'\omega} T_r(\mathrm{i}\nu_1, \ldots, \mathrm{i}\nu_4; \mathrm{i}\nu, \mathrm{i}\nu', \mathrm{i}\omega) \\ &\quad \times \sum_{l,l',m=0}^{L-1} \hat{U}_l^{\mathrm{F}}(\mathrm{i}\nu)\hat{U}_{l'}^{\mathrm{F}}(\mathrm{i}\nu')\hat{U}_m^{\bar{\mathrm{B}}}(\mathrm{i}\omega) G_{r,ll'm} + \epsilon_L, \end{aligned} \tag{12}$$

where $\hat{U}_l^{\mathrm{F}}(\mathrm{i}\nu)$ is defined in Eq. (4) and $\hat{U}_l^{\bar{\mathrm{B}}}$ are the singular functions of the bosonic kernel, augmented by the additional contributions in Eq. (11). Since $G_{r,ll'm}$ is then given by projection of Eq. (10):

$$G_{r,ll'm} = S_l^{\mathrm{F}} S_{l'}^{\mathrm{F}} S_m^{\bar{\mathrm{B}}} \sum_{ijk} A_{ijk}^{(r)} V_l^{\mathrm{F}}(\omega_i) V_{l'}^{\mathrm{F}}(\omega_j) V_m^{\bar{\mathrm{B}}}(\omega_k), \tag{13}$$

one again has faster than exponential convergence in Eq. (12), $\epsilon_L = \mathcal{O}(S_L)$. Storing G in the IR basis thus only requires $12L^3$ numbers, where L is $\mathcal{O}(\log(\beta\omega_{\max}\epsilon^{-1}))$ and ϵ is the desired accuracy.

Extracting $G_{r,ll'm}$ from G requires us to construct a fitting problem. We choose the set of sampling frequencies as

$$\mathcal{W} = \bigcup_{r=1}^{12} \{(\mathrm{i}\nu_1, \ldots, \mathrm{i}\nu_4) : (\mathrm{i}\nu, \mathrm{i}\nu', \mathrm{i}\omega) \in \mathcal{W}^{\mathrm{F}} \times \mathcal{W}^{\mathrm{F}} \times \mathcal{W}^{\bar{\mathrm{B}}}, \quad \text{where } T_r(\mathrm{i}\nu_1, \ldots, \mathrm{i}\nu_4; \mathrm{i}\nu, \mathrm{i}\nu', \mathrm{i}\omega) \neq 0\}, \tag{14}$$

TABLE I. Our choice of frequency translation tensor T_r in Eq. (9), with representations $r = 1, \ldots, 12$ corresponding to representations $5, \ldots, 16$ of Ref. [34] (representations 1 to 4 can be absorbed into the others by means of a partial fraction decomposition).

	$\mathrm{i}\nu$	$\mathrm{i}\nu'$	$\mathrm{i}\omega$
$T_1(\ldots) =$	$\delta(\mathrm{i}\nu, \mathrm{i}\nu_1)$	$\delta(\mathrm{i}\nu', +\mathrm{i}\nu_4)$	$\delta(\mathrm{i}\omega, \mathrm{i}\nu_1 - \mathrm{i}\nu_2)$
$T_2(\ldots) =$	$\delta(\mathrm{i}\nu, \mathrm{i}\nu_1)$	$\delta(\mathrm{i}\nu', -\mathrm{i}\nu_3)$	$\delta(\mathrm{i}\omega, \mathrm{i}\nu_1 - \mathrm{i}\nu_2)$
$T_3(\ldots) =$	$\delta(\mathrm{i}\nu, \mathrm{i}\nu_1)$	$\delta(\mathrm{i}\nu', +\mathrm{i}\nu_4)$	$\delta(\mathrm{i}\omega, \mathrm{i}\nu_1 + \mathrm{i}\nu_3)$
$T_4(\ldots) =$	$\delta(\mathrm{i}\nu, \mathrm{i}\nu_1)$	$\delta(\mathrm{i}\nu', +\mathrm{i}\nu_2)$	$\delta(\mathrm{i}\omega, \mathrm{i}\nu_1 + \mathrm{i}\nu_3)$
$T_5(\ldots) =$	$\delta(\mathrm{i}\nu, \mathrm{i}\nu_1)$	$\delta(\mathrm{i}\nu', -\mathrm{i}\nu_3)$	$\delta(\mathrm{i}\omega, \mathrm{i}\nu_1 - \mathrm{i}\nu_4)$
$T_6(\ldots) =$	$\delta(\mathrm{i}\nu, \mathrm{i}\nu_1)$	$\delta(\mathrm{i}\nu', +\mathrm{i}\nu_2)$	$\delta(\mathrm{i}\omega, \mathrm{i}\nu_1 - \mathrm{i}\nu_4)$
$T_7(\ldots) =$	$\delta(\mathrm{i}\nu, -\mathrm{i}\nu_2)$	$\delta(\mathrm{i}\nu', +\mathrm{i}\nu_4)$	$\delta(\mathrm{i}\omega, -\mathrm{i}\nu_2 + \mathrm{i}\nu_1)$
$T_8(\ldots) =$	$\delta(\mathrm{i}\nu, -\mathrm{i}\nu_2)$	$\delta(\mathrm{i}\nu', -\mathrm{i}\nu_3)$	$\delta(\mathrm{i}\omega, -\mathrm{i}\nu_2 + \mathrm{i}\nu_1)$
$T_9(\ldots) =$	$\delta(\mathrm{i}\nu, -\mathrm{i}\nu_2)$	$\delta(\mathrm{i}\nu', +\mathrm{i}\nu_4)$	$\delta(\mathrm{i}\omega, -\mathrm{i}\nu_2 + \mathrm{i}\nu_3)$
$T_{10}(\ldots) =$	$\delta(\mathrm{i}\nu, -\mathrm{i}\nu_2)$	$\delta(\mathrm{i}\nu', -\mathrm{i}\nu_3)$	$\delta(\mathrm{i}\omega, -\mathrm{i}\nu_2 - \mathrm{i}\nu_4)$
$T_{11}(\ldots) =$	$\delta(\mathrm{i}\nu, \mathrm{i}\nu_3)$	$\delta(\mathrm{i}\nu', +\mathrm{i}\nu_4)$	$\delta(\mathrm{i}\omega, \mathrm{i}\nu_3 + \mathrm{i}\nu_1)$
$T_{12}(\ldots) =$	$\delta(\mathrm{i}\nu, \mathrm{i}\nu_3)$	$\delta(\mathrm{i}\nu', +\mathrm{i}\nu_4)$	$\delta(\mathrm{i}\omega, \mathrm{i}\nu_3 - \mathrm{i}\nu_2)$

i.e., the outer product of the one-particle sampling frequencies $\mathcal{W}^\alpha$ from Eq. (6) according to the one-particle kernels in Eq. (9), transformed from the "native" frequencies $(i\nu, i\nu, i\omega)$ of each representation to the all-fermionic convention $(i\nu_1, \ldots, i\nu_4)$. We show an example of sampling frequencies later in Fig. 7(a).

This choice ensures that the sampling points required to construct each $r, ll'm$ are present in $\mathcal{W}$. However, since the IR basis is overcomplete, every representation projects to the same set of frequencies and thus the ordinary least-squares problem:

$$\min_G \left[\sum_{(i\nu_i)\in\mathcal{W}} \left| \hat{G}(i\nu_1, \ldots, i\nu_4) - \sum_{r,ll'm} E^{\nu_1\nu_2\nu_3\nu_4}_{r,ll'm} G_{r,ll'm} \right|^2 + \lambda^2 \sum_{r,ll'm} \left| \Gamma^{\text{Tikh.}}_{rll'm} G_{r,ll'm} \right|^2 \right], \tag{15}$$

where E is the transformation tensor from the IR to the sampling frequencies from Eq. (12), is ill-conditioned for $\lambda = 0$. A way to mitigate this problem is by using Tikhonov regularization, where one adds a term to the cost function penalizing large fitting parameters $G_{r,ll'm}$ ($\lambda \neq 0$): one, e.g., can use the *a priori* knowledge of the decay of basis coefficients from Eq. (13) and enforce this by choosing $(\Gamma^{\text{Tikh.}}_{rll'm})^{-1} = S^{\text{F}}_l S^{\text{F}}_{l'} S^{\text{B}}_m$ (for a more detailed discussion see Sec. IV of Ref. [33]).

As for the number of sampling frequencies, one finds $L^3 \leqslant |\mathcal{W}| \leqslant 12L^3$, as sampling frequencies coming from multiple representations may coincide, and one typically has $|\mathcal{W}| \approx 8L^3$.

III. DIAGRAMMATIC EQUATIONS

We will now discuss how to extend sparse modeling in order to solve two-particle diagrammatic equations. Diagrammatic equations are an algebraic way to sum up whole classes of diagrams, usually by invoking a topological argument. They are the bread and butter of the diagrammatic technique and at the heart of renormalization methods and embedding techniques. As in Sec. II, we will start with a one-particle example (the Dyson equation) and then go to the two-particle case.

A. Self-energy and vertex basis

For simplicity, we again start with the one-particle case, where we will introduce the tools that we later use for the two-particle case.

Let $\hat{G}_0$ be the noninteracting one-particle Green's function, i.e., Eq. (2) with $\langle \ldots \rangle$ replaced by the average $\langle \ldots \rangle_0$ over the noninteracting system. It is related to the full Green's function via

$$\hat{G}(i\nu) = \hat{G}_0(i\nu) + \hat{G}_0(i\nu)\,\hat{M}(i\nu)\,\hat{G}_0(i\nu), \tag{16}$$

where $\hat{M}$ is the full one-particle vertex encoding all interactions in the system, and $i\nu$ is the fermionic Matsubara frequency [52]. (We restrict ourselves to fermionic systems for simplicity.) Diagrammatically, Eq. (16) is shown in Fig. 2(a). The full vertex $\hat{M}$ is related to its irreducible counterpart, the

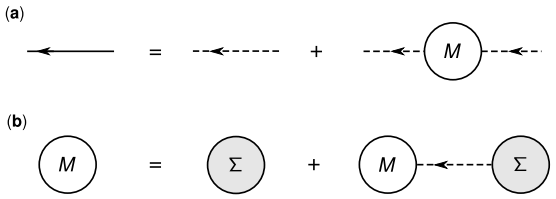


FIG. 2. One-particle diagrammatic equations: (a) relation of the full Green's function G (solid line), noninteracting Green's function G_0 (dashed line) and full one-particle vertex M; (b) Dyson equation in the vertex form, where Σ is the self-energy (irreducible one-particle vertex).

self-energy $\hat{\Sigma}$, via the Dyson equation:

$$\hat{M}(i\nu) = \hat{\Sigma}(i\nu) + \hat{M}(i\nu)\,\hat{G}_0(i\nu)\,\hat{\Sigma}(i\nu), \tag{17}$$

which is shown diagrammatically in Fig. 2(b).

Neither vertex, $\hat{M}$ nor $\hat{\Sigma}$, can be modelled by the same basis as the Green's function, because they contain the Hartree-Fock term, which is a constant in frequency. To model it we need to augment the fermionic kernel in Eq. (3) to

$$K^{\bar{\text{F}}}(i\nu, \nu') = \frac{1}{i\nu - \nu'} + S^{\bar{\text{F}}}_0, \tag{18}$$

where $S^{\bar{\text{F}}}_0$ is an arbitrary nonzero constant. As the remaining terms in $\hat{\Sigma}$ and $\hat{M}$ behave like a scaled Green's function, one has truncated expansions analogous to Eq. (5):

$$\hat{\Sigma}(i\nu) = \sum_{l=0}^{L-1} \hat{U}^{\bar{\text{F}}}_l(i\nu)\Sigma_l + \epsilon_L, \tag{19}$$

$$\hat{M}(i\nu) = \sum_{l=0}^{L-1} \hat{U}^{\bar{\text{F}}}_l(i\nu)M_l + \epsilon'_L, \tag{20}$$

where Σ_l and M_l are again basis coefficients and ϵ_L is an error term associated with truncating the series, which is guaranteed to drop quickly.

In principle, $S^{\bar{\text{F}}}_l$ and $\hat{U}^{\bar{\text{F}}}_l$ are the singular values and left singular functions, respectively, of the augmented kernel (18). However, the kernel is not compact, which means one is unable to compute the singular value expansion numerically. Instead, one chooses $\hat{U}^{\bar{\text{F}}}_0(i\omega) = 1$ and shifts the remaining terms of the unaugmented kernel by one, i.e., $S^{\bar{\text{F}}}_{l+1} = S^{\text{F}}_l$ and $\hat{U}^{\bar{\text{F}}}_{l+1} = \hat{U}^{\text{F}}_l$. We plot the lowest order basis functions in Figs. 3(a) and 3(b).

Fitting Σ_l from Matsubara data for the self-energy requires us to choose a set of fitting frequencies. Since the associated basis functions for $l = 1, \ldots, L-1$ are identical to the underlying left singular functions, the sampling frequencies $\mathcal{W}^{\text{F}}$ for order $L' \geqslant L-1$ allow stable and compact fitting. The Hartree-Fock term M_0, on the other hand, is given by the limit $i\nu \to \infty$, and corrections to this asymptotic constant only decay as $1/i\nu$. Fitting Σ_0 from $\hat{\Sigma}(i\nu)$ is thus a somewhat delicate procedure: On the one hand we would like to have an extra sampling frequency for $i\nu$ large, yet on the other hand one often has unfavorable scaling of the uncertainty in $\hat{\Sigma}(i\nu)$ with increasing frequency [53], which discourages us from doing so. We empirically observe that the distribution of $\mathcal{W}^{\text{F}}$ for order L' extends to higher frequencies as L' is increased.

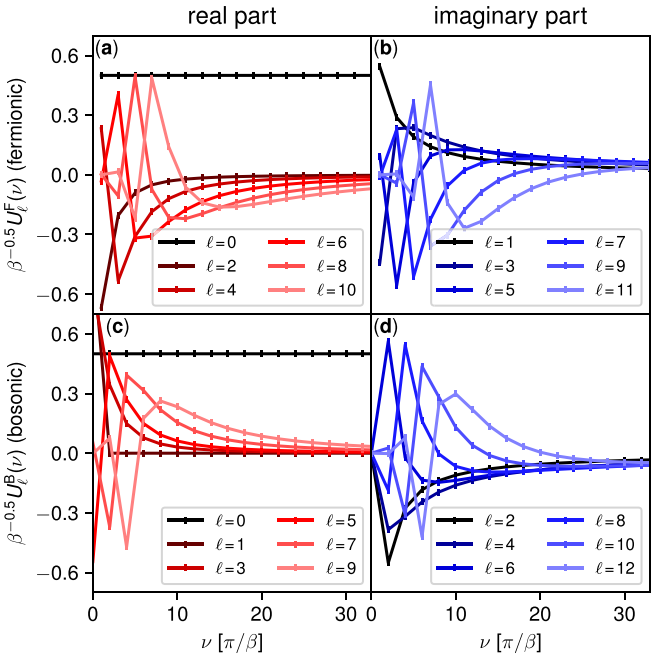


FIG. 3. Lowest-order IR basis functions $\hat{U}_l(\mathrm{i}\nu)$: [(a),(b)] for the augmented fermionic kernel (18); [(c),(d)] for the doubly augmented bosonic kernel (30). Symmetry forces these functions to be either purely real [(a) and (c)] or purely imaginary [(b) and (d)] and we always plot the respective nonzero part. The constant term $\hat{U}_0$ is scaled by a factor $1/2$.

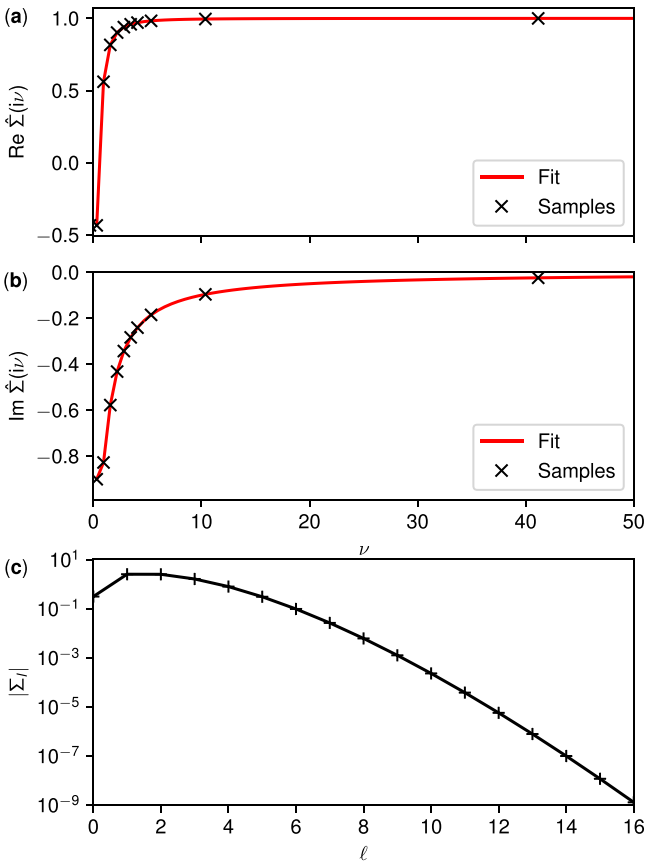


FIG. 4. Expansion of the self-energy in the augmented fermioinc basis: (a) Real part and (b) imaginary part of the self-energy (21), (c) expansion coefficients (19). The crosses in (a) and (b) denote the data on the sampling frequencies.

Therefore, a reasonable choice is to use $\mathcal{W}^{\mathrm{F}}$ for order $L' = L$ as $\mathcal{W}^{\bar{\mathrm{F}}}$ for order L.

Let us demonstrate the fitting for a model self-energy:

$$\hat{\Sigma}(\mathrm{i}\nu) = \Sigma_0 + \frac{1}{\mathrm{i}\nu - \epsilon_0}, \tag{21}$$

where $\Sigma_0 = 1$ and $\epsilon_0 = 0.5$ at $\beta = 10$ ($\omega_{\max} = 1$ and $S_L^{\bar{\mathrm{F}}}/S_0^{\bar{\mathrm{F}}} \simeq 10^{-15}$). The result is shown in Fig. 4. The augmented basis fits the self-energy accurately including the Hartree-Fock term Σ_0 from low to high frequencies. As seen in Fig. 4(c), the expansion coefficient Σ_l decay faster than exponentially [54].

We are now in a position to tackle a diagrammatic equation for the full one-particle vertex $\hat{M}$. We note that by introducing the prefactor $(1 - \hat{\Sigma}\hat{G}_0)$, Eq. (17) becomes a linear equation for $\hat{M}$. We can then insert Eq. (20) to arrive at a least-squares problem for the Dyson equation:

$$\min_{M_l} \sum_{\mathrm{i}\nu \in \mathcal{W}^{\mathrm{F}}} \left| \hat{\Sigma}(\mathrm{i}\nu) - \sum_{l=0}^{L-1} \underbrace{[1 - \hat{\Sigma}(\mathrm{i}\nu)\hat{G}_0(\mathrm{i}\nu)]\hat{U}_l^{\bar{\mathrm{F}}}(\mathrm{i}\nu)}_{A_l(\mathrm{i}\nu)} M_l \right|^2. \tag{22}$$

Let us note that solving the Dyson equation by solving the least-squares problem (22) is obviously not optimal: Since Eq. (17) is diagonal in frequency, one can first solve the equation on the sampling frequencies and then fit M_l from $M(\mathrm{i}\nu)$ in a second step [37]. However, in the two-particle case, this ceases to be an option, since it involves convolutions over frequencies.

B. Bethe-Salpeter equation

After having developed the necessary tools for the sparse modeling of one-particle vertices and the rewriting of diagrammatic equations as fitting problems for that case, we are ready to tackle the two-particle case.

The two-particle analog of Eq. (16) reads

$$\begin{aligned}\hat{G}(\mathrm{i}\nu_1, \mathrm{i}\nu_2, \mathrm{i}\nu_3, \mathrm{i}\nu_4) &= \beta^2 \hat{G}(\mathrm{i}\nu_1)\hat{G}(\mathrm{i}\nu_3)\delta(\nu_1, \nu_2)\delta(\nu_3, \nu_4) \\ &\quad - \beta^2 \hat{G}(\mathrm{i}\nu_1)\hat{G}(\mathrm{i}\nu_3)\delta(\nu_1, \nu_4)\delta(\nu_3, \nu_2) \\ &\quad + \hat{G}(\mathrm{i}\nu_1)\hat{G}(\mathrm{i}\nu_2)\hat{F}(\mathrm{i}\nu_1, \mathrm{i}\nu_2, \mathrm{i}\nu_3, \mathrm{i}\nu_4)\hat{G}(\mathrm{i}\nu_3)\hat{G}(\mathrm{i}\nu_4), \end{aligned} \tag{23}$$

where $\hat{F}$ is the full two-particle vertex [55]. Diagrammatically, Eq. (23) is shown in Fig. 5(a). There are now three different notions of two-particle reducibility in F: with respect to severing frequencies 1,2 from 3,4 (particle-hole channel), frequencies 1,4 from 3,2 (particle-hole transverse channel), and frequencies 1,3 from 2,4 (particle-particle channel). Consequently, there is an irreducible vertex and a corresponding diagrammatic equation for each of these channels [15,21].

Without loss of generality, we will restrict ourselves to the particle–hole channel for now. The Bethe-Salpeter equation, which relates the full vertex $\hat{F}$ to the irreducible vertex $\hat{\Gamma}$,

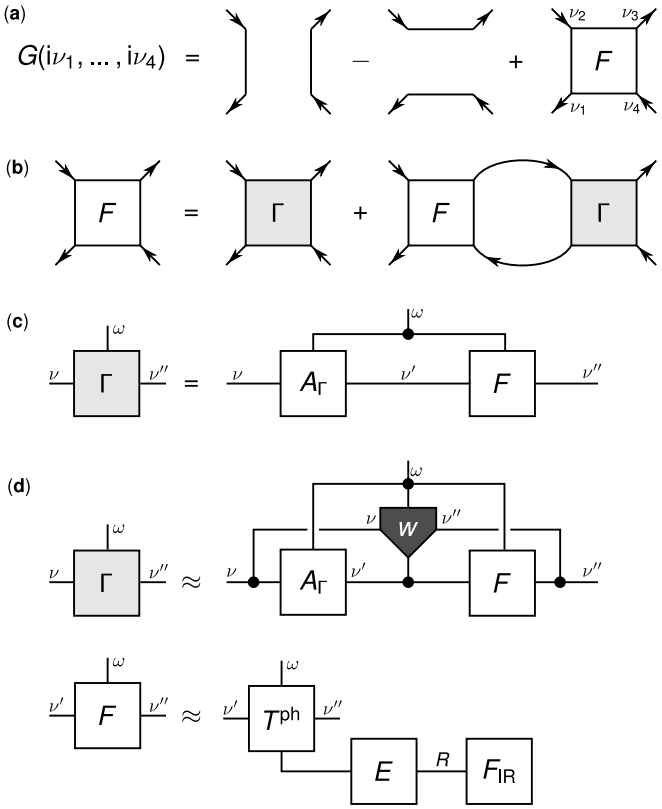


FIG. 5. Sparse modeling for the Bethe-Salpeter equation. (a) Diagrammatic representation the two-particle Green's function (23), where F is the full two-particle vertex. (b) BSE in the particle-hole channel (26), where Γ is the particle-hole irreducible two-particle vertex. (c) Tensor network for the BSE in natural frequencies (27), where A_Γ is defined in Eq. (28). (d) Sparse modeling and quadrature for the BSE (33), where w is the tensor of summation weights, T^{ph} changes from the fermionic to the particle–hole frequency convention, E is the expansion tensor (cf. Fig. 8), and F_{IR} are the IR basis coefficients of the full vertex. —•— only gives a contribution if all connected indices have the same value.

reads [cf. Fig. 5(b)]

$$\hat{F}(\mathrm{i}\nu_1, \mathrm{i}\nu_2, \mathrm{i}\nu_3, \mathrm{i}\nu_4) = \hat{\Gamma}(\mathrm{i}\nu_1, \mathrm{i}\nu_2, \mathrm{i}\nu_3, \mathrm{i}\nu_4) + \frac{1}{\beta^2}\sum_{\nu,\nu'} \hat{F}(\mathrm{i}\nu_1, \mathrm{i}\nu_2, \mathrm{i}\nu, \mathrm{i}\nu')\hat{G}(\mathrm{i}\nu)\hat{G}(\mathrm{i}\nu') \times \hat{\Gamma}(\mathrm{i}\nu', \mathrm{i}\nu, \mathrm{i}\nu_3, \mathrm{i}\nu_4). \tag{24}$$

Solving the BSE in this convention is cumbersome, as it requires a sum over two inner frequencies, one of which is fixed by conservation of energy. This can be eliminated by switching into the "natural" frequency convention for the particle-hole channel:

$$\hat{F}(\mathrm{i}\omega; \mathrm{i}\nu, \mathrm{i}\nu') := \frac{1}{\beta}\hat{F}(\mathrm{i}\nu + \mathrm{i}\omega, \mathrm{i}\nu', \mathrm{i}\nu' - \mathrm{i}\omega, \mathrm{i}\nu), \tag{25}$$

where $\mathrm{i}\omega$ is a bosonic transfer frequency and $\mathrm{i}\nu, \mathrm{i}\nu'$ are fermionic frequencies, and from now on the three-argument version of any quantity indicates the particle-hole convention [56]. Equation (24) then reads

$$\hat{F}(\mathrm{i}\omega; \mathrm{i}\nu', \mathrm{i}\nu'') = \hat{\Gamma}(\mathrm{i}\omega; \mathrm{i}\nu, \mathrm{i}\nu'') + \frac{1}{\beta}\sum_{\nu'} \hat{\Gamma}(\mathrm{i}\omega; \mathrm{i}\nu, \mathrm{i}\nu')\hat{G}(\mathrm{i}\nu')\hat{G}(\mathrm{i}\nu' + \mathrm{i}\omega) \times \hat{F}(\mathrm{i}\omega; \mathrm{i}\nu', \mathrm{i}\nu''). \tag{26}$$

This equation can be rewritten into the following form:

$$\hat{\Gamma}(\mathrm{i}\omega; \mathrm{i}\nu, \mathrm{i}\nu'') = \frac{1}{\beta}\sum_{\nu'} A_\Gamma(\mathrm{i}\omega; \mathrm{i}\nu, \mathrm{i}\nu')\hat{F}(\mathrm{i}\omega; \mathrm{i}\nu', \mathrm{i}\nu''), \tag{27}$$

where we have defined a "Bethe-Salpeter operator" A_Γ:

$$A_\Gamma(\mathrm{i}\omega; \mathrm{i}\nu, \mathrm{i}\nu') = \beta\delta_{\nu\nu'} - \hat{\Gamma}(\mathrm{i}\omega; \mathrm{i}\nu, \mathrm{i}\nu')\hat{G}(\mathrm{i}\nu')\hat{G}(\mathrm{i}\nu' + \mathrm{i}\omega). \tag{28}$$

1. Modelling of the two-particle vertex

Analogous to the expansion of the two-particle Green's function (12), we expand the two-particle vertex in an overcomplete basis:

$$\hat{F}(\mathrm{i}\nu_1, \ldots, \mathrm{i}\nu_4) = \sum_{r=1}^{12}\sum_{\nu\nu'\omega} T_r(\mathrm{i}\nu_1, \ldots, \mathrm{i}\nu_4; \mathrm{i}\nu, \mathrm{i}\nu', \mathrm{i}\omega) \times \sum_{l,l',m=0}^{L-1} \hat{U}_l^{\bar{\mathrm{F}}}(\mathrm{i}\nu)\hat{U}_{l'}^{\bar{\mathrm{F}}}(\mathrm{i}\nu')\hat{U}_m^{\bar{\bar{\mathrm{B}}}}(\mathrm{i}\omega)F_{r,ll'm} + \epsilon_L, \tag{29}$$

and similarly for the irreducible vertex Γ.

Like the self-energy, F and Γ have an asymptotically constant background. Therefore, as in Eq. (20), the fermionic kernel has to be augmented using Eq. (18). At the same time, the bosonic kernel for the two-particle Green's function $K^{\bar{\mathrm{B}}}$ has to be further augmented using a bosonic analog of Eq. (18). This defines an augmented bosonic kernel $K^{\bar{\bar{\mathrm{B}}}}$:

$$K^{\bar{\bar{\mathrm{B}}}}(\mathrm{i}\omega, \omega') = K^{\bar{\mathrm{B}}}(\mathrm{i}\omega, \omega') + S_0^{\bar{\bar{\mathrm{B}}}} = \frac{\omega'}{\mathrm{i}\omega - \omega'} + S_0^{\bar{\bar{\mathrm{B}}}} + S_1^{\bar{\bar{\mathrm{B}}}}\delta_{\mathrm{i}\omega,0} + \frac{S_2^{\bar{\bar{\mathrm{B}}}}(1 - \delta_{\mathrm{i}\omega,0})}{\mathrm{i}\omega}. \tag{30}$$

(We again shift the remaining singular values by $S_l^{\mathrm{B}} \to S_{l+3}^{\bar{\bar{\mathrm{B}}}}$). We plot the corresponding one-particle basis functions in Figs. 3(c) and 3(d).

In addition to the background, the vertex has a rich asymptotic structure, consisting of "lines" and "planes" running horizontally, vertically, and diagonally through the frequency domain and extending to infinity [11]. These structures are captured by the augmented basis functions when combined with the frequency translations in the different representations r. For F, we can prove this by applying the same arguments as for Σ (cf. Sec. III A), but for the dependence on each of the outer frequencies $\mathrm{i}\nu_1, \ldots, \mathrm{i}\nu_4$ separately. F differs from the connected two-particle propagator (23) by removal of four one-particle Green's function lines (one for each outer frequency). Similar to Σ, the dependence of F on each of the frequencies then amounts to a constant term plus a scaled one-particle Green's function, translated through the frequency translation tensor T_r.

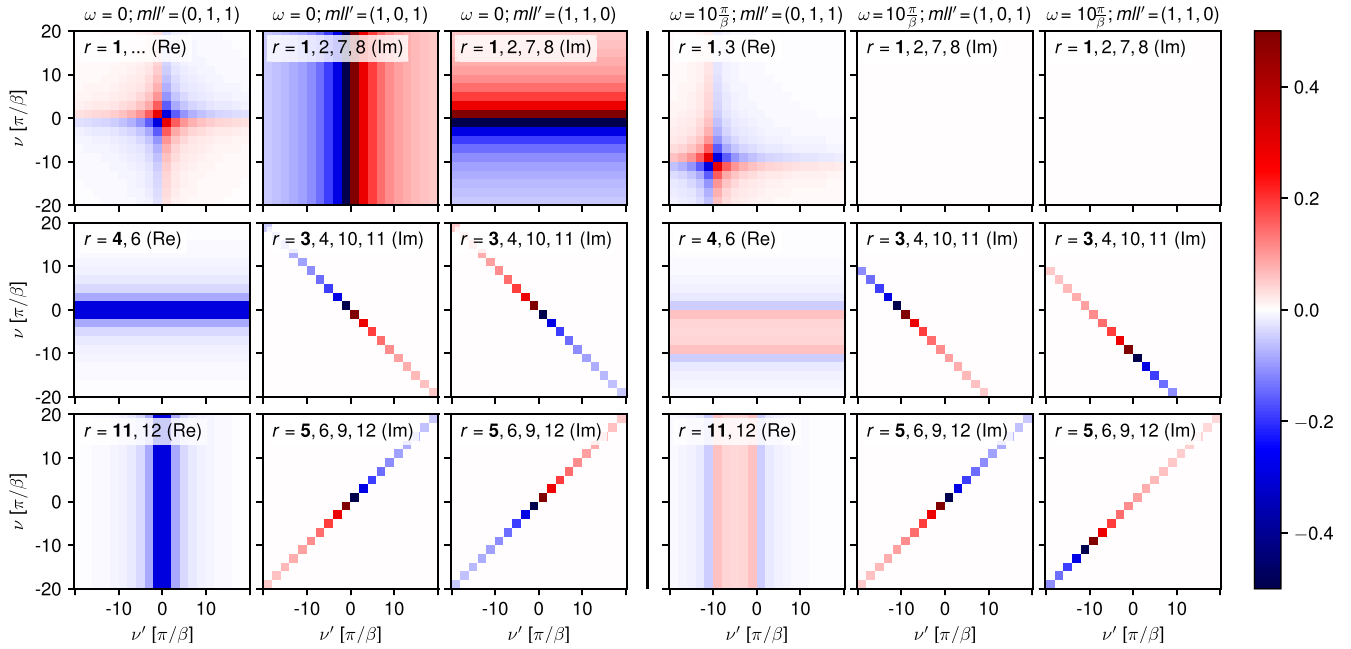


FIG. 6. Four-point vertex basis functions $\hat{U}^{\mathrm{F}}_{l}\hat{U}^{\mathrm{F}}_{l'}\hat{U}^{\bar{\mathrm{B}}}_{m}$ for selected r, l, l', m, in Matsubara frequencies and projected back into the particle-hole convention $(i\omega, i\nu, i\nu')$. We plot over a fermionic frequency box (ν, ν') for two given bosonic frequencies: $\omega = 0$ (left side) and $\omega = 10\pi/\beta$ (right side). Columns show different expansion orders (l, l', m), while the rows show different representations r (the first number, indicated in bold, is the representation plotted; other numbers given denote structurally similar terms). Symmetry dictates that the functions must be purely real (Re) or imaginary (Im), and we plot only the respective part.

For the irreducible vertex Γ, one can show that the asymptotic part has a similar fundamental form as for F away from the divergence lines (cf. Sec. VI A). We thus conjecture that the expansion (23) remains compact also for Γ, and offer numerical evidence in Secs. VI and VII.

Let us illustrate the effect of the augmented two-particle basis by plotting the low-order basis functions in Fig. 6 in Matsubara frequencies in the particle–hole convention, cf. Eq. (25). The results for $\omega = 0$ as given on the left side. We see that, e.g., $F_{r,110}$ translates to the horizontal line structure for $r = 4$ and to the vertical line for $r = 11$. As expected, these structures shift away from the center for $\omega = 10$ (right side).

2. Solving the Bethe-Salpeter equation

Equation (27) is a set of linear equations, which can be solved independently for each $i\omega$. However, unlike Eq. (17), it involves the (infinite) sum over an inner fermionic line [cf. Fig. 5(c)]. As vertices, F and Γ are asymptotically constant, which means the summand in Eq. (27) decays only by virtue of the Green's function lines in Eq. (28) as $O(1/\nu'^2)$. Asymptotic expansions of the vertex [10,57–59] improve upon this, but require additional knowledge of a set of two- and three-point correlation functions, together with a (usually uncontrolled) method of connecting asymptotic and nonasymptotic region.

Storing enough frequencies to reliably perform the fermionic sum quickly exceeds available memory. Fortunately, using sparse modeling (Sec. II B) we can solve this: Analogous to Eq. (22), we rewrite Eq. (27) as a least-squares problem for the IR coefficients $F_{r,ll'm}$:

$$\min_{F} \sum_{(i\omega,i\nu,i\nu'')\in\mathcal{W}'} \left| \Gamma(i\omega; i\nu, i\nu'') - \frac{1}{\beta}\sum_{\nu'}\sum_{r,ll'm} \times A_\Gamma(i\omega; i\nu, i\nu')\, E^{i\nu'+i\omega,i\nu'',i\nu''-i\omega,i\nu'}_{r,ll'm} F_{r,ll'm} \right|^2 . \tag{31}$$

Due to the augmentated kernels entering Eq. (29), $\mathcal{W}'$ is the set of sampling frequencies generated from Eq. (14), but with $\mathcal{W}^{\mathrm{F}}$ $(\mathcal{W}^{\bar{\mathrm{B}}})$ replaced by $\mathcal{W}^{\bar{\mathrm{F}}}$ $(\mathcal{W}^{\bar{\mathrm{B}}})$ and transformed in the particle-hole basis.

Equation (31) solves the storage problems, but it still requires a sum over an infinite set of frequencies. A naive truncation of that sum to the innermost N frequencies will only converge as $\mathcal{O}(1/N)$. In Sec. IV, we will improve on this by developing an algorithm, which replaces the full infinite sum by a weighted finite sum:

$$\frac{1}{\beta}\sum_{\nu'} A_\Gamma(i\omega; i\nu, i\nu')\,\hat{F}(i\omega; i\nu', i\nu'') \approx \sum_{\nu'\in W_{\omega\nu\nu''}} w_{\omega\nu\nu''}(i\nu')\, A_\Gamma(i\omega; i\nu, i\nu')\,\hat{F}(i\omega; i\nu', i\nu''), \tag{32}$$

where $W_{\omega\nu\nu''}$ are the set of quadrature frequencies for the outer frequencies $i\omega, i\nu, i\nu''$ and $w_{\omega\nu\nu''}(i\nu')$ are the corresponding integration weights. The summation rule in Eq. (32) is observed to converge exponentially to the true result with respect to the number of quadrature points.

Inserting Eq. (32) into Eq. (31), we arrive at

$$\min_F \sum_{(\mathrm{i}\omega,\mathrm{i}\nu,\mathrm{i}\nu'')\in\mathcal{W}'} \Bigg| \Gamma(\mathrm{i}\omega;\mathrm{i}\nu,\mathrm{i}\nu'') - \sum_{\nu'\in W_{\omega\nu\nu''}} w_{\omega\nu\nu''}(\mathrm{i}\nu') \times \sum_{r,ll'm} A_\Gamma(\mathrm{i}\omega;\mathrm{i}\nu,\mathrm{i}\nu')\, E_{r,ll'm}^{\mathrm{i}\nu'+\mathrm{i}\omega,\mathrm{i}\nu'',\mathrm{i}\nu''-\mathrm{i}\omega,\mathrm{i}\nu'} F_{r,ll'm} \Bigg|^2, \quad (33)$$

where we have omitted a, usually necessary, regularization term for brevity, cf. Eq. (15). Diagrammatically, Eq. (33) is shown in Fig. 5(d). By using this equation, we have both the truncation error and the quadrature error under control. Therefore we can expect exponential convergence in F in both time and memory. We will discuss details of the fitting in Sec. V.

Let us note that the other direction of the Bethe-Salpeter equation (26), i.e., computing Γ from F, can be done by introducing

$$A_F(\mathrm{i}\omega;\mathrm{i}\nu,\mathrm{i}\nu') = \beta\delta_{\nu\nu'} + \hat{F}(\mathrm{i}\omega;\mathrm{i}\nu,\mathrm{i}\nu')\,\hat{G}(\mathrm{i}\nu')\,\hat{G}(\mathrm{i}\nu'+\mathrm{i}\omega), \quad (34)$$

and switching F and Γ in Eqs. (27) to (33).

IV. SPARSE CONVOLUTION

Ultimately, we want to perform convolutions of vertices in a specific channel, cf. Eq. (27), which requires an (infinite) sum over a fermionic Matsubara frequency. As outlined in Sec. III B, truncating this sum to the innermost N frequencies will only converge as $\mathcal{O}(1/N)$, which is why we are seeking to replace it with a weighted sum over a finite set of frequencies instead, cf. Eq. (32).

A. Simpler case: The Lindhard bubble

To motivate our sparse convolution scheme, let us first consider a simpler problem, which we will again use to develop the necessary tools: Let A and B be one-particle Green's functions and let C be defined as

$$C(\mathrm{i}\omega) = \frac{1}{\beta}\sum_\nu A(\mathrm{i}\nu)B(\mathrm{i}\nu+\mathrm{i}\omega), \quad (35)$$

where $\mathrm{i}\omega$ is a bosonic or fermionic Matsubara frequency and, correspondingly, B and C are bosonic or fermionic Green's functions. The sum converges without a convergence factor as $A(\mathrm{i}\nu)B(\mathrm{i}\nu+\mathrm{i}\omega)$ decays as $\mathcal{O}(1/\nu^2)$. Using the residual calculus, it is straightforward to show that the product (35) can be decomposed as follows:

$$A(\mathrm{i}\nu)B(\mathrm{i}\nu+\mathrm{i}\omega) = A'_\omega(\mathrm{i}\nu) + B'_\omega(\mathrm{i}\nu+\mathrm{i}\omega), \quad (36)$$

where A' and B' are families of auxiliary Green's functions in $\mathrm{i}\nu$ and $\mathrm{i}\nu+\mathrm{i}\omega$, respectively, and both families are parameterized by $\mathrm{i}\omega$. (For completeness, we show this relation in Appendix A.)

Equation (36) again admits an overcomplete representation of the integrand in terms of two sets of coefficients, $A'_{\omega,l}$ and $B'_{\omega,l}$:

$$A(\mathrm{i}\nu)B(\mathrm{i}\nu+\mathrm{i}\omega) = \sum_{l=0}^{L-1}\left[\hat{U}_l^{\mathrm{F}}(\mathrm{i}\nu)A'_{\omega,l} + \hat{U}_l^{\alpha}(\mathrm{i}\nu+\mathrm{i}\omega)B'_{\omega,l}\right] + \epsilon_L. \quad (37)$$

Since each constituent can be represented compactly with the IR basis, there exists a compact representation for the product and the error drops superexponentially, $\epsilon_L \sim S_l$. Using Eq. (37), we can compute the Matsubara sum:

$$C(\mathrm{i}\omega) = \sum_{l=0}^{L-1}\left[U_l^{\mathrm{F}}(0^-)A'_{\omega,l} + U_l^{\alpha}(0^-)B'_{\omega,l}\right] + \epsilon_L, \quad (38)$$

where $U_l^\alpha(0^-)$ are the Fourier transform of the l-th bosonic or fermionic basis function $\hat{U}_l$, evaluated at $\tau = 0^-$.

We proceed in a manner similar to the overcomplete representation of the two-particle function, but for each value of $\mathrm{i}\omega$ separately. We first generate a set $\mathcal{W}_\omega$ of fermionic sampling frequencies for $A(\mathrm{i}\nu)B(\mathrm{i}\nu+\mathrm{i}\omega)$: expanding A' in $\mathrm{i}\nu$ corresponds to the standard set $\mathcal{W}_A = \mathcal{W}^{\mathrm{F}}$ of fermionic sampling frequencies (6), and expanding B' in $\mathrm{i}\nu+\mathrm{i}\omega$ corresponds to a shifted set $\mathcal{W}_B$ of fermionic sampling frequencies. The full set is then just the union of both individual sets:

$$\mathcal{W}_\omega = \mathcal{W}_A \cup \mathcal{W}_B = \mathcal{W}^{\mathrm{F}} \cup \{\mathrm{i}\nu - \mathrm{i}\omega : \mathrm{i}\nu \in \mathcal{W}^\alpha\}. \quad (39)$$

Using the frequency set (39), we can turn Eq. (37) into a least-squares problem:

$$\min_{A',B'} \left\| [AB]_\omega - \begin{pmatrix}\hat{U}_0^{\mathrm{F}} & \hat{U}_\omega^\alpha\end{pmatrix}\begin{pmatrix}A'_\omega \\ B'_\omega\end{pmatrix}\right\|, \quad (40)$$

where the data vector is $[AB]_{\omega,i} = A(\mathrm{i}\nu_i)B(\mathrm{i}\nu_i+\mathrm{i}\omega)$, $\mathrm{i}\nu_i$ runs over frequencies in V_ω, the design matrix is given block-wise as $[\hat{U}_\omega^\alpha]_{il} = \hat{U}_l^\alpha(\mathrm{i}\nu_i+\mathrm{i}\omega)$, and the fitting vector is just the IR coefficients of A' and B', stacked vertically. The Matsubara sum (38) is then given by

$$C(\mathrm{i}\omega) = \sum_{\mathrm{i}\nu\in V_\omega} w_\omega(\mathrm{i}\nu)A(\mathrm{i}\nu)B(\mathrm{i}\nu+\mathrm{i}\omega), \quad (41)$$

i.e., the full sum is replaced by a weighted sum over a small subset of frequencies. The vector of integration weights w is determined by the solution of the following least-squares problem:

$$\min_{w_\omega} \left\| \begin{pmatrix}u^{\mathrm{F}} \\ u^\alpha\end{pmatrix} - \begin{pmatrix}\left[U_0^{\mathrm{F}}\right]^{\mathrm{T}} \\ \left[\hat{U}_\omega^\alpha\right]^{\mathrm{T}}\end{pmatrix} w_\omega \right\|, \quad (42)$$

where $w_{\omega,i} = w_\omega(\mathrm{i}\nu_i)$, the evaluation vector is given block-wise as $[u^\alpha]_l = U_l^\alpha(0^-)$, and where $\hat{U}^{\mathrm{T}}$ denotes the transpose of the design matrix in Eq. (40). If Eq. (42) is underdetermined, we take its least-norm solution.

B. Full two-particle convolution

Now we turn to the case (27) of multiplying two two-particle functions. For simplicity, we focus on the particle-hole channel. Similar relations can be inferred for the transverse channel and for the particle-particle channel.

By transforming Eq. (9) into the particle–hole convention through Eq. (25) and focusing on the dependence on $\mathrm{i}\nu'$, one

has

$$A(\mathrm{i}\omega;\mathrm{i}\nu,\mathrm{i}\nu') = A^{(1)}_{\omega\nu}(\mathrm{i}\nu') + A^{(2)}_{\omega\nu}(\mathrm{i}\nu'+\mathrm{i}\omega) + A^{(3)}_{\omega\nu}(\mathrm{i}\nu'-\mathrm{i}\nu) + A^{(4)}_{\omega\nu}(\mathrm{i}\nu'+\mathrm{i}\nu+\mathrm{i}\omega), \tag{43}$$

where $A^{(1)}$ to $A^{(4)}$ are a family of auxiliary objects. With $\mathrm{i}\omega$ and $\mathrm{i}\nu$ held fixed, $A^{(1)}$ and $A^{(2)}$ have the structure of a fermionic Green's function, while $A^{(3)}$ and $A^{(4)}$ are bosonic Green's functions. Similarly, for the dependence on the other fermionic frequency, one finds:

$$F(\mathrm{i}\omega;\mathrm{i}\nu',\mathrm{i}\nu'') = F^{(1)}_{\omega\nu''}(\mathrm{i}\nu') + F^{(2)}_{\omega\nu''}(\mathrm{i}\nu'+\mathrm{i}\omega) + F^{(3)}_{\omega\nu''}(\mathrm{i}\nu'-\mathrm{i}\nu'') + F^{(4)}_{\omega\nu''}(\mathrm{i}\nu'+\mathrm{i}\nu''+\mathrm{i}\omega). \tag{44}$$

Similar to Eq. (36), we insert Eqs. (43) and (44) into the Eq. (27) and use the residual calculus to obtain a model for the summand:

$$\begin{aligned} &A(\mathrm{i}\omega;\mathrm{i}\nu,\mathrm{i}\nu')F(\mathrm{i}\omega;\mathrm{i}\nu',\mathrm{i}\nu'') \\ &= X^{(1)}_{\omega\nu\nu''}(\mathrm{i}\nu') + X^{(2)}_{\omega\nu\nu''}(\mathrm{i}\nu'+\mathrm{i}\omega) \\ &\quad + X^{(3)}_{\omega\nu\nu''}(\mathrm{i}\nu'-\mathrm{i}\nu) + X^{(4)}_{\omega\nu\nu''}(\mathrm{i}\nu'+\mathrm{i}\nu+\mathrm{i}\omega) \\ &\quad + X^{(5)}_{\omega\nu\nu''}(\mathrm{i}\nu'-\mathrm{i}\nu'') + X^{(6)}_{\omega\nu\nu''}(\mathrm{i}\nu'+\mathrm{i}\nu''+\mathrm{i}\omega), \end{aligned} \tag{45}$$

where $X^{(1)}, \ldots, X^{(6)}$, with ω, ν, ν'' held fixed, are again Green's functions.

This means we can generate an overcomplete representation consisting of six terms, and the set of sampling frequencies becomes

$$\begin{aligned} \mathcal{W}_{\omega\nu\nu''} = \{&\mathrm{i}\nu'+\mathrm{i}\omega_\mathrm{s} : \mathrm{i}\nu' \in \mathcal{W}^\mathrm{F}, \mathrm{i}\omega_\mathrm{s} \in \{0, -\mathrm{i}\omega\}\} \\ &\cup \{\mathrm{i}\nu'+\mathrm{i}\nu_\mathrm{s} : \mathrm{i}\nu' \in \mathcal{W}^\mathrm{B}, \\ &\mathrm{i}\nu_\mathrm{s} \in \{\mathrm{i}\nu, \mathrm{i}\nu'', -\mathrm{i}\nu-\mathrm{i}\omega, -\mathrm{i}\nu''-\mathrm{i}\omega\}\}, \end{aligned} \tag{46}$$

where the "mixing" of fermionic and bosonic models in Eq. (45) is reflected in the presence of both fermionic and bosonic sampling frequencies, shifted by a bosonic and fermionic shift frequency, $\mathrm{i}\omega_\mathrm{s}$ and $\mathrm{i}\nu_\mathrm{s}$, respectively, to create a grid of fermionic frequencies [60].

This fixes the quadrature frequencies in Eq. (32). What remains to be determined are the weights. Analogous to Eq. (42), $w_{\omega\nu\nu''}$ is given by the solution to the least-squares problem:

$$\min_{w_{\omega\nu\nu''}} \left\| \begin{pmatrix} u^\mathrm{F} \\ u^\mathrm{F} \\ u^\mathrm{B} \\ \vdots \\ u^\mathrm{B} \end{pmatrix} - \begin{pmatrix} [U^\mathrm{F}]^\mathrm{T} \\ [U^\mathrm{F}_\omega]^\mathrm{T} \\ [U^\mathrm{B}_\nu]^\mathrm{T} \\ \vdots \\ [U^\mathrm{B}_{\mathrm{i}\nu''+\mathrm{i}\omega}]^\mathrm{T} \end{pmatrix} w_{\omega\nu\nu''} \right\|, \tag{47}$$

where again the evaluation vector is given block-wise by $[u^\alpha]_l = U^\alpha_l(0^-)$ and the design matrix is formed using blocks of $[\hat{U}^\alpha_\omega]_{il} = \hat{U}^\alpha_l(\mathrm{i}\nu_i + \mathrm{i}\omega)$.

Figure 7(a) shows sampling frequencies $\mathcal{W}'$ for a vertex function for $\beta = 1$ and $\omega_\mathrm{max} = 10$ with a cutoff of 10^{-5}. We define quadrature points for the left and right

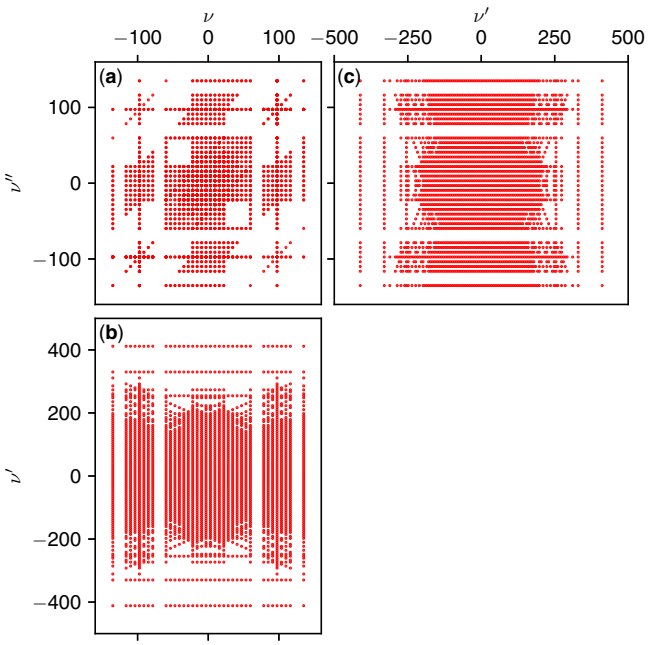


FIG. 7. Sparse frequency sets for $\beta = 1$ and $\omega_\mathrm{max} = 10$ with a cutoff of 10^{-5}: (a) sparse sampling frequency set $\mathcal{W}'$; (b) left-side quadrature frequencies $\mathcal{W}_\mathrm{L}$ for the BSE; (c) right-side quadrature frequencies $\mathcal{W}_\mathrm{R}$. We show only points for zero bosonic frequency, $\mathrm{i}\omega = 0$.

objects as $\mathcal{W}_\mathrm{L} \equiv \{(\mathrm{i}\nu, \mathrm{i}\nu', \mathrm{i}\omega) : \mathrm{i}\nu' \in \mathcal{W}_{\omega\nu\nu''} \wedge (\mathrm{i}\nu, \mathrm{i}\nu'', \mathrm{i}\omega) \in \mathcal{W}'\}$ and $\mathcal{W}_\mathrm{R} \equiv \{(\mathrm{i}\nu', \mathrm{i}\nu'', \mathrm{i}\omega) : \mathrm{i}\nu' \in \mathcal{W}_{\omega\nu\nu''} \wedge (\mathrm{i}\nu, \mathrm{i}\nu'', \mathrm{i}\omega) \in \mathcal{W}'\}$, respectively. We plot $\mathcal{W}_\mathrm{L}$ and $\mathcal{W}_\mathrm{R}$ in Figs. 7(b) and 7(c), respectively. The sampling frequencies and quadrature points are distributed sparsely especially at high frequencies.

With both quadrature points and weights specified, let us discuss computational cost and scaling. From Eq. (46), we have that for each choice of "outer" frequencies, $2L \leqslant |\mathcal{W}_{\omega\nu\nu''}| \leqslant 6L$, with values for typical outer frequencies close to $|\mathcal{W}_{\omega\nu\nu''}| \approx 6L$. The size of the design matrix in Eq. (47) is $6L \times |\mathcal{W}_{\omega\nu\nu''}|$, thus the weights require $\mathcal{O}(L^3)$ time to compute for each outer frequency.

In solving the BSE (33), the quadrature has to be computed for each of the sampling frequencies in $\mathcal{W}$ (14). Since one has $|\mathcal{W}| \approx 8L^3$ (cf. Sec. II B), this implies we in total have to store $\approx 50L^4$ quadrature weights and frequencies. The weights should be precomputed at a cost of $\mathcal{O}(L^6)$. The quadrature (convolution) then takes $\mathcal{O}(L^4)$ time to compute, as it is merely a weighted sum over $\mathcal{O}(L)$ frequencies for each of the $\mathcal{O}(L^3)$ sampling frequencies.

V. BASIS EXPANSION AND FITTING

We will now discuss the solution of the least-squares problem (15). For this, it is useful to first "flatten" the tensor E into a matrix form. We thus impose an ordering on the sampling frequency set (14) and on the set of basis coefficients:

$$\mathcal{W} = \{(\mathrm{i}\nu_{1V}, \mathrm{i}\nu_{2V}, \mathrm{i}\nu_{3V}, \mathrm{i}\nu_{4V})\}_{V=1,\ldots,|\mathcal{W}|}, \tag{48}$$

$$\mathcal{R} = \{(r_R, l_R, l'_R, m_R)\}_{R=1,\ldots,|\mathcal{R}|}, \tag{49}$$

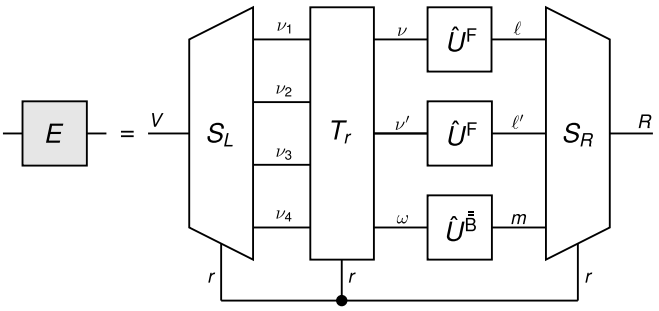


FIG. 8. Tensor network representation of the expansion matrix E (51). S_L and S_R are tensors, which select indices from the inner side and "flatten" them to a single index, T_r is the frequency translation tensor (cf. Table I), $\hat{U}^\alpha$ are one-particle transformation matrices from IR to frequencies with their elements given by Eq. (4), and —•— only gives a contribution if all connected indices have the same value.

where V is an index into the sampling point grid and R is a flat index into $r, ll'm$. Correspondingly, we define $\hat{G}_V := G(\mathrm{i}\nu_{1V}, \ldots, \mathrm{i}\nu_{4V})$ and $G_R := G_{r_R, l_R, l'_R, m_R}$. With these definitions, we arrive at the matrix form of Eq. (15):

$$\min_{\rho} \left\| \hat{G}_V - \sum_{R\in\mathcal{R}} E_{VR} G_R \right\|^2 + \lambda \|\Gamma_R G_R\|^2, \tag{50}$$

where E_{VR} is the flattened version of $E^{\nu_1\nu_2\nu_3\nu_4}_{r,ll'm}$ (we will discuss its explicit form shortly).

After constructing the matrix E_{VR}, Eq. (50) can be fed directly to a ordinary least-squares solver. However, for large L, the cost can be prohibitive: from Eq. (49), one has $|\mathcal{R}| = 12L^3$. Constructing E thus requires storing $\approx 100L^6$ numbers and solving the least-squares problem requires $\mathcal{O}(L^9)$ flops. Even though one has $L = \mathcal{O}(\log(\beta\omega_{\max}\epsilon^{-1}))$, this will only be viable for small values of L.

For larger L, we would like to construct EG and $E^\dagger\hat{G}$ on the fly and use an iterative least-squares solver. We start with the explicit form of E by combining Eqs. (12) and (50):

$$E_{VR} = \sum_{\nu\nu'\omega} T_{r_R}(\mathrm{i}\nu_{1V}, \mathrm{i}\nu_{2V}, \mathrm{i}\nu_{3V}, \mathrm{i}\nu_{4V}; \mathrm{i}\nu, \mathrm{i}\nu', \mathrm{i}\omega) \times \hat{U}^{\mathrm{F}}_{l_R}(\mathrm{i}\nu)\hat{U}^{\mathrm{F}}_{l'_R}(\mathrm{i}\nu')\hat{U}^{\bar{\mathrm{B}}}_{m_R}(\mathrm{i}\omega). \tag{51}$$

The tensor network representation of Eq. (51) is given in Fig. 8. Exploiting that internal structure, one can compute EG and $E^\dagger\hat{G}$ at a cost of $\mathcal{O}(L^5)$ with a negligible memory overhead. We discuss this algorithm in Appendix B.

We can now solve Eq. (50) efficiently using a conjugate gradient solver based on Lanczos bidiagonalization, as these solvers only require us to compute matrix-vector products EG and $E^\dagger\hat{G}$ for given G and $\hat{G}$ instead of creating the full E. These solvers come with a number of guarantees, in particular exponential convergence with the number of matrix–vector products, with the convergence rate depending on how E is conditioned [61]. Apart from pathological cases, they also guarantee success after constructing the "full" matrix, which implies a worst-case runtime scaling of $\mathcal{O}(L^8)$ of the fitting procedure. In practice, we choose the LSMR solver [62] and find that for cutoffs not too low, it converges in relatively few iterations, typically around 100.

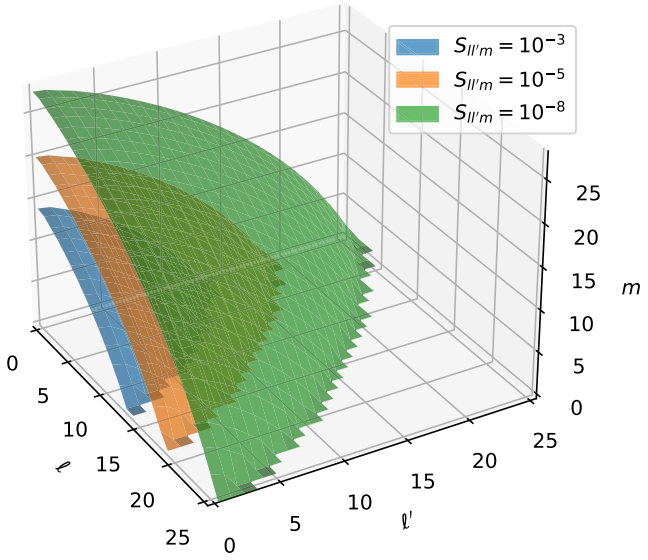


FIG. 9. Truncation of singular values $S_{ll'm} = S^{\mathrm{F}}_l S^{\mathrm{F}}_{l'} S^{\mathrm{B}}_m$ in order to combat overfitting: smoothened isosurfaces of $S_{ll'm}$ for $\beta\omega_{\max} = 100$ for different cutoffs of 10^{-3} (innermost), 10^{-5}, 10^{-8} (outermost).

Although L scales only logarithmically with $\beta\omega_{\max}$, the power in the scaling may become problematic in calculations with large L, e.g., at low T with a small cutoff ϵ_L. We may be able to improve on this scaling using the low-rank approximation of an IR basis vector (tensor network representations) [34].

Let us mention one complication in solving Eq. (50): examining Eq. (13), we see that $G_{r,ll'm}$ must be decay as $S_{ll'm} = S^{\mathrm{F}}_l S^{\mathrm{F}}_{l'} S^{\bar{\mathrm{B}}}_m$. This implies that for a given cutoff $\epsilon_L \sim S_L/S_0$, we are including terms $\rho_{ll'm}$ that are dampened below the level of ϵ_L. We illustrate in Fig. 9, where we plot the isosurfaces of $S_{ll'm}$ for different error levels. Basis coefficients outside of the isosurface cannot be reliably fitted by empirical $\hat{G}$ with errors at the same level, and including them may thus lead to overfitting.

One can however remedy this by restricting $\mathcal{R}$ to

$$\mathcal{R} = \left\{(r, l, l', m) : S_{ll'm}/S_{000} \geqslant S^\alpha_L/S^\alpha_0\right\}, \tag{52}$$

i.e., only to those coefficients, which are not dampened below the tolerance. Since one-particle singular values S^α_l decay faster than $\exp(-cl)$ but slower than $\exp(-cl^2)$ for all real coefficients c, we have $2L^3 < |\mathcal{R}| < 6L^3$, and typically $|\mathcal{R}| \approx 4L^3$. In addition to mitigating oversampling, we have thus also reduced the number of coefficients needed for modeling G by a factor of three. Since in practice $|\mathcal{W}| > 6L^3$, we have also turned Eq. (50) from a formally underdetermined to an overdetermined system.

VI. NUMERICAL BENCHMARKS

We now move to benchmark the method on physical examples and provide numerical evidence for the exponential convergence. One of the simplest test cases is the Anderson

impurity model. Its Hamiltonian reads

$$H = \sum_{ab=1}^{N_{\text{orb}}} E_{ab} \mathrm{c}_a^\dagger \mathrm{c}_b + \frac{1}{4} \sum_{abcd=1}^{N_{\text{orb}}} U_{abcd} \mathrm{c}_a^\dagger \mathrm{c}_b^\dagger \mathrm{c}_d \mathrm{c}_c + \sum_{k=1}^{N_{\text{bath}}} \sum_{a=1}^{N_{\text{orb}}} (V_{ka} \mathrm{f}_k^\dagger \mathrm{c}_a + V_{ka}^* \mathrm{c}_a^\dagger \mathrm{f}_k) + \sum_{k=1}^{N_{\text{bath}}} E'_k \mathrm{f}_k^\dagger \mathrm{f}_k, \tag{53}$$

where c_a annihilates an electron on the impurity spin orbital a, $a = 1, \ldots, N_{\text{orb}}$, f_k annihilates an electron on the bath spin orbital k, $k = 1, \ldots, N_{\text{bath}}$, E_{ab} parameterizes the impurity levels, U_{abcd} is the (antisymmetrized) two-body interaction strength, V_{ka} are the hybridization strengths, and E'_k are the bath level energies.

For $N_{\text{orb}} + N_{\text{bath}}$ small enough, one can compute the full vertex F (23) to arbitrary precision using exact diagonalization. However, even with the full vertex exact, only one or two digits of accuracy in the irreducible vertices Γ are achievable with existing methods before the computational resources are exhausted.

For a convergence analysis across several orders of magnitude, we thus resort to limiting cases of the Anderson impurity model for which analytical results are available: the atomic limit (Sec. VI A) and the weak-coupling limit (Sec. VI B).

A. Hubbard atom

We first consider the Hubbard atom, which is the zero-hybridization limit, $V \to 0$, of the half-filled single impurity Anderson model (53). Its Hamiltonian reads:

$$H = U \mathrm{c}_\uparrow^\dagger \mathrm{c}_\downarrow^\dagger \mathrm{c}_\downarrow \mathrm{c}_\uparrow - \frac{U}{2} (\mathrm{c}_\downarrow^\dagger \mathrm{c}_\downarrow + \mathrm{c}_\uparrow^\dagger \mathrm{c}_\uparrow), \tag{54}$$

where c_σ annihilates an electron of spin σ and U is the strength of the electron-electron interaction. The spectral function is given by $\rho(\omega) = \frac{1}{2}[\delta(\omega + \frac{U}{2}) + \delta(\omega - \frac{U}{2})]$.

Despite the simplicity of Eq. (54), the irreducible and reducible vertices of the Hubbard atom have highly nontrivial structures in the Matsubara frequency domain: there are sharp, "δ-like" planes running horizontally, vertically, and diagonally through the frequency box, structures, which do not decay asymptotically. Moreover, the proximity of a family of poles on the imaginary frequency axis [63,64] as well as channel coupling of the dominant spin susceptibility and the exponentially suppressed charge susceptibility [65] causes the irreducible vertex to vary across several order of magnitudes. As a result, solving the BSE for the atomic limit presents a formidable challenge to solvers, which truncate the Matsubara frequency domain to a finite frequency box.

Fortunately, in Ref. [66] analytical results are derived for F and Γ for the Hubbard atom. We are thus able to perform an absolute convergence analysis of our sparse modeling approach to solving the BSE: we construct A_Γ in Eq. (28) using the analytical expressions for Γ, use Eq. (33) to solve the BSE, and finally compare the resulting F to the analytical expression F_{ex}. In the following we (arbitrarily) choose $\beta = 1.55$, $U = 2.3$ [67] (for results at other values of β see Appendix C). We use the IR basis for $\beta\omega_{\max} = 10$.

Figure 10 shows the fitting error in F for different choices of IR basis truncation L, which corresponds to different cutoffs ϵ for the singular values. The fitting was performed using the LSMR iterative solver (cf. Sec. V) and the system was regularized by imposing an accuracy goal of $\|F - F_{\text{ex}}\|_2 = 5\epsilon\|F\|_2$ (black thin line).

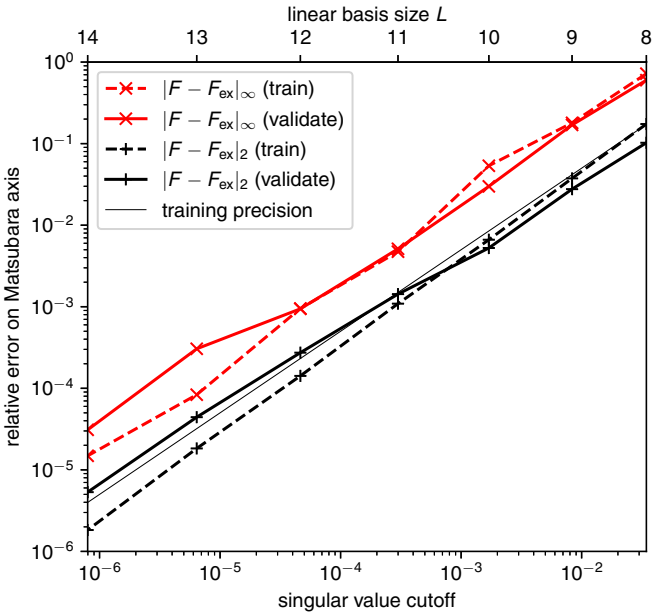


FIG. 10. Sparse modeling of the Bethe-Salpeter equation for the Hubbard atom with $U = 2.3$, $\beta = 1.55$. We plot the error of the irreducible vertex F on the Matsubara axis (normalized by the largest value of F) over the singular value cutoff (bottom axis) or, equivalently, the IR basis cutoff L (top axis). Dotted lines mark the training error (error on the sampling frequencies $\mathcal{W}'$), solid lines indicate the validation error, computed on a frequency box $\mathcal{W}''$ (55). Black plusses denote the least-squares deviation, while red crosses mark the maximum deviation from the analytical result.

The black plusses indicate the least-squares deviation $\|F - F_{\text{ex}}\|_2$, while red crosses indicate the maximum deviation $\|F - F_{\text{ex}}\|_\infty$. Both values are normalized by $\|F\|_\infty$, the largest value of F. We see that the "training" error indicated by dotted lines, i.e., the error on the sampling frequencies $\mathcal{W}'$, closely tracks the desired singular value cutoff. This shows that the IR representation (12) is able to fully capture features of the vertex across multiple orders of magnitude without any underfitting.

To check our fitting we construct a set of validation frequencies:

$$\mathcal{W}'' = \{(\mathrm{i}\omega, \mathrm{i}\nu, \mathrm{i}\nu') : \nu, \nu' \in \left\{-29\frac{\pi}{\beta}, \ldots, 29\frac{\pi}{\beta}\right\}; \omega \in \left\{-28\frac{\pi}{\beta}, \ldots, 28\frac{\pi}{\beta}\right\}\right\} \setminus \mathcal{W}', \tag{55}$$

i.e., a dense frequency box of 30 fermionic and 29 bosonic frequencies centered around the origin, with the sampling frequencies removed. Figure 10 shows that the validation error (solid lines) follows the training error (dotted lines) closely, both for the maximum and average deviation, which implies there is no significant overfitting and the basis has predictive

power at the accuracy level specified by the training (for results obtained for data with statistical noise see Appendix D).

Let us finally direct our attention to the error scaling with the truncation L of the IR basis, shown as top axis. We see that by doubling L, we gain more than four orders of magnitude in terms of precision.

Next, we compare the error scaling of sparse modeling with the conventional (dense) approach: in the latter, one constructs the operator A_Γ (28) on a box of $N \times N$ fermionic frequencies centered around the origin and then solves Eq. (27) by matrix inversion. Fig. 1, shown in the Introduction, compares the validation error on $\mathcal{W}''$ of sparse modeling with cutoff L and of the conventional approach with box size N. Let us note that this is not a fair comparison in terms of computational time, rather, the point is the *scaling* of the error: we see that the error improves as $1/N$, which together with the factor that one requires for storage of N^2 frequencies, makes it difficult to add precision once N becomes sufficiently large. On the other hand, sparse modeling converges quickly with cutoff.

B. Multiorbital weak coupling

We will now consider the opposite limit of the Anderson impurity model (53): the limit of weak coupling. There, it is reasonable to approximate the irreducible vertex with the bare vertex U:

$$\hat{\Gamma}_{abcd}(\mathrm{i}\omega, \mathrm{i}\nu, \mathrm{i}\nu') \approx U_{abcd}, \tag{56}$$

where we now consider more than one orbital and thus Γ acquires spin-orbital indices a, b, c, d. Similarly, the one-particle Green's function can be approximated by its non-interacting counterpart:

$$\hat{G}^{-1}_{ab}(\mathrm{i}\nu) \approx \hat{G}^{-1}_{0,ab}(\mathrm{i}\nu) = \mathrm{i}\nu\delta_{ab} - E_{ab} - \underbrace{\sum_{k=1}^{N_\mathrm{bath}} \frac{V^*_{ka} V_{kb}}{\mathrm{i}\nu - E'_k}}_{\Delta_{ab}(\mathrm{i}\nu)}, \tag{57}$$

where E, E', and V are defined in Eq. (53) and we have introduced the hybridization function $\Delta(\mathrm{i}\nu)$.

The Bethe-Salpeter equation (27) in this multiorbital case then takes the form:

$$\begin{aligned}\hat{\Gamma}_{abgh}(\mathrm{i}\omega; \mathrm{i}\nu, \mathrm{i}\nu'') + \frac{1}{\beta}\sum_{\nu'}\sum_{cdef} \hat{\Gamma}_{abcd}(\mathrm{i}\omega; \mathrm{i}\nu, \mathrm{i}\nu')\hat{G}_{de}(\mathrm{i}\nu') \\ \times \hat{G}_{fc}(\mathrm{i}\nu' + \mathrm{i}\omega)\hat{F}_{efgh}(\mathrm{i}\omega; \mathrm{i}\nu', \mathrm{i}\nu'') \\ = \hat{F}_{abgh}(\mathrm{i}\omega; \mathrm{i}\nu, \mathrm{i}\nu'').\end{aligned} \tag{58}$$

By iterating Eq. (58) with $F_0 = \Gamma = U$ and using the residual calculus (cf. Appendix A), one can show that F is given as the solution to the following algebraic equation:

$$U_{abgh} + \sum_{cdef} U_{abcd}\hat{\chi}_{0,cdef}(\mathrm{i}\omega)\hat{F}_{efgh}(\mathrm{i}\omega) = \hat{F}_{abgh}(\mathrm{i}\omega), \tag{59}$$

where we have introduced

$$\hat{\chi}_{0,cdef}(\mathrm{i}\omega) = \sum_{ij} \frac{f(\gamma_i) - f(\gamma_j)}{\mathrm{i}\omega + \gamma_i - \gamma_j} g^*_{id} g_{ie} g^*_{jf} g_{jc} \tag{60}$$

and $f(x)$ is the Fermi function, and γ_i and g_j are the eigenvalues and eigenvectors, respectively, of the one-body matrix

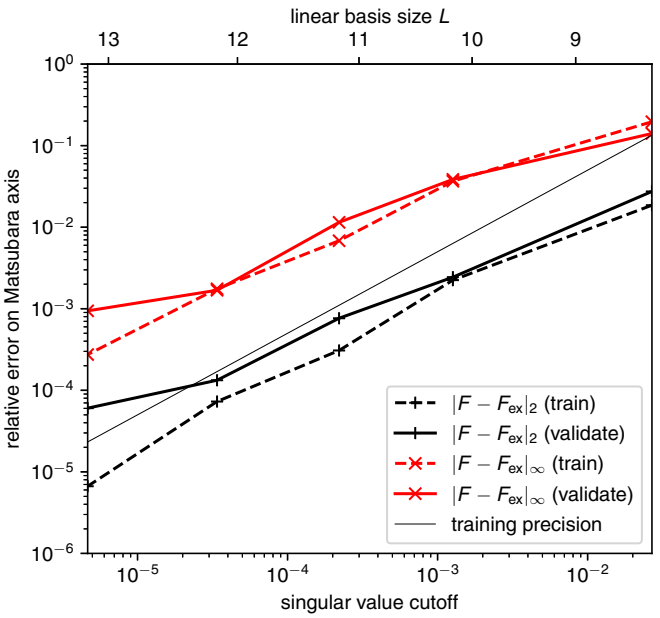


FIG. 11. Sparse modeling of the Bethe-Salpeter equation for the multiorbital weak-coupling limit with $U = 0.3$, $\beta = 1.55$ ($\beta\omega_{\max} = 10$). We plot the error of the full vertex F on the Matsubara axis (normalized by the largest value of F) over the singular value cutoff (bottom axis) or, equivalently, the IR basis cutoff L (top axis). Dotted lines mark the training error (error on the sampling frequencies $\mathcal{W}'$), solid lines indicate the validation error, computed on a frequency box $\mathcal{W}''$ (55). Black plusses denote the least-squares deviation, while red crosses mark the maximum deviation from the analytical result.

formed blockwise by $(E, V, V^\dagger, E')$. (For $\gamma_i = \gamma_j$, the corresponding term in Eq. (60) has to be understood in the limit $\gamma_i \to \gamma_j$.)

We note that in this approximation F has no dependence on fermionic frequencies. By combining a pair of spin-orbital indices into a single index, Eq. (59) can be transformed into a system of linear equations and solved exactly. It thus provides an ideal benchmark for solving Eq. (58) with our sparse modeling approach.

Figure 11 shows the fitting error in F for different choices of IR basis truncation L, which corresponds to different cutoffs ϵ for the singular values. We considered three impurity spin orbitals ($N_\mathrm{orb} = 3$) and four bath sites ($N_\mathrm{bath} = 4$): $U_{abba} = 0.3$ ($a \neq b$) or 0 (otherwise), $\{\gamma_i\} = \{-0.25, -0.1, 0.1, 0.25\}$, $g_{ia} = \cos(i + 3a/2 + 1/10)$ for $a = 1, 2, 3$ and $i = 1, 2, 3, 4$. We used $\beta = 1.55$. Similarly to Fig. 10, Fig. 11 shows the least-squares deviation $\|F - F_\mathrm{ex}\|_2$ and the maximum deviation $\|F - F_\mathrm{ex}\|_\infty$ for the same validation frequencies. One can see that there is neither significant overfitting nor underfitting. The result shows that the present method works for multiorbital systems.

VII. RESULTS FOR THE ANDERSON IMPURITY MODEL

In this section we apply the method to solve the BSE for a fully-fledged Anderson impurity model (53), where analytical expressions for the vertices are not known. As already mentioned in Sec. VI, although this model can be solved exactly, it is not possible to construct benchmark two-particle vertex

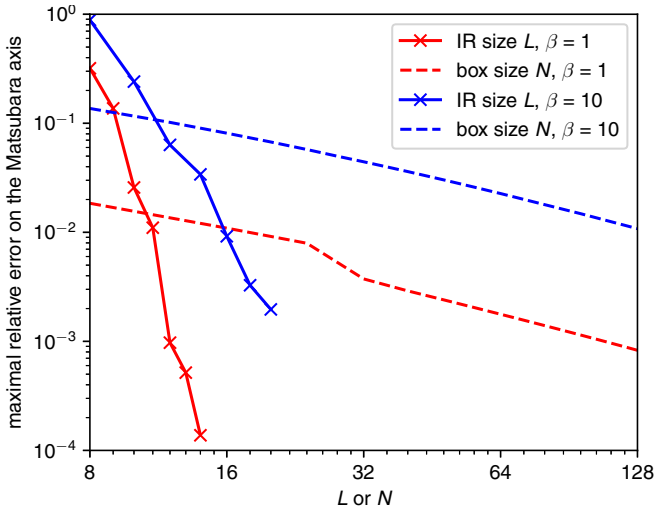


FIG. 12. Comparison between the maximal relative deviation $\|F - F_{ex}\|_\infty / \|F\|_\infty$ of sparse modeling of the Bethe-Salpeter equation with the dense (box) calculation for the impurity model with constant hybridization function $\Delta = \pi/5$ for $U/\Delta = 1.59$ and two values of the inverse temperature $\beta = 1$ ($\beta\omega_{\max} = 10$) and $\beta = 10$ ($\beta\omega_{\max} = 100$).

functions with arbitrary accuracy. The vertices can be however obtained numerically with the precision of several digits from the parquet equations method [21,68].

In the following we will also use the parquet approximation (PA). PA is not exact but gives excellent results in the weak coupling regime [69,70]. It has the advantage, that the two-particle vertices do not have statistical errors. Contrary to the weak-coupling model used for benchmarking in Sec. VI, both the irreducible vertex Γ and the full vertex F are dependent on two fermionic and one bosonic frequency and have nontrivial structure coming from channel mixing in the parquet equations.

The Anderson impurity model is characterized by: (i) The strength of impurity-bath hybridization function $\Delta_{ab}(\nu)$ [cf. Eq. (57)]. We choose it to be spin-orbital and energy independent with $\Delta = \pi/5$ [71]. We restrict ourselves to two spin orbitals (spins) on the impurity and two spins in the bath. (ii) The interaction U on the impurity between electrons with different spins, which is here $U/\Delta = 1.59$, corresponding to weak coupling. (iii) The impurity filling n (here $n = 1$, i.e., half-filling). For these parameters the estimated Kondo temperature is $T_K \approx 0.36$ but due to small value of U/Δ we are far from vertex divergencies present in this model [72,73]. We consider two temperatures $T = 1$ and $T = 0.1$.

In Fig. 12 we show a comparison between maximal relative deviation $\|F - F_{ex}\|_\infty / \|F\|_\infty$ of dense (with linear size of the frequency box N) and sparse (with IR basis size L) evaluations of the BSE for two different inverse temperatures (red: $\beta = 1$, blue: $\beta = 10$). The input irreducible vertex Γ and the benchmark vertex F_{ex} were obtained from a PA solution on a frequency box with linear size $N = 1024$ [74]. The precision of this benchmark calculation, as estimated from box-size convergence, is 10^{-5} for $\beta = 1$ and 10^{-3} for $\beta = 10$, which limits our comparison to only a couple of orders of magnitude for the lower temperature. It is however already visible, that also in this case the error drops quickly with L as compared to the $\mathcal{O}(1/N)$ scaling of the dense calculation. We show here only the larger, maximal relative deviation on the validation set $\mathcal{W}''$ (55). The average relative deviation $\|F - F_{ex}\|_2$ is, similarly to the atomic limit, smaller and has the same scaling behavior as the maximal one.

Comparing the results for the two different inverse temperatures, we see that reaching the same precision for an order of magnitude lower temperature requires only twice the IR basis size L, whereas N needs to be at least an order of magnitude larger.

VIII. CONCLUSIONS AND OUTLOOK

We proposed an efficient method for solving the Bethe-Salpeter equation based on sparse modeling and the intermediate representation (IR). Our algorithm is based on a sparse convolution method, which allows us to perform summation over frequencies of inner propagators needed in solving the Bethe-Salpeter equation. All intermediate objects, such as vertices, are stored in compressed form. We numerically demonstrated the exponential convergence of the algorithm with respect to the basis size for the Hubbard atom, the weak-coupling limit of a multiorbital impurity, and a realistic impurity problem. In the present study, we focused on the particle-hole channel, however, the proposed method can be straightforwardly applied to the particle-particle channel as well.

In Ref. [34], some of the authors and coworkers introduced a tensor network representation of the two-particle Green's functions, which allows us to compress further multidimensional data with many indices for the IR basis: spins and orbitals. Combining the present method and the tensor network representation will open a new avenue to efficient calculation of two-particle response functions of correlated materials in, e.g., DFT(GW)+BSE [3–8], DFT+dynamical mean-field theory [9–12], and nonlocal extensions thereof [13–15], or functional renormalization group [16–18].

The natural connection of the IR basis to the analytic continuation kernel charts a course to obtaining two-particle response functions on the real-frequency axis from Matsubara data, though we expect challenges of regularization and bias similar to the one-particle case will have to be overcome first. Ultimately, this may allow the interpretation of experimental spectroscopy, optical conductivity, and neutron scattering data for models and parameter regimes where direct calculation in real frequencies is infeasible.

Also diagrammatic calculations based on more numerically involving equations, such as parquet equations [19–21], and calculations involving higher-order vertices can be made possible by combining the IR basis for frequencies with the form-factor basis for momenta [75,76].

It was recently shown that irreducible vertices diverge on specific lines in the parameter space [63,64]. This causes numerical instability in solving diagrammatic equations near the divergence line. Such divergence is characterized by emergent poles on an imaginary frequency axis, which the original IR basis does not take into account. Further augmentation of the IR basis may provide a controllable way to numerically handle effects of the vertex divergence.

The unit-tested implementation, with which the data in this paper has been generated, is available from the authors upon request. We expect to release it as an open-source package in the near future.

ACKNOWLEDGMENTS

We are indebted to F. Krien, K. Held, S. Huber, E. Gull, J. Kuneš, T. Nomoto, E. Maggio, and J. Tomczak for fruitful discussions and careful review of the manuscript. H.S. was supported by JSPS KAKENHI Grants No. 18H01158, No. 21H01003, and No. 21H01041. M.W. and A.K. acknowledge support from the Austrian Science Funds (FWF) through project P30997. The calculations were done in part on the Vienna Scientific Cluster (VSC). We used the irbasis library [77] for computing IR basis functions.

APPENDIX A: RESIDUAL CALCULUS

For completeness, this Appendix is deriving Eq. (36) from Eq. (35). We begin by restating Eq. (35): let A and B be one-particle Green's functions and let C be defined as follows:

$$C(\mathrm{i}\omega) = \frac{1}{\beta}\sum_{\nu} A(\mathrm{i}\nu)B(\mathrm{i}\nu + \mathrm{i}\omega). \tag{A1}$$

We expand A and B, respectively, into a set of poles a_i and b_i with expansion coefficients A_i and B_i:

$$C(\mathrm{i}\omega) = \frac{1}{\beta}\sum_{\nu}\sum_{i,j}\frac{A_i B_j}{(\mathrm{i}\nu - a_i)(\mathrm{i}\nu + \mathrm{i}\omega - b_j)}. \tag{A2}$$

The sum (A2) can be performed explicitly via residual calculus, yielding the Lindhard bubble

$$C(\mathrm{i}\omega) = \sum_{i,j}\frac{A_i B_j}{\mathrm{i}\omega + a_i - b_j}(f(a_i) - f(b_j + \mathrm{i}\omega)), \tag{A3}$$

where $f(x)$ is the Fermi function. We note that for a bosonic Matsubara frequency, $f(x + \mathrm{i}\omega) = f(x)$, while for a fermionic Matsubara frequency, $f(x + \mathrm{i}\omega) = b(x)$, where $b(x)$ is the Bose function. By expanding $f(x)$ in its Matsubara sum, we find

$$C(\mathrm{i}\omega) = \frac{1}{\beta}\sum_{\nu}\sum_{i,j}\frac{A_i B_j}{\mathrm{i}\omega + a_i - b_j}\left[\frac{\mathrm{e}^{\mathrm{i}\nu 0^-}}{\mathrm{i}\nu - a_i} - \frac{\mathrm{e}^{\mathrm{i}\nu 0^-}}{\mathrm{i}\nu + \mathrm{i}\omega - b_j}\right]. \tag{A4}$$

Let us briefly comment on the inclusion of Eq. (A3): This may seem like a detour, as a partial fraction decomposition of Eq. (A2) yields Eq. (A4) for each $\mathrm{i}\nu$. However, splitting the terms into two sums, there is an ambiguity in the convergence factor $\exp(\mathrm{i}\nu 0^{\pm})$. This ambiguity must be spurious as the series (A1) is convergent, yet a convergence factor of $\exp(\mathrm{i}\nu 0^+)$ will give an overall sign in Eq. (A3). The proper way to split up this sum is using the residual calculus, which fixes the convergence factor to be $\exp(\mathrm{i}\nu 0^-)$.

Finally, comparing coefficients in Eqs. (A1) and (A4) yields Eq. (36), which we restate here for convenience:

$$C(\mathrm{i}\omega) = A'_{\omega}(\mathrm{i}\nu) + B'_{\omega}(\mathrm{i}\nu + \mathrm{i}\omega), \tag{A5}$$

where A' and B' are now auxiliary Green's functions.

APPENDIX B: FAST ON-THE-FLY EXPANSION

In order to solve the fitting problem (50) using a sparse least-squares solver, we have to apply E, defined in Eq. (51) to an arbitrary IR basis vector as well as $E^\dagger$ to a sampling frequency vector in an efficient manner.

The core part of applying E is the construction of the following intermediate object in each channel:

$$G_n = \sum_{l=0}^{L-1} U_l^{\mathrm{F}}(\mathrm{i}\nu_n)\sum_{l'=0}^{L-1} U_{l'}^{\mathrm{F}}(\mathrm{i}\nu'_n)\sum_{m=0}^{L-1} U_m^{\bar{\mathrm{B}}}(\mathrm{i}\omega_n)\rho_{ll'm}, \tag{B1}$$

i.e., the application of the transformation matrices $\hat{U}$ followed by the projection to those frequencies, which, after translation using T_r, will end up in the sampling frequency set.

We note that this structure (cf. Fig. 8), is in principle well-suited for on-the-fly application, as the $\hat{U}$ tensors can be applied separately one after the other and we then simply select elements. The problem is that the internal frequencies $\mathrm{i}\nu, \mathrm{i}\nu', \mathrm{i}\omega$ not only contain the L one-particle sampling frequencies in $\mathcal{W}^\alpha$, but also any shifts of these frequencies due to T_r. In total one has about $4L^2$ unique frequencies for $\mathrm{i}\nu$ and similarly for $\mathrm{i}\nu'$ and $\mathrm{i}\omega$. Evaluating Eq. (B1) from right to left, we construct an intermediate tensor of size $\mathcal{O}(L^6)$ before selecting $\mathcal{O}(L^3)$ elements from it. This puts the total cost at $\mathcal{O}(L^9)$ time.

We can improve upon this by discarding the block structure and simply compute G_n for each n separately. This involves contracting $\rho_{ll'm}$ along each axis with three vectors $U_l^{\mathrm{F}}(\mathrm{i}\nu_n)$, $U_{l'}^{\mathrm{F}}(\mathrm{i}\nu'_n)$, and $U_m^{\bar{\mathrm{B}}}(\mathrm{i}\omega_n)$ at a cost of $\mathcal{O}(L^3)$, $\mathcal{O}(L^2)$, and $\mathcal{O}(L)$, respectively. Since this has to be done for each sampling frequency, the total cost is $\mathcal{O}(L^6)$.

There is still room for improvement: If we order the sampling frequencies lexicographically, we can reuse partial contraction results from one sampling point to the next. Since one observes only $\mathcal{O}(L^2)$ unique $\mathrm{i}\nu_n$, this brings down the total cost to $\mathcal{O}(L^5)$, while incurring a memory overhead of $\mathcal{O}(L^2)$ for storing the partial results. We list as function `apply` in the algorithm in Fig. 13.

The core of the reverse direction, i.e., the application $E^\dagger$ to a sampling frequency vector, is the construction of the following object in each channel:

$$\rho_{ll'm} = \sum_n \left[U_l^{\mathrm{F}}(\mathrm{i}\nu_n)U_{l'}^{\mathrm{F}}(\mathrm{i}\nu'_n)U_m^{\bar{\mathrm{B}}}(\mathrm{i}\omega_n)\right]^* G_n, \tag{B2}$$

where U^* denotes the complex conjugate. A similar idea applies to Eq. (B2) as to Eq. (B1): We perform the outer product for a sequence of vectors and cache the intermediate results from frequency to frequency. This again yields a $\mathcal{O}(L^5)$ cost and $\mathcal{O}(L^2)$ auxiliary memory requirement.

APPENDIX C: DIVERGENCES OF THE IRREDUCIBLE VERTEX

In Refs. [63,64] it has been shown that the irreducible vertex Γ diverges for certain values of temperature and interaction. These divergencies are also present in the Hubbard atom for a series of T/U values (the ratio T/U is the only energy scale in the atomic limit). The results presented in Sec. VI A were obtained for the irreducible vertex in the density channel for $T/U = 0.28$, which is very close to,

Common input:
– Tuple $I = ((\nu_1, \nu_1', \omega_1), \ldots, (\nu_N, \nu_N', \omega_N))_{n=1,\ldots,N}$, where $i\nu$ and $i\nu'$ are fermionic frequencies, $i\omega$ is a bosonic frequency, and the elements of I are in strict lexicographical order. $I_0 := (-\infty, 0, 0)$.

Function `apply`(ρ) **is**
  Input: $\rho_{ll'm}$, a $L \times L \times L$ complex tensor
  Result: G_n through Eq. (B1)
  Data: A, a $L \times L$ tensor, and B, and L vector
  for $n = 1, \ldots, N$ **do**
    if $\nu_n \neq \nu_{n-1}$ **then**
      $A_{l'm} \leftarrow \sum_{l=0}^{L-1} U_l^{\mathrm{F}}(i\nu_n)\rho_{ll'm}$
    if $\nu_n \neq \nu_{n-1} \vee \nu_n' \neq \nu_{n-1}'$ **then**
      $B_m \leftarrow \sum_{l'=0}^{L-1} U_{l'}^{\mathrm{F}}(i\nu_n')A_{l'm}$
    $G_n \leftarrow \sum_{m=0}^{L-1} U_m^{\mathrm{B}}(i\omega_n)B_m$

Function `adjoint`(G) **is**
  As in apply, but with matrices replaced by their adjoints and steps done in reverse order.

FIG. 13. Algorithm for fast-on-the fly expansion of $E\rho$ using `apply` and $E^\dagger G$ using `adjoint`, each in $\mathcal{O}(L^5)$ time with $\mathcal{O}(L^2)$ auxiliary memory.

but slightly above the first divergence point at $T/U = \frac{\sqrt{3}}{2\pi} \approx 0.276$. In Fig. 14 we show the scaling of the relative average error $\|F - F_{ex}\|_2/\|F\|_\infty$ on the validation set of Matsubara frequencies (cf. Fig. 10) for several values of the inverse temperature $\beta = 1/T$ (keeping the value of $U = 2.3$ constant), that lie in the vicinity of vertex divergencies [64]: slightly

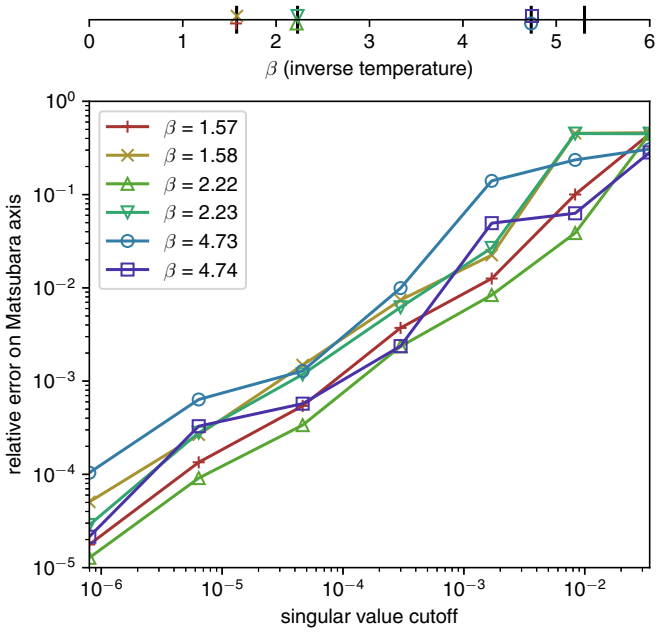


FIG. 14. Relative error $\|F - F_{ex}\|_2/\|F\|_\infty$ on the validation set of Matsubara frequencies for sparse modeling of the Bethe-Salpeter equation in the density channel for the Hubbard atom with $U = 2.3$, and different values of β in the vicinity of the divergencies of Γ that occur for $U = 2.3$ at $\beta = \{1.577, 2.229, 4.732, \ldots\}$ as shown on the inverse-temperature axis above the plot. The rest of the calculation conditions are the same as in Fig. 10.

above and slightly below the first, second, and third divergence point. We observe that the closeness to a divergency does not change the exponential convergence of the sparse solver. It does however influence the overall magnitude of the error.

APPENDIX D: STABILITY AGAINST NOISE

In practical calculations, the propagators and vertex functions may sometimes only be known from a stochastic method: e.g., $F(i\omega; i\nu, i\nu')$ can be obtained from continuous-time quantum Monte Carlo [78], provided the sign problem is not too strong, while $\Gamma(i\omega, i\nu, i\nu')$ is available from bold diagrammatic Monte Carlo [79], provided that the diagrammatic series is convergent and converges to a physical solution. Vertices obtained in this way contain significant statistical noise.

Before we discuss the two-particle case, let us briefly review the effect of noise on fitting one-particle propagators and vertices. There the fitting problem translates to an ordinary least-squares problem, $\min_G \|AG - \hat{G}\|^2$. If we now add white noise on $\hat{G}$, $\hat{G} \to \hat{G} + \delta\|G\|$, the relative error on G on the fitting problem is bounded by

$$\frac{\|\Delta G\|}{\|G\|} \leqslant \kappa[A] \cdot \|\delta\| + \frac{S_L}{S_0}, \tag{D1}$$

where S_l are the singular values and $\kappa[A]$ is the condition number of the fitting matrix A. This means that as we increase the size of the basis L, we expect the error to drop like the singular values until we reach the κ times the noise level, at which point the error flattens out. κ is thus a measure of "noise amplification". Since $\kappa \sim 1$ for Σ and G, we do not observe any noise amplification there [37].

For the two-particle vertices and propagators, the picture is similar, with the complications that (a) the fitting matrix in Eq. (33) is generated from a quadrature rule, which involves another fitting problem (47) and (b) overcompleteness causes the condition number κ to deteriorate. This makes it considerably more complicated to prove the stability of our method against noise in all possible cases.

In order to investigate the stability of the present method against noise for a challenging benchmark, we construct noisy input by adding Gaussian noise to the exact irreducible vertex of the Hubbard atom. We add noise to the atomic Γ as

$$\Gamma(i\omega; i\nu, i\nu') \to \Gamma(i\omega; i\nu, i\nu') + \delta \cdot r(i\omega; i\nu, i\nu')\|\Gamma\|_\infty, \tag{D2}$$

where r are independent identically distributed Gaussian random variables of mean zero and unit variance, and $\delta\,(> 0)$ denotes the level of the noise.

Figure 15 shows the computed results for the sparse modeling of the BSE for the Hubbard atom for $U = 2.3$ and $\beta = 1.8$ and different noise levels. The rest of the parameters in the calculations are the same as those in Fig. 10. One can see that the relative error of the BSE on the Matsubara axis $\|F - F_{ex}\|_2/\|F\|_\infty$ vanishes exponentially until it hits the (amplified) noise level in the input $\Gamma(i\omega; i\nu, i\nu')$, following the form of Eq. (D1). While the noise is thus somewhat amplified, the method seems to be fundamentally numerically stable.

Since F comes with approximately white noise when generated from, e.g., continuous-time quantum Monte Carlo

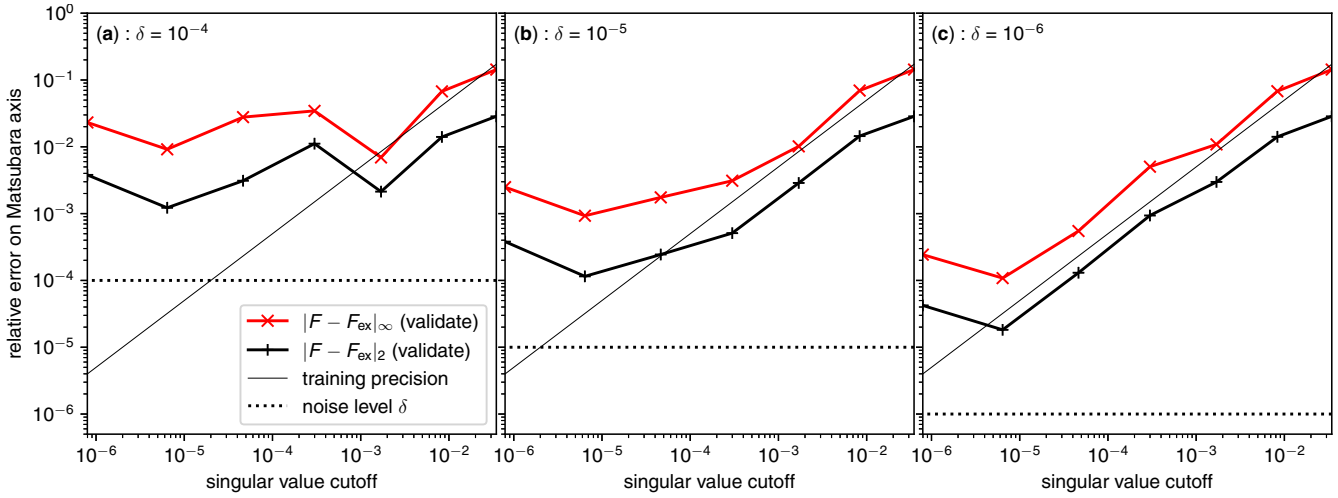


FIG. 15. Sparse modeling of the Bethe-Salpeter equation for the Hubbard atom with $U = 2.3$ and $\beta = 1.8$ with random Gaussian noise in the input data Γ. The noise level δ was set to $\delta = 10^{-4}$, 10^{-5}, and 10^{-6} for panels (a), (b), and (c), respectively. The rest of the calculation conditions are the same as those in Fig. 10. The horizontal dotted lines denote the noise levels, respectively.

in the interaction expansion, auxiliary-field expansion, and also in the hybridization expansion when using symmetric improved estimators [53], we expect the picture to remain qualitatively similar there.

While thorough performance analysis of solving the BSE for noisy input data is beyond the scope of this paper, the numerical results presented suggest the robustness of the exponential scaling of the error against noise.

[1] E. E. Salpeter and H. A. Bethe, A relativistic equation for bound-state problems, Phys. Rev. **84**, 1232 (1951).
[2] M. Gell-Mann and F. Low, Bound states in quantum field theory, Phys. Rev. **84**, 350 (1951).
[3] X. Blase, I. Duchemin, D. Jacquemin, and P.-F. Loos, The Bethe-Salpeter equation formalism: From physics to chemistry, J. Phys. Chem. Lett. **11**, 7371 (2020).
[4] G. Onida, L. Reining, and A. Rubio, Electronic excitations: Density-functional versus many-body Green's-function approaches, Rev. Mod. Phys. **74**, 601 (2002).
[5] F. Aryasetiawan and O. Gunnarsson, The *GW* method, Rep. Prog. Phys. **61**, 237 (1998).
[6] Y. Takada, Inclusion of Vertex Corrections in the Self-Consistent Calculation of Quasiparticles in Metals, Phys. Rev. Lett. **87**, 226402 (2001).
[7] E. Maggio and G. Kresse, *GW* vertex corrected calculations for molecular systems, J. Chem. Theory Comput. **13**, 4765 (2017).
[8] Y. Pavlyukh, G. Stefanucci, and R. van Leeuwen, Dynamically screened vertex correction to *GW*, Phys. Rev. B **102**, 045121 (2020).
[9] A. Georges, G. Kotliar, W. Krauth, and M. J. Rozenberg, Dynamical mean-field theory of strongly correlated fermion systems and the limit of infinite dimensions, Rev. Mod. Phys. **68**, 13 (1996).
[10] J. Kuneš, Efficient treatment of two-particle vertices in dynamical mean-field theory, Phys. Rev. B **83**, 085102 (2011).
[11] G. Rohringer, A. Valli, and A. Toschi, Local electronic correlation at the two-particle level, Phys. Rev. B **86**, 125114 (2012).
[12] H. Hafermann, E. G. C. P. van Loon, M. I. Katsnelson, A. I. Lichtenstein, and O. Parcollet, Collective charge excitations of strongly correlated electrons, vertex corrections, and gauge invariance, Phys. Rev. B **90**, 235105 (2014).
[13] N. Lin, E. Gull, and A. J. Millis, Two-Particle Response in Cluster Dynamical Mean-Field Theory: Formalism and Application to the Raman Response of High-Temperature Superconductors, Phys. Rev. Lett. **109**, 106401 (2012).
[14] J. Otsuki, H. Hafermann, and A. I. Lichtenstein, Superconductivity, antiferromagnetism, and phase separation in the two-dimensional Hubbard model: A dual-fermion approach, Phys. Rev. B **90**, 235132 (2014).
[15] G. Rohringer, H. Hafermann, A. Toschi, A. A. Katanin, A. E. Antipov, M. I. Katsnelson, A. I. Lichtenstein, A. N. Rubtsov, and K. Held, Diagrammatic routes to nonlocal correlations beyond dynamical mean field theory, Rev. Mod. Phys. **90**, 025003 (2018).
[16] W. Metzner, M. Salmhofer, C. Honerkamp, V. Meden, and K. Schönhammer, Functional renormalization group approach to correlated fermion systems, Rev. Mod. Phys. **84**, 299 (2012).
[17] C. Taranto, S. Andergassen, J. Bauer, K. Held, A. Katanin, W. Metzner, G. Rohringer, and A. Toschi, From Infinite to Two Dimensions Through the Functional Renormalization Group, Phys. Rev. Lett. **112**, 196402 (2014).
[18] F. B. Kugler and J. von Delft, Multiloop Functional Renormalization Group that Sums Up All Parquet Diagrams, Phys. Rev. Lett. **120**, 057403 (2018).
[19] C. De Dominicis and P. C. Martin, Stationary entropy principle and renormalization in normal and superfluid systems. II. Diagrammatic formulation, J. Math. Phys. **5**, 31 (1964).
[20] A. N. Vasil'ev, *Functional Methods in Quantum Field Theory and Statistical Physics*, 1st ed. (Taylor and Francis Group, London, 1998), Vol. 1.
[21] N. E. Bickers, Self-consistent many-body theory for condensed matter systems, in *Theoretical Methods for Strongly Correlated*

Electrons. CRM Series in Mathematical Physics, edited by D. Sénéchal, A.-M. Tremblay, and C. Bourbonnais (Springer, New York, 2004).

[22] P. Maris and P. Tandy, QCD modeling of hadron physics, Nucl. Phys. B–Proc. Suppl. **161**, 136 (2006).

[23] N. Brambilla, E. Montaldi, and G. M. Prosperi, Bethe-Salpeter equation in QCD in a Wilson loop context, Phys. Rev. D **54**, 3506 (1996).

[24] P. Ring and P. Schuck, *The Nuclear Many-Body Problem* (Springer-Verlag, Berlin, 1980).

[25] E. Bergli and M. Hjorth-Jensen, Summation of parquet diagrams as an ab initio method in nuclear structure calculations, Ann. Phys. (New York) **326**, 1125 (2011).

[26] The number of orbital indices (by which we mean spin, orbital and for lattice systems also momentum indices here) can often be reduced by using symmetry relations. The number of momentum indices can be further reduced to three using momentum conservation. Eq. (1) is in the particle-hole channel. For details see Sec. III B and Ref. [15].

[27] G. Esirgen and N. E. Bickers, Fluctuation-exchange theory for general lattice Hamiltonians, Phys. Rev. B **55**, 2122 (1997).

[28] Y. M. Vilk and A.-M. S. Tremblay, Non-perturbative many-body approach to the Hubbard model and single-particle pseudogap, J. Phys. I France **7**, 1309 (1997).

[29] K. Zantout, S. Backes, and R. Valentí, Effect of Nonlocal Correlations on the Electronic Structure of LiFeAs, Phys. Rev. Lett. **123**, 256401 (2019).

[30] J. Lichtenstein, S. A. Maier, C. Honerkamp, C. Platt, R. Thomale, O. K. Andersen, and L. Boeri, Functional renormalization group study of an eight-band model for the iron arsenides, Phys. Rev. B **89**, 214514 (2014).

[31] A. Galler, P. Thunström, P. Gunacker, J. M. Tomczak, and K. Held, Ab initio dynamical vertex approximation, Phys. Rev. B **95**, 115107 (2017).

[32] A. Kauch, P. Pudleiner, K. Astleithner, P. Thunström, T. Ribic, and K. Held, Generic Optical Excitations of Correlated Systems: π-tons, Phys. Rev. Lett. **124**, 047401 (2020).

[33] H. Shinaoka, J. Otsuki, K. Haule, M. Wallerberger, E. Gull, K. Yoshimi, and M. Ohzeki, Overcomplete compact representation of two-particle Green's functions, Phys. Rev. B **97**, 205111 (2018).

[34] H. Shinaoka, D. Geffroy, M. Wallerberger, J. Otsuki, K. Yoshimi, E. Gull, and J. Kuneš, Sparse sampling and tensor network representation of two-particle Green's functions, SciPost Phys. **8**, 012 (2020).

[35] H. Shinaoka, J. Otsuki, M. Ohzeki, and K. Yoshimi, Compressing Green's function using intermediate representation between imaginary-time and real-frequency domains, Phys. Rev. B **96**, 035147 (2017).

[36] N. Chikano, J. Otsuki, and H. Shinaoka, Performance analysis of a physically constructed orthogonal representation of imaginary-time Green's function, Phys. Rev. B **98**, 035104 (2018).

[37] J. Li, M. Wallerberger, N. Chikano, C.-N. Yeh, E. Gull, and H. Shinaoka, Sparse sampling approach to efficient ab initio calculations at finite temperature, Phys. Rev. B **101**, 035144 (2020).

[38] T. Wang, T. Nomoto, Y. Nomura, H. Shinaoka, J. Otsuki, T. Koretsune, and R. Arita, Efficient ab initio Migdal-Eliashberg calculation considering the retardation effect in phonon-mediated superconductors, Phys. Rev. B **102**, 134503 (2020).

[39] T. Nomoto, T. Koretsune, and R. Arita, Formation Mechanism of the Helical Q Structure in Gd-Based Skyrmion Materials, Phys. Rev. Lett. **125**, 117204 (2020).

[40] T. Nomoto, T. Koretsune, and R. Arita, Local force method for the ab initio tight-binding model: Effect of spin-dependent hopping on exchange interactions, Phys. Rev. B **102**, 014444 (2020).

[41] Y. Nomura, T. Nomoto, M. Hirayama, and R. Arita, Magnetic exchange coupling in cuprate-analog $d9$ nickelates, Phys. Rev. Res. **2**, 043144 (2020).

[42] S. Iskakov, C.-N. Yeh, E. Gull, and D. Zgid, Ab initio self-energy embedding for the photoemission spectra of NiO and MnO, Phys. Rev. B **102**, 085105 (2020).

[43] N. Witt, E. G. C. P. van Loon, T. Nomoto, R. Arita, and T. O. Wehling, Efficient fluctuation-exchange approach to low-temperature spin fluctuations and superconductivity: From the Hubbard model to $Na_x\ CoO_2 \cdot yH_2O$, Phys. Rev. B **103**, 205148 (2021).

[44] J. Otsuki, M. Ohzeki, H. Shinaoka, and K. Yoshimi, Sparse modeling approach to analytical continuation of imaginary-time quantum Monte Carlo data, Phys. Rev. E **95**, 061302(R) (2017).

[45] H. Shinaoka, N. Chikano, E. Gull, J. Li, T. Nomoto, J. Otsuki, M. Wallerberger, T. Wang, and K. Yoshimi, Efficient ab initio many-body calculations based on sparse modeling of Matsubara Green's function, arXiv:2106.12685.

[46] J. Otsuki, M. Ohzeki, H. Shinaoka, and K. Yoshimi, Sparse modeling in quantum many-body problems, J. Phys. Soc. Japan **89**, 012001 (2020).

[47] G. D. Mahan, *Many-Particle Physics* (Kluwer Academic/Plenum Publishers, New York, 2000).

[48] P. C. Hansen, *Discrete Inverse Problems: Insights and Algorithms* (SIAM, Philadelphia, 2010).

[49] In general, the complex conjugate of $V_l^\alpha(\omega')$ appears in Eq. (4). For convenience, we choose $V_l^\alpha(\omega')$ as real functions. In this gauge, the left singular functions, when transformed to τ, are real because the Fredholm kernel is real in τ.

[50] C.-W. Ha, Eigenvalues of differentiable positive definite kernels, SIAM J. Math. Anal. **17**, 415 (1982).

[51] N. Yarvin and V. Rokhlin, Generalized Gaussian quadratures and singular value decompositions of integral operators, SIAM J. Sci. Comput. **20**, 699 (1996).

[52] $\hat{M}$ is usually known as scattering matrix or improper self-energy [80].

[53] J. Kaufmann, P. Gunacker, A. Kowalski, G. Sangiovanni, and K. Held, Symmetric improved estimators for continuous-time quantum Monte Carlo, Phys. Rev. B **100**, 075119 (2019).

[54] At lower T, one may have to impose compactness explicitly on the expansion coefficients in the fitting as the augmented basis is overcomplete. See discussions in Sec. V.

[55] Note that this convention is somewhat uncommon in literature, and especially differs from Ref. [15]. There, usually a three-frequency convention is used.

[56] Note, that there is an additional prefactor β coming from the evaluation of a δ for enforcing the energy conservation.

[57] N. Wentzell, G. Li, A. Tagliavini, C. Taranto, G. Rohringer, K. Held, A. Toschi, and S. Andergassen, High-frequency asymptotics of the vertex function: Diagrammatic parametrization

and algorithmic implementation, Phys. Rev. B **102**, 085106 (2020).
[58] F. Krien, Efficient evaluation of the polarization function in dynamical mean-field theory, Phys. Rev. B **99**, 235106 (2019).
[59] G. Li, N. Wentzell, P. Pudleiner, P. Thunström, and K. Held, Efficient implementation of the parquet equations: Role of the reducible vertex function and its kernel approximation, Phys. Rev. B **93**, 165103 (2016).
[60] Let us emphasize that vertex augmentation of the basis, cf. Sec. III A, is not necessary here for the quadrature as the summand behaves as a Green's function like by virtue of Eq. (27).
[61] Y. Saad, *Iterative Methods for Sparse Linear Systems* (SIAM, Philadelphia, 2003).
[62] D. C.-L. Fong and M. Saunders, LSMR: An iterative algorithm for sparse least-squares problems, SIAM J. Sci. Comput. **33**, 2950 (2011).
[63] T. Schäfer, G. Rohringer, O. Gunnarsson, S. Ciuchi, G. Sangiovanni, and A. Toschi, Divergent Precursors of the Mott–Hubbard Transition at the Two-Particle Level, Phys. Rev. Lett. **110**, 246405 (2013).
[64] T. Schäfer, S. Ciuchi, M. Wallerberger, P. Thunström, O. Gunnarsson, G. Sangiovanni, G. Rohringer, and A. Toschi, Nonperturbative landscape of the Mott–Hubbard transition: Multiple divergence lines around the critical endpoint, Phys. Rev. B **94**, 235108 (2016).
[65] F. Krien, A. Valli, and M. Capone, Single-boson exchange decomposition of the vertex function, Phys. Rev. B **100**, 155149 (2019).
[66] P. Thunström, O. Gunnarsson, S. Ciuchi, and G. Rohringer, Analytical investigation of singularities in two-particle irreducible vertex functions of the Hubbard atom, Phys. Rev. B **98**, 235107 (2018).
[67] We show the results for the density vertex. For this choice of parameters the only energy scale of the Hubbard atom is $T/U = 0.28$ and lies slightly above the first divergence line at $T/U = \sqrt{3}/(2\pi) \approx 0.276$ [64].
[68] G. Li, A. Kauch, P. Pudleiner, and K. Held, The victory project v1.0: An efficient parquet equations solver, Comput. Phys. Commun. **241**, 146 (2019).
[69] C. Hille, F. B. Kugler, C. J. Eckhardt, Y.-Y. He, A. Kauch, C. Honerkamp, A. Toschi, and S. Andergassen, Quantitative functional renormalization group description of the two-dimensional Hubbard model, Phys. Rev. Res. **2**, 033372 (2020).
[70] T. Schäfer, N. Wentzell, F. Šimkovic, Y.-Y. He, C. Hille, M. Klett, C. J. Eckhardt, B. Arzhang, V. Harkov, F. m. c.-M. Le Régent, A. Kirsch, Y. Wang, A. J. Kim, E. Kozik, E. A. Stepanov, A. Kauch, S. Andergassen, P. Hansmann, D. Rohe, Y. M. Vilk *et al.*, Tracking the Footprints of Spin Fluctuations: A MultiMethod, MultiMessenger Study of the Two-Dimensional Hubbard Model, Phys. Rev. X **11**, 011058 (2021).
[71] The same hybridization function was used in Refs. [72,73].
[72] P. Chalupa, P. Gunacker, T. Schäfer, K. Held, and A. Toschi, Divergences of the irreducible vertex functions in correlated metallic systems: Insights from the Anderson impurity model, Phys. Rev. B **97**, 245136 (2018).
[73] P. Chalupa, T. Schäfer, M. Reitner, D. Springer, S. Andergassen, and A. Toschi, Fingerprints of the Local Moment Formation and its Kondo Screening in the Generalized Susceptibilities of Many-Electron Problems, Phys. Rev. Lett. **126**, 056403 (2021).
[74] Concretely, we show the results for the magnetic channel.
[75] C. J. Eckhardt, C. Honerkamp, K. Held, and A. Kauch, Truncated unity parquet solver, Phys. Rev. B **101**, 155104 (2020).
[76] C. J. Eckhardt, G. A. H. Schober, J. Ehrlich, and C. Honerkamp, Truncated-unity parquet equations: Application to the repulsive Hubbard model, Phys. Rev. B **98**, 075143 (2018).
[77] N. Chikano, K. Yoshimi, J. Otsuki, and H. Shinaoka, irbasis: Open-source database and software for intermediate-representation basis functions of imaginary-time Green's function, Comput. Phys. Commun. **240**, 181 (2019).
[78] E. Gull, A. J. Millis, A. I. Lichtenstein, A. N. Rubtsov, M. Troyer, and P. Werner, Continuous-time Monte Carlo methods for quantum impurity models, Rev. Mod. Phys. **83**, 349 (2011).
[79] N. V. Prokof'ev and B. V. Svistunov, Bold diagrammatic Monte Carlo: A generic sign-problem tolerant technique for polaron models and possibly interacting many-body problems, Phys. Rev. B **77**, 125101 (2008).
[80] A. Fetter and J. Walecka, *Quantum Theory of Many-Particle Systems* (Dover Publications, New York, 2003).